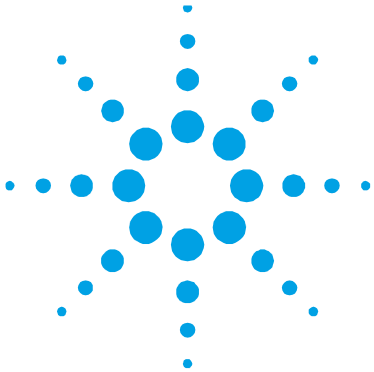


Agilent GeneSpring

User Manual

For Research Use Only. Not for use in diagnostic procedures.

Agilent Technologies

Notices

© Agilent Technologies, Inc. 2016

No part of this manual may be reproduced in any form or by any means (including electronic storage and retrieval or translation into a foreign language) without prior agreement and written consent from Agilent Technologies, Inc. as governed by United States and international copyright laws.

Edition

Revised March 2016

Printed in USA

Agilent Technologies, Inc.
5301 Stevens Creek Blvd.
Santa Clara, CA 95051 USA

Java™ is a U.S. trademark of Sun Microsystems, Inc.

Windows ® is a U.S. registered trademark of Microsoft Corporation.

Software Revision

This guide is valid for 14.5 revisions of the Agilent GeneSpring software.

Software Revision

The material contained in this document is provided "as is," and is subject to being changed, without notice, in future editions. Further, to the maximum extent permitted by applicable law, Agilent disclaims all warranties, either express or implied, with regard to this manual and any information contained herein, including but not limited to the implied warranties of merchantability and fitness for a particular purpose. Agilent shall not be liable for errors or for incidental or consequential damages in connection with the furnishing, use, or performance of this document or of any information contained herein. Should Agilent and the user have a separate written agreement with warranty terms covering the material in this document that conflict with these terms, the warranty terms in the separate agreement shall control.

Technology Licenses

The hardware and/or software described in this document are furnished under a license and may be used or copied only in accordance with the terms of such license.

Restricted Rights Legend

U.S. Government Restricted Rights. Software and technical data rights granted to the federal government include only those rights customarily provided to end user customers. Agilent provides this customary commercial license in Software and technical data pursuant to FAR 12.211 (Technical Data) and 12.212 (Computer Software) and, for the Department of Defense, DFARS 252.227-7015 (Technical Data - Commercial Items) and DFARS 227.7202-3 (Rights in Commercial Computer Software or Computer Software Documentation).

Contents

Contents i

List of Figures iii

List of Tables v

1 GeneSpring Installation 1

1.1 Supported and Tested Platforms 1

1.2 Installation on Windows 2

1.2.1 Installation and Usage Requirements 3

1.2.2 **GeneSpring** Installation Procedure for Microsoft Windows 3

1.2.3 Activating your **GeneSpring** 4

1.2.4 Uninstalling **GeneSpring** from Windows 5

1.3 Installation on Linux 6

1.3.1 Installation and Usage Requirements 6

1.3.2 **GeneSpring** Installation Procedure for Linux 7

1.3.3 Activating your **GeneSpring** 7

1.3.4 Uninstalling **GeneSpring** from Linux 8

1.4 Installation on Apple Macintosh . . . 9
1.4.1 Installation and Usage Requirements . . . 9
1.4.2 **GeneSpring** Installation Procedure for Macintosh . . . 9
1.4.3 Activating your **GeneSpring** . . . 10
1.4.4 Uninstalling **GeneSpring** from Mac . . . 11
1.5 **GeneSpring** Licensing . . . 11
1.6 License Manager . . . 13
1.6.1 Utilities of the License Manager . . . 14
1.6.2 Viewer Mode . . . 17
1.7 Upgrade . . . 17
1.8 System Requirements for Copy Number Experiments, Association Experiments, and Correlation Analysis . . . 17
1.8.1 Copy Number and Association Experiments . . . 17
1.8.2 Correlation Analysis . . . 18

2 GeneSpring Quick Tour **21**

2.1 Introduction . . . 21
2.2 Launching **GeneSpring** . . . 21
2.3 **GeneSpring** User Interface . . . 21
2.3.1 **GeneSpring** Desktop . . . 22
2.3.2 Project Navigator . . . 23
2.3.3 The Workflow Browser . . . 24
2.3.4 My Lists . . . 25

2.3.5 The Legend Window 25

2.3.6 Status Bar 25

2.4 Organizational Elements and Terminology in **GeneSpring** 26

2.4.1 Project 26

2.4.2 Experiment 27

2.4.3 Sample 28

2.4.4 Experiment Grouping, Parameters and Parameter Values 29

2.4.5 Conditions and Interpretations 29

2.4.6 Entity List 32

2.4.7 Region List 34

2.4.8 Report 35

2.4.9 Entity Tree, Condition Tree, Combined Tree and Classification 36

2.4.10 Class Prediction Model 37

2.4.11 Script 37

2.4.12 Pathway 37

2.4.13 Inspectors 38

2.4.14 Hierarchy of objects 39

2.4.15 Right-click operations 40

2.4.16 Search 51

2.4.17 Saving Projects 56

2.4.18 Sharing Projects - Export and Import utilities 57

2.4.19 Software Organization 59

2.5 Exporting Views 59
2.5.1 Export As Report 59
2.6 Utilities under Tools Menu 62
2.6.1 Scripting and R Editor 62
2.6.2 Import Pathways From WikiPathways 63
2.6.3 Import Pathways From BioCyc 63
2.6.4 Import Pathways From KEGG 63
2.6.5 Import Pathways From File 63
2.6.6 Import from Pathway Architect 63
2.6.7 Import BROAD GSEA Genesets 64
2.6.8 Import NCBI GEO Experiment 64
2.6.9 Export ArrayExpress MAGE-TAB 64
2.6.10 Import Report 65
2.6.11 Import Report from Dropbox 65
2.6.12 Import Report from Googledrive 65
2.6.13 Migration 66
2.6.14 Memory Monitor 66
2.6.15 Edit Pathway Theme 66
2.6.16 Options 66
2.6.17 Create Custom Reference 67
2.6.18 Backup and Restore 67
2.7 Update Utility 70

2.7.1 Product Updates . . . 71

2.8 Getting Help . . . 72

3 Technology and Biological Genome **75**

3.1 Technology . . . 75

3.1.1 Standard Technology Creation . . . 76

3.1.2 Agilent eArray Technology Creation . . . 78

3.1.3 Custom Technology Creation . . . 80

3.1.4 Creating Custom CGH Technology . . . 80

3.1.5 Technology creation on the fly . . . 80

3.1.6 Inspection of Technology . . . 81

3.1.7 Technology Deletion . . . 81

3.2 Update Technology . . . 81

3.2.1 Automatic Query of Update Server . . . 82

3.2.2 Update Technology Annotations . . . 83

3.3 Annotations Manager . . . 86

3.3.1 Create New Organism . . . 88

3.3.2 Create New Build . . . 89

3.4 Translation . . . 95

3.4.1 Implementation . . . 97

3.4.2 Explicit Translation mapping . . . 101

3.4.3 Translation in Copy number and Association experiments . . . 101

3.5 Biological Genome 101

4 Data Migration **105**

4.1 **GeneSpring** Data Migration from GeneSpring GX 7 105

4.1.1 Migration Steps 105

4.1.2 Migrated Objects - Genomes, Projects, Experiments 109

4.1.3 Migrated Objects - Experiments 110

4.1.4 Migrated Objects - Samples 110

4.1.5 Entity Lists, Gene Trees, Condition Trees and Classifications 112

4.1.6 Migrated Objects - Other objects and miscellaneous rules 112

4.1.7 Potential causes of Migration failure and Known Issues 112

5 Data Visualization **115**

5.1 View 115

5.1.1 The View Framework in **GeneSpring** 115

5.1.2 View Operations 116

5.2 The Spreadsheet View 124

5.2.1 Spreadsheet Operations 126

5.2.2 Spreadsheet Properties 126

5.3 MvA plot 129

5.4 The Scatter Plot 130

5.4.1 Scatter Plot for 2 color experiments 131

5.4.2 Scatter Plot Operations 132

5.4.3 Scatter Plot Properties 132

5.5 The Profile Plot View 137

5.5.1 Time Series Plot 138

5.5.2 Profile Plot Operations 142

5.5.3 Profile Plot Properties 142

5.6 The Heatmap View 145

5.6.1 Heatmap Operations 146

5.6.2 Heatmap Toolbar 147

5.6.3 Heatmap Properties 148

5.6.4 Heatmap for viewing Copy Number Analysis Results 150

5.7 The Histogram View 150

5.7.1 Histogram Operations 152

5.7.2 Histogram Properties 152

5.8 The Bar Chart 155

5.8.1 Bar Chart Operations 156

5.8.2 Bar Chart Properties 157

5.9 The Matrix Plot View 159

5.9.1 Matrix Plot Operations 161

5.9.2 Matrix Plot Properties 161

5.10 Summary Statistics View 163

5.10.1 Summary Statistics Operations 165

5.10.2 Summary Statistics Properties 165

5.11 The Box Whisker Plot . . . 168
5.11.1 Box Whisker Operations . . . 169
5.11.2 Box Whisker Properties . . . 170
5.12 The Venn Diagram . . . 173
5.12.1 Venn Diagram Operations . . . 175
5.12.2 Venn Diagram Properties . . . 178
5.13 The Violin Plot . . . 180
5.13.1 Violin Plot Operations . . . 181
5.13.2 Violin Plot Properties . . . 181
5.13.3 Formula for drawing a Violin plot in **GeneSpring** . . . 184
5.14 LD Plot . . . 185
5.14.1 LD Plot Toolbar . . . 186
5.14.2 LD Measure Options . . . 186
5.14.3 LD Plot Properties . . . 187
5.15 Genome Browser . . . 188
5.16 Plot Options . . . 188
5.16.1 Plot Log10/Linear Values . . . 188
5.16.2 Plot List Associated Values . . . 188
5.17 Miscellaneous operations . . . 191
5.17.1 Save Current view . . . 191
5.17.2 Find Entity . . . 192
5.17.3 Inspect Entities . . . 192

5.17.4 Properties . . . 192
5.17.5 Exporting Views . . . 192
5.17.6 Configuring Views . . . 192

6 Genome Browser **195**

6.1 Introduction . . . 195
6.2 Getting started with the Genome Browser . . . 196
6.2.1 Launching the Genome Browser . . . 196
6.3 Visualization in Genome Browser . . . 198
6.3.1 Region View . . . 199
6.3.2 Genome View . . . 205
6.3.3 Chromosome View . . . 206
6.3.4 Annotations Tab . . . 207
6.3.5 Spreadsheet . . . 207
6.3.6 Search Result . . . 209
6.4 View Customizations . . . 209
6.4.1 Edit Track Properties . . . 210
6.4.2 Other Right Click Options of a Selected Track . . . 213
6.5 Track Operations . . . 214
6.6 Elastic Genome Browser . . . 214
6.7 A few utilities in Copy Number Experiments . . . 218

7 Analyzing Expression Data **219**

7.1 Overview . . . 219
7.1.1 Summary of experiment types . . . 219
7.2 Running the Expression Experiment Workflow . . . 220
7.2.1 Creating the experiment . . . 220
7.2.2 Data Processing . . . 223
7.2.3 Analysis: Biological Significance . . . 223
7.3 Data Import Wizard . . . 235
7.3.1 Experiment Setup . . . 236
7.3.2 Quality Control . . . 247
7.3.3 Analysis . . . 254
7.3.4 Class Prediction . . . 286
7.3.5 Results Interpretation . . . 287
7.3.6 Pathway Analysis . . . 288
7.3.7 Utilities . . . 289

8 Analyzing Agilent Single Color Expression Data 295

8.1 Running the Agilent Single Color Workflow . . . 295
8.1.1 Analyzing Agilent Two Color data in Agilent Single Color Experiment Type 298
8.2 Data Processing for Agilent Single Color arrays . . . 298
8.3 Analysis: Biological Significance . . . 300
8.4 Data Import Wizard . . . 302
8.4.1 Experiment Setup . . . 307

8.4.2 Quality Control . . . 307
8.4.3 Analysis . . . 312
8.4.4 Class Prediction . . . 312
8.4.5 Results . . . 312
8.4.6 Utilities . . . 313

9 Analyzing Agilent Two Color Expression Data 315

9.1 Running the Agilent Two Color Workflow . . . 315
9.2 Data Processing for Agilent Two Color arrays . . . 318
9.3 Analysis: Biological Significance . . . 319
9.4 Data Import Wizard . . . 321
9.4.1 Experiment Setup . . . 327
9.4.2 Quality Control . . . 327
9.4.3 Analysis . . . 331
9.4.4 Class Prediction . . . 332
9.4.5 Results . . . 332
9.4.6 Utilities . . . 333
9.5 Custom Agilent Arrays . . . 333

10 Analyzing Affymetrix Expression Data 335

10.1 Running the Affymetrix Workflow . . . 335
10.2 Data Processing . . . 337
10.3 Analysis: Biological Significance . . . 338

10.4 Data Import Wizard . . . 340
10.4.1 Creating an Affymetrix Expression Experiment . . . 341
10.4.2 Experiment Setup . . . 351
10.4.3 Quality Control . . . 351
10.4.4 Analysis . . . 356
10.4.5 Class Prediction . . . 357
10.4.6 Results . . . 357
10.4.7 Utilities . . . 358
10.4.8 Affymetrix Technology creation using Custom CDF . . . 358

11 Affymetrix Summarization Algorithms **361**

11.0.1 Probe Summarization Algorithms . . . 361
11.0.2 Computing Absolute Calls . . . 365

12 Analyzing Illumina Data **367**

12.1 Running the Illumina Workflow: . . . 367
12.2 Data Processing for Illumina arrays . . . 369
12.3 Analysis: Biological Significance . . . 370
12.4 Data Import Wizard . . . 372
12.4.1 Experiment Setup . . . 376
12.4.2 Quality control . . . 376
12.4.3 Analysis . . . 380
12.4.4 Class Prediction . . . 381

12.4.5 Results . . . 381
12.4.6 Utilities . . . 382
12.4.7 Illumina Custom Technology creation . . . 382

13 Analyzing Generic Single Color Expression Data **383**

13.1 Creating Technology . . . 383
13.1.1 Project and Experiment Creation . . . 393
13.2 Data Processing for Generic Single Color Experiment . . . 396
13.3 Data Import Wizard . . . 397
13.3.1 Experiment Setup . . . 402
13.3.2 Quality Control . . . 402
13.3.3 Analysis . . . 407
13.3.4 Class Prediction . . . 408
13.3.5 Results . . . 408
13.3.6 Utilities . . . 409

14 Analyzing Generic Two Color Expression Data **411**

14.1 Creating Technology . . . 411
14.1.1 Creation of Custom Technology-Non gpr files . . . 411
14.1.2 GenePix Result Technology creation . . . 418
14.1.3 Project and Experiment Creation . . . 419
14.2 Data Import Wizard . . . 422
14.2.1 Data Processing for Generic Two Color Data . . . 426

14.2.2 Experiment Setup . . . 427
14.2.3 Quality Control . . . 427
14.2.4 Analysis . . . 430
14.2.5 Class Prediction . . . 431
14.2.6 Results . . . 431
14.2.7 Utilities . . . 432

15 Analyzing Agilent miRNA Data **433**

15.1 Running the Agilent miRNA Workflow . . . 433
15.1.1 Selecting technology for Agilent miRNA . . . 438
15.1.2 Sample validation . . . 441
15.1.3 Backward Compatibility with old miRNA experiments . . . 442
15.2 Data Processing for miRNA experiments . . . 442
15.3 Analysis: Biological Significance . . . 446
15.3.1 Summary Report (Step 1 of 8) . . . 446
15.3.2 Experiment Grouping (Step 2 of 8) . . . 447
15.3.3 Quality Control (QC) on samples (Step 3 of 8) . . . 449
15.3.4 Filter Probesets (Step 4 of 8) . . . 451
15.3.5 Significance Analysis (Step 5 of 8) . . . 453
15.3.6 Fold Change (Step 6 of 8) . . . 459
15.4 Data Import Wizard Workflow . . . 460
15.5 Workflow Options for a miRNA experiment . . . 467

15.5.1 Experiment Setup . . . 468
15.5.2 Quality Control . . . 468
15.5.3 Analysis . . . 474
15.5.4 Class Prediction . . . 475
15.5.5 Results Interpretation . . . 475
15.5.6 Pathway Analysis . . . 481
15.5.7 NLP Networks . . . 482
15.5.8 Utilities . . . 482

16 Analyzing Agilent Exon Data **483**

16.1 Running the Agilent Exon Workflow . . . 484
16.1.1 Overview . . . 484
16.1.2 Data Processing for Agilent Exon Experiments . . . 485
16.1.3 Creating an Agilent Exon Single Color Experiment . . . 488
16.1.4 Creating an Agilent Exon Two Color Experiment . . . 495
16.1.5 Experiment Setup . . . 496
16.1.6 Quality Control . . . 496
16.1.7 Gene Level Analysis . . . 503
16.1.8 Exon Level Analysis . . . 503
16.1.9 Class Prediction . . . 527
16.1.10 Results Interpretation . . . 528
16.1.11 Utilities . . . 528

16.2 Detailed Descriptions . . . 529

16.2.1 Agilent Exon Technology . . . 529

17 Analyzing Affymetrix Exon Expression Data **533**

17.1 Running the Affymetrix Exon Workflow . . . 533

17.2 Data Processing . . . 536

17.3 Analysis: Biological Significance . . . 537

17.4 Data Import Wizard . . . 539

17.4.1 Creating an Affymetrix ExonExpression Experiment . . . 539

17.4.2 Experiment setup . . . 548

17.4.3 Quality Control . . . 548

17.4.4 Analysis . . . 551

17.4.5 Class Prediction . . . 551

17.4.6 Results . . . 551

17.4.7 Utilities . . . 552

17.4.8 Algorithm Technical Details . . . 552

18 Analyzing Affymetrix Exon Splicing Data **555**

18.1 Running the Affymetrix Exon Splicing Workflow . . . 555

18.1.1 Creating an Affymetrix Exon Splicing Experiment . . . 556

18.1.2 Data Processing for Exon arrays . . . 565

18.1.3 Experiment setup . . . 566

18.1.4 Quality Control . . . 566

18.1.5 Analysis . . . 569
18.1.6 Exon Splicing Analysis . . . 569
18.1.7 Class Prediction . . . 585
18.1.8 Results . . . 586
18.1.9 Utilities . . . 586
18.1.10 Algorithm Technical Details . . . 587
18.2 Affymetrix Exon Splicing Technology . . . 587
18.2.1 Inputs for Technology Creation . . . 588
18.2.2 Technology creation . . . 588
18.3 Tutorial for Exon Splicing Analysis . . . 589

19 Analyzing Real Time PCR Data **593**

19.1 Running the Real Time PCR Workflow . . . 593
19.1.1 Technology Creation in RT-PCR experiments . . . 596
19.1.2 Data Processing . . . 596
19.1.3 Experiment Setup . . . 597
19.1.4 Quality Control . . . 597
19.1.5 Analysis . . . 599
19.1.6 Class Prediction . . . 600
19.1.7 Results . . . 600
19.1.8 Utilities . . . 601

20 Multi-Omic Analysis in GeneSpring **603**

20.1 Introduction to Multi-Omic Analysis in GeneSpring 604
20.1.1 Prerequisites for Multi-Omic Analysis 604
20.1.2 Supported experiment types 604
20.1.3 Supported analysis workflows 605
20.2 Creating a Multi-omic Analysis Experiment 606
20.2.1 How to create an MOA experiment using the toolbar icon 606
20.2.2 How to create an MOA experiment using the Project menu option 607
20.2.3 How to create an MOA experiment by performing Multi-Omic Pathway Analysis 607
20.3 Working in a Multi-omic Analysis Experiment 609
20.4 Multi-Omic Analysis Screen Reference 610
20.4.1 New Multi-Omic Analysis Experiment Window 610
20.4.2 MOA Experiment Navigator 611
20.4.3 MOA Experiment Toolbar 611

21 Importing Strand NGS experiments into GeneSpring **613**

21.1 Introduction to Importing Strand NGS Experiments for Multi-Omic Analysis in GeneSpring 614
21.1.1 Prerequisites for importing Strand NGS experiments 614
21.1.2 Supported Strand NGS experiment types for importing into GeneSpring 614
21.1.3 Supported MOA analysis workflows for imported Strand NGS experiments 615
21.2 Importing a Strand NGS Experiment and creating an MOA experiment 616
21.2.1 How to Export a NGS Experiment from Strand NGS 616

21.2.2 **How to Import a Strand NGS Experiment into GeneSpring** 617

21.2.3 **How to create an MOA experiment with the imported Strand NGS Experiment** 618

21.3 **Working in an imported Strand NGS Experiment** 619

21.3.1 **Workflow Browser** 619

21.3.2 **Toolbar** 620

21.3.3 **Right-click options on objects in an imported NGS experiment** . . 620

21.4 **Imported Strand NGS Experiment Screen Reference** 621

21.4.1 **Imported Strand NGS Experiment Navigator** 621

21.4.2 **Imported Strand NGS Experiment Toolbar** 621

21.4.3 **Right-click options on an imported Strand NGS experiment** 622

21.4.4 **Right-click options on a Region List** 623

21.4.5 **Right-click options on an Entity List** 624

21.4.6 **Right-click options on a Read Lists** 624

22 Loading Experiment from NCBI GEO **625**

22.1 Introduction 625

22.1.1 Load a GSE dataset 625

22.1.2 Experiment Parameters 627

22.2 Possible Error Messages 630

22.3 Experiment Parameters and Sample Attributes 631

22.3.1 Create Experiment Parameters from Sample Attributes 631

23 Exporting Expression Experiments in MAGE-TAB Format **633**

23.1 Introduction . . . 633
23.2 IDF file . . . 634
23.2.1 IDF Model file . . . 634
23.2.2 IDF Preferences File . . . 634
23.2.3 IDF Mappings . . . 634
23.2.4 IDF Defaults . . . 635
23.2.5 Ontology mappings . . . 636
23.3 SDRF file . . . 636
23.3.1 SDRF Model File . . . 636
23.3.2 SDRF Preferences . . . 637
23.3.3 SDRF Mappings . . . 637
23.3.4 SDRF Rows . . . 638
23.3.5 Array Design REF . . . 638
23.3.6 SDRF Defaults . . . 638
23.3.7 Ontology mappings . . . 640
23.4 Array Design Format . . . 640
23.5 Data files . . . 640
23.5.1 Raw Files . . . 640
23.5.2 Raw Array Data Matrix File . . . 641
23.5.3 Derived Array Data Matrix File . . . 641
23.6 Validations . . . 641
23.7 Points to note . . . 641

24 Normalization, Statistical Hypothesis Testing, and Differential Expression Analysis 643

24.1 Threshold . . . 643

24.2 Normalization Algorithms . . . 644

24.2.1 Percentile Shift . . . 644

24.2.2 Scale . . . 644

24.2.3 Quantile . . . 645

24.2.4 Normalize to Control Genes . . . 645

24.2.5 Normalize to External Value . . . 646

24.2.6 Lowess Normalization . . . 646

24.3 Details of Statistical Tests in **GeneSpring** . . . 647

24.3.1 The *t*-Test against Zero for a Single Group . . . 648

24.3.2 The Unpaired *t*-Test for Two Groups . . . 648

24.3.3 The Paired *t*-Test for Two Groups . . . 649

24.3.4 The Moderated *t*-Test for Two Groups . . . 649

24.3.5 The Unpaired Unequal Variance *t*-Test (Welch *t*-Test) for Two Groups . . . 650

24.3.6 The Mann-Whitney Test against Zero for a Single Group . . . 651

24.3.7 The Unpaired Mann-Whitney Test for Two Groups . . . 652

24.3.8 The Paired Mann-Whitney Test for Two Groups . . . 652

24.3.9 The One-Way ANOVA . . . 653

24.3.10 Post Hoc Testing of ANOVA Results . . . 654

24.3.11 The Unequal Variance (Welch) ANOVA . . . 655

24.3.12 The Kruskal-Wallis Test . . . 655
24.3.13 The Repeated Measures ANOVA . . . 656
24.3.14 The Repeated Measures Friedman Test . . . 656
24.3.15 The N-way ANOVA . . . 657
24.4 Obtaining *p*-Values . . . 658
24.4.1 *p*-values via Permutative Tests . . . 659
24.5 Adjusting for Multiple Comparisons - Corrected *p*-values . . . 660
24.5.1 Bonferroni FWER . . . 661
24.5.2 Bonferroni Holm FWER (Holm step-down) . . . 661
24.5.3 Westfall Young Permutative . . . 662
24.5.4 Benjamini-Yekutieli method . . . 662
24.5.5 Benjamini Hochberg FDR . . . 663
24.5.6 The Storey method . . . 663
24.5.7 Recommendations . . . 667
24.5.8 FAQ . . . 667

25 Clustering Analysis **669**

25.1 **Getting Started** . . . 670
25.1.1 **Dendrogram visualization** . . . 670
25.1.2 **Default options for Clustering Algorithms** . . . 670
25.2 **Clustering Analysis in GeneSpring** . . . 671
25.2.1 **How to Perform Hierarchical Clustering Analysis** . . . 671

25.2.2 How to Perform K-Means Clustering Analysis 673
25.2.3 How to Perform Self-Organizing Map Clustering Analysis 675
25.2.4 How to Perform Clustering Analysis on Region Lists 678
25.2.5 How to Explore a Cluster Tree Result in the Dendrogram View . 679
25.2.6 How to Visualize Sample Metadata in the Dendrogram View . . . 680
25.2.7 Working with a Cluster Tree in the Dendrogram View 683
25.2.8 Clustering result visualizations in GeneSpring 696
25.3 Clustering Analysis Screen Reference 699
25.3.1 Clustering Window Common Features 699
25.3.2 Clustering Window - Hierarchical Clustering Parameters 701
25.3.3 Clustering Window - K-Means Clustering Parameters 703
25.3.4 Clustering Window - SOM Clustering Parameters 704
25.3.5 Result Preview Pane 706
25.3.6 The Dendrogram View 707
25.3.7 Dialog Boxes 712
25.4 Clustering Analysis - Statistical Algorithms 742
25.4.1 Distance Metric 742
25.4.2 Linkage Rule 744
25.4.3 Missing Values Handling 745

26 Correlation Analysis **747**

26.1 Correlation Analysis in GeneSpring 749

26.1.1 **How to Perform Correlation Analysis on Entities** 749

26.1.2 **How to Perform Correlation Analysis on Samples** 750

26.1.3 **How to Perform Correlation Analysis in a MOA experiment** . . . 752

26.1.4 **How to Explore the Correlation Heatmap View** 756

26.1.5 **Working with a Correlation Heatmap View** 757

26.1.6 **Correlation analysis result visualizations in GeneSpring** 769

26.2 **Correlation Analysis Screen Reference** 772

26.2.1 **Correlation Analysis Window Features** 772

26.2.2 **The Correlation Heatmap View** 774

26.2.3 **Dialog Boxes** 780

26.3 **Correlation Analysis - Statistical Algorithms** 798

26.3.1 **Correlation computation in GeneSpring** 798

26.3.2 **Similarity Metric** 799

27 Class Prediction: Learning and Predicting Outcomes **801**

27.1 General Principles of Building a Prediction Model 801

27.2 Prediction Pipeline 802

27.2.1 Validate 803

27.2.2 Prediction Model 803

27.3 Running Class Prediction in **GeneSpring** 804

27.3.1 Build Prediction Model 804

27.3.2 Run Prediction 810

27.4 Partial Least Squares Discrimination . . . 812
27.4.1 PLSD Model and Parameters . . . 812
27.5 Support Vector Machines . . . 813
27.5.1 SVM ModelParameters . . . 814
27.6 Naive Bayesian . . . 816
27.6.1 Naive Bayesian Model Parameters . . . 817
27.6.2 Naive Bayesian Model View . . . 817
27.7 Decision Trees . . . 818
27.7.1 Decision Tree Model Parameters . . . 819
27.7.2 Decision Tree Model . . . 820
27.8 Neural Network . . . 821
27.8.1 Neural Network Model Parameters . . . 821
27.8.2 Neural Network Model . . . 822
27.9 Viewing Classification Results . . . 823
27.9.1 Confusion Matrix . . . 824
27.9.2 Classification Report . . . 824
27.9.3 Lorenz Curve . . . 825

28 Gene Ontology Analysis **827**

28.1 Working with Gene Ontology Terms . . . 827
28.2 Introduction to GO Analysis in **GeneSpring** . . . 828
28.3 GO Analysis . . . 829

28.4 GO Analysis Views . . . 831
28.4.1 GO Spreadsheet . . . 831
28.4.2 The GO Tree View . . . 832
28.4.3 The Pie Chart . . . 834
28.5 GO Enrichment Score Computation . . . 837

29 Gene Set Enrichment Analysis **839**

29.1 Introduction to GSEA . . . 839
29.2 Gene sets . . . 839
29.3 Performing GSEA in **GeneSpring** . . . 840
29.4 GSEA Computation . . . 846
29.5 Import BROAD GSEA Genesets . . . 847

30 Gene Set Analysis **849**

30.1 Introduction to GSA . . . 849
30.2 Gene sets . . . 849
30.3 Performing GSA in **GeneSpring** . . . 850
30.4 GSA Computation . . . 856

31 Pathway Analysis **859**

31.1 **Getting Started** . . . 862
31.1.1 **Pathway Module Licensing** . . . 862
31.1.2 **KEGG Pathways Licensing** . . . 862
31.1.3 **Importing Pathways into GeneSpring** . . . 863

31.1.4 **Creating and Editing Pathways** 878

31.1.5 **Exporting and Transferring Pathways** 885

31.1.6 **Supporting Databases for Pathway Analysis** 887

31.2 **Pathway Analysis in GeneSpring** 896

31.2.1 **What is Multi-Omic Pathway Analysis (MOA)?** 896

31.2.2 **What is Single Experiment Analysis (SEA)?** 897

31.2.3 **Prerequisites for Multi-Omic Pathway Analysis** 897

31.2.4 **Starting and Navigating the SEA and MOA Workflows** 898

31.3 **Single Experiment and Multi-Omic Analysis** 900

31.3.1 **Step 1: Specifying Experiments, Organisms, and Pathway Sources** 900

31.3.2 **Step 2: Selecting Entity Lists and Interpretations** 907

31.3.3 **Step 3: Multi-Omic Analysis Results** 909

31.3.4 **Step 4: Saving Pathway List** 911

31.3.5 **Working with Pathway Lists** 912

31.3.6 **Working in an MOA Experiment** 931

31.4 **Multi-Omic Analysis Workflow Reference** 932

31.4.1 **The Wizard-driven Multi-Omic Analysis (MOA) Workflow** 932

31.4.2 **Multi-Omic Analysis (Step 1 of 4)** 932

31.4.3 **Multi-Omic Analysis (Step 2 of 4)** 934

31.4.4 **Multi-Omic Analysis (Step 3 of 4)** 936

31.4.5 **Multi-Omic Analysis (Step 4 of 4)** 938

31.4.6 **The Pathway View** 940

31.4.7 **Dialog Boxes** . . . 966

31.4.8 **Right-click Options** . . . 1006

31.5 **Mapping and Statistical Algorithms** . . . 1011

31.5.1 **HomoloGene Translation for Gene/Protein Identifiers** . . . 1011

31.5.2 **BridgeDb - ID Mapping** . . . 1012

31.5.3 **Supported Data Sources** . . . 1012

31.5.4 **p-value Computation** . . . 1013

31.5.5 **Heatstrip Error Bar Computation** . . . 1014

31.6 **Analyzing Pathways Using External Tools Accessible from GeneSpring** 1015

31.6.1 **Ingenuity Pathways Analysis (IPA) Connector** . . . 1015

31.6.2 **MetaCore™ Connector** . . . 1031

31.6.3 **Cytoscape Connector** . . . 1035

32 NLP Networks **1037**

32.1 **Natural Language Processing (NLP) in GeneSpring** . . . 1038

32.1.1 **Entity Recognition** . . . 1038

32.1.2 **Syntax Analysis** . . . 1038

32.1.3 **Semantic Analysis** . . . 1039

32.1.4 **Semantic Inference** . . . 1039

32.2 **Interaction Databases** . . . 1040

32.2.1 **Interaction Database Organization Overview** . . . 1040

32.2.2 **Database Entities** . . . 1041

32.2.3 **Relations** 1045

32.2.4 **Database Statistics** 1050

32.3 **NLP Network Discovery** 1051

32.3.1 **The NLP Network Discovery Workflow** 1051

32.4 **Extract Relations via NLP** 1061

32.4.1 **NLP Settings** 1065

32.5 **MeSH Network Builder** 1067

32.5.1 **Launching MeSH Pathway Builder** 1067

32.6 **Creating a Pathway Experiment** 1072

32.6.1 **Launching a Pathway Experiment** 1072

32.6.2 **Lassoing** 1078

32.6.3 **Simple Analysis** 1078

32.6.4 **Advanced Analysis** 1078

33 Analyzing CGH Data **1079**

33.1 **Creating a CGH Experiment** 1080

33.1.1 **Supported data files** 1080

33.1.2 **Technology Creation in CGH Experiments** 1080

33.1.3 **Data Processing** 1081

33.1.4 **Experiment Creation** 1083

33.1.5 **Experiment Setup** 1085

33.1.6 **Quality Control** 1086

33.2 **Analysis** . . . 1090

33.2.1 **Translate Regions to Genes** . . . 1090

33.2.2 **Translate Regions to Probes** . . . 1091

33.2.3 **Find Common Aberrations** . . . 1092

33.2.4 **Clustering** . . . 1094

33.2.5 **Correlation** . . . 1096

33.3 **Results Interpretations** . . . 1097

33.3.1 **GO Analysis** . . . 1097

33.3.2 **Genome Browser Visualization** . . . 1097

33.4 **Pathway Analysis** . . . 1099

33.4.1 **Quilt plots in Pathway View of CGH Experiment** . . . 1099

33.5 **Utilities** . . . 1100

33.5.1 **Filter on Entity List** . . . 1100

33.5.2 **Region List Operations** . . . 1101

33.5.3 **Import Region List** . . . 1102

33.5.4 **Annotate Region List** . . . 1103

33.5.5 **Compare Region Lists** . . . 1104

33.5.6 **Import Probe-level Reports** . . . 1105

34 Copy Number Analysis **1107**

34.1 Introduction . . . 1107

34.1.1 Terminology in Copy Number analysis . . . 1108

34.2 Technologies supported by **GeneSpring** 1108
34.2.1 Experiment Creation 1108
34.2.2 Reference 1112
34.2.3 Copy Number Analysis 1114
34.2.4 Special mention for Affymetrix Mapping 100k Array 1117
34.3 Workflow description for Affymetrix files 1117
34.3.1 Create Technology 1117
34.3.2 Creating a Copy Number experiment 1118
34.3.3 Experiment Setup 1121
34.3.4 Quality Control 1122
34.3.5 Copy Number Analysis 1128
34.3.6 Common Genomic Variant Regions 1131
34.3.7 Filters 1135
34.3.8 Views 1142
34.3.9 Results Analysis 1144
34.3.10 Utilities 1146
34.4 Copy Number analysis of Illumina 1146
34.4.1 Obtaining Data from Illumina 1146
34.4.2 Handling Missing Values 1148
34.4.3 Workflow description for Illumina Outputs 1148
34.5 Create Custom Reference 1149
34.6 Useful information 1150

34.6.1 Using disc cache . . . 1150
34.6.2 Entity Lists and Translation rules in Copy Number . . . 1150
34.6.3 Configuration options for Copy Number analysis . . . 1150
34.6.4 Performance Statistics for Copy Number Analysis . . . 1152
34.7 Copy Number Algorithms . . . 1152
34.7.1 BRLMM . . . 1152
34.7.2 Hidden Markov Model (HMM) . . . 1154
34.7.3 Canary algorithm . . . 1157
34.7.4 Birdseed algorithm . . . 1157
34.7.5 CBS for segmenting genome with respect to Copy Number . . . 1160
34.7.6 Post Processing to assign Copy Numbers to segments created by CBS . . . 1160
34.7.7 Fawkes algorithm . . . 1161
34.8 Tutorials for Copy Number Analysis . . . 1162

35 Association Analysis **1163**

35.1 Introduction . . . 1163
35.2 Technology . . . 1164
35.3 Experiment Creation . . . 1164
35.3.1 Illumina Association Analysis Experiment . . . 1165
35.3.2 Affymetrix Association Analysis Experiment . . . 1167
35.4 Quality Control . . . 1167
35.4.1 Filter Samples by Missing Values . . . 1168

35.4.2 Birdseed Report 1168
35.4.3 EIGENSTRAT Filter on Samples 1169
35.5 Filters 1172
35.5.1 Filter SNPs by Missing Value 1173
35.5.2 Identify SNPs with Differential Missingness 1173
35.5.3 Filter SNPs by HWE p-value 1173
35.5.4 Filter SNPs by MAF 1174
35.6 Analysis 1175
35.6.1 EIGENSTRAT Correction on Samples 1175
35.6.2 Statistical Analysis 1178
35.6.3 SNP Tagging 1191
35.6.4 SNP Regression 1193
35.6.5 Haplotype Trend Regression (HTR) 1195
35.6.6 LD Analysis 1196
35.7 Views 1198
35.7.1 Genome Browser 1198
35.8 Results Interpretations 1200
35.8.1 Identify Overlapping Genes 1200
35.9 Utilities 1200
35.9.1 Using disc cache 1200

36 GeneSpring WorkGroupPlus **1203**

36.1 **GeneSpring WorkGroupPlus** setup . . . 1203
36.1.1 **GeneSpring WorkGroupPlus** Administrator . . . 1203
36.1.2 **GeneSpring** Client License Manager . . . 1204
36.1.3 Login . . . 1206
36.2 **GeneSpring** Manager . . . 1206
36.2.1 Users and Groups . . . 1207
36.2.2 **GeneSpring Manager** User Interface . . . 1207
36.3 Operations on **GeneSpring** Objects . . . 1209
36.3.1 Object ownership . . . 1209
36.3.2 Object permissions . . . 1209
36.3.3 Conflicts with permissions . . . 1212
36.3.4 Propagating permissions . . . 1212
36.3.5 Inheriting Permissions . . . 1212
36.4 Run in background . . . 1213
36.4.1 Task Manager . . . 1213
36.4.2 Operations with an option to Run in background . . . 1214
36.4.3 Interpreting Task Logs . . . 1215

37 Writing Scripts in GeneSpring **1217**

37.1 The Script Editor . . . 1217
37.2 Hierarchy of data organization in **GeneSpring** . . . 1218
37.2.1 Accessing Projects, Experiments and their Constituent Elements . . . 1218

37.2.2 Accessing the Experiment Dataset 1220
37.2.3 Some More Useful Functions 1222
37.2.4 Some Common Marks 1224
37.2.5 Creating UI Components 1225
37.2.6 Example Scripts 1228
37.3 The R Editor 1235
37.3.1 Setting up R Editor in **GeneSpring** 1235
37.3.2 Commands related to R input from **GeneSpring** 1235
37.3.3 Commands related to R output to **GeneSpring** 1236
37.3.4 Debugging a Script 1237
37.3.5 Example R scripts 1238
37.4 GeneSpring-R Integration 1241
37.4.1 Old paradigm 1241
37.4.2 GeneSpring-R Integration : New paradigm 1241
37.4.3 Package Installation Instructions 1241
37.4.4 R Application Programming Interface 1242
37.4.5 Examples 1244

38 Table of Key Bindings and Mouse Clicks **1245**

38.1 Mouse Clicks and their actions 1245
38.1.1 Global Mouse Clicks and their actions 1245
38.1.2 Some View Specific Mouse Clicks and their Actions 1246

38.1.3 Mouse Click Mappings for Mac 1246

38.2 Key Bindings 1246

38.2.1 Global Key Bindings 1246

Bibliography **1249**

List of Figures

1.1 Microsoft Visual C++ 2005 RunTime Libraries Installation 3

1.2 Activation Failure 5

1.3 Database error Dialog 6

1.4 Activation Failure 8

1.5 Activation Failure 11

1.6 The License Activation Dialog 12

1.7 The License Manager Dialog 13

1.8 License Expiry Message 14

1.9 Confirm Surrender Dialog 15

1.10 Manual Surrender Dialog 15

1.11 Change License Dialog 16

1.12 License Re-activation Dialog 16

2.1 **GeneSpring** Layout 22

2.2 The Workflow Browser 24

2.3 The Legend 25

2.4 The Status Bar . . . 25

2.5 A region list of an imported NGS experiment . . . 34

2.6 Options for Search All objects in **GeneSpring** . . . 56

2.7 MAGE-TAB Exporter Window . . . 64

2.8 Tools - Options . . . 67

2.9 Confirmation Dialog . . . 71

2.10 Product Update Dialog . . . 72

3.1 Create Technology . . . 76

3.2 Technology Creation . . . 79

3.3 Technology Update . . . 79

3.4 Data Library Updates Dialog . . . 82

3.5 Automatic Download Confirmation Dialog . . . 82

3.6 Update Technology Annotations . . . 83

3.7 Input Parameters . . . 84

3.8 Format data file . . . 85

3.9 Choose Annotation Columns . . . 86

3.10 Annotations Manager . . . 87

3.11 Create New Organism . . . 89

3.12 Create New Build . . . 90

3.13 Add Annotations to a build . . . 91

3.14 Add Gene and Transcript annotations . . . 92

3.15 Input Parameters 98

3.16 Translation Table 99

3.17 Save Entity List 100

3.18 Create Biological Genome 102

4.1 Experiment Exporter 106

4.2 Confirmation Window 106

4.3 Migrate GS7 Data 107

4.4 Partially Migrated Genomes 108

5.1 Menu accessible by Right-Click on the plot views 117

5.2 Export submenus 118

5.3 Export Image Dialog 119

5.4 Tools - Options Dialog for Export as Image 120

5.5 Error Dialog on Image Export 121

5.6 Menu accessible by Right-Click on the table views 122

5.7 Filter Dialog for numerical values 123

5.8 Filter Dialog for categorical values 123

5.9 Spreadsheet 125

5.10 Spreadsheet Properties Dialog 127

5.11 MvA Plot 130

5.12 Scatter Plot 131

5.13 Scatter Plot Properties 133

5.14 Viewing Profiles and Error Bars using Scatter Plot 135

5.15 Scatter plot with Fold Change lines 136

5.16 Profile Plot 138

5.17 Interpretation for which a Time Series Plot will be launched (Parameters) 139

5.18 Interpretation for which a Time Series Plot will be launched (Conditions) 140

5.19 Time Series Plot 141

5.20 Time Series Plot 143

5.21 Heatmap 146

5.22 Heatmap Export submenu 147

5.23 Heatmap Toolbar 148

5.24 Heatmap Properties 149

5.25 Histogram 151

5.26 Histogram Viewing Options 152

5.27 Histogram Properties 153

5.28 Bar Chart 156

5.29 Matrix Plot 160

5.30 Matrix Plot Properties 161

5.31 Summary Statistics View 164

5.32 Summary Statistics Properties 166

5.33 Box Whisker Plot 169

5.34 Box Whisker Properties 171

5.35 The Venn Diagram 175

5.36 Create New Entity List from Venn Diagram 177

5.37 The Venn Diagram Properties 179

5.38 Violin Plot 180

5.39 Violin Plot Properties 182

5.40 LD Plot Toolbar 186

5.41 Plot List Associated Values 189

5.42 Plot List Associated Values-Scatter plot 189

5.43 Plot List Associated Values-Profile plot 190

5.44 Plot List Associated Values-Histogram 191

5.45 Configuration options for Views 193

6.1 Genome Browser- Select Build 197

6.2 Update Genome Browser Annotations 197

6.3 Genome Browser- on launch 198

6.4 Genome Browser Organization 199

6.5 Selecting a track 200

6.6 Genome Browser- Select Data 201

6.7 View expression data 202

6.8 View expression data- Normalized and Raw signal values 202

6.9 Copy Number Analysis- Select Data 203

6.10 Genome Browser Organization 206

6.11 Chromosome View-Selecting the region 207

6.12 Spreadsheet View 208

6.13 Search Result 209

6.14 Edit Track Properties Window 210

6.15 Customize Features Layout 213

6.16 Novel splice junction discovered by **Strand NGS** in the ISYNA1 gene 215

6.17 An elastic view of the ISYNA1 gene in which exons 4,5 alone are shown in great detail. The elastic band located at the top of the browser contains the pin controls that govern the resolution at which different genomic regions are displayed in the viewing area 216

6.18 Gene Fusion discovered by **Strand NGS** between DENND4B and CRTC2 genes . 217

6.19 Panel (a) shows an EGB view in which two pin-pairs have been added. Panel (b) shows the view after one pin has been moved to a new screen location. Panel (c) shows the view after the second pin has also been moved 218

7.1 Welcome Screen 220

7.2 Create New Project 221

7.3 Experiment Selection 221

7.4 Summary Report 224

7.5 Experiment Grouping 226

7.6 Edit or Delete Parameters 227

7.7 Filter Probesets-Single Parameter 228

7.8 Re-run Filter 228

7.9 Significance Analysis 232

7.10 Fold Change 233

7.11 Experiment Grouping 237

7.12 Experiment Grouping-selecting parameter type 238

7.13 Edit or Delete of Parameters 240

7.14 Create Interpretation (Step 1 of 4) 241

7.15 Create Interpretation (Step 2 of 4) 242

7.16 Create Interpretation (Step 3 of 4) 243

7.17 Create Interpretation (Step 4 of 4) 244

7.18 Gene-level Experiment - Entity List Chooser 246

7.19 Gene-level Experiment Creation 246

7.20 Filter probesets by expression (Step 1 of 4) 248

7.21 Filter probesets by expression (Step 2 of 4) 249

7.22 Filter probesets by expression (Step 3 of 4) 250

7.23 Filter probesets by expression (Step 4 of 4) 251

7.24 Statistical Analysis (Step 1 of 9) - Input Parameters 255

7.25 Statistical Analysis (Step 2 of 9) - Select Test 256

7.26 Statistical Analysis (Step 3 of 9) - Select Posthoc test 256

7.27 Statistical Analysis (Step 4 of 9) - Select Posthoc test 257

7.28 Statistical Analysis (Step 5 of 9) - p-value Computation 258

7.29 Statistical Analysis (Step 6 of 9) - p-value Computation 259

7.30 Statistical Analysis (Step 7 of 9) - Pairing Options 260

7.31 Statistical Analysis (Step 8 of 9) - Results 261

7.32 Statistical Analysis (Step 8 of 9) - Results for 2-way ANOVA 262

7.33 Statistical Analysis (Step 9 of 9) - Save Entity List 263

7.34 Filter on Volcano Plot (Step 6 of 7) - Results 270

7.35 Fold Change (Step 1 of 4) - Input Parameters 271

7.36 Fold Change (Step 2 of 4) - Pairing Options 272

7.37 Fold Change (Step 3 of 4) - Fold Change Results 273

7.38 Configuration Menu for Fold Change Colour Settings 274

7.39 Fold Change (Step 4 of 4) - Object Details 275

7.40 Find Similar Entities (Step 1 of 3) - Input Parameters 277

7.41 Find Similar Entities (Step 2 of 3) - Output View of Find Similar Entities 278

7.42 Find Similar Entities (Step 3 of 3) - Save Entity List 279

7.43 Filter on Parameters (Step 1 of 3) - Input Parameters 280

7.44 Filter on Parameters (Step 2 of 3) - Output View of Filter on Parameters 281

7.45 Filter on Parameters (Step 3 of 3) - Save Entity List 282

7.46 PCA (Step 1 of 4) - Entity List and Interpretation 283

7.47 PCA (Step 2 of 4) - Input Parameters 284

7.48 PCA (Step 3 of 4) - Output Views 286

7.49 Filter on Entity List (Step 1 of 4) - Input Parameters 291

7.50 Filter on Entity List (Step 2 of 4) - Add/Remove Filter Conditions 291

7.51 Filter on Entity List (Step 3 of 4) - Filter Results 292

7.52 Filter on Entity List (Step 4 of 4) - Save Entity List 293

8.1 Experiment Description 296

8.2 Load Data 296

8.3 Choose Samples . . . 297
8.4 Reordering Samples . . . 297
8.5 Confirmation Dialog Box . . . 298
8.6 Quality Control on Samples . . . 301
8.7 Load Data . . . 303
8.8 Advanced flag Import . . . 304
8.9 Preprocess Options . . . 305
8.10 Normalize to control genes . . . 306
8.11 Baseline Transformation Options . . . 307
8.12 Quality Control Dialog . . . 308
8.13 Entity List and Interpretation . . . 310
8.14 Input Parameters . . . 310
8.15 Output Views of Filter by Flags . . . 311
8.16 Save Entity List . . . 311

9.1 Experiment Description . . . 316
9.2 Load Data . . . 316
9.3 Choose Samples . . . 317
9.4 Reordering Samples . . . 317
9.5 Dye Swap . . . 318
9.6 Quality Control on Samples . . . 320
9.7 Load Data . . . 322

9.8 Samples Validation 323
9.9 Choose Dye-Swaps 324
9.10 Advanced flag Import 325
9.11 Preprocess Options 326
9.12 Quality Control 328
9.13 Entity list and Interpretation 329
9.14 Input Parameters 330
9.15 Output Views of Filter by Flags 330
9.16 Save Entity List 331

10.1 Experiment Description 336
10.2 Load Data 336
10.3 Choose Samples 337
10.4 Reordering Samples 337
10.5 Quality Control on Samples 339
10.6 Load Data 342
10.7 Choose Technology and Template 343
10.8 Select Row Scope for Import 344
10.9 Choose Identifier and Signal Column 345
10.10Single Color Multiple Samples in one File Selection 346
10.11Select ARR files 347
10.12Summarization Algorithm 348

10.13Normalization and Baseline Transformation 349

10.14Normalize to control genes 350

10.15Baseline Transformation 351

10.16Quality Control Dialog 352

10.17Entity List and Interpretation 354

10.18Input Parameters 354

10.19Output Views of Filter by Flags 355

10.20Save Entity List 356

10.21Confirmation Dialog Box 359

10.22Choose Input Files 359

12.1 Experiment Description 368

12.2 Load Data 368

12.3 Choose Samples 369

12.4 Quality Control on Samples 371

12.5 Load Data 373

12.6 Identify Calls Range 373

12.7 Preprocess Options 374

12.8 Choose Entities 375

12.9 Preprocess Baseline Options 376

12.10Quality Control Dialog 377

12.11Entity List and Interpretation 378

12.12Input Parameters 379

12.13Output Views of Filter by Flags 379

12.14Save Entity List 380

13.1 Technology Name 384

13.2 Format data file 385

13.3 Select Row Scope for Import 386

13.4 Single Color one sample in one file selections 387

13.5 Single Color-Multiple Samples Per File-Keyword Selection 388

13.6 Single Color-Multiple Samples Per File-Custom Selection 389

13.7 Annotation Column Options 392

13.8 Annotation Mark Colors 393

13.9 Welcome Screen 394

13.10Create New Project 394

13.11Experiment Selection 395

13.12Experiment Description 396

13.13Load Data 398

13.14Preprocess Options 400

13.15Choose Entities 401

13.16Preprocess Baseline Options 402

13.17Quality Control Dialog 403

13.18Entity List and Interpretation 405

13.19Input Parameters 405

13.20Output Views of Filter by Flags 406

13.21Save Entity List 407

14.1 Technology Name 412

14.2 Format data file 413

14.3 Select Row Scope for Import 414

14.4 Two Color Selections 415

14.5 Annotation Column Options 417

14.6 Annotation Mark Colors 418

14.7 Technology Creation 419

14.8 Welcome Screen 420

14.9 Create New Project 420

14.10Experiment Selection 421

14.11Experiment Description 421

14.12Load Data 423

14.13Choose Dye-Swaps 424

14.14Preprocess Options 424

14.15Preprocess Baseline Options 425

14.16Quality Control 427

14.17Entity List and Interpretation 428

14.18Input Parameters 429

14.19Output Views of Filter by Flags 429

14.20Save Entity List 430

15.1 Welcome Screen 434

15.2 Create New Project 434

15.3 Experiment Selection 435

15.4 Experiment Selection 437

15.5 Load Data 438

15.6 Select Technology in miRNA Experiment 439

15.7 Creating Technology: From Agilent eArray 440

15.8 Creating miRNA Technology from GEML File 440

15.9 Sample processing based on miRBase version - Autoselected 443

15.10Sample processing based on miRBase version - Select from list 444

15.11Summary Report 446

15.12Experiment Grouping 448

15.13Add/Edit Parameters 449

15.14Quality Control on Samples 450

15.15Filter Probes-Single Parameter 452

15.16Filter Probes-Two Parameters 453

15.17Significance Analysis T-Test 457

15.18Significance Analysis ANOVA 458

15.19Fold Change 460

15.20Load Data 462

15.21Normalization Options 464

15.22Choose entities 465

15.23Baseline Transformation 466

15.24Selection of Controls 467

15.25Quality Control 469

15.26Entity list and Interpretation 471

15.27Input Parameters 472

15.28Output Views of Filter by Flags 473

15.29Save Entity List 474

15.30Find Targeted Genes: Select Inputs 477

15.31Find Targeted Genes: Object Details 478

15.32Step 3 of Update Annotations 481

16.1 New Experiment for Agilent Exon 485

16.2 Sequence of events while creating Agilent Exon Experiment 487

16.3 Load Data 489

16.4 Error message if technology is not found 490

16.5 Chose Corresponding Build for the Technology 490

16.6 Advanced Flag Import 491

16.7 Normalization Options 492

16.8 Preprocess Baseline Options 494

16.9 Choose Dye Swaps 495
16.10Quality Control on Samples for Agilent Exon Experiment 497
16.11Filter by Expression - Input Parameters 499
16.12Filter by Flags - Input Parameters 500
16.13Filter by Error - Input Parameters 501
16.14Filter - Output Views 502
16.15Filter On Splicing Index - Pairing Options 506
16.16Filter On Splicing Index - Filter On Splicing Index 507
16.17Filter On Splicing Index - Object Details 508
16.18Statistical Splicing Analysis - Select Test 510
16.19Statistical Splicing Analysis - p-value Computation 511
16.20Statistical Splicing Analysis - p-value Computation 512
16.21Statistical Splicing Analysis - Filter Probes 513
16.22Statistical Splicing Analysis - Results 515
16.23Statistical Splicing Analysis - Object Details 516
16.24Results of Statistical Splicing Analysis 517
16.25Splicing Visualization - Input Parameters 518
16.26Splicing Visualization - Splicing Index 519
16.27Splicing Visualization for Agilent Exon 520
16.28Splicing Visualization - Save Entity List 522
16.29Export Probe Data 524
16.30Export Probe Level Data - Choose Entity list and Interpretation 526

16.31Export p-values 527

17.1 Experiment Description 534

17.2 Load Data 535

17.3 Choose Samples 535

17.4 Reordering Samples 536

17.5 Quality Control on Samples 538

17.6 Load Data 540

17.7 Select ARR files 541

17.8 Summarization Algorithm 542

17.9 Normalization 543

17.10Search Entities 545

17.11Output Views 546

17.12Choose Entities 547

17.13Normalization and Baseline Transformation 548

17.14Quality Control Dialog 549

18.1 Select Technology 557

18.2 Load Data 558

18.3 Error Message 558

18.4 Select ARR files 559

18.5 Pairing of CHP files 560

18.6 Summarization Algorithm 561

18.7 Normalization . . . 562
18.8 Normalize to control genes . . . 564
18.9 Normalization and Baseline Transformation . . . 565
18.10Quality Control Dialog . . . 567
18.11Input Data . . . 570
18.12Filtering Options . . . 570
18.13Output Views . . . 571
18.14Save Entity List . . . 572
18.15Input Data . . . 573
18.16Filtering of Probesets . . . 574
18.17Multiple Testing Correction . . . 575
18.18Results . . . 576
18.19Save Entity List . . . 577
18.20Input Data . . . 578
18.21Pairing Options . . . 579
18.22Results . . . 580
18.23Save Entity List . . . 581
18.24Input Data . . . 582
18.25Visualization . . . 583
18.26Visualization . . . 584
18.27Save Entity List . . . 585
18.28Gene Normalized Variance Plot . . . 591

18.29Gene Normalized Profile Plot 591

19.1 Experiment Creation 594

19.2 Baseline Transformation Options 595

19.3 Quality Control 598

19.4 Input Parameters 601

19.5 Choose Annotation Columns 602

22.1 GEO Identifier Entry Dialog 626

22.2 Create New Experiment Dialog 627

22.3 Experiment Grouping Information is automatically copied over 628

22.4 Duplicate Experiment Parameters 628

22.5 Duplicate Parameters 629

22.6 Final Experiment Grouping 630

22.7 Sample attributes that can be chosen as Experiment Parameters 632

24.1 Anova result showing 'Excluded Entities' because of missing values 658

24.2 Histogram of p-values 664

24.3 Asymptote Plot 664

25.1 Dendrogram View of a combined tree 707

25.2 Legend of a combined cluster tree 709

25.3 Legend of a combined cluster tree with sample parameter plots 710

25.4 Chose Entity List dialog box 712

25.5 Choose Interpretation dialog box 713

25.6 Classification Inspector dialog box 714

25.7 Combined Tree Inspector dialog box 715

25.8 Condition Tree Inspector dialog box 716

25.9 Entity Tree Inspector dialog box 717

25.10Choose Parameters dialog box 718

25.11Add/Remove/Re-order Parameters dialog box 719

25.12Export options dialog box 720

25.13Export as Report options dialog box 721

25.14Export as Image options dialog box 722

25.15Export as Image with sample parameter plots options dialog box 723

25.16GO analysis on entity cluster tree dialog box 724

25.17Inspector dialog box 725

25.18Properties dialog box - Colors tab 727

25.19Properties dialog box - Colors tab - Create New Colors Theme 729

25.20Properties dialog box - Columns tab 730

25.21Properties dialog box - Labels tab 731

25.22Properties dialog box - Rulers tab 733

25.23Properties dialog box - Description tab 734

25.24Sample parameters Heat Map Properties dialog box 735

25.25Sample parameters Scatter Plot Properties dialog box 736

25.26Sample parameters Profile Plot Properties dialog box 737

25.27Sample parameters Bar Chart Properties dialog box 738

25.28Sample parameters Label Plot Properties dialog box 739

25.29Select Columns dialog box 740

26.1 Correlation Analysis in a Multi-Omic Analysis experiment - Step 1 of 2 753

26.2 Correlation Analysis in a Multi-Omic Analysis experiment - Step 2 of 2 754

26.3 MOA Correlation View 774

26.4 Clustered Correlation Heatmap view of an MOA experiment 775

26.5 Legend of a MOA correlation analysis heatmap 778

26.6 Choose Entity List dialog box 780

26.7 Choose Interpretation dialog box 781

26.8 Correlation Node Inspector dialog box 782

26.9 Correlation Analysis Clustering Inputs dialog box 783

26.10Correlation Analysis Create Entity List from selection in view dialog box 784

26.11MOA Correlation Analysis Create Entity List from selection in view dialog box . . 785

26.12Select Annotation Columns dialog box 786

26.13Correlation Analysis Entity List Chooser dialog box in a SEA experiment 787

26.14Correlation Analysis Entity List Chooser dialog box in a MOA experiment 788

26.15Export options dialog box to export correlation coefficients 789

26.16Export as Report options dialog box 790

26.17Export as Image options dialog box 791

26.18Properties dialog box - Colors tab 793

26.19Properties dialog box - Colors tab - Create New Colors Theme 794

26.20Properties dialog box - Labels tab 795

26.21Properties dialog box - Rulers tab 796

26.22Properties dialog box - Description tab 797

26.23Computation of pair-wise correlation coefficients between entities in a SEA experiment 798

26.24Computation of pair-wise correlation coefficients between entities of two different experiments in a MOA experiment 799

27.1 Classification Pipeline 802

27.2 Build Prediction Model: Input Parameters 805

27.3 Build Prediction Model: Validation Parameters for PLSD 806

27.4 Build Prediction Model: R2Q2 output for PLSD 807

27.5 Build Prediction Model: Validation Output 808

27.6 Build Prediction Model: Training Output 809

27.7 Build Prediction Model: Model Object 810

27.8 Run Prediction: Prediction Output 811

27.9 Model Parameters for Support Vector Machines 816

27.10Model Parameters for Naive Bayesian Model 818

27.11Axis Parallel Decision Tree Model 820

27.12Neural Network Model 823

27.13Confusion Matrix for Training with Decision Tree 824

27.14Decision Tree Classification Report 825

27.15Lorenz Curve for Neural Network Training 826

28.1 Input Parameters . . . 829

28.2 Output Views of GO Analysis . . . 830

28.3 Spreadsheet view of GO Terms . . . 832

28.4 The GO Tree View . . . 833

28.5 Properties of GO Tree View . . . 834

28.6 Pie Chart View . . . 835

28.7 Pie Chart Properties . . . 837

29.1 Step1: Input Parameters . . . 841

29.2 Step 2: Pairing Options . . . 842

29.3 Step 3: Choose Gene Sets . . . 843

29.4 Step 4: Choose Entity Lists as Gene Sets . . . 844

29.5 Step 5: Results . . . 845

30.1 Input Parameters . . . 851

30.2 Pairing Options . . . 852

30.3 Choose Gene Sets . . . 853

30.4 Choose Gene Sets . . . 854

30.5 Choose Gene Sets . . . 855

31.1 BioCyc Database Collection . . . 860

31.2 Validation to import KEGG pathways . . . 863

31.3 Select Organism for WikiPathways download . . . 864

31.4 Import Statistics dialog - no duplicates . . . 864

31.5 Import Statistics dialog - duplicates are present 865

31.6 Rename Pathways Dialog 865

31.7 Options dialog to select BioCyc pathway version to use for analysis 868

31.8 Options dialog to select KEGG pathway version to use for analysis 870

31.9 Import Pathways From File options 870

31.10Import BioPAX pathways 873

31.11Input Relation Properties dialog 875

31.12Choose pathway files in the GPML format 876

31.13Set the organism for the imported pathway 877

31.14Choose the pathway provider for the imported pathway 878

31.15Pathway View Toolbar with enabled icons for creating a new pathway 879

31.16Entities Legend visible in the pathway viewer 881

31.17Input Node Properties dialog box 882

31.18Input Relation Properties dialog 883

31.19Export options for Pathways 886

31.20Annotations available from the Annotations Manager 889

31.21Downloaded annotations as shown in the Annotations Manager 889

31.22Download Pathway Infrastructure Database 890

31.23Automatic Software Update dialog for Interaction Databases 891

31.24Update Pathways dialog 892

31.25View obsolete entities and relations 893

31.26Automatic Software Update dialog for BridgeDb and ID Mapper databases 894

31.27SEA/MOA workflow behavior based on experiment types and non-availability of at least one of the organisms' BridgeDb or HomoloGene database 903

31.28Multi-Omic Analysis workflow - Step 1 of 4 932

31.29Multi-Omic Analysis workflow - Step 2 of 4 934

31.30Multi-Omic Analysis workflow - Step 3 of 4 936

31.31Multi-Omic Analysis workflow - Step 4 of 4 938

31.32Pathway View of a pathway list resulting from MOA 940

31.33Pathway Viewer pane of the Pathway View 943

31.34Interactive pathway behavior and display 944

31.35Reference KEGG pathway (from `http://www.genome.jp/kegg/kegg3a.html`) . . 945

31.36Organism specific KEGG pathway derived from reference pathway (from `http://www.genome.jp/kegg/kegg3a.html`) 946

31.37Organism specific KEGG pathway in **GeneSpring** (without experiment matching colors) 947

31.38Interactive pathway behavior and display of KEGG pathways in **GeneSpring** . . 948

31.39Tool tip information for a KEGG pathway entity representing multiple entities that match with both experiments as well as just one of the two experiments in an MOA experiment 949

31.40Entity Properties 950

31.41BioPAX Level 3 Entities Legend 951

31.42Relations Legend 952

31.43Edges Legend 953

31.44Relation Properties 954

31.45Relations Legend for BioPAX Level 3 Pathways 955

31.46Notation Legend for KEGG pathways 956

31.47KEGG Entity Properties 957

31.48KEGG Entity Properties with Enzyme information 958

31.49Style Theme Dialog 959

31.50Twopi layout 960

31.51Quilt plots as displayed in the pathway viewer of a KEGG pathway in a MOA experiment 961

31.52A sample heatstrip as displayed in the pathway viewer 962

31.53A Heatmap Pane containing data from a Metabolomics experiment 963

31.54A pathway list panel from the Pathway View of a pathway list resulting from Multi-Omic Pathway Analysis 965

31.55A sample filter panel from the Pathway View of a pathway list resulting from Multi-Omic Pathway Analysis 965

31.56Legend displaying the Color Settings for the Pathway View of an MOA experiment 966

31.57Chose Entity List dialog box 967

31.58Choose Interpretation dialog box 968

31.59Choose Quilt Plot Parameters dialog box 969

31.60Confirm BioCyc Pathways Download dialog box 970

31.61Confirm KEGG Pathways Download dialog box 971

31.62Confirm WikiPathways Download dialog box 972

31.63Entity List Chooser dialog box 973

31.64Entity Properties dialog box - GPML pathways 974

31.65Entity Properties dialog box - non-GPML pathways 975

31.66Entity and Enzyme Properties dialog box - KEGG pathways 977

31.67Entity and Enzyme Properties dialog box - KEGG pathways 978

31.68Export to eArray dialog box 979

31.69Export to eArray/SureDesign dialog box 980

31.70Export to File dialog box 981

31.71Input Node Properties dialog box 982

31.72Input Relation Properties dialog box 983

31.73Interpretation Chooser dialog box 984

31.74Multiple Objects Inspector dialog box 985

31.75New Multi-Omic Analysis Experiment dialog box 987

31.76Organism Chooser dialog box 989

31.77Pathway Inspector dialog box 990

31.78Pathway List Clipboard Inspector dialog box 991

31.79Pathway List Inspector dialog box 992

31.80Pathway Properties Dialog dialog box - Data tab 995

31.81Pathway Properties Dialog dialog box - Entity tab 996

31.82Pathway Properties Dialog dialog box - Heatstrip tab 998

31.83Pathway Properties Dialog dialog box - Quilt Plot Layout tab 999

31.84Pathway Properties Dialog dialog box - Select Highlight tab 1000

31.85Profile Plot dialog box 1001

31.86Relation Properties dialog box 1002

31.87Select Entities to Export dialog box 1003

31.88Venn Diagram dialog box 1004

31.89IPA options in the workflow browser 1015
31.90The Export to IPA dialog 1016
31.91Create New Pathway 1017
31.92IPA Pathway Creation - security warning 1018
31.93Java Startup 1018
31.94IPA Login Dialog 1019
31.95Pathway Analysis in IPA 1019
31.96Creation of Entity List 1020
31.97Creation of Entity List 1021
31.98Save List 1022
31.99**GeneSpring** suitable list creation 1022
31.10Saved List Location 1023
31.10Selection of Folder 1023
31.10Entity List Creation 1024
31.10Import Entity List from IPA - Error Message 1024
31.10Perform Data Analysis on Experiment 1025
31.10Perform Data Analysis on Experiment 1026
31.10Create Analysis 1027
31.10Data Analysis on Entity List 1028
31.10Perform Data Analysis on Entity List 1030
31.10Export to MetaCore - Choose Parameters 1032
31.11Export to MetaCore - Choose Data Column 1033

31.11Export to MetaCore - Login Details 1033

31.11Export to MetaCore - Processing 1034

31.11Export to MetaCore - Experiments View 1035

32.1 Entities Legend 1042

32.2 Algorithm options for the Simple Analysis workflow 1053

32.3 Algorithm options for the Advanced Analysis workflow 1054

32.4 Matching Statistics in Step 2 of the Advanced Analysis workflow 1055

32.5 Analysis Filters for Direct Interactions in the Advanced Analysis workflow 1056

32.6 Analysis Filters for Expand Interactions in the Advanced Analysis workflow 1057

32.7 Analysis Filters for Shortest Connect in the Advanced Analysis workflow 1058

32.8 Analysis Result of the Simple/Guided and Advanced Analysis workflows 1059

32.9 Save Pathway List 1060

32.10Extract Relations via NLP 1061

32.11Input Data 1061

32.12View Tagged Content 1062

32.13Pathway View 1063

32.14Extract Relations via NLP - Save Pathway List 1064

32.15Extract Relations via NLP - Pathway Inspector 1065

32.16Step 1: Input Page 1068

32.17Step 2: Select Relevant MeSH Terms 1068

32.18Step 3: MeSH Pathway 1070

32.19MeSH Network Builder Step 4: Pathway created from MeSH 1071

32.20Pathway Experiment Creation 1072

32.21Input Parameters 1073

32.22Import List from File 1074

32.23Choose signal columns 1075

32.24Choose extra columns 1075

32.25Pathway Experiment Workflow 1077

33.1 Experiment Creation - Rename samples and assign AMADID number 1084

33.2 Spreadsheet View of CGH experiment 1085

33.3 Filter by Samples 1087

33.4 Filter by Aberrations Step 2 - Select Calls 1088

33.5 Find Common Aberrations 1093

33.6 Genomer Browser view of a CGH experiment displaying a Multi-sample Interval and Scatter Plot(s). 1098

33.7 Pathway View of a CGH Experiment illustrating the Quilt Plots with different type of Aberrations. 1099

34.1 Experiment Creation for Affy CEL files 1111

34.2 Affymetrix Genome-Wide Human SNP Array 6.0, Genome-wide Human SNP array 5.0, and Human Mapping 500K Array Set - Reference Creation 1113

34.3 Reference Creation for Affy 100K array set 1114

34.4 Create Technology for Copy Number Analysis - Affymetrix technology 1118

34.5 Step 1: Load Data 1119

34.6 Step 2: Pair CEL files 1120

34.7 Step 3: Choose Copy Number/LOH Analysis Type 1121

34.8 QC views for Copy Number Experiment 1123

34.9 Batch Effect Correction - Step 1 1126

34.10Batch Effect Correction - Step 2 1127

34.11Copy Number Analysis - Paired Normal Method 1130

34.12Copy Number Analysis - Against Reference Method 1131

34.13Common Genomic Variant Regions - Step 3 1134

34.14Common Genomic Variant Regions - Step 4 1135

34.15Step 2: Filter Conditions for Filter by Region 1138

34.16Step 2: Input parameters for PSCN Filter 1141

34.17Heap Map View for a Copy Number Experiment 1143

34.18BRLMM-Flow chart 1153

34.19Transition Probabilities for LOH analysis against Reference HMM 1156

34.20The Paired Normal HMM 1156

35.1 EIGENSTRAT Filter 1171

35.2 EIGENSTRAT Correction View 1176

35.3 EIGENSTRAT Correction Results 1177

35.4 LD Plot 1197

35.5 Allele Frequencies on Genome Browser 1199

36.1 Shared License from server window 1205

36.2 Select License Modules to connect to **WorkGroupPlus** 1205

36.3 Permission Dialog 1210

36.4 Permission Dialog for multiple objects 1211

36.5 Confirmation Dialog for sharing descendants of an object 1211

36.6 Task Manager 1214

List of Tables

1.1 Platform Compatibility 19

1.2 Windows Platform Compatibility 20

1.3 Linux Platform Compatibility 20

1.4 Mac OS X Platform Compatibility 20

2.1 Interpretations and Views 31

2.2 Interpretations and Workflow Operations 32

2.3 Menu options available on right-click over a region list 45

2.4 Menu options available on right-click for a Report 46

2.5 Menu options available on right-click for a Merged Report 48

2.6 Menu options available on right-click for a Merged Report folder 48

3.1 Gene Annotation file format 93

3.2 Transcript Annotation file format 94

3.3 HomoloGene Table 96

4.1 Migration Rate 108

4.2 Migration Rate on Windows OS 108

4.3 Migration Rate on Debian OS 109

7.1 Sample Grouping and Significance Tests I 229

7.2 Sample Grouping and Significance Tests II 229

7.3 Sample Grouping and Significance Tests III 230

7.4 Sample Grouping and Significance Tests IV 230

7.5 Sample Grouping and Significance Tests V 230

7.6 Sample Grouping and Significance Tests VI 231

7.7 Sample Grouping and Significance Tests VII 231

7.8 Table of Default parameters for Analysis: Biological Significance workflow 235

7.9 Sample Grouping and Significance Tests I 264

7.10 Sample Grouping and Significance Tests II 265

7.11 Sample Grouping and Significance Tests III 265

7.12 Sample Grouping and Significance Tests IV 266

7.13 Sample Grouping and Significance Tests V 266

7.14 Sample Grouping and Significance Tests VI 267

7.15 Sample Grouping and Significance Tests VII 267

7.16 Sample Grouping and Significance Tests VIII 268

7.17 Normalized Intensities for Probes X and Y 276

8.1 Agilent Flag Logic: Example 1 299

8.2 Agilent Flag Logic: Example 2 300

15.1 Probe Mappings from tags in GEML(XML) file 441

15.2 Lookup Table 443

15.3 Example: Probe Summarization 445

15.4 Sample Grouping and Significance Tests I 454

15.5 Sample Grouping and Significance Tests II 454

15.6 Sample Grouping and Significance Tests III 455

15.7 Sample Grouping and Significance Tests IV 455

15.8 Sample Grouping and Significance Tests V 455

15.9 Sample Grouping and Significance Tests VI 456

15.10 Sample Grouping and Significance Tests VII 456

15.11 Table of Default parameters for Analysis: Biological Significance 461

16.1 Agilent Flag Logic: Example 1 486

16.2 Agilent Flag Logic: Example 2 486

21.1 Availability of analysis options available from the Workflow Browser by experiment type (applicable to both Targeted Sequencing and Whole Genome experiments . . 620

21.2 Experiment Toolbar icons available for an imported **Strand NGS** experiment . . 622

21.3 Menu options available on right-click of an imported **Strand NGS** experiment . . 623

21.4 Menu options available on right-click over a region list of an imported **Strand NGS** experiment 623

23.1 IDF Defaults 635

23.2 SDRF Defaults 639

24.1 Available statistical tests by experimental conditions and grouping 647

25.1 Result output views by clustering parameters 706

25.2 Availability of Dendrogram View Toolbar icons by clustering algorithm 708

25.3 Menu options available on right-click over a cluster tree view 711

26.1 Analysis Types in **GeneSpring** for Correlation Analysis 748

26.2 Function and Availability of Corrrelation Heatmap View Toolbar icons 777

26.3 Menu options available on right-click over a correlation analysis heatmap 779

27.1 Validation Parameters 813

27.2 Decision Tree Table 818

31.1 Available Annotations by Omic Type 909

31.2 Toolbar Icons available in the Pathway View depending on pathway type/format . 941

31.3 Toolbar Icons available in the Pathway View for existing non-GPML pathways *vs* creating new pathways 942

31.4 Right-click options on the Heatmap Pane 1006

31.5 Right-click options on pathway list name in the Experiment Navigator 1007

31.6 Right-click options on the pathway list panel 1007

31.7 Right-click options for Pathway Viewer depending on pathway source 1009

31.8 Right-click options for KEGG pathways in the Pathway Viewer 1010

32.1 Process, Function, and Complex Properties 1045

32.2 Participant Roles 1047

32.3 Protein Modification Mechanisms 1049

32.4 Type of relationship and Frequency 1069

33.1 Imported Columns by Source File Type 1082

33.2 Aberration Type Call Determination by Data Type 1082

33.3 Call Determination for Translated Genes and Probes 1091

33.4 Clustering on Calls - Numerical Value Assignment by Call Type 1094

34.1 Terminology in Copy Number Analysis 1109

34.2 Mapping Fawkes state to LOH 1116

34.3 Batch Effect Correction 1125

34.4 Snapshot of 'Common Genomic Variant Region' Detection Algorithm 1132

34.5 Filter by CGVs 1142

34.6 Identify Overlapping Genes 1145

34.7 Utilities in Copy Number Analysis 1146

34.8 Workflow for Illumina output files 1149

34.9 Additional notes on BRLMM 1154

34.10Snap-shot of Birdseed Algorithm 1158

34.11Snap-shot of CBS Algorithm 1160

34.12Snap-shot of Fawkes Algorithm 1162

35.1 Technologies and Genotype Call Algorithms for Association Analysis Experiments 1164

35.2 Summary of Steps: Filter Samples by Missing Values 1168

35.3 Birdseed Report 1184

35.4 Summary of Steps: EIGENSTRAT Filter 1185

35.5 Summary of Steps: Filter SNPs by Missing Value 1185

35.6 Summary of Steps: Identify SNPs by Differential Missingness 1186

35.7 Contingency Table for Differential Missingness 1186

35.8 Summary of Steps: Filter SNPs by HWE p-value 1186

35.9 Summary of Steps: Filter SNPs by MAF 1187

35.10Summary of Steps: EIGENSTRAT Correction on Samples 1187

35.11EIGENSTRAT Correction Result Screen 1187

35.12Mode of Inheritance 1187

35.13Summary of Steps: Pearson's χ^2 Test 1188

35.14Contingency Table for Pearson's χ^2 Test 1188

35.15Summary of Steps: Fisher's Exact Test 1189

35.16Summary of Steps: Cochran-Armitage Test 1189

35.17Contingency Table for Cochran-Armitage Test 1190

35.18Weights (d_i) for Cochran-Armitage Test 1190

35.19Summary of Steps: χ^2 Correlation Test 1190

35.20Statistic Computed for SNP blocks and individual SNPs 1194

35.21Statistic Computed For Haplo blocks and individual Haplotypes. 1196

35.22Mapping for Genotype Calls 1198

35.23Utilities in Association Analysis 1201

37.1 Accessing Projects and Experiments 1219

37.2 Accessing Experiment Dataset 1222

37.3 Some Useful Functions 1224

37.4 Some Common Marks 1224

37.5 Creating UI Components 1227

37.6 R API calls 1243

38.1 Mouse Clicks and their Action 1245

38.2 Scatter Plot Mouse Clicks 1246

38.3 3D Mouse Clicks 1246

38.4 Mouse Click Mappings for Mac 1246

38.5 Global Key Bindings 1247

Chapter 1

GeneSpring Installation

This version of **GeneSpring** is available for Windows, Mac OS X, and Linux. This chapter describes how to install **GeneSpring** on Windows, Mac OS X and Linux. Note that this version of **GeneSpring** can coexist with **GeneSpring** 7.x on the same machine.

1.1 Supported and Tested Platforms

The table below gives the platforms on which **GeneSpring** has been tested.

For experiments with more than 100 samples, a 64-bit QuadCore Desktop with more than 4 GB RAM is recommended. Ensure that there is enough free space in the Hard Disk drive of the installation. The temporary Hard Disk usage while running various analyses in Copy Number and Association experiments may be several times the total size of the samples used in the experiment. A free space of 5 times the total size of the samples should suffice under most conditions, but the more the better.

For performance statistics on Copy Number Experiments, see section Performance Statistics for Copy Number Analysis.

1.2 Installation on Windows

Supported Windows Platforms

1.2.1 Installation and Usage Requirements

- Processor with 1.5 GHz and 4 GB RAM.
- Disk space required: 2.5 GB
- At least 16 MB Video Memory. Check this via Start → Settings → Control Panel → Display → Settings tab → Advanced → Adapter tab → Memory Size field. 3D graphics may require more memory. Also changing Display Acceleration settings may be needed to view 3D plots.
- Administrator privileges are required for installation. Once installed, other users can use **GeneSpring** as well.

1.2.2 GeneSpring Installation Procedure for Microsoft Windows

GeneSpring can be installed on any of the Microsoft Windows platforms listed above. To install **GeneSpring**, follow the instructions given below:

- You must have the installable for your particular platform `genespring_windows.exe`.
- Run the `genespring_windows.exe` installable file.
- The wizard will guide you through the installation procedure.
- By default, **GeneSpring** will be installed in the `C:\Program Files\Agilent\GeneSpring\` directory. You can specify any other installation directory of your choice during the installation process.
- During installation process or when you upgrade **GeneSpring**, a window will open for installing Microsoft Visual C++ 2015 RunTime Libraries (See Figure 1.1). Click **Yes** to accept and proceed with the installation.

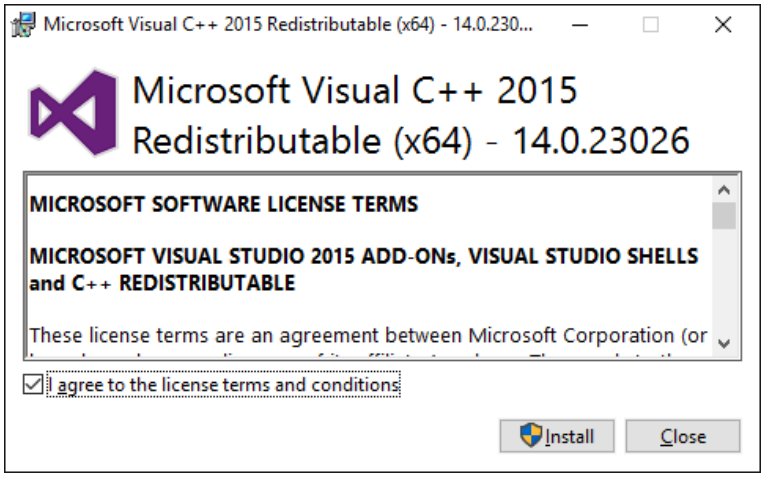


Figure 1.1: Microsoft Visual C++ 2005 RunTime Libraries Installation

NOTE If you do not install these libraries during the initial installation, **GeneSpring** functionalities can be effected. If you did not install these libraries during initial installation of **GeneSpring**, go to **GeneSpring** installation folder → app → msvc_redist folder and select **vcredist_x86.exe** for a 32 bit machine or **vcredist_x64.exe** for a 64 bit machine to install manually.

- At the end of the installation process, a browser is launched with the documentation index, showing all the documentation available with the tool.
- Following this, **GeneSpring** is installed on your system. By default the **GeneSpring** icon appears on your desktop and in the programs menu.
- To start using **GeneSpring**, you will have to activate your installation by following the steps detailed in the Activation step.

By default, **GeneSpring** is installed in the programs group with the following utilities:

- **GeneSpring**, for starting up the **GeneSpring** tool.
- Documentation, leading to all the documentation available online in the tool.
- Uninstall, for uninstalling the tool from the system.

1.2.3 Activating your GeneSpring

Your **GeneSpring** installation has to be activated for you to use **GeneSpring**. **GeneSpring** imposes a node-locked license, so it can be used only on the machine that it was installed on.

- You should have a valid OrderID to activate **GeneSpring**. If you do not have an OrderID, register at http://genespring.com. An OrderID will be e-mailed to you to activate your installation.
- Auto-activate **GeneSpring** by connecting to **GeneSpring** website. The first time you start up **GeneSpring** you will be prompted with the '**GeneSpring** License Activation' dialog-box. Enter your OrderID in the space provided. This will connect to the **GeneSpring** website, activate your installation and launch the tool. If you are behind a proxy server, then provide the proxy details in the lower half of this dialog-box.
- The license is obtained by contacting the licenses server over the Internet and obtaining a node-locked, fixed duration license. If your machine date and time settings are different and cannot be matched with the server date and time settings you will get an *Clock Skew Detected* error and will not be able to proceed. If this is a new installation, you can change the date and time on your local machine and try to activate again.

- **Manual activation.** If the auto-activation step has failed due to any other reason, you will have to manually get the activation license file to activate **GeneSpring**, using the instructions given below:
  - Locate the activation key file `manualActivation.txt` in the `\bin\license\` folder in the installation directory.
  - Go to `http://lcosgens.cos.agilent.com/gsLicense/MultipleActivateNew.html`, enter the OrderID, upload the activation key file, `manualActivation.txt` from the file-path mentioned above, and click Submit. This will generate an activation license file (strand.lic) that will be e-mailed to your registered e-mail address. If you are unable to access the website or have not received the activation license file, send a mail to `informatics_support@agilent.com` with the subject `Registration Request`, with `manualActivation.txt` as an attachment. We will generate an activation license file and send it to you within one business day.
  - Once you have got the activation license file, strand.lic, copy the file to your `\bin\license\` subfolder.
  - Restart **GeneSpring**. This will activate your **GeneSpring** installation and will launch **GeneSpring**.
  - If **GeneSpring** fails to launch and produces an error, please send the error code to `informatics_support@agilent.com` with the subject Activation Failure. You should receive a response within one business day.

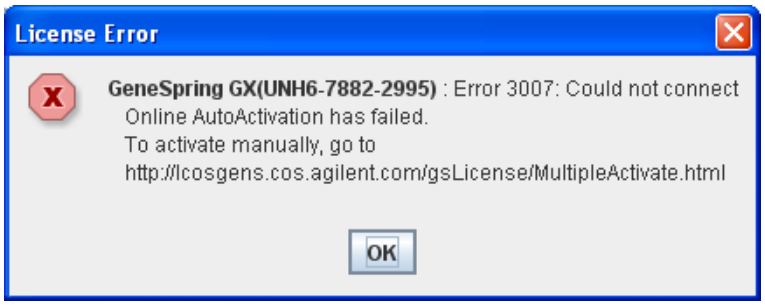


Figure 1.2: Activation Failure

1.2.4 Uninstalling GeneSpring from Windows

The Uninstall program is used for uninstalling **GeneSpring** from the system. Before uninstalling **GeneSpring**, make sure that the application and any open files from the installation directory are closed.

To start the **GeneSpring** uninstaller, click Start, choose the Programs option, and select `GeneSpring`. Click Uninstall. Alternatively, click Start, select the Settings option, and click Control Panel. Double-click the Add/Remove Programs option. Select `GeneSpring` from the list of products. Click Uninstall. The Uninstall **GeneSpring** wizard displays the features that are to be removed. Click Done to close the Uninstall Complete wizard. **GeneSpring** will be successfully uninstalled from the Windows system. Some files and folders like log files and data, samples and templates folders that have been created after the installation of **GeneSpring** would not be removed.

1.3 Installation on Linux

Supported Linux Platforms

1.3.1 Installation and Usage Requirements

- In addition certain run-time libraries are required for activating and running **GeneSpring**. Ensure that the OS has a libc6 version of 2.15+. The required run-time libraries are `libpam0g:i386 libkrb5-3:i386 libgssapi-krb5-2:i386`. To confirm that the required libraries are available for activating the license, go to
`Agilent/GeneSpring/bin/packages/cube/license/x.x/lib(32/64)`
and run the following command
`ldd liblicense.so`
Check that all required linked libraries are available on the system.
- Processor with 1.5 GHz and 4 GB RAM.
- Disk space required: 2.5 GB
- At least 16 MB Video Memory.
- Administrator privileges are NOT required. Only the user who has installed **GeneSpring** can run it. Multiple installs with different user names are permitted.

NOTE

It is important to note that, in Linux 64 bit machines, packages specified above are required to start **GeneSpring** database server. If you see an error message similar to Figure 1.3, this could be because of the missing packages. Check with your system administrator to install the package or contact **GeneSpring** support

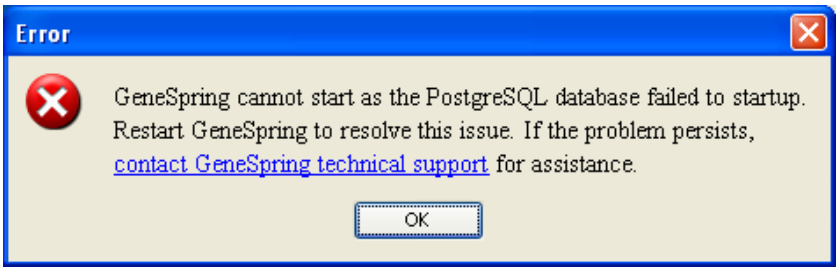


Figure 1.3: Database error Dialog

1.3.2 GeneSpring Installation Procedure for Linux

GeneSpring can be installed on most distributions of Linux. To install **GeneSpring**, follow the instructions given below:

- You must have the installable for your particular platform `genespring_linux.bin` or `genespring_linux.sh`.
- Run the `genespring_linux.bin` or `genespring_linux.sh` installable.
- The program will guide you through the installation procedure.
- By default, **GeneSpring** will be installed in the `$HOME/Agilent/GeneSpring$` directory. You can specify any other installation directory of your choice at the specified prompt in the dialog box.
- At the end of the installation process, a browser is launched with the documentation index, showing all the documentation available with the tool.
- **GeneSpring** should be installed as a normal user and only that user will be able to launch the application.
- Following this, **GeneSpring** is installed in the specified directory on your system. However, it will not be active yet. To start using **GeneSpring** , you will have to activate your installation by following the steps detailed in the Activation step.

By default, **GeneSpring** is installed with the following utilities in the **GeneSpring** directory:

- **GeneSpring**, for starting up the **GeneSpring** tool.
- Documentation, leading to all the documentation available online in the tool.
- Uninstall, for uninstalling the tool from the system

1.3.3 Activating your GeneSpring

Your **GeneSpring** installation has to be activated for you to use **GeneSpring**. **GeneSpring** imposes a node-locked license, so it can be used only on the machine that it was installed on.

- You should have a valid OrderID to activate **GeneSpring**. If you do not have an OrderID, register at `http://genespring.com`. An OrderID will be e-mailed to you to activate your installation.
- Auto-activate **GeneSpring** by connecting to **GeneSpring** website. The first time you start up **GeneSpring** you will be prompted with the '**GeneSpring** License Activation' dialog-box. Enter your OrderID in the space provided. This will connect to the **GeneSpring** website, activate your installation and launch the tool. If you are behind a proxy server, then provide the proxy details in the lower half of this dialog-box.

- The license is obtained by contacting the licenses server over the Internet and obtaining a node-locked, fixed duration license. If your machine date and time settings are different and cannot be matched with the server date and time settings you will get an *Clock Skew Detected* error and will not be able to proceed. If this is a new installation, you can change the date and time on your local machine and try activate again.
- **Manual activation.** If the auto-activation step has failed due to any other reason, you will have to manually get the activation license file to activate **GeneSpring**, using the instructions given below:
  - Locate the activation key file `manualActivation.txt` in the `\bin\license\` folder in the installation directory.
  - Go to `http://lcosgens.cos.agilent.com/gsLicense/MultipleActivateNew.html`, enter the OrderID, upload the activation key file, `manualActivation.txt` from the file-path mentioned above, and click Submit. This will generate an activation license file (strand.lic) that will be e-mailed to your registered e-mail address. If you are unable to access the website or have not received the activation license file, send a mail to `informatics_support@agilent.com` with the subject `Registration Request`, with `manualActivation.txt` as an attachment. We will generate an activation license file and send it to you within one business day.
  - Once you have got the activation license file, strand.lic, copy the file to your `\bin\license\` subfolder.
  - Restart **GeneSpring**. This will activate your **GeneSpring** installation and will launch **GeneSpring**.
  - If **GeneSpring** fails to launch and produces an error, please send the error code to `informatics_support@agilent.com` with the subject `Activation Failure`. You should receive a response within one business day.

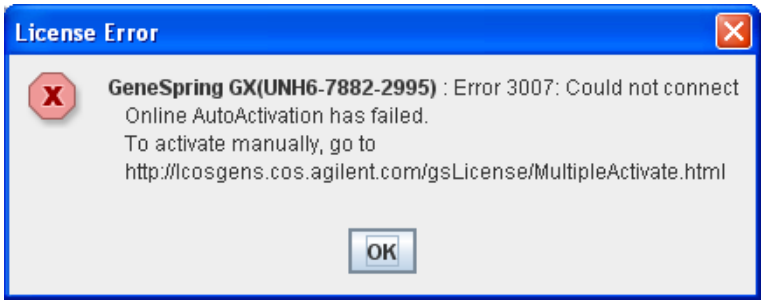


Figure 1.4: Activation Failure

1.3.4 Uninstalling GeneSpring from Linux

Before uninstalling **GeneSpring**, make sure that the application is closed. To uninstall **GeneSpring**, run Uninstall from the **GeneSpring** home directory and follow the instructions on screen.

1.4 Installation on Apple Macintosh

Supported Mac Platforms

1.4.1 Installation and Usage Requirements

- Processor with 1.5 GHz and 4 GB RAM.
- Disk space required: 2.5 GB
- At least 16 MB Video Memory. (Refer section on 3D graphics in FAQ)
- **GeneSpring** should be installed as a normal user and only that user will be able to launch the application.

1.4.2 GeneSpring Installation Procedure for Macintosh

- You must have the installable for your particular platform `genespring_mac.zip`.
- **GeneSpring** should be installed as a normal user and only that user will be able to launch the application.
- Uncompress the executable by double clicking on the .zip file. This will create a .app file at the same location. Make sure this file has executable permission.
- Double click on the .app file and start the installation. This will install **GeneSpring** on your machine. By default **GeneSpring** will be installed in
  `$HOME/Applications/Agilent/GeneSpring$` or
  You can install **GeneSpring** in an alternative location by changing the installation directory.
- To start using **GeneSpring**, you will have to activate your installation by following the steps detailed in the Activation step.
- At the end of the installation process, a browser is launched with the documentation index, showing all the documentation available with the tool.
- Note that **GeneSpring** is distributed as a node locked license. For this the hostname of the machine should not be changed. If you are using a DHCP server while being connected to be net, you have to set a fixed hostname. To do this, give the command `hostname` at the command prompt during the time of installation. This will return a hostname. And set the `HOSTNAME` in the file /etc/hostconfig to `your_machine_hostname_during_installation`

  For editing this file you should have administrative privileges. Give the following command:
  `sudo vi /etc/hostconfig`
  This will ask for a password. You should give your password and you should change the following line
  from
  `HOSTNAME=-AUTOMATIC-`

to
`HOSTNAME=your_machine_hostname_during_installation`

- You need to restart the machine for the changes to take effect.

By default, **GeneSpring** is installed with the following utilities in the **GeneSpring** directory:

- **GeneSpring**, for starting up the **GeneSpring** tool.
- Documentation, leading to all the documentation available online in the tool.
- Uninstall, for uninstalling the tool from the system

GeneSpring uses left, right and middle mouse-clicks. On a single button Macintosh mouse, here is how you can emulate these clicks.

- Left-click is a regular single button click.
- Right-click is emulated by Control + click.
- Control-click is emulated by Apple + click.

1.4.3 Activating your GeneSpring

Your **GeneSpring** installation has to be activated for you to use **GeneSpring**. **GeneSpring** imposes a node-locked license, so it can be used only on the machine that it was installed on.

- You should have a valid OrderID to activate **GeneSpring**. If you do not have an OrderID, register at http://genespring.com An OrderID will be e-mailed to you to activate your installation.
- Auto-activate **GeneSpring** by connecting to **GeneSpring** website. The first time you start up **GeneSpring** you will be prompted with the '**GeneSpring** License Activation' dialog-box. Enter your OrderID in the space provided. This will connect to the **GeneSpring** website, activate your installation and launch the tool. If you are behind a proxy server, then provide the proxy details in the lower half of this dialog-box.
- The license is obtained by contacting the licenses server over the internet and obtaining a node-locked, fixed duration license. If your machine date and time settings are different cannot be matched with the server date and time settings you will get an *Clock Skew Detected* error and will not be able to proceed. if this is a new installation, you can change the date and time on your local machine and try activate again.
- **Manual activation.** If the auto-activation step has failed due to any other reason, you will have to manually get the activation license file to activate **GeneSpring**, using the instructions given below:

- Locate the activation key file `manualActivation.txt` in the `\bin\licence` subfolder of the installation directory.
- Go to http://lcosgens.cos.agilent.com/gsLicense/MultipleActivateNew.html, enter the OrderID, upload the activation key file, `manualActivation.txt` from the file-path mentioned above, and click Submit. This will generate an activation license file (strand.lic) that will be e-mailed to your registered e-mail address. If you are unable to access the website or have not received the activation license file, send a mail to `informatics_support@agilent.com` with the subject `Registration Request`, with `manualActivation.txt` as an attachment. We will generate an activation license file and send it to you within one business day.
- Once you have got the activation license file, strand.lic, copy the file to your `\bin\license\` subfolder of the installation directory.
- Restart **GeneSpring**. This will activate your **GeneSpring** installation and will launch **GeneSpring**.
- If **GeneSpring** fails to launch and produces an error, please send the error code to `informatics_support@agilent.com` with the subject `Activation Failure`. You should receive a response within one business day.

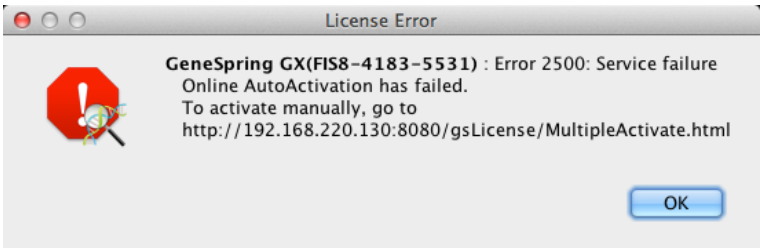


Figure 1.5: Activation Failure

1.4.4 Uninstalling GeneSpring from Mac

Before uninstalling **GeneSpring**, make sure that the application is closed. To uninstall **GeneSpring**, run Uninstall from the **GeneSpring** home directory and follow the instructions on screen.

1.5 GeneSpring Licensing

You can activate any of the following modules in the same installation of **GeneSpring** by obtaining the corresponding licenses:

- GeneSpring GX - This enables all gene expression, exon expression, copy number and genome-wide association analysis types (see License Manager for the complete list of analysis workflows).
- Mass Profiler Professional - This enables metabolomics and proteomics analysis types (see the MassProfiler Manual accessible via '**Help** $\longrightarrow$ **Documentation Index**' for more details).

- Pathway Architect - This provides native access to curated pathways from the WikiPathways portal, the BioCyc Pathway Collection, as well as KEGG pathways (with a valid end user license) and provides the ability to import pathways from file in the .gpml, Biopax Level 2 and 3, and .txt formats. These imported pathways can then be used for pathway directed analysis (see Pathway Analysis) including multi-omic pathway analysis in addition to any hand created and NLP derived pathways.

After installing **GeneSpring** select the applicable modules in the **License Activation Dialog** and enter the corresponding licenses to activate **GeneSpring**.

If you are installing **GeneSpring** on a cloud server or a virtual machine (VM), select **Cloud or VM Installation**.

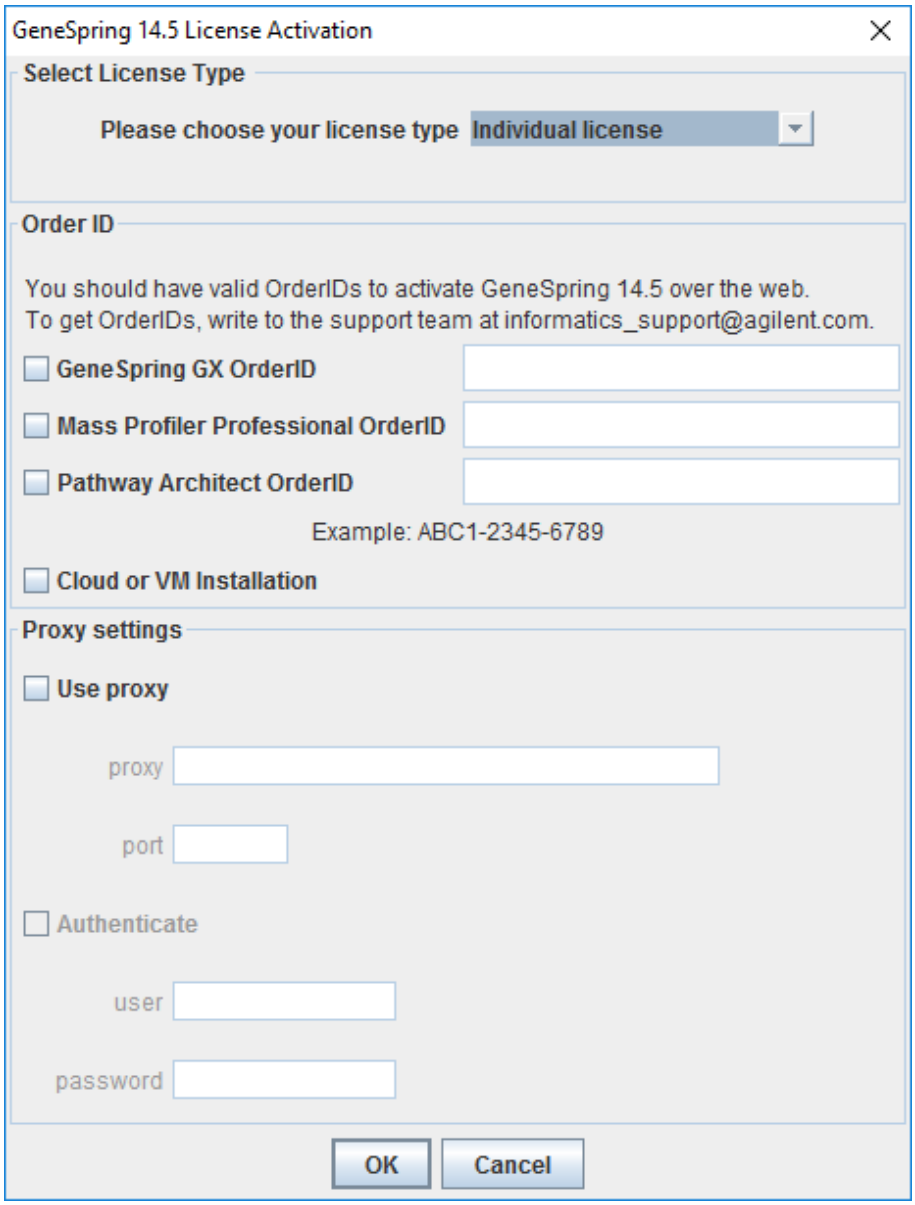


Figure 1.6: The License Activation Dialog

1.6 License Manager

After successful installation and activation of **GeneSpring**, you will be able to use certain utilities to manage the license. These utilities are available from *Help* ⟶ *License Manager* on the top menu bar of the tool. Choosing *Help* ⟶ *License Manager* from the top menu will launch the *License Description* dialog.

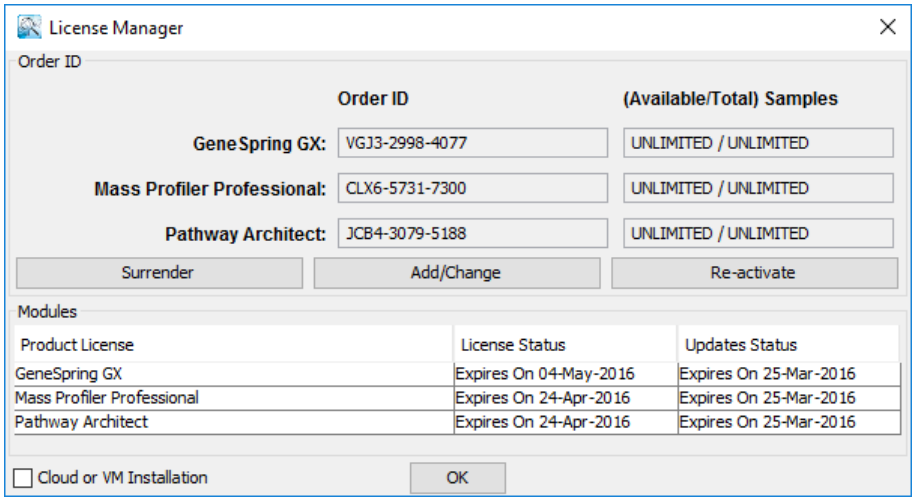


Figure 1.7: The License Manager Dialog

The corresponding boxes of the *License Manager* show the `Order ID` that was used to activate the license. If you are using a floating server to activate and license **GeneSpring**, you will see the port and the host name of the license server. You may need to note the license `Order ID` to change the installation, or to refer to your installation at the time of support.

Indicate that **GeneSpring** is installed on a cloud server or a virtual machine (VM) by selecting **Cloud or VM Installation**.

GeneSpring is licensed as a set of module bundles that allow various functionalities. The table in the dialog shows the modules available in the current installation along with their status. Currently the modules are bundled into the above mentioned four categories and based on the modules licensed, appropriate menu items will be enabled or disabled.

The **GeneSpring GX** module enables the following analysis workflows (excluding pathway-related functionality):

- Affymetrix® Expression,
- Affymetrix Exon Expression,
- Affymetrix Exon Splicing,
- Affymetrix Copy Number Analysis,

- Affymetrix Association Analysis,
- Agilent Expression Single Color,
- Agilent Expression Two Color,
- Agilent Exon Single Color,
- Agilent Exon Two Color,
- Agilent miRNA,
- Illumina® Single Color,
- Illumina Copy Number Analysis,
- Illumina Association Analysis,
- Generic Single Color,
- Generic Two Color,
- RealTime-PCR.

1.6.1 Utilities of the License Manager

The *License Manager* contains your Order IDs for the activated modules and the corresponding expiration dates as well as the Updates Status for each module. A **License Expiry Message** will appear 45 days before any of your licenses expries.

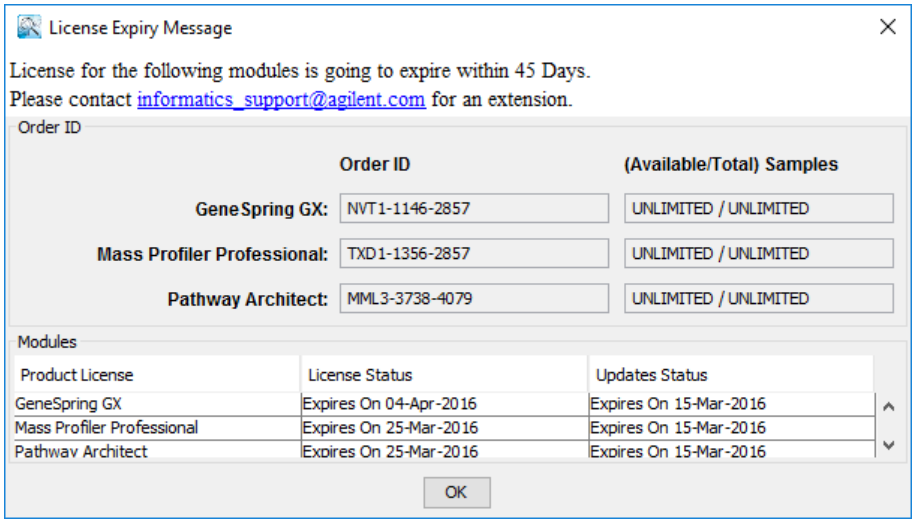


Figure 1.8: License Expiry Message

In addition, the License Manager Dialog provides the following utilities:

Surrender: Click on this button to surrender the license to the license server. You must be connected to the internet for surrender to operate. The surrender utility is used if you want to check-in or surrender the license into the license server and check out or activate the license on another machine. This utility is useful to transfer licenses from one machine to another, like from an office desktop machine to a laptop machine.

Note that the license can be activated from only one installation at any time. Thus, when you surrender the license, the current installation will be in-activated. You will be prompted to confirm your intent to surrender the license and clicking *OK* will surrender the license and shut the tool. If you want to activate your license on another machine, or on the same machine, you will need to store the Order ID and enter the Order ID in the *License Activation Dialog*.

If you are not connected to the Internet, or if you are unable to reach the license server, you can do a manual surrender. You will be prompted with a dialog to confirm manual surrender. If you confirm, then the current installation will be deactivated. Follow the on screen instructions. Upload the file
`<install_dir>/Agilent/GeneSpring/bin/license/surrender.bin`
to `http://lcosgens.cos.agilent.com/gsLicense/MultipleSurrenderNew.html` This will surrender the license which can be reused on another machine.

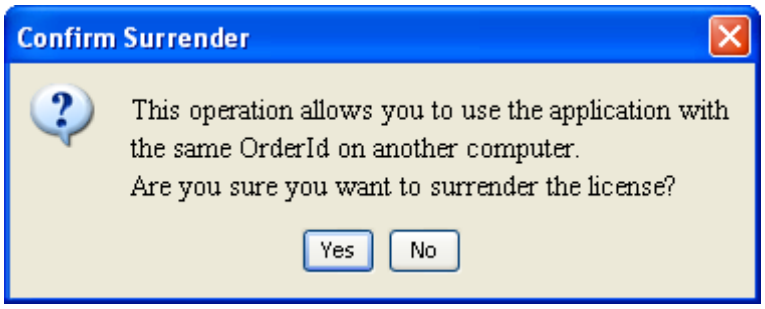


Figure 1.9: Confirm Surrender Dialog

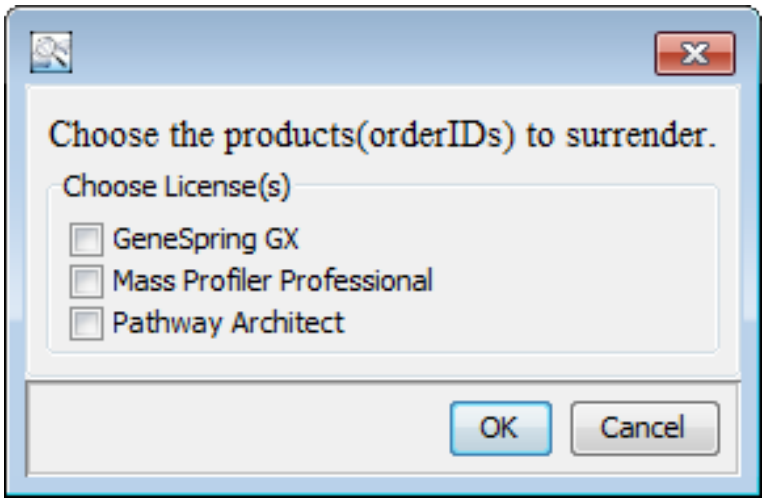


Figure 1.10: Manual Surrender Dialog

Add/ Change: This utility allows you to add or change the Order ID of the product and activate the product with a new Order ID. This utility is used to procure a different set of modules or

change the module status and module expiry of the current installation. If you had a limited duration trial license and would like to purchase and convert the license to an annual license, click on the *Change* button. This will launch a dialog for Order ID. Enter the new Order ID obtained from Agilent. This will activate **GeneSpring** with the new Order ID and all the modules and module status will confirm to the new Order ID.

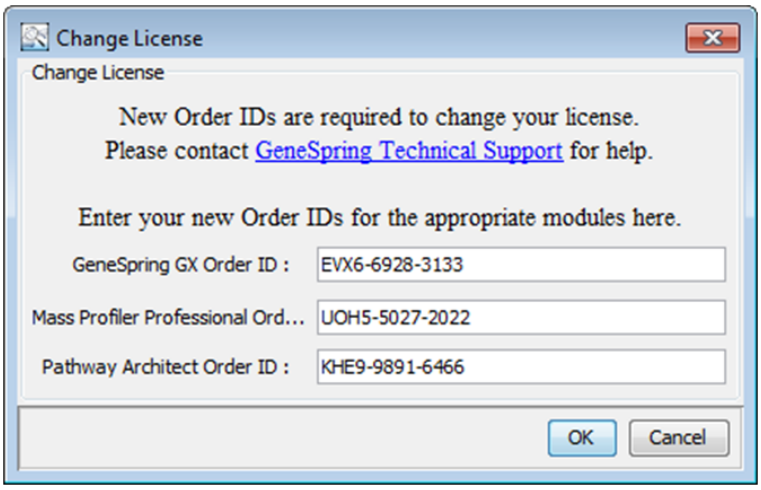


Figure 1.11: Change License Dialog

Re-activate: To reactivate the license, click on the Re-activate button on the **License Description Dialog**. This will reactivate the license from the license server with the same Order ID and on the same machine. The operation will prompt a dialog to confirm the action, after which the license will be reactivated and the tool will be shut down. When the tool is launched again, the tool will be launched again with the license obtained for the same Order ID. Note that reactivation can be done only on the same machine with the same Order ID. This utility may be necessary if the current installation is and license have been corrupted and you would like to reactivate and get a fresh license on the same Order ID on the same machine. Or you have Order ID definition and corresponding modules have changed and you have been advised by support to re-activate the license.

If you are not connected to the Internet, or if you are unable to reach the license server, you can re-activate manually. You will be prompted with a dialog stating that the reactivation failed and if you want to reactivate manually. If you confirm, then the current installation will be deactivated. Follow the on screen instructions to re-activate your tool.
`<install_dir>/Agilent/GeneSpring/bin/license/manualActivation.txt`
to `http://lcosgens.cos.agilent.com/gsLicense/MultipleActivateNew.html`

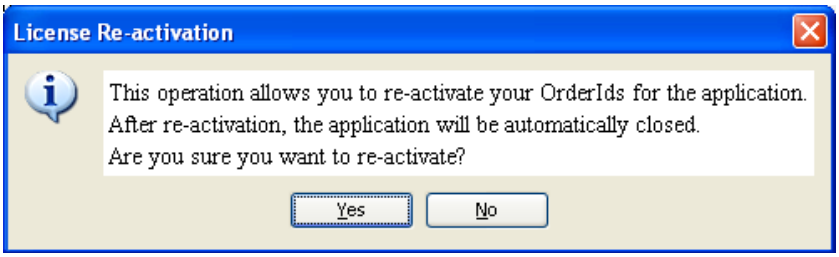


Figure 1.12: License Re-activation Dialog

1.6.2 Viewer Mode

GeneSpring viewer mode is activated on the expiry of the module license. In **GeneSpring** viewer mode, the available options are:

- Open existing experiments and change views for selected entity lists.
- Export and Import of Project is enabled.
- You can connect to external resources and for performing downstream analyses.

In viewer mode, you cannot:

- Create new experiments.
- Perform new analyses as all **GeneSpring Workflow** options are disabled.

1.7 Upgrade

For upgrading **GeneSpring**, go to the menu **Help** → **Update Product** → **From Agilent Server or From File** in the tool and follow the instructions thereon.

1.8 System Requirements for Copy Number Experiments, Association Experiments, and Correlation Analysis

1.8.1 Copy Number and Association Experiments

1.8.1.1 Minimum Specifications

As a guideline, the minimum specifications for, say, 75 samples of Affymetrix Genome-Wide Human SNP Array 6.0 are:

- 32-bit system with 4 GB RAM
- 25 GB of free disk space

1.8.1.2 Recommended Specifications

- A 64-bit, Quadcore platform with 4 GB or higher RAM is recommended.
- Free disk space required will be proportional to the number and size of samples. An approximation can be made based on the guidelines provided in the 'Minimum Specifications' section above.

1.8.2 Correlation Analysis

To perform correlation analysis on large entity lists please use the following specifications as guidance on whether your system specifications are sufficient:

- 64-bit system with 4 GB RAM: **GeneSpring** performs correlation analysis on up to 7,000 entities comfortably.
- 64-bit system with 8 GB RAM: **GeneSpring** performs correlation analysis on up to 10,000 entities comfortably.

Operating System	Hardware Architecture	Installer
Microsoft Windows 10	x86 compatible architecture	genespring_windows32.exe
Microsoft Windows 10	x86_64 compatible architecture	genespring_windows64.exe
Microsoft Windows 8.1	x86 compatible architecture	genespring_windows32.exe
Microsoft Windows 8.1	x86_64 compatible architecture	genespring_windows64.exe
Microsoft Windows 8	x86 compatible architecture	genespring_windows32.exe
Microsoft Windows 8	x86_64 compatible architecture	genespring_windows64.exe
Microsoft Windows 7	x86 compatible architecture	genespring_windows32.exe
Microsoft Windows 7	x86_64 compatible architecture	genespring_windows64.exe
Microsoft Windows Vista	x86 compatible architecture	genespring_windows32.exe
Microsoft Windows Vista	x86_64 compatible architecture	genespring_windows64.exe
Red Hat Enterprise Linux 6	x86 compatible architecture	genespring_linux32.bin
Red Hat Enterprise Linux 6	x86_64 compatible architecture	genespring_linux64.bin
Debian GNU/Linux 8	x86 compatible architecture	genespring_linux32.bin
Debian GNU/Linux 8	x86_64 compatible architecture	genespring_linux64.bin
Ubuntu GNU/Linux 12.04	x86 compatible architecture	genespring_linux32.bin
Ubuntu GNU/Linux 12.04	x86_64 compatible architecture	genespring_linux64.bin
Apple Mac OS X v10.11 (El Capitan)	x64 compatible architecture	genespring_mac.zip
Apple Mac OS X v10.10 (Yosemite)	x64 compatible architecture	genespring_mac.zip
Apple Mac OS X v10.9 (Mavericks)	x64 compatible architecture	genespring_mac.zip
Apple Mac OS X v10.8 (Mountain Lion)	x64 compatible architecture	genespring_mac.zip
Apple Mac OS X v10.7.3 (Lion)	x64 compatible architecture	genespring_mac.zip

Table 1.1: Platform Compatibility

Operating System	Hardware Architecture	Installer
Microsoft Windows 10	x86 compatible architecture	genespring_windows32.exe
Microsoft Windows 10	x86_64 compatible architecture	genespring_windows64.exe
Microsoft Windows 8.1	x86 compatible architecture	genespring_windows32.exe
Microsoft Windows 8.1	x86_64 compatible architecture	genespring_windows64.exe
Microsoft Windows 8	x86 compatible architecture	genespring_windows32.exe
Microsoft Windows 8	x86_64 compatible architecture	genespring_windows64.exe
Microsoft Windows 7	x86 compatible architecture	genespring_windows32.exe
Microsoft Windows 7	x86_64 compatible architecture	genespring_windows64.exe
Microsoft Windows Vista	x86 compatible architecture	genespring_windows32.exe
Microsoft Windows Vista	x86_64 compatible architecture	genespring_windows64.exe

Table 1.2: Windows Platform Compatibility

Operating System	Hardware Architecture	Installer
Red Hat Enterprise Linux 6	x86 compatible architecture	genespring_linux32.bin
Red Hat Enterprise Linux 6	x86_64 compatible architecture	genespring_linux64.bin
Debian GNU/Linux 8	x86 compatible architecture	genespring_linux32.bin
Debian GNU/Linux 8	x86_64 compatible architecture	genespring_linux64.bin
Ubuntu GNU/Linux 12.04	x86 compatible architecture	genespring_linux32.bin
Ubuntu GNU/Linux 12.04	x86_64 compatible architecture	genespring_linux64.bin

Table 1.3: Linux Platform Compatibility

Operating System	Hardware Architecture	Installer
Apple Mac OS X v10.11 (El Capitan)	x64 compatible architecture	genespring_mac.zip
Apple Mac OS X v10.10 (Yosemite)	x64 compatible architecture	genespring_mac.zip
Apple Mac OS X v10.9 (Mavericks)	x64 compatible architecture	genespring_mac.zip
Apple Mac OS X v10.8 (Mountain Lion)	x64 compatible architecture	genespring_mac.zip
Apple Mac OS X v10.7.3 (Lion)	x64 compatible architecture	genespring_mac.zip

Table 1.4: Mac OS X Platform Compatibility

Chapter 2

GeneSpring Quick Tour

2.1 Introduction

This chapter gives a brief introduction to **GeneSpring**, explains the terminology used to refer to various organizational elements in the user interface, and provides a high-level overview of the data and analysis paradigms available in the application. The description here assumes that **GeneSpring** has already been installed and activated properly. To install and get **GeneSpring** activated, see **GeneSpring** Installation.

2.2 Launching GeneSpring

To launch **GeneSpring**, you should have activated your license and your license must be valid. Launch the tool from the start menu or the desktop icon. On first launch, **GeneSpring** opens up with the demo project. On subsequent launches, the tool is initialized and shows a startup dialog. This dialog allows you to create a new project, open an existing project or open a recent project from the drop-down list. If you do not want the startup dialog uncheck the box on the dialog. You can restore the startup dialog by going to *Tools* $\longrightarrow$ *Options* $\longrightarrow$ *Miscellaneous* $\longrightarrow$ *Startup Dialog*

2.3 GeneSpring User Interface

The layout of **GeneSpring** is shown in Figure 2.1.

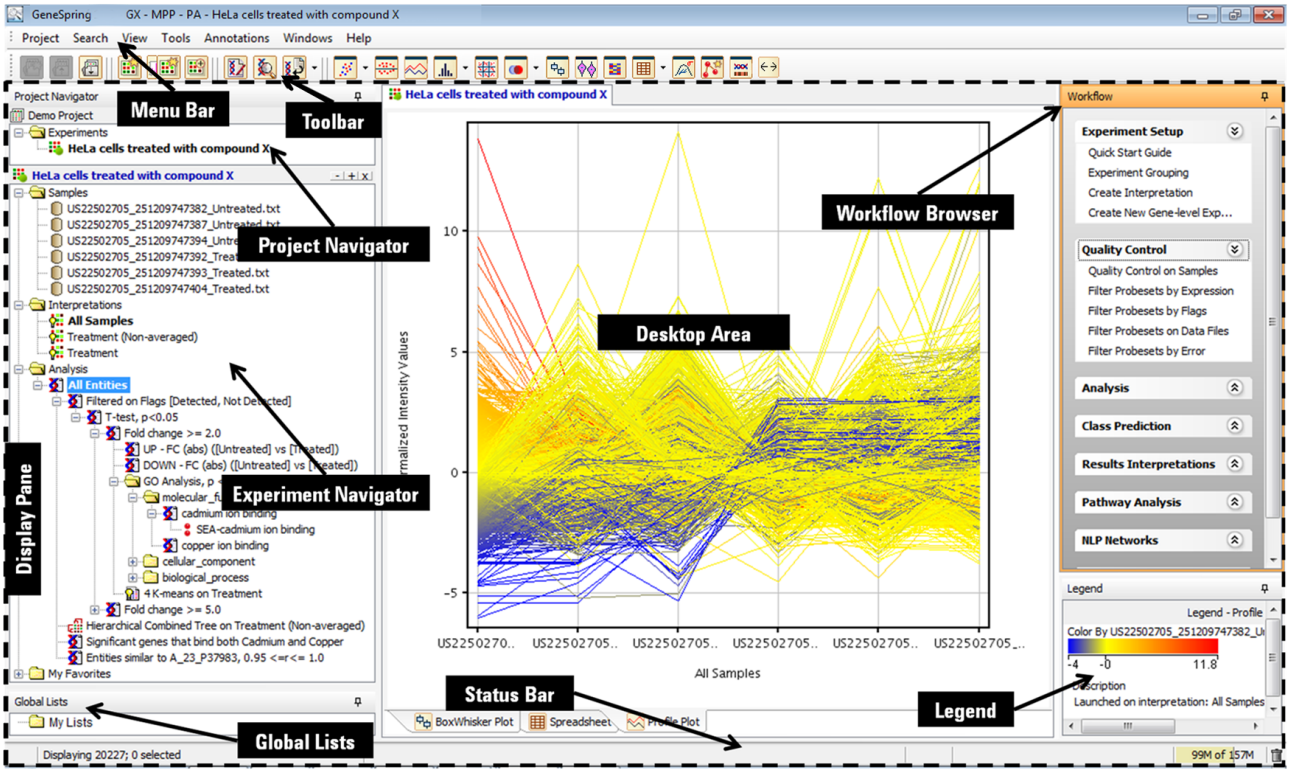


Figure 2.1: **GeneSpring** Layout

The main window consists of four parts - the Menu Bar, the Toolbar, the Display Pane and the Status Bar. The Display Pane contains several graphical views of the dataset, as well as algorithm results. The Display Pane is divided into three parts:

- The main **GeneSpring** Desktop in the center,
- The Project Navigator and the Experiment Navigator on the left,
- The **GeneSpring** Workflow Browser, and the Legend Window on the right.

2.3.1 GeneSpring Desktop

The desktop accommodates all the views pertaining to each experiment loaded in **GeneSpring**. Each window can be manipulated independently to control its size. Less important windows can be minimized. Windows can be tiled or cascaded in the desktop using the *Windows* menu. One of the views in the desktop is the active view.

2.3.2 Project Navigator

The Project Navigator on the left displays the project and all the experiments within it. Each experiment has its own navigator windows. The Project Navigator window shows all the experiments in the project. The experiment navigator window shows by default a **Samples** folder, an **Interpretation** folder, an **Analysis** folder and a **My Favorites** folder. The **My Favorites** folder can be populated with entity lists, Hierarchical trees, pathways or any other analysis objects that have been generated within the experiment., by copying (***Right-Click*** $\longrightarrow$ ***Copy***) and pasting it (***Right-Click*** $\longrightarrow$ ***Paste***), on to the appropriate sub-folder of ***My Favorites***. New sub-folders can be created by going to ***My Favorites*** $\longrightarrow$ ***Right-Click*** $\longrightarrow$ ***New Folder***.

2.3.3 The Workflow Browser

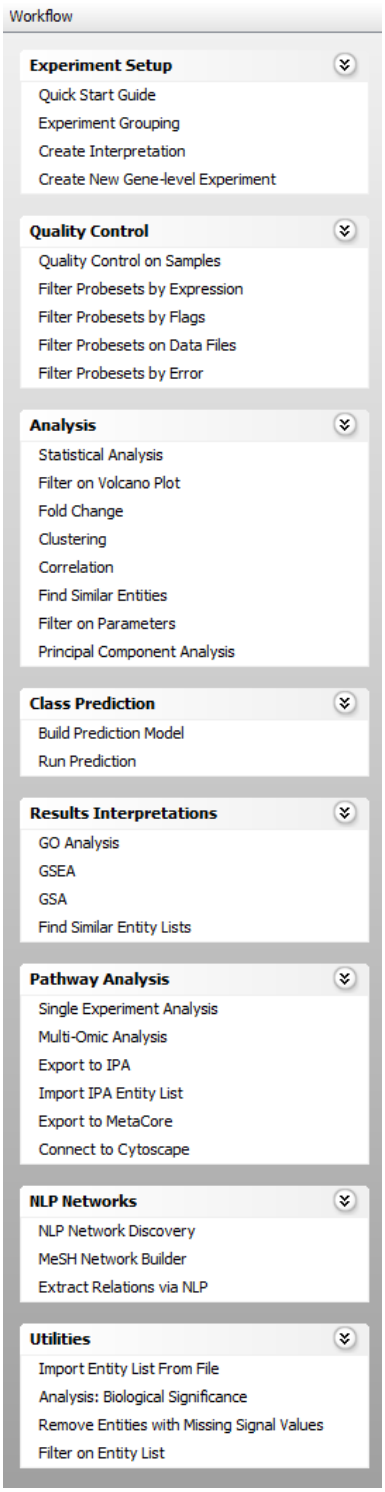


Figure 2.2: The Workflow Browser

The Workflow Browser shows the list of operations available in the experiment. It is organized into sequential groups of operations to help in the analysis of microarray data. The links in the Workflow Browser change according to the type of experiment being analyzed.

2.3.4 My Lists

My Lists in **Global Items** enables users to tag entity lists as public and make it available across projects. This is different from the 'My favourites' utility under the experiment navigator, as Global Items appear across projects while 'My favourites' appear under each experiment in a particular project only.

See Section Entity List for details on operations possible on the 'My Lists in 'Global Items'.

2.3.5 The Legend Window

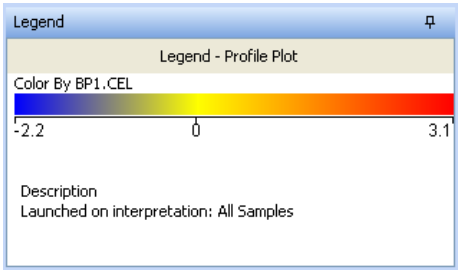


Figure 2.3: The Legend

The Legend window shows the legend for the current view in focus. Right-Click on the legend window shows options to *Copy* or *Export* the legend. Copying the legend will copy it to the Windows clipboard enabling pasting into any other Windows application using Control-V. Export will enable saving the legend as an image in one of the standard image file formats.

2.3.6 Status Bar

The status bar is divided into four informative areas as depicted below. See Figure 2.4

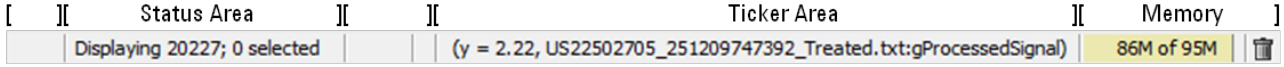


Figure 2.4: The Status Bar

Status Area This area displays high-level information about the current view. If a view is

selection enabled, the status area shows the total number of rows or columns displayed and the number of entities / conditions selected. If the view is limited to selection, it will show that the view is limited to selection.

Ticker Area This area displays transient messages about the current view based upon the cursor location, eg., indicates the row and column indices for a spreadsheet or the X-Y co-ordinates of a scatter plot.

Memory This displays the total memory allocated to the Java process and the amount of memory currently used. You can clear memory running the Garbage Collector by Left-Click on the *Garbage Can* icon on the left. This will reduce the memory currently used by the tool.

2.4 Organizational Elements and Terminology in GeneSpring

Work in **GeneSpring** is organized into *projects*. A project comprises one or more related *experiments*. An experiment comprises *samples* (i.e., data sources), *interpretations* (i.e., groupings of samples based on experimental parameters), and *analyses* (i.e., statistical steps and associated results, typically entity lists). Statistical steps and methods of analysis are driven by a *workflow* which finds prominent mention on the right side of **GeneSpring**. These concepts are expanded below.

2.4.1 Project

A project is the key organizational element in **GeneSpring**. It is a container for a collection of experiments. For instance, researcher John might have a project on Lung Cancer. As part of this project, John might run several experiments. One experiment measures gene expression profiles of individuals with and without lung cancer, and one experiment measures the gene expression profiles of lung cancer patients treated with various new drug candidates. A single ''Lung Cancer'' project comprises both of these experiments. The ability to combine experiments into a project in **GeneSpring** allows for easy interrogation of ''cross-experimental facts'', e.g., how do genes which are differentially expressed in individuals with lung cancer react to a particular drug.

A project can be created , viewed, deleted or closed using the following commands:

- ***Project***$\longrightarrow$ ***New Project*** creates a new project where the name and user notes can be specified.
- ***Project***$\longrightarrow$ ***Open Project*** opens an already created project.
- ***Project***$\longrightarrow$ ***Recent Projects*** allows access to recently opened projects.
- ***Project***$\longrightarrow$ ***Close Project***
- ***Project***$\longrightarrow$ ***Delete Project*** deletes all the experiments and samples contained within the project.

You can export out Projects using ***Project*⟶ *Export Project*** (See Export Project). Similarly, you can import projects using ***Import Project*** (See Import Project).

A project could have multiple experiments that are run on different technology types, and possibly different organisms as well.

2.4.2 Experiment

An experiment in **GeneSpring** represents a collection of samples for which arrays have been run in order to answer a specific scientific question. A new experiment is created from *Project* ⟶ *New Experiment* by loading samples of a particular technology and performing a set of customary pre-processing steps like, normalization, summarization, baseline transform, etc., that will convert the raw data to a state where it is ready for analysis. An already created experiment can be opened and added to the open project from *Project* ⟶ *Add Experiment*.

A **GeneSpring** project could have many experiments. You can choose to selectively open/close each experiment. Each open experiment has its own section in the Navigator. **GeneSpring** allows exactly one of the open experiments to be *active* at any given point in time. The name of the active experiment is reflected in the title bar of the **GeneSpring** application. Also, the active experiment is highlighted with a broad orange line, letters in purple and a green icon.

An experiment consists of multiple samples, with which it was created, multiple interpretations, which group these samples by user-defined experimental parameters, and all other objects created as a result of various analysis steps in the experiment.

2.4.2.1 Datasets from GEO

Users can import datasets obtained from Gene Expression Omnibus (GEO) from `http://www.ncbi.nlm.nih.gov/geo/` into **GeneSpring**. Only expression datasets will be supported; exon, splicing data will not be imported. During the import, if **GeneSpring** detects that the technology of the datasets are not part of the standard technologies supported within the product, a message will be shown that "Unable to create experiment because matching Technology is not available. Do you want to import it as a Generic Experiment?" The user can then decide to create a generic experiment with such datasets.

Chapter Loading Experiment from NCBI GEO gives details on how to work with datasets from GEO.

2.4.2.2 MOA Experiments

GeneSpring allows you to create multi-omic analysis (MOA) experiments that contain interpretations and entity lists from two different experiment types in order to perform one of the multi-omic analyses enabled in **GeneSpring**, Pathway Analysis and Correlation Analysis. Unlike

in an ordinary experiment, the results of either of these analyses are saved in dedicated folders in chronological order, rather than as a child of the entity list that was selected for analysis. Translation of entities is enabled for the entity lists in MOA experiments (see Translation on page 95, but functionality accessible from the right-click menus, the Tool bar, and the Workflow Browser is limited.

See "Multi-Omic Analysis in **GeneSpring** " on page 603 for more details.

2.4.2.3 Imported Strand NGS experiments

From **GeneSpring** version 13.0 onwards it is possible to import NGS experiments from the **Strand NGS** software into **GeneSpring** in order to perform multi-omic analyses with gene expression or metabolomics experiments. To be able to be imported into **GeneSpring**, the experiments from **Strand NGS** have to be exported using the "Export Experiment for **GeneSpring**" option available on right-click over any **Strand NGS** experiment in the Experiment Navigator (except Alignment experiments). This type of exported **Strand NGS** experiment contains:

- user-selected read lists, region lists, and data sets,
- all sample names, experiment grouping and interpretation information,
- all the entity lists and their associated data from analyses like SNP Effect Analysis, for example.

Once imported into **GeneSpring** using the **Import Strand NGS Experiment** option from the Project menu, the following analysis and visualization options are available for these **Strand NGS** experiments:

- Multi-Omic Pathway Analysis and Correlation Analysis (see "Multi-Omic Analysis in **GeneSpring**" on page 603),
- Creation of new experiment groupings and interpretations,
- Venn Diagram on pathway lists,
- Viewing of the imported read and region lists in the Genome Browser.

2.4.3 Sample

An experiment comprises a collection of samples. **GeneSpring** differentiates between a data file and a sample. A data file refers to the hybridization data obtained from a scanner. A sample on the other hand, is created within **GeneSpring**, when it associates the data files with its appropriate technology. Thus, when an experiment is created with the raw hybridization data files, they get registered as samples of the appropriate technology in **GeneSpring**. Once registered, samples are available for use in other experiments as well. Thus an experiment can be created with new raw data files as well as samples already registered and available with **GeneSpring**.

2.4.4 Experiment Grouping, Parameters and Parameter Values

Samples in an experiment have associated experiment parameters and corresponding parameter values. For instance, if an experiment contains 6 samples, 3 treated with Drug X and 3 not treated, you would have one experimental parameter which you could call ''Treatment Type''. Each sample needs to be given a value for this parameter. So you could call the 3 no treatment samples ''Control'' and the 3 treated samples ''Drug X''. ''Treatment Type'' is the experimental parameter and ''Control''/''Drug X'' are the values for this parameter.

An experiment can be defined by multiple experimental parameters. For instance, the samples could be divided into males and females, and each of these could have ages 1, 2, 5 etc. With this experimental design, there would be 2 experimental parameters, ''Gender'' and ''Age''. ''Gender'' takes values ''male'' and ''female'' and ''Age'' takes the values ''1'', ''2'' etc.

Experimental parameters and values can be assigned to each sample from the Experiment Grouping link in the workflow browser. These can either be entered manually, or can be imported from a text file, or can be imported from sample attributes. Once these values are provided, you could also shift the parameters from left to right and vice versa. Parameter values within each parameter can also be ordered. All views in **GeneSpring** will automatically reflect this order. Suppose you have experimental parameters ''Gender'' and ''Age'' and you want your profile plots to show all females first and then all males. Furthermore you would like all females to appear in order of increasing age from left to right and likewise for males. To achieve this, you will need to do the following. First, order the experimental parameters so ''Gender'' comes first and ''Age'' comes next. Then order the parameter values for parameter ''Gender,'' so ''Female'' comes first and ''Male'' comes next. Finally, order the parameter values for parameter ''Age'' so that these are in increasing numeric order.

2.4.5 Conditions and Interpretations

An interpretation defines a particular way of grouping samples into experimental conditions for both data visualization and analysis. When a new experiment is created, **GeneSpring** automatically creates a default interpretation for the experiment called ''All Samples''. This interpretation just includes all the samples that were used in the creation of the experiment. New interpretations can be created using the ''Create New Interpretation'' link in the workflow browser. Once a new interpretation is created, the interpretation will be added to the Interpretations folder within the Navigator.

First, identify the experimental parameters by which you wish to group samples. **GeneSpring** will now show you a list of conditions that would result from such grouping. For example, if you choose two parameters, ''Gender'' and ''Age'', and each sample is associated with parameter values Female or Male, and Young or Old, **GeneSpring** will take all unique combinations of parameter values to create the following conditions: Female,Old; Female,Young; Male,Old; and Male,Young. Samples that have the same Gender and Age values will be grouped in the same experimental condition. Samples within the same experimental conditions are referred to as ''replicates''.

You can choose to ignore certain conditions in the creation of an interpretation. Thus, if you want to analyze only the conditions Female,Old and Female,Young, you can do that by excluding the conditions Male,Old and Male,Young in the creation of the interpretation.

GeneSpring creates both interpretations by default. In the averaged interpretation, the mean intensity value for each entity across the replicates will be used for display and for analysis when the interpretation is chosen. In the non-averaged interpretation, the intensity value for each entity in each sample will be used for display and for analysis when the interpretation is chosen.

Every open experiment has one active interpretation at any given point in time. The active interpretation of each experiment is shown in **bold** in the navigator for that experiment. By default, when an experiment is opened, the ''All Samples'' interpretation shows active. You can make a different interpretation active, by simply clicking on it in the Navigator. Invoking a view from the View menu will open the view and automatically customize it to the current active interpretation wherever applicable. Most steps in the Workflow browser also take the active interpretation as default and automatically customize analysis to this interpretation, wherever applicable.

An interpretation can be visualized graphically by double-clicking on it. This will launch a profile plot which shows expression profiles corresponding to the chosen interpretation, i.e., the x-axis shows conditions in the interpretation ordered based on the ordering of parameters and parameter values provided in the Experiment Grouping. If the parameter values in the interpretation are numerical, then the tool attempts to show a Time Series plot (see section Time Series Plot for details).

2.4.5.1 Interpretations and Views

Most views in **GeneSpring** change their behavior depending on the current active interpretation of the experiment. The table below lists these changes. Refer Table 2.1.

2.4.5.2 Interpretations and Workflow Operations

Most of the analysis steps in the workflow browser depend on the current active interpretation of the experiment. These dependencies are tabulated below. The steps not mentioned in the table do not depend on the active interpretation. Refer Table 2.2.

2.4.5.3 Changes in Experiment Grouping and Impact on Interpretations

Note that Experiment Grouping can change via creation of new parameters or edits/deletions of existing parameters and parameter values. Such changes made to Experiment Grouping will have an impact on already-created interpretations. The following cases arise:

- Deleting a parameter: If all parameters used in an interpretation have been subsequently deleted, or even renamed, the interpretation's behavior defaults to that of the ''All Samples'' interpretation. If however, only a part of the parameters used in an interpretation have been changed, for e.g., if an interpretation uses parameters Gender and Age, and say, Age has been deleted, then the interpretation behaves as if it was built using only the Gender

View	Behavior on active Interpretation
Scatter Plot Matrix Plot Histogram	Axes show only conditions in this interpretation for averaged interpretations, and individual samples for each condition in the interpretation, for non-averaged interpretations.
Profile Plot Box Whisker Plot	Axes show only conditions in this interpretation for averaged interpretations, and individual samples for each condition in the interpretation, for non-averaged interpretations. Parameter markings are shown on the x-axis.
Venn Diagram	Interpretation does not apply.
Spreadsheet Heat Map	Columns show only conditions in this interpretation for averaged interpretations, and individual samples for each condition in the interpretation, for non-averaged interpretations.
Entity Trees	When constructing entity trees, only conditions in this interpretation are considered for averaged interpretations, and individual samples for each condition in this interpretation are considered for non-averaged interpretations. When double-clicking on an entity tree object in the Navigator, the conditions corresponding to the current interpretation show in the tree.
Condition Trees	When constructing condition trees, only conditions in the chosen interpretation are considered for averaged interpretations, and individual samples for each condition in this interpretation are considered for non-averaged interpretations. When double-clicking on a condition tree object in the Navigator, the current interpretation is ignored and the view launches with the interpretation used when constructing the tree. If the conditions of the original interpretation and their associated samples are no longer valid, a warning message to that effect will be shown.
Entity Classification	When constructing entity classifications, only conditions in a chosen interpretation are considered for averaged interpretations, and individual samples for each condition in this interpretation are considered for non-averaged interpretations. When double-clicking on an entity classification object in the Navigator, the columns corresponding to the current interpretation show up.

Table 2.1: Interpretations and Views

parameter. If the interpretation had any excluded conditions, they are now ignored. If at a later stage, the Age parameter is restored, the interpretation will again start functioning the way it did when it was first created.

- Change in parameter order: The order of parameters relative to each other can be changed from the Experiment Grouping workflow step. If for e.g., Age is ordered before Gender, then the conditions of an interpretation which includes both Gender and Age, will automatically become Old,Female; Young,Female; Old,Male and Young,Male.

- Deleting a parameter value: The interpretation only maintains the conditions that it needs to exclude. So, if for example, the parameter value Young is changed to Adolescent, an interpretation on the parameter Age without any excluded conditions will have Adolescent and Old as its conditions. Another interpretation on the parameter Age, that excluded the condition Young will also have as its new conditions - Adolescent and Old.

- Change in order of parameter values: If the order of parameter values is changed, the conditions of the interpretation are also accordingly re-ordered. Thus for parameter Age, if value Young is ordered before Old, the conditions of an interpretation with both Gender and Age, will likewise become Female,Young; Female,Old; Male,Young and Male,Old.

Workflow Step	Action on Interpretation
Filter probesets by Expression	Runs on all samples involved in all the conditions in the chosen interpretation; averaging is ignored except for purposes of showing the profile plot after the operation finishes. If the parameter values are numerical, the tool attempts to show a Time series plot.
Filter probesets by Flags	Runs on all samples involved in all the conditions in the chosen interpretation; averaging is ignored except for purposes of showing the profile plot after the operation finishes. If the parameter values are numerical, the tool attempts to show a Time series plot.
Statistical Analysis	The statistical test options shown depend on the interpretation selected. For instance, if the selected interpretation has only one parameter and two conditions then a T-Test option is shown, if the selected interpretation has only one parameter and many conditions then an ANOVA option is shown, and if the selected interpretation has more than one parameter then a multi-way ANOVA is run; averaging in the interpretation is ignored.
Fold Change	All conditions involved in the chosen interpretation are shown and the user can choose which pairs to find fold change between; averaging in the interpretation is ignored.
GSEA	All conditions involved in the chosen interpretation are shown and the user can choose which pairs to perform GSEA on; averaging in the interpretation is ignored.
Clustering	Only conditions in this interpretation are considered for averaged interpretations, and individual samples for each condition in this interpretation are considered for non-averaged interpretations.
Find Similar Entities	Only conditions in this interpretation are considered for averaged interpretations, and individual samples for each condition in this interpretation are considered for non-averaged interpretations.
Filter on Parameters	All samples involved in conditions in the chosen interpretation are considered irrespective of whether or not the interpretation is an averaged one. Next, the parameter to be matched is restricted to values on only these samples. Once the calculations have been performed, entities passing the threshold are displayed in a Profile plot (or Time Series Plot) that reflects the chosen interpretation.
Build Prediction Model	All conditions involved in the chosen interpretation are used as class labels for building a model; averaging in the interpretation is ignored.

Table 2.2: Interpretations and Workflow Operations

The key point to note is that an interpretation internally only maintains the names of the parameters that it was created with and the conditions that were excluded from it. Based on any changes in the Experiment Grouping, it logically recalculates the set of conditions it represents.

2.4.6 Entity List

An Entity List comprises a subset of entities (i.e., genes, exons, genomic regions, etc.) associated with a particular technology. When a new experiment is created, **GeneSpring** automatically creates a default entity list called the ‘‘All Entities’’ entity list. This entity list includes all the entities that the experiment was created with. In most cases, all entities present in the samples loaded into the experiment will also be the same as the entities of the technology associated with the samples. In the case of an Exon Expression experiment however, it contains the Core/Full/Extended

transcript cluster IDs depending on which option was chosen to create the experiment. Entity list cannot appear in a pathway experiment.

New entity lists are typically created in **GeneSpring** as a result of analysis steps like ''Filter probesets by Flags'' for example. One could also manually create a new entity list by selecting a set of entities in any of the views and then using the *Create Entity List* toolbar button. Note that entities selected in one view will also show selected in all other views as well.

Existing entity lists can be added to a non-pathway experiment of the same technology via ***Search*⟶ *Entity Lists***. Please note that entity lists cannot be cut and pasted across experiments.

Entity lists can be translated implicitly across experiments with possibly differing technologies. Implicit translation happens when you click on an entity list in the analysis navigator of an experiment which is not currently active. Data views in the currently active experiment are restricted to the entities in the above list, after translation is performed silently behind the scenes, possibly via a homologene cross organism map using Entrez gene ID to go across.

This implicit translation works across most experiment types with two notable exceptions.

1. Translation into pathway experiments is not performed
2. Translation of miRNA lists into non-miRNA experiments is not allowed. Translation of gene lists from non-miRNA experiments to miRNA experiments happens implicitly but will lead to nothing being visible. TargetScan translated genelist in miRNA experiments do participate in implicit translation though.

Entity lists alone can be translated explicitly across experiments with possibly differing technologies; this happens on the entity list right click menu and results in a new list. The same exceptions apply as for implicit translation. For more details on how this is executed, refer to the section on Translation.

Every open project only has one active entity list at any given point in time. When an experiment of the project is opened, the ''All Entities'' entity list of that experiment becomes the active entity list of the project. You can make a different entity list active, simply by clicking on it in the Navigator. The user experience key to **GeneSpring** is the fact that clicking on an entity list restricts all open views to just the entities in that list, making for fast exploration.

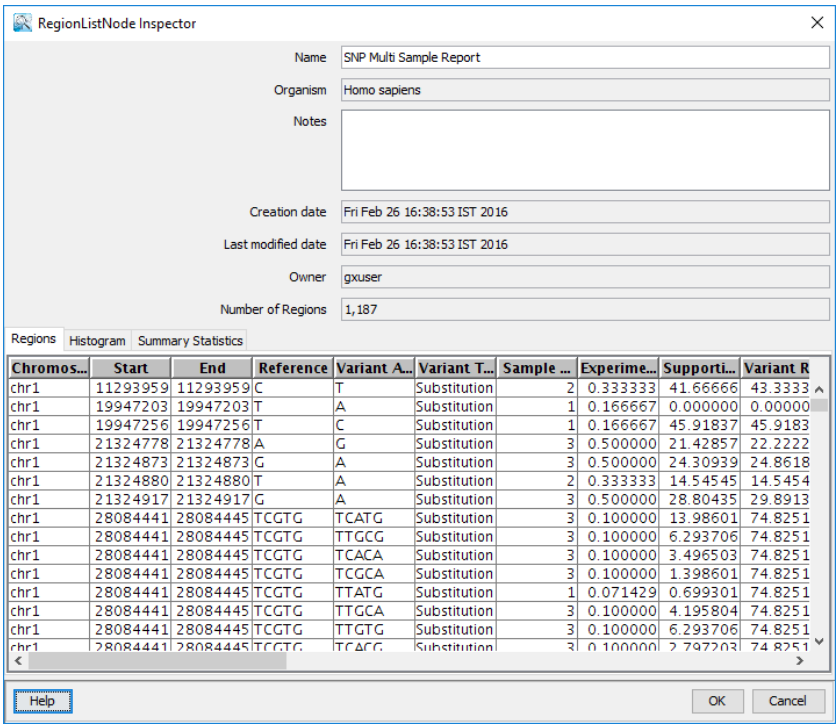
Any entity list in **GeneSpring** can be made universally available by right clicking on that list in the navigator and clicking the option 'Mark as Global List'. This list would then appear under **My Lists** in **Global Items** and will be available across experiments in a project and across projects too. Actions possible on 'My Lists' are:

1. Highlight List - Makes it active entity list and does the translation into the active experiment (This option is only available only when the active experiment is of the same Analysis type).
2. Inspect List - Brings up Entity List Inspector.
3. Export List - Brings up a window to enable exporting the entity list to a local folder.

4. Translate List - Within the same project or across projects. Obeys general rules of translation.
5. Share List - Active only for **WorkGroupPlus** version of **GeneSpring**.
6. Unmark as Global List - Will remove the list from the **My Lists** in **Global Items**.

2.4.7 Region List

The **Region list** is a list of chromosome, start, end, strand and other annotation information, such as zygocity, score, type, etc. Region Lists are present for imported NGS and CGH experiments.



Chromos...	Start	End	Reference	Variant A...	Variant T...	Sample ...	Experime...	Supporti...	Variant R
chr1	11293959	11293959	C	T	Substitution	2	0.333333	41.66666	43.3333
chr1	19947203	19947203	T	A	Substitution	1	0.166667	0.000000	0.00000
chr1	19947256	19947256	T	C	Substitution	1	0.166667	45.91837	45.9183
chr1	21324778	21324778	A	G	Substitution	3	0.500000	21.42857	22.2222
chr1	21324873	21324873	G	A	Substitution	3	0.500000	24.30939	24.8618
chr1	21324880	21324880	T	A	Substitution	2	0.333333	14.54545	14.5454
chr1	21324917	21324917	G	A	Substitution	3	0.500000	28.80435	29.8913
chr1	28084441	28084445	TCGTG	TCATG	Substitution	3	0.100000	13.98601	74.8251
chr1	28084441	28084445	TCGTG	TTGCG	Substitution	3	0.100000	6.293706	74.8251
chr1	28084441	28084445	TCGTG	TCACA	Substitution	3	0.100000	3.496503	74.8251
chr1	28084441	28084445	TCGTG	TCGCA	Substitution	3	0.100000	1.398601	74.8251
chr1	28084441	28084445	TCGTG	TTATG	Substitution	1	0.071429	0.699301	74.8251
chr1	28084441	28084445	TCGTG	TTGCA	Substitution	3	0.100000	4.195804	74.8251
chr1	28084441	28084445	TCGTG	TTGTG	Substitution	3	0.100000	6.293706	74.8251
chr1	28084441	28084445	TCGTG	TCACG	Substitution	3	0.100000	2.797203	74.8251

Figure 2.5: A region list of an imported NGS experiment

The Region List Inspector shows some information, such as name (editable), organism, creation and modification date, and number of regions. The Regions section contains the actual location information, with chromosome, start and end coordinates, in base pairs, counted from base 1 of the chromosome.

Many algorithms also calculate a score, P value or other information for each region that is also shown in the tables. Which extra columns are saved with the region is dependent on the analysis step, but **chromosome**, **start** and **end** are always part of a region.

Region lists can be displayed in the Genome Browser and can also be compared to each other. These Region list objects can also be exported and imported from an external source (BED files).

Export & Import a region list: A region list can be exported as a *.tsv, *.bed, or *.fa file. To export a region list, select **Export Region List** from the right-click menu option on the selected region list.

A region list can also be imported into the experiment, by selecting the **Import Region List** workflow step from the **Utilities** section.

2.4.8 Report

In **GeneSpring**, you can export plot views and pathways including matched pathway lists as a Report. A report can be downloaded as a *.pdf* file or saved locally or on the cloud as a *.gsreport* file. Reports can be created to include;

- legend
- annotation columns to include experiment data.
- customized paper size
- customized page orientation
- customized page margins

Multiple reports can me merged to create a single report. For a report, the maximum number of pages to be displayed in **GeneSpring** can be modified from **Tools** $\longrightarrow$**Options** $\longrightarrow$**Miscellaneous** $\longrightarrow$**Report**. If a report contains more than the specified number of pages, you can be download it as a *.pdf* to the local system.

Any report in **GeneSpring** can be made universally available by right clicking on that report in the navigator and clicking the option 'Mark as Global Report'. This report would then appear under **My Reports** in **Global Items** and will be available across experiments in a project and across projects too. A report marked as a Global Report is displayed in Yellow. Actions possible on 'My Reports' are:

1. Inspect Report - Opens the Report Inspector.
2. Download Report - Allows you to download the report as **.pdf* file to a local folder.
3. Export Report - Use this option to enable exporting the report as **.gsreport* to a local folder. This report can be imported back into **GeneSpring**.
4. Export Report to DropBox - Use this option to export a report to Dropbox.
5. Export Report to GoogleDrive - Use this option to export a report to Google Drive.
6. Share Report - This operation is disabled in the desktop mode of **GeneSpring**. In the **WorkGroupPlus** mode this operation can be used to share the report with other users of **GeneSpring WorkGroupPlus**.

7. Unmark as Global Report - Removes the report from the folder **My Reports** in **Global Items**.

In **Global Items** $\longrightarrow$**My Reports**, you can merge a minimum of two or more reports by;

1. **Ctrl-select** the reports.
2. Right-click $\longrightarrow$**Merge Reports**.

Merged reports will be saved in **Global Items** $\longrightarrow$**My Reports** folder.

2.4.9 Entity Tree, Condition Tree, Combined Tree and Classification

Clustering methods are used to identify co-regulated genes. Trees and classifications are the result of clustering algorithms. All clustering algorithms require a choice of an entity list and an interpretation, and allow for clustering on entities, conditions or both.

Performing hierarchical clustering on entities results in an entity tree, on conditions results in a condition tree and on both entities and conditions results in a combined tree. Performing KMeans or SOM on entities results in a classification, on conditions results in a condition tree, and on both entities and conditions result in a classification and condition tree.

A classification is just a collection of disjoint entity lists. Double-clicking on a classification from the navigator results in the current active view to be split up based on the entity lists of the classification. If the active view does not support splitting up, for e.g., if it is already split, or if it is a Venn Diagram view, etc., then the classification is displayed using split up profile plot views. The classification is displayed according to the conditions in the active interpretation of the experiment. A classification can also be expanded into its constituent entity lists, by right-clicking on the classification and using the *Expand as Entity list* menu item.

Double-clicking on the trees will launch the dendrogram view for the corresponding tree. For entity trees, the view will show all the entities and the corresponding tree, while the columns shown will correspond to the conditions in the active interpretation. For condition trees and combined trees, the same tree as was created will be reproduced in the view. However, it may be that the conditions associated with the samples of the tree are now different, due to changes in the experiment grouping. In this case a warning message will be shown. If any of the samples that were used to create the tree are no longer present in the experiment, after performing a Add/Remove Samples operation for e.g., then an error message will be shown and the tree cannot be launched.

Refer to chapter Clustering Analysis for details on clustering algorithms.

2.4.10 Class Prediction Model

Class prediction methods are typically used to build prognostics for disease identification. For instance, given a collection of normal samples and tumor samples with associated expression data, **GeneSpring** can identify expression signatures and use these to predict whether a new unknown sample is of the tumor or normal type. Extending this concept to classifying different types of possibly similar tumors, class prediction provides a powerful tool for early identification and tailored treatment.

Running class prediction involves three steps, validation, training and prediction. The process of learning expression signatures from data automatically is called training. Clearly, training requires a dataset in which class labels of the various samples are known. Performing statistical validation on these signatures to cull out signal from noise is called validation. Once validated these signatures can be used for prediction on new samples.

GeneSpring supports four different class prediction algorithms namely, Decision Tree, Neural Network, Support Vector Machine and Naive Bayes. These can be accessed from the ''Build Prediction Model'' workflow step. Each of these algorithms create a class prediction model at the end of the training. These models can be used for prediction on a potentially different experiment using the ''Run Prediction'' workflow step.

Refer to chapter Class Prediction: Learning and Predicting Outcomes for details on the class prediction algorithms.

2.4.11 Script

Python and R scripts can be created and saved in **GeneSpring** for performing custom tasks and to easily add and enhance features.

To create a new python script, launch the *Tools* $\longrightarrow$ *Script Editor*, refer the chapter Writing Scripts in **GeneSpring** on scripting to implement the script, and then save the script using the Save button on the toolbar of the Script Editor. This script can later be invoked on a potentially different experiment by launching a new Script Editor and clicking on the Open toolbar button to search for all existing scripts and load the already saved script.

R scripts can be created and saved similarly using the *Tools* $\longrightarrow$ *R Editor*. Refer to the chapter Writing Scripts in **GeneSpring** on R scripts for details on the R API provided by **GeneSpring**.

2.4.12 Pathway

In **GeneSpring**, you can visualize your experimental data on imported curated pathways and interaction networks, or create your own interaction networks using **GeneSpring**'s proprietary Natural Language Processing (NLP) algorithm and searches based on Medical Subject Headings (MeSH) terms. You can import curated pathways from the WikiPathways [46] portal (`http:`

//www.wikipathways.org), the BioCyc Pathway/Genome databases (http://www.biocyc.org/) provided by the Bioinformatics Research Group at SRI International® (used under license) [9], the KEGG (Kyoto Encyclopedia of Genes and Genomes) Pathway Databases (http://www.genome.jp/kegg/pathway.html) developed by Kanehisa Laboratories® (can be used with a valid end user license) [37], in the BioPAX (Level 2 and Level 3) format, the GPML format, or in a Text format.

Once pathways have been imported into **GeneSpring**, or created in **GeneSpring**, they can be added to different experiments to highlight matching experimental data in the pathway. Matching entities between the pathway and the active entity list in the Experiment Navigator are highlighted by a blue border. You can also perform Multi-Omic Analysis (MOA) or Single Experiment Analysis (SEA) to find significant pathways for your research. Both SEA and MOA allow you to compare your data with pathways from different organisms than your experiment. MOA, as the name suggests, even allows you to compare data from different omics experiments (for example Expression Analysis and Metabolomics) to find significant pathways.

In addition, **GeneSpring** allows you to connect to other popular pathway analysis tools like MetaCore, Ingenuity Pathways Analysis (IPA), or Cytoscape to further analyze your data.

You can find more information on how to get a license for the Pathway Analysis Module and all the available features under the **Pathway Analysis** and **NLP Networks** chapters in this manual.

2.4.13 Inspectors

All the objects mentioned above have associated properties. Some properties are generic like the name, date of creation and some creation notes, while others are specific to the object, e.g., entities in an entity list. The inspectors of the various objects can be used to view the important properties of the object or to change the set of editable properties associated with the object like Name, Notes, etc.

- The **Project Inspector** is accessible from *Project* $\longrightarrow$ *Inspect Project* and shows a snapshot of the experiments contained in the project along with their notes.
- The **Experiment Inspector** is accessible by right-clicking on the experiment and shows a snapshot of the samples contained in the experiment and the associated experiment grouping. It also has the notes that detail the pre-processing steps performed as part of the experiment creation.
- The **Sample Inspector** is accessible by double-clicking on the sample in the navigator or by right-clicking on the sample. It shows the experiment the sample belongs to, the sample attributes, attachments and parameters and parameter values from all experiments that it is part of. The name and parameters information associated with the sample are uneditable. Sample attributes can be added/changed/deleted from the inspector, as also the attachments to the sample.
- The **Technology Inspector** is accessible by right-clicking on the experiment and shows a snapshot of all the entities that belong to the technology. None of the properties of the technology inspector are editable. The set of annotations associated with the entities can be customized using the ''Configure Columns'' button, and can also be searched for using the search bar at the bottom. Further hyper-linked annotations can be double-clicked to launch a web browser with further details on the entity.

- The **Entity List Inspector** is accessible by double-clicking on the entity list in the navigator or right-clicking on the entity list. It shows the entities associated with the list, and user attributes if any. It also shows the technology of the entity list and the experiments that it belongs to. The set of displayed annotations associated with the entities can be customized using the ''Configure Columns'' button, and can also be searched for using the search bar at the bottom. Further, entities in the table can be double clicked to launch the Entity Inspector.
- The **Entity Inspector** is accessible in the following ways:
  - Double clicking on an entity in the entity list inspector described above
  - Double clicking on some of the views like Scatter plot, MvA plot, Profile plot, Heat map
  - Selecting an entity in any view and clicking on the toolbar button.
  - By using the key binding Ctrl-I or by using the menu *View* → *Inspect Entities*

  The entity inspector window shows the Id and the technology relevant to the selected entity. The inspector also has tabs to view the following:

  1. *Annotation:* Lists annotations. The set of default annotations associated with the entity can be customized by using the ''Configure Columns'' button at the bottom.
  2. *Data:* Shows the raw and normalized data associated with the entity in all the samples of the experiment, along with the flag.
  3. *Box whisker plot:* With the normalized data under the current active interpretation
  4. *Profile Plot:* With the normalized data under the current active interpretation.
- Inspectors for Entity Trees, Condition Trees, Combined Trees, Classifications, Class Prediction Models are all accessible by double-clicking or right-clicking on the object in the navigator, and provide basic information about it. The name and notes of all these objects can be changed from the inspector.
- The **Report Inspector** is accessible by right-clicking on a report. The name of the report can be edited from this inspector. Notes have details of the experiment the report belongs to and the Entity List used to create the report. Notes cannot be edited for a report.
- The **Pathway List Clipboard Inspector** is accessible via the corresponding icon in the Pathway View. The inspector allows you to review, export, or delete all or a selection of the pathway entities previously added to the Pathway List Clipboard.

2.4.14 Hierarchy of objects

All the objects described above have an inherent notion of hierarchy amongst them. The project is right at the top of the hierarchy, and is a parent for one or more experiments. Each experiment is a parent for one or more samples, interpretations and entity lists. Each entity list could be a parent for other entity lists, trees, classifications, class prediction models, pathways, or folders containing some of these objects. The only exceptions to this hierarchy are technologies and scripts that do not have any parentage.

Additionally, many of these objects are *first class objects* that can exist without any parent. This includes experiments, entity lists, samples, class prediction models and pathways. Interpretations, trees and classifications, however cannot exist independently without their parents. Finally, the

independent objects can have more than one parent as well. Thus an experiment can belong to more than one project, samples can belong to more than one experiment and so on.

Note that in the case of independent objects, only those that do have a valid parent show up in the navigator. However all objects with or without parents show up in search results.

2.4.15 Right-click operations

Each of the objects that show up in the navigator have several right-click operations. For each object, one of the right-click operations is the default operation and shows in bold. This operation gets executed if you double-click on the object.

The set of common operations available on all objects include the following:

- Inspect object : Most of the objects have an inspector that displays some of the useful properties of the object. The inspector can be launched by right-clicking on the object and choosing the inspect object link.
- Share object : This operation is disabled in the desktop mode of **GeneSpring**. In the **WorkGroupPlus** mode, this operation can be used to share the object with other users of the **GeneSpring WorkGroupPlus**.
- Change owner : This operation is disabled in the desktop mode of **GeneSpring**. In the **WorkGroupPlus** mode, this operation can be used by a group administrator to change the owner of the object.

The other operations available on each of the objects are described below:

2.4.15.1 Experiment

- Open Experiment : (default operation) This operation opens the experiment in **GeneSpring**. Opening an experiment opens up the experiment navigator in the navigator section of **GeneSpring**. The navigator shows all the objects that belong to the experiment, and the desktop shows the views of the experiment. This operation is enabled only if the experiment is not already open.
- Close Experiment : This operation closes the experiment, and is enabled only if the experiment is already open.
- Inspect Technology : This operation opens up the inspector for the technology of the experiment.
- Create New Experiment : This operation can be used to create a copy of the chosen experiment. The experiment grouping information from the chosen experiment is carried forward to the new experiment. In the process of creating the copy, some of the samples can be removed, or extra samples can be added if desired.

- Remove Experiment : This operation removes the experiment from the project. Note that the remove operation only disassociates the experiment with this project. The experiment could still belong to other projects in the system, or it could even not belong to any project.
- Delete Experiment : This operation will permanently delete the experiment from the system. All the children of the experiment will also be permanently deleted, irrespective of whether they are used in other experiments or not. The only exception to this is samples. So, if an experiment contains ten samples, two of which are used in another experiment, this operation will result in deleting all the eight samples that belong only to this experiment. The remaining two samples will be left intact.

2.4.15.2 Sample

- Inspect Sample : (default operation) This will open up the inspector for the sample.
- Download Sample : This operation enables downloading the sample to a folder of choice on the local file system.

2.4.15.3 Samples Folder

- Add Attachments : This operation can be used to upload attachments to all the samples in the folder. Multiple files can be chosen to be added as attachments. **GeneSpring** checks the files to see if the name of any of the file (after stripping its extension) matches the name of any sample (after stripping its extension) and uploads that file as an attachment to that sample. Files that do not match this condition are ignored. Note that if a file without a matching name needs to be uploaded as an attachment, it can be done from the sample inspector.
- Add Attributes : This operation can be used to upload sample attributes for all the samples in the folder. **GeneSpring** expects a comma or tab separated file in the following tabular format. The first column of the file should be the name of the samples. All the remaining columns will be considered as sample attributes. The column header of each column is taken as the names of the sample attribute. Each cell in this tabular format is assigned as the value for the corresponding sample (row header) and sample attribute (column header).
- Download Samples : This operation can be used to download all the raw files of the samples in bulk to a folder of choice on the local filesystem.

2.4.15.4 Interpretation

- Open Interpretation : (default operation) This opens a profile plot view of the interpretation. If the chosen interpretation has numerical parameters, the tool attempts to show a Time series plot.
- Edit Interpretation : This allows for editing the interpretation. The parameters of the interpretation, conditions to exclude, name and notes can all be edited.
- Delete Interpretation : This operation deletes the interpretation from the experiment. Note that there is no notion of removing an interpretation, since an interpretation is not an independent object and always exists only within the experiment.

2.4.15.5 Entity List

- Highlight List : This operation restricts all the views in the experiment to the entities of the chosen list.
- Inspect List: This launches the entity list inspector. For more details, refer to Inspectors.
- Export List : This option can be used to export the entity list and associated data and annotations as a plain text file. You can choose an interpretation according to which the raw and normalized data will be exported, if chosen. If the experiment has flags, you can also choose to export the flags associated with the entities of this list. If the entity list has data associated with it as a result of the analysis using which the list was created, these can also be exported. Finally, you can also choose which annotations to export with the entity list.

NOTE If the active interpretation is an averaged interpretation, the option to include **Raw signal values** in export entity list is not available for Agilent Two Color Expression, Genric Two Color Expression, and Agilent Two Color Exon Expression experiments.

- Export List to Dropbox: Use this option to export an entity list to Dropbox (Refer 2.4.15.5.1 for further details).
- Export List to Google Drive: Use this option to export an entity list to Google Drive (Refer 2.4.15.5.1 for further details).
- Copy: This allows copying of the entity list into the **My Favorites** folder of the active experiment only.
- Copy genes to clipboard: This allows copying of GeneSymbol of all the entities in the entity list.
- Copy list to clipboard: This allows copying of the entity list with all the associated properties.
- Remove List : This operation removes the entity list from the experiment. Note that the remove operation only disassociates this entity list and all its children with the experiment, and does not actually delete the list or its children. The entity list and its children could still belong to other experiments in the system, or they may even exist independently without belonging to any experiment.
- Delete List : This operation will permanently delete the list and all its children from the system.

NOTE Remove List and Delete List options are disabled for 'All Entities' entity list.

- Mark as Global List: This option adds the selected entity list to the Global Lists folder.

- Export to eArray: Using this option, GenBank Accession data or Genomic co-ordinates data from an entity list of a Gene Expression experiment or imported Strand NGS experiment (Target enrichment and RNA enrichment) or region lists of an imported Strand NGS experiment can be uploaded to Agilent's eArray portal. These lists should contain either GenBank Accession numbers or Genomic Co-ordinates.

NOTE

- Data pertaining to only 10,000 entities from a single entity list can be exported to the eArray portal. If the data exceeds the given threshold, an error message is displayed.
- For any given entity, if the genomic co-ordinates have only either of the start or stop index, both start and stop values will be considered equal.
- Chromosomal information is also allowed in the format where the chromosome number of an entity has only number 'n' instead of the standard format of 'chrn'. Chromosome information can also be present in the format of 'chrXY' or 'chrX—Y' or 'XY' or 'X—Y'

2.4.15.5.1 Import to and Export of Entity List from cloud storage (Dropbox and Google Drive)

Dropbox

Dropbox is a file hosting service by Dropbox, Inc., that offers cloud storage.

To export an entity list to Dropbox;

- right-click on any entity list and select the option, **Export to Dropbox**.
- select interpretation, data columns, and annotation columns as required, and click **OK**.
- select a local folder to temporarily save the entity list.
- specify the Dropbox folder path to save the file on your Dropbox account.
- enable **Get the share link after export** option, for Dropbox to generate a link for the exported entity list. You can share this link with others, and they can either download the file using the link or add it to their Dropbox account.

NOTE

Export to Dropbox is dependent on your bandwidth and server traffic. The process of upload and export is in progress until the progress bar is visible. If you see an error, please try the export again.

To import an entity list from Dropbox:

- click on the Import Entity List from File icon on the toolbar, and from the drop-down, select **From Dropbox**.
- select the entity list that you would like to import and click **OK**.
- select technology and columns as required, and click **OK**.
- the entity list will be downloaded to the imported lists folder of your experiment.

Google Drive

Google Drive is a file storage service by Google, Inc., that offers cloud storage.

To export an entity list to Google Drive:

- right-click on any entity list and select the option, **Export to Google Drive**.
- select interpretation, data columns, and annotation columns as required, and click **OK**.
- select a local folder to temporarily save the entity list.
- specify the Google Drive folder path to save the file on your Google Drive account.
- enter email ids (only one email per line) of those you would like to share the entity list with. Using this link, they can download the entity list or add it to their Google Drive account.

NOTE Export to Google Drive is dependent on your bandwidth and server traffic. The process of upload and export is in progress until the progress bar is visible. If you see an error, please try the export again.

To import an entity list from Google Drive:

- Click on the Import Entity List from File icon on the toolbar, and from the drop-down, select **From Google Drive**.
- Select the entity list that you would like to import and click **OK**.
- Select technology and columns as required, and click **OK**.
- The entity list will be downloaded to the imported lists folder of your experiment.

NOTE In cloud storage, all the exported entity lists are saved as *.gsel* extension files. If this extension is altered, the entity lists become unavailable for import into **GeneSpring**. In cloud storage, if any other files are changed to have a *.gsel* extension, they do not become visible in the options for import.

2.4.15.6 Right-click options on a Region List

The available right-click options on a region list of an imported **Strand NGS** experiment or CGH experiments are summarized in Table 2.3.

Table 2.3: Menu options available on right-click over a region list

Option	Description
Spreadsheet View	Opens the Region List as a Spreadsheet
Inspect Region List	Opens the Region List Node Inspector.
Show in Genome Browser	Opens the region List in the Genome Browser.
Navigate in Genome Browser	Displays the regions of a region list as a spreadsheet in the Navigate Spreadsheet tab in the Genome Browser.
Run Clustering	Allows you to perform Clustering Analysis
Export Region List	Allows you to export a region list as a *.tsv, *.bed, or *.fa files.
Export List as VCF	Exports the region list in the Variant Call Format (only available for region lists resulting from SNP-related workflows.
Remove Region List	Removes the region list from the active experiment.
Delete Region List	Deletes the region list from **GeneSpring**.
Export to eArray	Allows the export of GenBank Accession data or Genomic co-ordinates data from the region list to Agilent's eArray portal

2.4.15.7 Report

The available right-click options on a Report are summarized in Table 2.4.

Table 2.4: Menu options available on right-click for a Report

Option	Description
Open Report	Displays the report in **GeneSpring**.
Inspect Report	Opens the Report Inspector dialog which contains options to change the Name and Notes of the report.
Download PDF	Opens a Save Dialog which allows you to save the report as a *.pdf* file.
Export Report	Opens a Save Dialog which allows you to save the report as a *.gsreport* file in a local folder.
Export Report to Dropbox	Use this option to export a report to Dropbox (Refer 2.4.15.7.1 for further details).
Export Report to Google Drive	Use this option to export a report to Google Drive (Refer 2.4.15.7.1 for further details).
Mark as Global Report	Use this option to add the selected report to the Global items folder.
Unmark as Global Report	Use this option to remove the selected report from the Global Items folder.
Remove Report	Use this option to remove a report from the active experiment.
Delete Report	Use this option to delete a report from **GeneSpring**.
Share Report	This operation is disabled in the desktop mode of **GeneSpring**. In the **WorkGroupPlus** mode this operation can be used to share the report with other users of **GeneSpring WorkGroupPlus**.
Change owner	This operation is disabled in the desktop mode of **GeneSpring**. In the **WorkGroupPlus** mode, this operation can be used by a group administrator to change the owner of the object.

2.4.15.7.1 Import to and Export of Report from cloud storage (Dropbox and Google Drive)

Dropbox

Dropbox is a file hosting service by Dropbox, Inc., that offers cloud storage.

To export a report to Dropbox;

- right-click on any report and select the option, **Export Report to Dropbox**.
- select a local folder to temporarily save the report.
- specify the Dropbox folder path to save the file on your Dropbox account.
- enable **Get the share link after export** option, for Dropbox to generate a link for the exported entity list. You can share this link with others, and they can either download the file using the link or add it to their Dropbox account.

NOTE Export to Dropbox is dependent on your bandwidth and server traffic. The process of upload and export is in progress until the progress bar is visible. If you see an error, please try the export again.

To import an entity list from Dropbox:

- Go to **Tools** ⟶**Import Report from Dropbox**.
- select the report that you would like to import and click **OK**.
- the entity list will be downloaded to the **Imported Reports** folder of your experiment.

Google Drive

Google Drive is a file storage service by Google, Inc., that offers cloud storage.

To export a report to Google Drive:

- right-click on any report and select the option, **Export Report to Google Drive**.
- select a local folder to temporarily save the report.
- specify the Google Drive folder path to save the file on your Google Drive account.
- enter email ids (only one email per line) of those you would like to share the report with. Using this link, they can download the entity list or add it to their Google Drive account.

NOTE Export to Google Drive is dependent on your bandwidth and server traffic. The process of upload and export is in progress until the progress bar is visible. If you see an error, please try the export again.

To import a report from Google Drive:

- Go to **Tools** ⟶**Import Report from Google Drive**.
- Select the report that you would like to import and click **OK**.
- The report will be downloaded to the **Imported Reports** folder of your experiment.

NOTE In cloud storage, all the exported entity lists are saved as *.gsreport* extension files. If this extension is altered, the entity lists become unavailable for import into **GeneSpring**. In cloud storage, if any other files are changed to have a *.gsreport* extension, they do not become visible in the options for import.

2.4.15.8 Merged Reports

Right-click options available on a merged report are summarized in Table 2.5.

Table 2.5: Menu options available on right-click for a Merged Report

Option	Description
Inspect Report	Opens the Report Inspector dialog which contains options to change the Name and Notes of the report.
Download PDF	Opens a Save Dialog which allows you to save the report as a *.pdf* file.
Export Report	Opens a Save Dialog which allows you to save the report as a *.gsreport* file in a local folder.
Export Report to Dropbox	Use this option to export a report to Dropbox (Refer 2.4.15.7.1 for further details).
Export Report to Google Drive	Use this option to export a report to Google Drive (Refer 2.4.15.7.1 for further details).
Share Report	This operation is disabled in the desktop mode of **GeneSpring**. In the **WorkGroupPlus** mode this operation can be used to share the report with other users of **GeneSpring WorkGroupPlus**.
Unmark as Global Report	Use this option to remove the selected report from the Global Items folder.

2.4.15.9 Merged Reports folder

The available right-click options on a Merged Report folder are summarized in Table 2.6.

Table 2.6: Menu options available on right-click for a Merged Report folder

Option	Description
Rename Folder	Use this option to rename the folder.
Remove Folder	Use this option to remove the folder from the active experiment.
Delete Folder	Use this option to delete the folder from **GeneSpring**.
Share lists	This operation is disabled in the desktop mode of **GeneSpring**. In the **WorkGroupPlus** mode, this operation can be used to share all the reports in this folder with other users of **GeneSpring WorkGroupPlus**.
Change owner	This operation is disabled in the desktop mode of **GeneSpring**. In the **WorkGroupPlus** mode, this operation can be used by a group administrator to change the owner of the object.

2.4.15.10 Context Menu

You can perform common tabular operations using the context (right-click) menu options, which are listed hereunder:

- Select All Rows: Allows you to select all the rows from the list, and then export the view as an image or html file.
- Invert Row Selection: Allows you to invert the row selection, and then use the "Limit to Row Selection" option to launch the selected rows in the view.
- Clear Row Selection: Allows you to clear the existing row selection.
- Limit to Row Selection: Allows you to launch the list with only the selected rows.
- Copy View: Allows you to copy the view to the clipboard.
- Print:Allows you to launch the view in the web browser, which
- Export As: Allows you to Export the view as an Image or HTML file:
  - **Image:** Exports the view as an image in .tiff, .png, .bmp, .jpg, .jpeg, .bmp, .svg, or .eps formats.
  - **HTML:** Exports the view as an HTML file.
  - **Text:** Exports the view as a Text file.
- Properties: Allows you to add a Title and Description for the view.

NOTE HTR entity List is a different entity list that can be exported only as text file. Contact **GeneSpring** support for a script to convert an HTR list to a 'regular' entity list.

2.4.15.11 Entity List Folder

- Rename Folder : This operation can be used to rename the folder.
- Remove Folder : This operation will remove the folder and all its children from the experiment. Note that the remove operation will delete the folder itself, but will only disassociate all the children from the experiment. The children could still belong to zero or more experiments in the system.
- Delete Folder : This operation will permanently delete the folder and all its children from the system.

2.4.15.12 Classification

- Open Classification : (default operation) This operation results in the current active view to be split up based on the entity lists of the classification. If the active view does not support splitting up, for e.g., if it is already split, or if it is a Venn Diagram view, etc., then the classification is displayed using split up profile plot views.
- Expand as Entity List : This operation results in creating a folder with entity lists that each correspond to a cluster in the classification.
- Delete Classification : This operation will permanently delete the classification from the experiment. Note that there is no notion of removing a classification, since a classification is not an independent object and always exists only within the experiment.

2.4.15.13 Entity/Condition/Combined Tree

- Open Tree : (default operation) This operation opens up the tree view for this object. In the case of entity trees, the tree shows columns corresponding to the active interpretation. In the case of condition and combined trees, the tree shows the conditions that were used in the creation of the tree.
- Delete Tree : This operation will permanently delete the tree from the experiment. Note that there is no notion of removing a tree, since a tree is not an independent object and always exists only within the experiment.

2.4.15.14 Class Prediction Model

- Remove Model : This operation removes the model from the experiment. Note that this operation only disassociates the model with the experiment and does not actually delete the model. The model could still belong to other experiments in the system, or may even exist without being part of any other experiment.
- Delete Model : This operation permanently deletes the model from the system.

2.4.15.15 Pathway List

- Open Pathway List: (default operation) This operation opens up the pathway view. Protein nodes in the pathway view that have an Entrez id matching with an entity of the current experiment have a blue halo around them.
- Inspect Pathway List: This operation opens the Pathway List Inspector. In the Inspector you can view the name and description of the pathway list as well as the list of all the pathways that the pathway list contains.
- Remove Pathway List: This operation removes the pathway list from the experiment. Note that this operation only disassociates the pathway list with the experiment and does not actually delete the pathway list. The pathway list could still belong to other experiments in the system, or may even exist without being part of any other experiment.

- Delete Pathway List : This operation permanently deletes the pathway list from the system.

2.4.15.16 Correlation Node

- Open Correlation Node (default operation): This operation opens the correlation heatmap view.
- Delete Correlation Node : This operation permanently deletes the correlation node from the system.

2.4.16 Search

GeneSpring could have many projects, experiments, samples, region lists, entity lists, entities, pathway lists, pathways, prediction models, scripts, samples, technologies, and reports. All of these carry searchable annotations and can be searched from **Search** option in the main toolbar. **GeneSpring** supports two types of search:

- **Simple Search**

  Features of the Simple Search are:

  - A case insensitive keyword search for the selected object(s).
  - Search is performed over all the annotations associated with the object including its name, notes, etc.
  - A blank keyword will result in all the available objects of that type in the results.

- **Advanced Search**

  Features of the Advanced Search are:

  - Search options to include multiple **Search Fields**, **Conditions**, and **Search Value**.
  - Option to search by combining search conditions by OR or AND conditions. For e.g., search all entity lists that contain the **Name** ''Fold change'' and created on a certain **Technology**.
  - The maximum number of search results to display is set at 100 by default and can be changed from **Search Parameters** wizard(Step 1). This can also be changed from **Tools** $\longrightarrow$ **Options** $\longrightarrow$ **Miscellaneous** $\longrightarrow$ **Search Results**.

NOTE

In the **Search Field** if **Creation Date** or **Modified Date** are used with any other available option, the search criteria is applied only for **AND** condition. For example, a search performed for an Entity List with a **Name** All Entities and created after a specified date using the **Combine search conditions by** option **OR** will only result in Entity Lists with name All Entities and created after the given date.

Depending on the type of object being searched for, a variety of operations can be performed on the results of Search. The **Search Results** wizard (Step 3) displays a message about the total number of results obtained for that search as well as the number of results on that page. The total number of pages are also given and you can navigate to the page of choice by entering the page number in the box provided. All the toolbar buttons on the search results page are applicable only on the set of selected objects in the result.

Each search options results has multiple options. All the options are described and their respective icons are described in the following sections.

2.4.16.1 Search Experiments

- Inspect experiments ( ): This operation opens up the inspector for all the selected experiments.
- Delete experiments ( ): This operation permanently deletes the selected experiments and their children from the system. The only exception to this is samples, and samples will be deleted only if they are not used by another experiment in the system. If the experiment being deleted also belongs to the currently open project and it is currently open, it will be closed and will show with a grey font in the project navigator. Also, at a later stage, on opening a project that contains some of these deleted experiments, the experiments will show in grey in the navigator, as a feedback of the delete operation.
- Add experiments to project ( ): This operation adds the selected experiments to the current project, if one is open. If any of the selected experiments already belong to the project, then they are ignored.
- Change permissions: This operation is disabled in the desktop mode of **GeneSpring**. In the **WorkGroupPlus** mode, this operation allows sharing the experiment with other users of the **WorkGroupPlus**.

2.4.16.2 Search Samples

- Inspect samples ( ): This operation opens up the inspector for all the selected samples.
- Create new experiment ( ): This operation creates a new experiment with the set of selected samples. If the selected samples do not belong to the same technology an error message will be shown. This operation will close the search wizard and launch the new experiment creation wizard with the set of selected samples.
- Change permissions: This operation is disabled in the desktop mode of **GeneSpring**. In the **WorkGroupPlus** mode, this operation allows sharing the samples with other users of the **WorkGroupPlus**.
- View containing experiments( ): This operation shows a dialog with the list of experiments that the selected samples belong to. This dialog also shows an inverse view with the list of all samples grouped by the experiments that they belong to. One can select and add experiments to the current project from this view.

2.4.16.3 Search Entity Lists

- Inspect entity lists ( ): This operation opens up the inspector for all the selected entity lists.
- Delete entity lists( ): This operation will permanently delete the selected entity lists from the system. Note that only the selected entity lists will be deleted, and if they belong to any experiments, their children in each of those experiments will remain intact. If the entity lists being deleted belong to one or more of the currently open experiment, the navigator of the experiment will refresh itself and the deleted entity lists will show in grey.
- Change permissions : This operation is disabled in the desktop mode of **GeneSpring**. In the **WorkGroupPlus** mode, this operation allows sharing the entity lists with other users of the **WorkGroupPlus**.
- View containing experiments ( ): This operation shows a dialog with the list of experiments that the selected entity lists belong to. This dialog also shows an inverse view with the list of all entity lists grouped by the experiments that they belong to. One can select and add experiments to the current project from this view.
- Add entity lists to experiment( ): This operation adds the selected entity lists to the active experiment. The entity lists get added to a folder called ''Imported Lists'' under the All Entities entity list. Entity lists that do not belong to the same technology as the active experiment are ignored.

2.4.16.4 Search Entities

The search entities wizard enables searching entities from the technology of the active experiment. The first page of the wizard allows choosing the annotations to search on, and the search keyword. The second page of the wizard shows the list of entities that match the search criterion. A subset of entities can be selected here to create a custom list. On clicking next and then finish, an entity list gets created with all the entities that match the search criterion. This entity list is added under the All Entities entity list.

2.4.16.5 Search Pathway Lists

- Inspect pathway lists ( ): This operation opens up the inspector for all the selected pathway lists.
- Delete pathway lists ( ): This operation will permanently delete the selected pathway lists from the system. If the pathway lists being deleted belong to one or more of the currently open experiments, these become non-functional. Later, on re-opening or refreshing an experiment that contained some of these deleted pathway lists, the pathway lists are not displayed.
- Add pathway list in the experiment ( ): This operation adds the selected pathway lists to the active experiment. The pathway list gets added to a folder called '' **Imported Pathway Lists**'' under the All Entities entity list.

- Change permissions: This operation is disabled in the desktop mode of **GeneSpring**. In the **WorkGroupPlus** mode, this operation allows sharing the pathway lists with other users of the **WorkGroupPlus**.

2.4.16.6 Search Pathways

- Inspect pathways ( ): This operation opens up the inspector for all the selected pathways.
- Delete pathways ( ): This operation will permanently delete the selected pathways from the system. If the pathways being deleted belong to one or more of the currently open experiment or any pathway list, these become non-functional. Later, on re-opening or refreshing an experiment/ pathway list that contained some of these deleted pathways, the pathways are displayed in grey in the navigator, as a feedback of the delete operation.
- Add pathways to experiment ( ): This operation adds the selected pathways to the active experiment as a pathway list. The pathway list gets added to a folder called " **Imported Pathway Lists**" under the All Entities entity list.
- Change permissions: This operation is disabled in the desktop mode of **GeneSpring**. In the **WorkGroupPlus** mode, this operation allows sharing the pathways with other users of the **WorkGroupPlus**.

2.4.16.7 Search Prediction Model

- Inspect models ( ): This operation opens up the inspector for all the selected models.
- Delete models ( ): This operation will permanently delete the selected models from the system. If the models being deleted belong to one or more of the currently open experiment, the navigator of the experiment will refresh itself and the deleted models will show in grey. Also, at a later stage, on opening an experiment that contains some of these deleted models, the models will show in grey in the navigator, as a feedback of the delete operation.
- Add models to experiment ( ): This operation adds the selected models to the active experiment. The models get added to a folder called "Imported Models" under the All Entities entity list. Models that do not belong to the same technology as the active experiment are ignored.

2.4.16.8 Search Scripts

- Inspect scripts ( ): This operation opens up the inspector for all the selected scripts.
- Delete scripts ( ): This operation will permanently delete the selected scripts from the system.
- Open scripts ( ): This operation opens the selected scripts in Python or R Script Editor in the active experiment. Multiple scripts are opened in invidual tabs.

2.4.16.9 Search Technology

- Inspect technologies ( ): This operation opens up the inspector for all the selected technologies.
- Delete selected technology ( ): This operation deletes the selected technology and will result in an improper functioning of the associated experiments.

2.4.16.10 Search Reports

- Inspect selected Reports ( ): This operation opens an inspector that displays all the details of the selected reports.
- Delete selected Reports ( ): This operation deletes all the selected reports from **GeneSpring**.
- Change permissions: This operation is disabled in the desktop mode of **GeneSpring**. In **WorkGroupPlus** mode, this operation allows sharing reports with other users of **WorkGroupPlus**.
- Add selected Reports to active experiment ( ): This operation adds the selected reports to the active experiment in the Experiment Navigator. Reports added to an experiment in this way are saved in the Imported Reports folder, rather than the view or entity list with which they are associated.

2.4.16.11 Search All

GeneSpring provides the ability to search for multiple objects at the same time using the Search All functionality and selecting or de-selecting the required objects (Refer Figure 2.6).

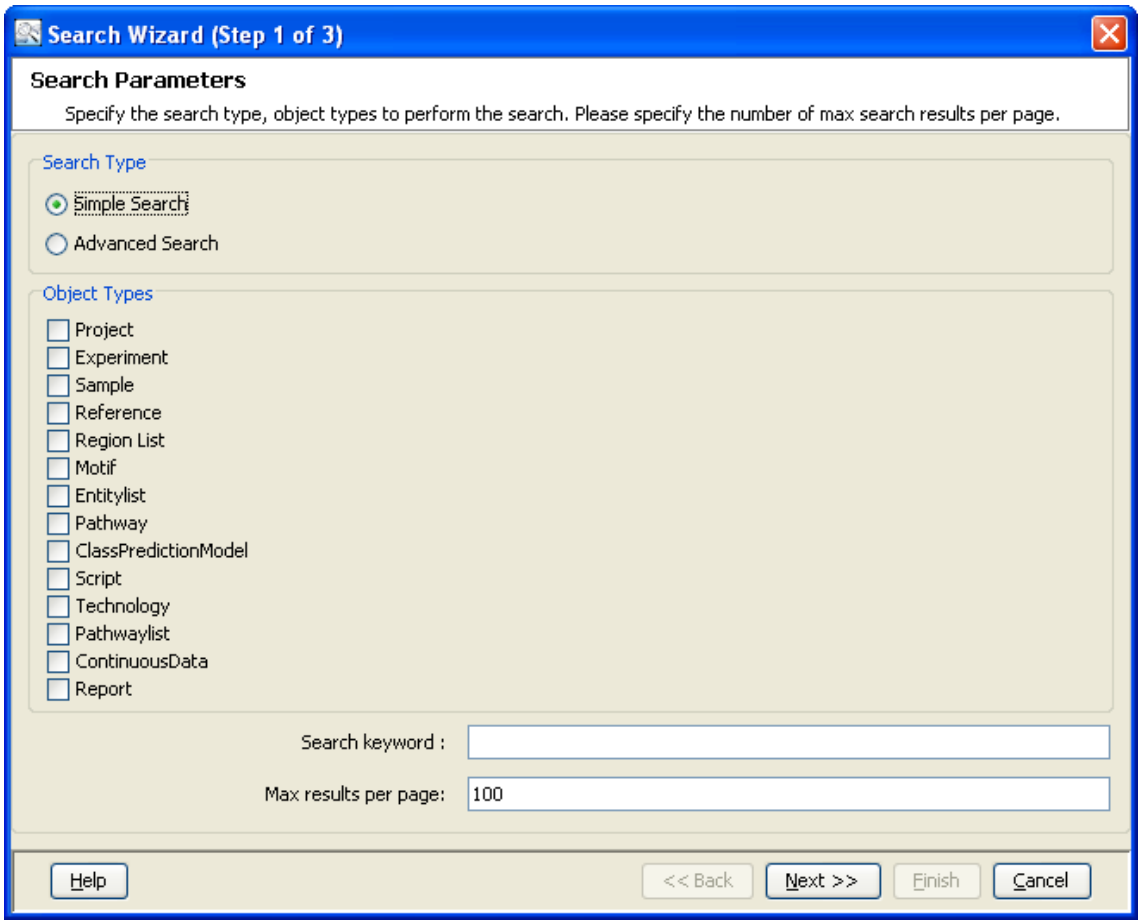


Figure 2.6: Options for Search All objects in **GeneSpring**

- Inspect objects ( ): This operation opens up the inspector for all the selected objects.
- Delete objects ( ): This operation will permanently delete the selected objects from the system. Samples that belong to any experiment will not be deleted.
- Change permissions : This operation is disabled in the desktop mode of **GeneSpring**. In the **WorkGroupPlus** mode, this operation allows sharing the objects with other users of the **WorkGroupPlus**.

2.4.17 Saving Projects

The state of an open project, i.e., all experiments and their respective navigators, are always auto-saved and therefore do not need to be saved explicitly. This is however not true of the open views, which unless saved explicitly are lost on shut-down. Save the views in existing state via the menu **View** → **Save Current View** option.

2.4.18 Sharing Projects - Export and Import utilities

Use these options to share your projects with others or move your projects from one machine to another.

Projects can be shared with other users using the **Export Project** and **Import Project** sub menus under the menu **Project**.

2.4.18.1 Export Project

You can export a project using either of these three options;

- **Export Project**

  A *.tar* file is created while exporting and it is portable across platforms i.e., Linux, Windows and Mac. This feature allows you to export the whole project along with experiments, in a tar format. Some/all experiments within a project can be exported. When a project tar archive is created, in case of Generic Single Color, Generic Two Color and the experiments migrated from GX7, the technologies are bundled along with the tar file. The standard technologies and Affymetrix Custom technologies are not bundled along with the archived project. These can be selected from the **Choose Technologies** window that appears after the experiments to be exported are chosen.

- **Export Project to Dropbox**:

  To export a project to Dropbox;

  - select the experiments of the project to be exported with the project.
  - select the technologies associated with the experiments of the project to export.
  - select a local folder to temporarily save the entity list
  - specify the Dropbox folder path to save the file on your Dropbox account.
  - enable **Get the share link after export** option, for Dropbox to generate a link for the exported project. You can share this link with others, and they can either download the file using the link or add it to their Dropbox account.

- **Export Project to Google Drive**

  To export a project to Google Drive;

  - select the experiments of the project to be exported with the project.
  - select the technologies associated with the experiments of the project to export.
  - select a local folder to temporarily save the project.
  - specify the Google Drive folder path to save the file on your Google Drive account.
  - enter email ids (only one email per line) of those you would like to share the project with. Using this link, they can download the project or add it to their Google Drive account.

NOTE Export to cloud storage is dependent on your bandwidth and server traffic. The process of upload and export is in progress until the progress bar is visible. If you see an error, please try the export again.

NOTE In cloud storage, all the exported projects are saved as *.gsp* extension files. If this extension is altered, the projects become unavailable for import into **GeneSpring**. In cloud storage, if any other files are changed to have a *.gsp* extension, they will not become visible in the options for import.

2.4.18.2 Import Project

You can import a project using either of these three options;

- **Import** - Exported projects can be imported by a different user or in a different machine from the menu **Project** → **Import**. This allows the import of the zipped or tar archived projects along with the experiments. If the standard technologies were not packaged with the project, a message will be prompted asking to download the technologies needed to open the project and experiments. For Affymetrix Custom experiments, the technology will have to be created prior to importing projects, as the technology associated with the custom experiment is not exported with the project.
- **Import from Dropbox**

  To import a project from Dropbox;

  - select a project for import from the list of available projects.
  - click **OK** and the project will be downloaded and opened in **GeneSpring**.
- **Import from Google Drive**

  To import a project from Google Drive;

  - select a project for import from the list of available projects.
  - click **OK** and the project will be downloaded and opened in **GeneSpring**.

NOTE If the project that is being imported is the same as an existing project in **GeneSpring** that was exported, this existing project will be overwritten with the imported project.

2.4.19 Software Organization

At this point, it may be useful to provide a software architectural overview of **GeneSpring**. **GeneSpring** contains three parts, a UI layer, a database and a file system. The file system is where all objects are stored physically; these are stored in the app/data subfolder in the installation folder. A PostegreSQL database carries all annotations associated with the various objects in the file system (i.e., properties like notes, names etc which can be searched on); a database is used to drive fast search. Finally, the UI layer displays relevant objects organized into projects, experiments, analysis etc.

2.5 Exporting Views

Views can be exported as a Report, image, or a HTML file. Right-Click on the view and use the Export As option, and choose either **Report**, **Image**, or **HTML**. Image format options include jpeg (compressed), png (high resolution), and vector graphic formats like svg and eps.

NOTE **Exporting Whole Images:** Exporting an image will export only the VISIBLE part of the image. Only the dendrogram view supports whole image export via the Print or Export as HTML options; you will be prompted for this. The Print option generates an HTML file with embedded images and pops up the default HTML browser to display the file. You need to explicitly print from the browser to get a hard copy.

Images can also be copied directly to the clipboard and then pasted into any application like PowerPoint or Word. Right-Click on the view, use the *Copy View* option and then paste into the target application. Further, columns in a dataset can be exported to the Windows clipboard. Select the columns in the spreadsheet and using Right-Click *Select Columns* and then paste them into other applications like Excel using Ctrl-V.

2.5.1 Export As Report

This section describes the procedure to save a view as a Report in a **GeneSpring** experiment and all the available options for generating a Report in **GeneSpring**.

NOTE Export As Report is not supported for the following views: Correlation plot in Quality Control on Samples, LD Plot of Association Analysis and Neural Network Model for Class Prediction.

2.5.1.1 How to Create a Report

Use the following steps to create and save a simple or advanced report from a view.

1. Right-Click on a view and select **Export as Report**.

   The **Report Exporter** dialog opens. Customize the report layout and view options before saving or downloading the report.

2. By default, the Legend for the view is included in the report. Deselect **Add Legend** to exclude the legend from the report.

3. Select the **Report Type** you want to export: **Simple** or **Advanced**.

   For a **Simple Report**:

   (a) Select one of the following **Page Orientation**: **Portrait, Landscape**.

   (b) Select a **Paper Size** for the report from: **A1, A2, A3, A4, A5, Letter, Legal, Ledger, Executive, B4, B5, B9,** or **Custom**. For custom paper size, specify Width and Height in *inch*.

   (c) Customize the margins (Left, Right, Top, or Bottom) of the report by entering a new value in the value field or by dragging the scroll bar to get a desired value for each of the respective margins.

   (d) From the active toolbar icons on the right side of the report view:

   i. **Full Screen** icon is not active in the report exporter dialog by default. You have to save and re-open a report and click icon to view the report in Full Screen mode.

   ii. Click **Best Fit** icon to view the report in a best vertical fit.

   iii. Click **Fit Page Width** icon to view the report in a best horizontal fit.

   iv. Click **100%** icon to set the zoom level of the report view at 100%.

   v. Click **Zoom In** icon to increase the zoom level.

   vi. Click **Zoom Out** icon to decrease the zoom level.

   vii. Click **Previous Page** icon and navigate to the previous page of the report.

   viii. Click **Next Page** icon and navigate to the next page of the report

   ix. Click **Download** icon to download and save the report to a local folder.

   For an **Advanced Report**:

   i. Select which items you want to include in the report by moving the desired items from **Available items** to the **Selected items** field or vice versa using the move right and move left buttons.

   ii. Reorder the selected items, or refresh the order using the up , down , and refresh order buttons.

   iii. Select the remaining parameters as described in Steps 3a - d of this section.

(e) Click **Preview**.
GeneSpring applies all the customized settings for the displayed report on the right side of the report view.

4. You can change any of the parameter settings in the pane on the left or the view setting from the toolbar and recreate the report by clicking **Preview**.
5. When you are ready to save your report, click **Save Report**.

NOTE If a report contains > 100 pages (default setting), it cannot be viewed in **GeneSpring**. You can download the report as a *.pdf* or modify report settings from **Tools** ⟶**Options** ⟶**Miscellaneous** ⟶**Report**.

The report is saved as an object in the experiment below the selected input entity list from which the view was generated.

NOTE You can export Cluster Tree, Heatmap, and Dendrogram views as a Summary, Detailed, or Combined View report.

NOTE Advanced Report options are not available for the following views: Venn Diagram, Multiple Histogram, Cluster Set view of Clustering by K-Means, and U-Matrix view of Clustering by Self-Organizing Map.

2.5.1.2 How to Merge Reports

Use the following steps to merge multiple reports and save them as a single report.

To Merge reports in the same experiment:

1. Select reports (> 1) you would like to merge and Right-Click on one of the selected report.
2. Select **Merge Reports**.

   The **Merge Reports Wizard** opens.
3. To **Reorder Reports**, select reports and reorder using the up and down buttons.
4. Click **Merge** to save the Merged Report to a **Merged Reports** folder in the experiment navigator.

To Merge reports across experiments from Global Items:

1. Select each report that you would like to merge and Right-Click on it and select as **Mark as Global Report**.
2. Go to **Global Items** $\longrightarrow$**My Reports** folder.
3. Select reports (> 1) you would like to merge and Right-Click on one of the selected report.
4. Select **Merge Reports**.

   The **Merge Reports Wizard** opens.
5. To **Reorder Reports**, select reports and reorder using the up and down buttons.
6. Click **Merge** to save the Merged Report to the **My Reports** folder in the **Global Items**.

To Merge reports across experiments from Experiment Navigator:

1. Go to **Search** $\longrightarrow$**Reports** and perform a **Simple Search** or an **Advanced Search** (Refer Search).
2. Select reports you would like to merge and add them to the active experiment using Add selected Reports to active experiment ( ) icon.
3. From the experiment navigator, select reports (> 1) you would like to merge and Right-Click on one of the selected report.
4. Select **Merge Reports**.

   The **Merge Reports Wizard** opens.
5. To **Reorder Reports**, select reports and reorder using the up and down buttons.
6. Click **Merge** to save the Merged Report to the **Merged Reports** folder in the experiment navigator.

2.6 Utilities under Tools Menu

Various utilities in **GeneSpring** are accessible from the **Tools** menu. These include scripting, various export/import options, a memory monitor, backup/restore and configurable options.

2.6.1 Scripting and R Editor

GeneSpring has a powerful scripting interface which allows for automation of tasks via flexible Jython scripts. In addition, R scripts can also integrated with **GeneSpring**. To run a Jython script, go to *Tools* $\rightarrow$ *Script Editor*. To run a R script go to *Tools* $\rightarrow$ *R Editor*. A few sample scripts are packaged with the demo project. For further details, refer to the Scripting chapter.

2.6.2 Import Pathways From WikiPathways

You can import pathways directly from the WikiPathways portal http://www.wikipathways.org using this option. See To import pathways from WikiPathways through the main menu.

2.6.3 Import Pathways From BioCyc

You can import pathways derived from the BioCyc Database Collection http://www.biocyc.org/ using this option. See To import pathways from BioCyc through the main menu.

2.6.4 Import Pathways From KEGG

You can import pathways derived from the KEGG Database Collection http://www.genome.jp/kegg/pathway.html using this option. See To import pathways from KEGG through the main menu.

2.6.5 Import Pathways From File

You can import pathways in three different file formats:

- BioPaX
  - Level 2
  - Level 3
- Text
- GPML

Imported pathways can then be used to perform *Single Experiment and Multi-Omic Analysis*. See To import pathways from file for additional details.

2.6.6 Import from Pathway Architect

See section Import from PathwayArchitect

2.6.7 Import BROAD GSEA Genesets

GSEA can be performed using the 7 genesets which are available from the BROAD Institute's website http://www.broadinstitute.org/gsea/index.jsp. These genesets can be downloaded and imported into **GeneSpring** to perform GSEA. Clicking on this option allows the user to navigate to the appropriate folder where the genesets are stored and select the set(s) of interest. The files should be present either in .xml or .grp or .gmt formats.

2.6.8 Import NCBI GEO Experiment

See Chapter Loading Experiment from NCBI GEO

2.6.9 Export ArrayExpress MAGE-TAB

Contents of **GeneSpring** Expression experiments can be exported in ArrayExpress MAGE-TAB format from the menu **Tools** → **Export ArrayExpress MAGE-TAB**. On clicking this menu, the MAGE-TAB Exporter Window (shown in Fig 2.7) opens.

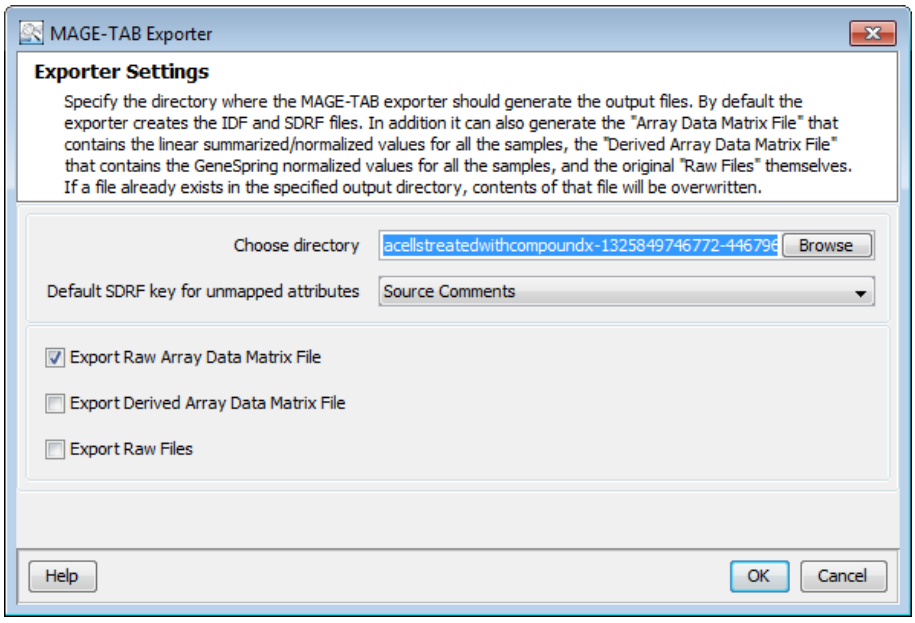


Figure 2.7: MAGE-TAB Exporter Window

The window gives the following options:

- **Choose Directory:** Choose a directory where the MAGE-TAB output files will be written in. By default, **GeneSpring** will create a folder with the name of the experiment ID in the

path
`Installation Directory/Agilent/GeneSpring/samples/` The user can specify a different folder and path for the outputs. By default, **GeneSpring** writes out the IDF and SDRF files.

- **Default SDRF key for unmapped attributes:** This option allows marking the fields in the SDRF file. The available fields/keys are shown in the drop down in Fig 2.7.
- **Export Raw Array Data Matrix File:** (Optional) This writes out the data just after the pre-processing step in **GeneSpring**.
- **Export Derived Array Data Matrix File:** (Optional) This writes out the normalized data.
- **Export Raw Files:** (Optional) This writes out the raw files.

Note that only the active experiment will be exported. A message 'MAGE-TAB Files exported successfully' along with the output file name and path is shown on completion.

In **GeneSpring WorkGroupPlus** Server , in addition to the UI mode of export explained above, a script can be run in command line mode to export all experiments in MAGE-TAB format inone go. The script (magetab.py) can be found in the folder:
`Installation Directory/Agilent/GeneSpring/app/MageTab/`. The script contains instructions on the arguments that should be passed to it. Note that this command line option is available only in **WorkGroupPlus** mode.

For complete technical details on MAGE-TAB format and customizable options, see chapter Exporting Expression Experiments in MAGE-TAB Format.

2.6.10 Import Report

You can import a report (*.gsreport* file) from the local drive into **GeneSpring** using this option.

2.6.11 Import Report from Dropbox

You can import a report directly from Dropbox using this option. See Import of Report from cloud storage (Dropbox and Google Drive).

2.6.12 Import Report from Googledrive

You can import a report directly from Google Drive using this option. See Import of Report from cloud storage (Dropbox and Google Drive).

2.6.13 Migration

The menu links **Prepare for GS7 Migration**, **Migrate from GS7** are for migrating data from **GeneSpring** version 7 to the latest version. See Chapter Data Migration for complete details.

2.6.14 Memory Monitor

Memory monitor helps to collect and clear unused memory.

2.6.15 Edit Pathway Theme

See section Themes

2.6.16 Options

Various views, algorithms etc can be configured from this menu. Figure 2.8 shows an image of the available options. Individual sections carry details on how to use these options.

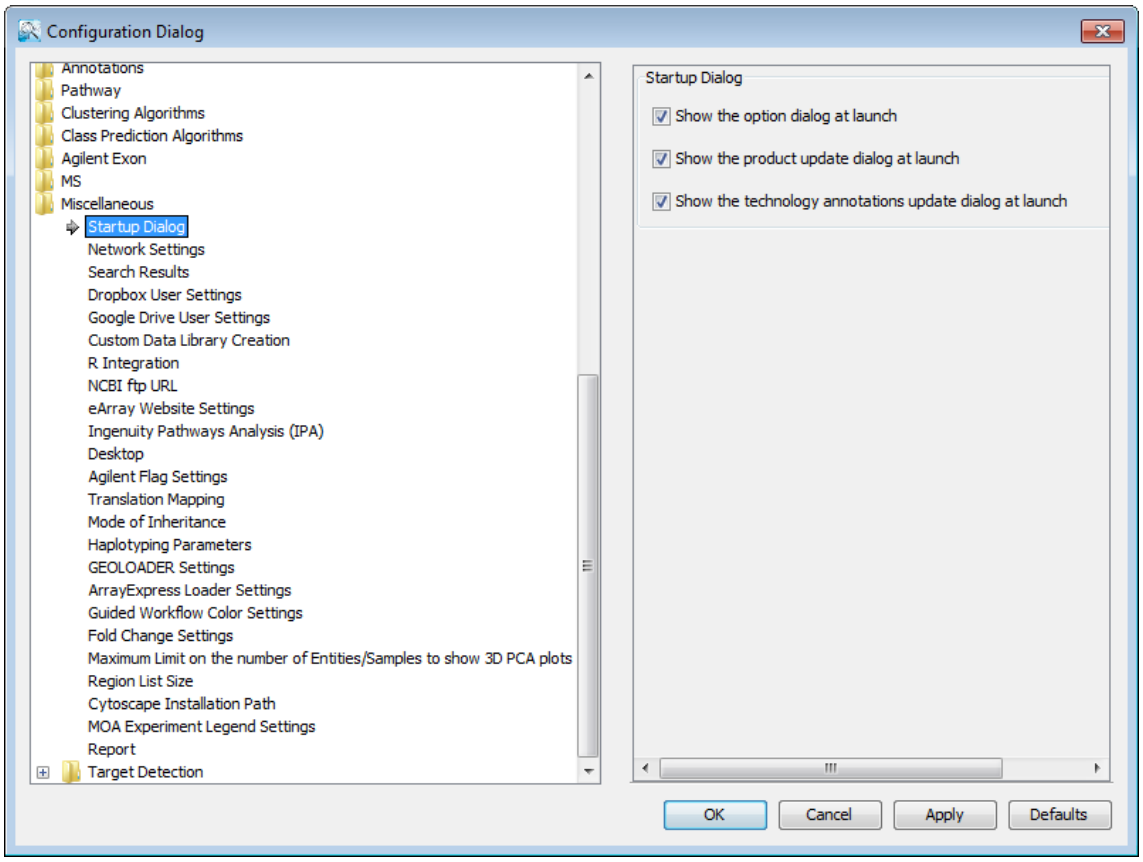


Figure 2.8: Tools - Options

2.6.17 Create Custom Reference

This is specific to Copy Number experiments. See Section Custom Reference Creation for complete details.

2.6.18 Backup and Restore

The backup utility in **GeneSpring** allows users to take a complete back up of their work in **GeneSpring** and restore it completely. The backup includes experimental data, configuration settings, technologies, ontologies, etc.

NOTE: The backup and restoration needs to be done within the same version of **GeneSpring** and in the same Operating System. For example, if you backup in a **Windows** system, this backup file can only be restored in another **Windows** system. Note that during restoration, all the existing **GeneSpring** files are cleaned up and only the backed up files are restored.

2.6.18.1 How to execute Backup

- Open the tool, click on the menu **Tools** → **Backup Repository**.
- A window will appear prompting the user for the backup folder name. (the default name is <**GeneSpring** _Inst_Dir>_backup) Click 'OK' to start the backup.
- A window appears prompting the user to shut down the tool in order to start the backup process. Click 'Yes' to continue. A progress bar is displayed until the backup process is complete.
- The backup process is incremental on folders that contain failed backups due to power failure or abrupt shut-down of the machine. A valid file check and clean-up is done before re-starting the backup.
- If you do not change the backup folder, and it contains previous backup data, **GeneSpring** will only copy the updated data (incremental backup).
- If the specified folder is empty, a full backup of all the data is initiated.

2.6.18.2 How to execute Restore

- For restoration, open the tool and click on the menu **Tools** → **Restore Backup**.
- A window will appear prompting the user to select a backup folder to restore from. Choose the required backup folder and click 'OK'.
- A window appears prompting the user to shut down the tool in order to begin the restoration. A Progress bar is displayed until the restoration is complete.
- If there are multiple backup folders, only the data present in the specified backup folder will be restored.
- Once restoration is complete, open the tool and all your files would have been restored.

2.6.18.3 How to change the repository path

GeneSpring stores all the analysis data generated by you in two separate locations. The so-called meta-data (like Name, Notes, Creation Date, etc., of an experiment for example) and the data hierarchy (parent-child relationships) is stored inside a PostgreSQL database. The actual data behind a **GeneSpring** object (eg. samples, experiments, entity lists, etc.) is stored as randomly numbered files on disk. This disk location is referred to as the ''repository path". Within the PostgreSQL database, **GeneSpring** maintains a mapping of the **GeneSpring** object and its corresponding file on disk. The default location of the PostgreSQL database is $< GENESPRING_INSTALLATION_DIRECTORY > /app/PostgreSQL$. The default repository path is $< GENESPRING_INSTALLATION_DIRECTORY > /app/Data/files$.

Note: Once the path is set to a new location, all **further** analyses and results generated from this time onwards are now stored in the new folder. All previously generated data will remain in its old location; all previous repository paths are listed under **Old Repository Paths** in the **Change Repository** dialog box the next time you open the dialog box. Please note that all meta-data information continues to be stored in the PostgreSQL database. The exceptions to this rule are the Interaction Databases for the Pathway Analysis Module which are stored and updated in the installation directory.

You can change the repository path as follows:

1. Select **Tools** → **Change Repository** from the main menu.

   The Change Repository dialog box opens.

2. Click **Browse** and determine the new location for the repository on your system.

3. Click **Open**.

   The new repository path is reflected in the Repository Path field.

4. Click **OK**.

   A Confirmation dialog box opens. This dialog box allows you to move existing pathway and experiment annotation data (eg. interaction databases, BridgeDb databases, pathways) from their current location in the installation directory to the new repository location that you just specified. The exact size requirement for moving this data is indicated in the dialog box.

5. Click **Yes** to move the pathway and experiment annotation data to the new repository location. This is useful to free up space in the system directory in which **GeneSpring** is installed.

   Click **No** to only change the repository path.

Points of Caution:

- **GeneSpring** requires all previous repository path locations to exist. Deleting any of the folders in older repository paths will result in irretrievable data loss.
- Any backup policy should take care to backup both the repository paths and the PostgreSQL database. Please note that taking backup of just one of the two will result in irrevocable data loss.
- To safely move data from all previous repository paths to a new location please refer to the How to execute Backup section on page 68 and the How to execute Restore section on page 68.

2.6.18.4 FAQ

1. What folders are backed up and restored?
   There are 3 sources for backing up data in **GeneSpring** :

- The repository path, which by default is `Installation Directory/Agilent/GeneSpring/app/Data/files` . If you changed the repository path at any time, **GeneSpring** maintains a record of all repository paths. All these paths are automatically identified and backed up. You can review the current repository path and old repository paths in the **Change Repository** dialog box by selecting **Tools** → **Change Repository**.
- The PostgreSQL data directory, which by default is located in `Installation Directory/Agilent/GeneSpring/app/PostgreSQL/data` , but could have been modified by the user.
- User configuration files in `Installation Directory/Agilent/GeneSpring/app/Configuration`
- DataLibrary folder from `Installation Directory/Agilent/GeneSpring/app/DataLibrary` (which contains, among other things, the Gene Ontology library files)

2. What happens during restoration?
   - If there are multiple repositories, all the repositories will be merged into one repository, i.e app/Data/files, or whatever is the currently specified repository path.
   - If the PostgreSQL data directory is different from the default (app/PostgreSQL/data), then the data is copied into that data directory.
3. What are the space requirements for Restoration?
   - If the PostgreSQL data directory and the repository path are on the same drive, there must be at least as much free space as the size of the backup tar file in that drive.
   - If the PostgreSQL data directory and the repository path are on different drives, the corresponding drives need to have as much free space as the corresponding sections of the backup tar file.
   - During restoration, the application prints out the space required in the corresponding drives and does not proceed if enough free space is not available.
4. Is there a size limit to the data being backed up?
   There is no size limit for the back up.
5. Can individual projects be backed up via this mechanism?
   No, this utility is for complete backup and complete restoration only. It does not provide any incremental capabilities.
6. Can users back their data up on a remote file server?
   Yes, users can back up their data on any disk that is accessible to the operating system.

2.7 Update Utility

GeneSpring has an update utility that can be used to update the product or get data libraries needed for creating an experiment. These data library updates and product updates are periodically deployed on the **GeneSpring** product site and is available online through the tool. The update utility is available from the *Annotations* and *Help*⟶ *Update Product*. This will launch the update utility that will contact the online update server, verify the license, query the sever and retrieve the update (if any) that are available. Note that you have to be connected to the Internet and should be able to access the **GeneSpring** update server to fetch the updates. In situations where

you are unable to connect to the update server, you can do an update form a file provided by Agilent support.

2.7.1 Product Updates

GeneSpring product updates are periodically deployed on the update server. These updates could contain bug fixes, feature enhancements and product enhancements. Choosing product update from *Help* $\longrightarrow$ *Update Product* $\longrightarrow$ *from Web* will prompt a dialog stating that the application will be terminated before checking for updates. Confirm to close the application. This will launch the update utility that will contact the online update server, verify the license, query the sever and retrieve the product update (if any) available. See Figure 2.9

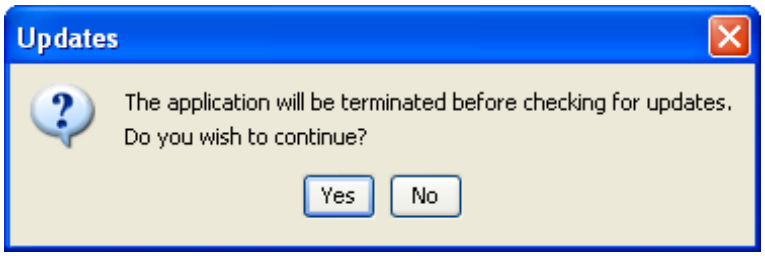


Figure 2.9: Confirmation Dialog

If updates are available, the dialog will show the available updates. Left-Click on the check box to select the update. If multiple updates are available, you can select multiple updates simultaneously. Details about the selected update(s) will be shown in the description box of the update dialog. Left-Click *OK* will download the update and execute the update to apply it on your productns. When you launch the tool, these updates will be available. To verify the update, you can check the version of build number from the *Help* $\longrightarrow$ *About* ***GeneSpring*** . See Figure 2.10

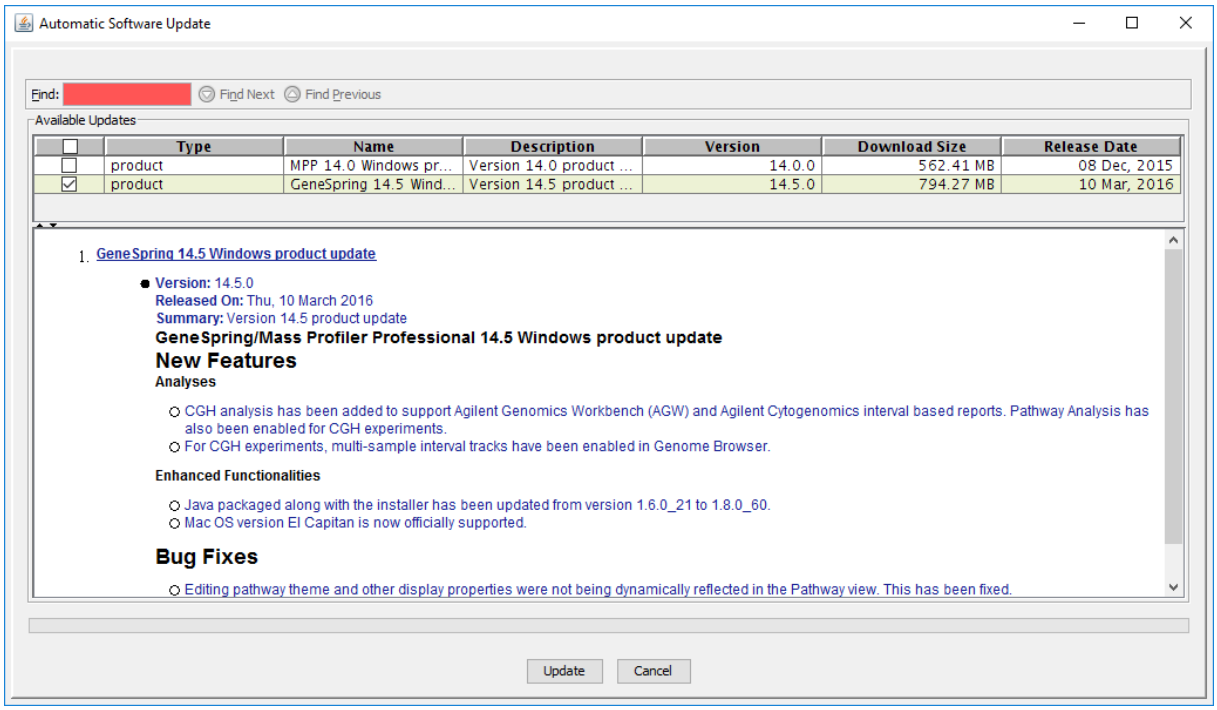


Figure 2.10: Product Update Dialog

2.8 Getting Help

Help is accessible from various places in **GeneSpring** and always opens up in an HTML browser.

Single Button Help. Context sensitive help is accessible by pressing F1 from anywhere in the tool.

All configuration utility and dialogs have a *Help* button. Left-Click on these takes you to the appropriate section of the help. All error messages with suggestions of resolution have a help button that opens the appropriate section of the online help. Additionally, hovering the cursor on an icon in any of the windows of **GeneSpring** displays the function represented by that icon as a tool tip.

Help is accessible from the drop down menu on the menubar. The Help menu provides access to all the documentation available in **GeneSpring**. These are listed below:

- **Help**: This opens the Table of Contents of the on-line **GeneSpring** user manual in a browser.
- **Documentation Index**: This provides an index of all documentation available in the tool.

- **About GeneSpring** : This provides information on the current installation, giving the edition, version and build number.

Chapter 3

Technology and Biological Genome

3.1 Technology

Technology in **GeneSpring** is defined as the package of data regarding array design, biological and other information about the entities, eg., Entrez gene ID, GO accession etc. Technology is available for each individual array type-i.e., the technology for Affymetrix HG-U133_plus_2 would contain information specific to its design and would thus differ from other technologies, like the Agilent 12097 (Human 1A). An experiment comprises samples which all belong to the same technology.

A technology initially must be installed for each new array type to be analyzed. For standard arrays from Affymetrix, Agilent and Illumina, technologies have been created beforehand and **GeneSpring** will automatically prompt for downloading these technologies from Agilent's server whenever required. For other array types, technologies can be created in **GeneSpring** via the custom technology creation wizard from ***Annotations*** $\longrightarrow$ ***Create Technology***. See Figure 3.1

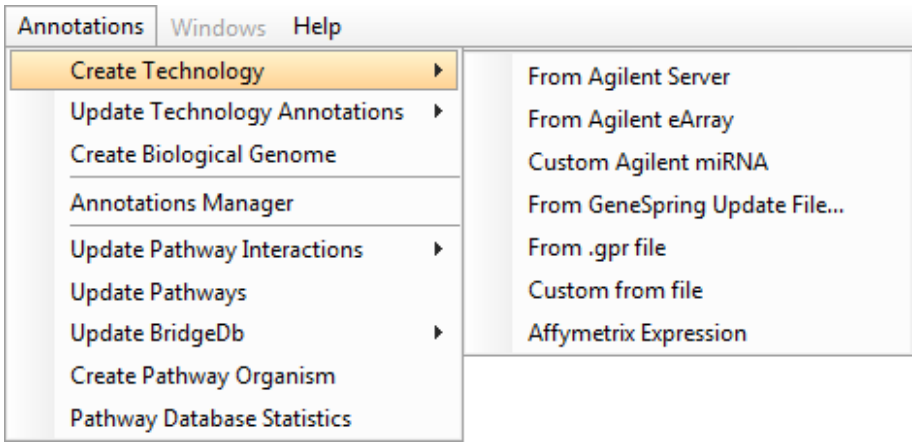


Figure 3.1: Create Technology

3.1.1 Standard Technology Creation

The creation of a Standard Technology involves processing of the information present in the annotation files into a standard internal format used in **GeneSpring**. This is done for greater efficiency while using functionalities such as GO Analysis.

The different files used for different technologies are detailed below:

- **Affymetrix Expression:**

  The files that are used for creating a Standard Technology are .csv, .cdf, .psi, .cif and probetab. The .cif file is needed when summarization is being performed using MAS5. Likewise the probetab file is required while doing summarization using GCRMA. Additional parsing of the data files goes on during technology creation, for eg., the chromosomal information pertaining to a probe (number, strand, position and orientation) in the Affymetrix annotation file is present in a single column and during the process of technology creation; this is split into 4 different columns. The information required for creating a Standard Technology is taken from the following site: http://www.affymetrix.com/analysis/index.affx

- **Affymetrix Exon Expression:**

  The files that are used for creating a Standard Technology are .clf, .pgf and the transcript level .csv annotation file. The meta probeset lists regarding the Core, Extended and Full transcripts are the same as Affymetrixs files and are packaged with the Standard Technology. The information required for creating a Standard Technology is taken from the following website: http://www.affymetrix.com/analysis/index.affx.

- **Affymetrix Exon Splicing:**

  The files that are used for creating a Standard Technology are .clf, .pgf, probeset level .csv annotation file and the transcript level .csv annotation file. The meta probeset and the probeset files regarding the Core, Extended and Full transcripts and exons are the same as Affymetrixs files and are packaged with the Standard Technology. The information required for creating a Standard Technology is taken from the following website: http://www.affymetrix.com/analysis/index.affx.

- **Agilent Exon Expression:**

  The files that are used for creating a Standard Technology are the GEML files which describe the actual chip lay out and the annotations files including probe, gene and transcript level annotations and GO annotations for genes. The technology name will be 'Agilent.SingleColor.Exon.XYZ' or 'Agilent.TwoColor.Exon.XYZ where XYZ is the AMADID number.

- **Illumina:**

  The creation of a Standard Technology for Illumina arrays uses the information content of the .bgx manifest file to associate the annotations with the probes. The information required for creating a Standard Technology is taken from the following website: http://support.illumina.com/documentation.ilmn.

- **Agilent Single and Two Colour:**

  The creation of a Standard Technology for Agilent arrays involves parsing the biological information present in the annotation file into a **GeneSpring** recognizable format. For eg., the chromosomal information pertaining to a probe (chromosome number, strand, position and orientation) is present in a single column in the annotation file while the GO annotations are present in 3 columns. During technology creation, the chromosomal information is parsed into 4 columns while the GO annotations are collapsed into 1 column. Annotations for Agilent arrays are available on the following website: http://www.chem.agilent.com

- **Agilent miRNA:**

  From **GeneSpring** version 11.5, the technology creation for miRNA experiments are no longer done on the fly; standard technologies are created for miRNA after parsing the GEML design file from earray for the given AMADID. The GEML file contains annotations which can be mapped to the mirBase Version Id. The standard technology name for miRNA in **GeneSpring** would be Agilent.mirna.$< AMADID >_< mirBaseId >$.

  The fields in the standard technology along with the actual field names shown in Technology Inspector would be:

  - miRNA name; shown as 'systematic_name' in Technology Inspector
  - mirBase Accession No; shown as 'mirbase accession No' in Technology Inspector
  - Chromosome; shown as 'chr' in Technology Inspector
  - Chromosome Start; shown as 'start' in Technology Inspector
  - Chromosome End; shown as 'end' in Technology Inspector
  - Sequence of the probes; shown as 'active_sequence' in Technology Inspector

  The Technology Inspector also shows a field on 'control_type'. The columns to be shown can be configured. Apart from these fields which are shown to the user while inspecting technology, the standard technology creates and stores a map for matching Agilent Probe IDs to the corresponding miRNA names. During the technology creation, the mirBase (MIMAT) accession numbers are extracted from the GEML files and are stored as a column in the technology as listed above. This field will be used to find the targets for miRNAs across mirBase versions for Targetscan.

  Before creating miRNA experiments, user should download the technology or technologies from the menu **Annotations → Create Technology → From Agilent Server or From Agilent eArray**.

- **Real Time PCR:**

  The technology creation is dependant on the samples given and each individual experiment has its own technology. This technology creation does not have annotations associated with

it. The user can update annotations after experiment creation from **Utilities**⟶ ***Update RTPCR Technology Annotations*** under the workflow navigator.

- **Copy Number Analysis:**

  For Copy Number Analysis, **GeneSpring** 11.0 supports the following standard technologies:

  1. **Affymetrix Genome-Wide Human SNP Array 6.0, Genome-wide Human SNP array 5.0, and Human Mapping 500K Array Set**
  2. **Affymetrix Human Mapping 100K Set**
  3. **Illumina Genotyping output files from GenomeStudio**

  Refer to Chapter Copy Number Analysis for details.

- **Association Analysis:**

  GeneSpring supports the following technologies for Association Analysis experiments:

  – Affymetrix Mapping 50K Xba240
  – Affymetrix Mapping 50K Hind240
  – Affymetrix Mapping 50K Xba240 and 50K Hind240
  – Affymetrix Mapping 250K Nsp
  – Affymetrix Mapping 250K Sty
  – Affymetrix Mapping 250K Nsp and 250K Sty
  – Affymetrix GenomeWide SNP5
  – Affymetrix GenomeWide SNP6
  – Genotyping Output files from Illumina GenomeStudio
  – Any file created in Illumina GenomeStudio output format (refer to Illumina File Format section for details).

  Refer to Technology section for details.

3.1.2 Agilent eArray Technology Creation

Agilent Single color and Two color technologies can also be created for arrays ordered through the eArray portal of Agilent. This can be accessed from ***Annotations***⟶ ***Create Technology***⟶ ***From Agilent eArray***. An account at eArray is required to create technology using this feature. Details such as user name, password for your eArray account, AMADID Number, organism and technology type are required for creating an eArray technology. See Figure 3.2. Once the details are provided, a technology is created along with annotation columns like Entrez-ID, GO etc (if available) from the tab delimited text (TDT) file of the specified eArray.

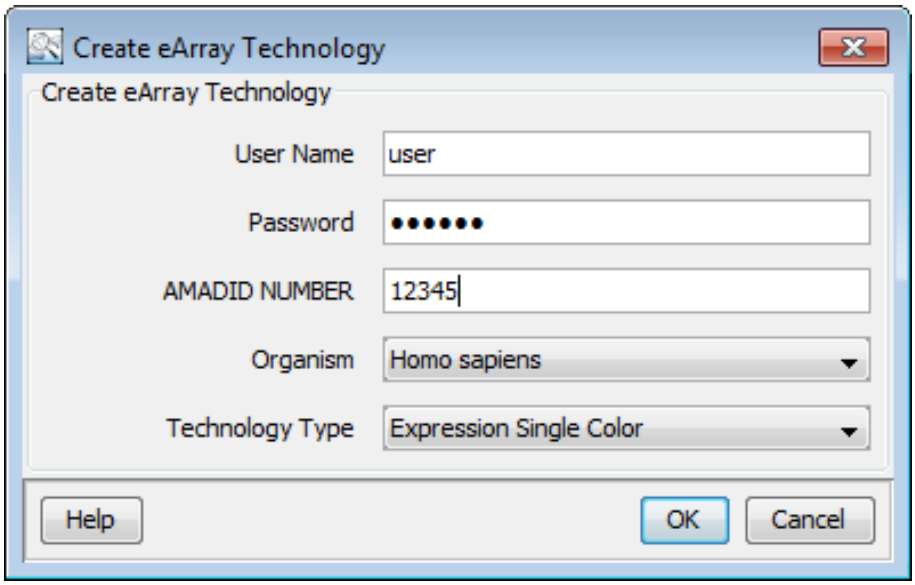


Figure 3.2: Technology Creation

If the user wants to populate the created technology with more annotations, then this can be done through ***Annotations*** $\longrightarrow$ ***Update Technology Annotations*** $\longrightarrow$ ***From Agilent eArray***. This opens up a window in which the user needs to key in information regarding the User Name, Password and the Technology Name. See Figure 3.3. This opens up a window which is similar to the step 3 of Update Technology Annotations from file. **GeneSpring** uses the information present in the 'AllAnnotations' file of the eArray to populate the technology with further annotations. In case this file is not available for the eArray, then it uses the TDT file to populate the technology.

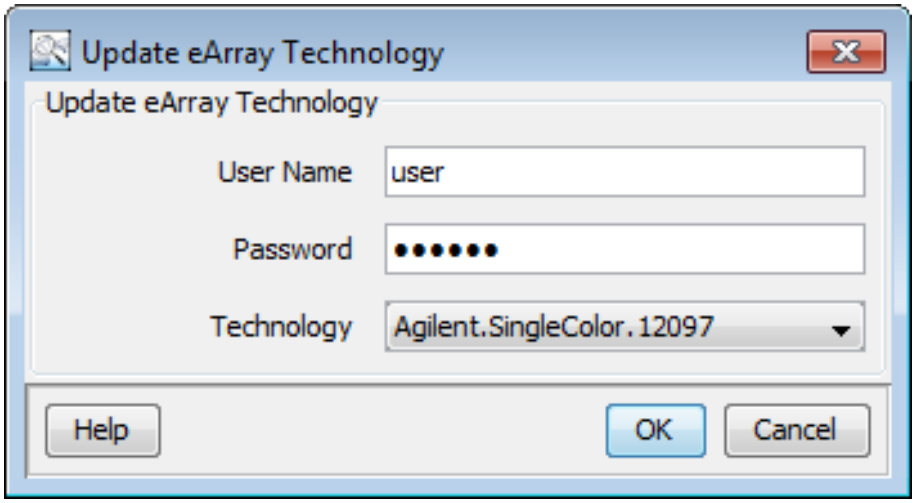


Figure 3.3: Technology Update

3.1.3 Custom Technology Creation

GeneSpring allows the user to create a custom technology. This is useful in cases where the user has a custom array from the vendors mentioned above or has an array from a different vendor. The option to create a custom technology Generic One or Two Color arrays can be availed from ***Annotations***⟶ ***Create Technology***⟶ ***Custom from file***. Custom Technology for Affymetrix Expression arrays (if a custom .cdf file is available) can be created using ***Annotations***⟶ ***Create Technology***⟶ ***Affymetrix Expression***. For GenePix Results (.gpr) format files use ***Annotations***⟶ ***Create Technology***⟶ ***From .gpr file***. Please note that .gpr technology creation is only supported for two color files.

You can create a Custom Technology to run an Illumina Association Analysis experiment on any file created in GenomeStudio output format (refer to Illumina File Format section for details).

3.1.4 Creating Custom CGH Technology

In **GeneSpring** you can use a Custom CGH Technology to create a new CGH expeiment. Follow these steps to create a custom CGH technology:

1. Download the *.tdt file for your custom CGH array from the SureDesign portal to your system.
2. Go to **Annotations** ⟶**Create Technology** ⟶**Custom Agilent CGH**.
   The **Create Custom CGH Technology** dialog opens.
3. Specify the AMADID NUMBER for your custom CGH array.
4. Specify the custom CGH array file by clicking on **Browse** and point to the file location on your system.
5. Click **OK** to complete the creation of custom CGH technology. This technology will be available for creating a new CGH experiment.

3.1.5 Technology creation on the fly

This option is used by the application when Agilent FE files are used to create an experiment and the technology for the FE file does not exist either in the **GeneSpring** application or on the Agilent server. It allows the user to proceed with the experiment creation and creates a technology. This technology is created with just the identifier column and without any annotations. The annotations can be updated later on as and when the annotation files are available. This can be done from ***Annotations***⟶ ***Update Technology Annotations***. This update can be done using either the ***From Agilent eArray***(Refer to section on eArray) or the ***From file or Biological Genome*** options.

An organism is needed for creating the technology and the user is prompted for the same during the workflow. Please note that technology creation on the fly will also come into picture when the

technology does not exist in **GeneSpring** and the application could not connect to the Agilent server to download for the technology.

3.1.6 Inspection of Technology

A technology once created or downloaded can be inspected at any time using the Technology Inspector. It is accessible by right-clicking on the experiment name in the project navigator and provides information regarding the organism, type (Single or Two Color), version (for Standard Technology) and the number of entities and the date of creation. Except for the organism name and notes, none of the other information can be edited. The set of annotations associated with the entities can be customized using the ''Configure Columns'' button, and can also be searched for using the search bar at the bottom. Further hyperlinked annotations can be double-clicked to launch a web browser with further details on the entity.

3.1.7 Technology Deletion

Technologies once created can be deleted if no longer in use. This can be done using *Search* $\longrightarrow$ *Technology*. The toolbar in the search wizard has an icon for deleting technology.

3.2 Update Technology

The available technologies in **GeneSpring** can be updated regularly. Updates can be carried out in a file-based format by using the necessary file (provided by **GeneSpring** support on request) or can be updated from the update server. Updates are available on the server whenever new data libraries are made available by the chip manufacturers.

Data libraries are also required for other applications in the tool. For example, Gene Ontology (GO) data is necessary for GO analysis. To see the available updates, go to ***Annotations*** $\longrightarrow$ ***Update Technology Annotations*** $\longrightarrow$ ***From Agilent Server***. This will contact the update server, validate the license and show the data libraries available for update. The user can look for the desired technology using the search functionality. The search is a alphanumeric, partial-keyword, case insensitive one performed across all the columns in the *Available Updates* table. Then the corresponding library can be selected from the list of technologies. Details of the selected libraries will appear in the text box below the data library list. See Figure 3.4.

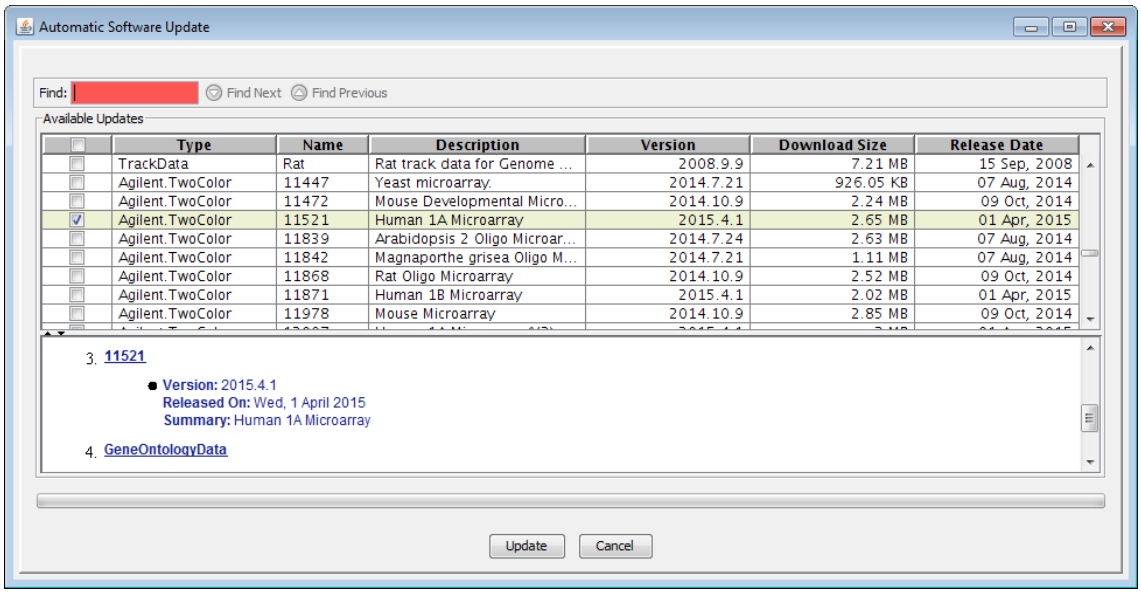


Figure 3.4: Data Library Updates Dialog

You can Left-Click on the check box header to select or unselect all the data libraries. Left-Click on a check box will toggle the selection. Thus if the check box is unselected, Left-Click on it will select the row. If the row is selected, Left-Click on the check box will unselect the row. Shift-Left-Click on the check box will toggle the selection of all rows between the last Left-Click and Shift-Left-Click .

You can sort the data library list on any column by Left-Click on the appropriate column header.

3.2.1 Automatic Query of Update Server

When experiments are created, if the appropriate libraries are not available, the tool will prompt the user to download the required data library before proceeding further. See Figure 3.5

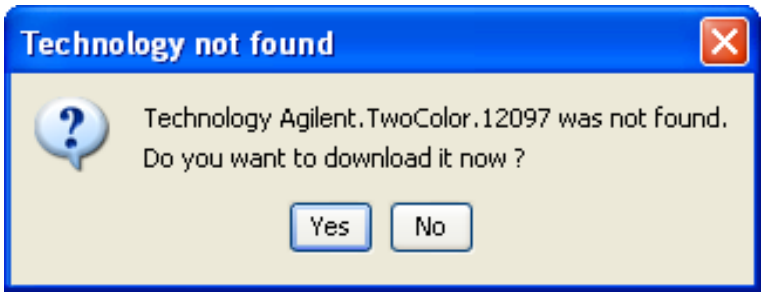


Figure 3.5: Automatic Download Confirmation Dialog

3.2.2 Update Technology Annotations

Update Technology Annotations, enables the user to update the annotations of an existing Standard or Generic technology. It is a particularly useful feature when newer information necessitates updating an existing technology. The Standard Technologies can typically be updated from the web whenever the chip vendor releases newer annotation. However, this particular feature of **Update Technology Annotations** for Standard Technologies is used when you would want to add additional information over and above that provided by the vendor. Different ways to update technology annotations can be accessed from the menu ***Annotations*** $\longrightarrow$ ***Update Technology Annotations***. See Figure 3.6

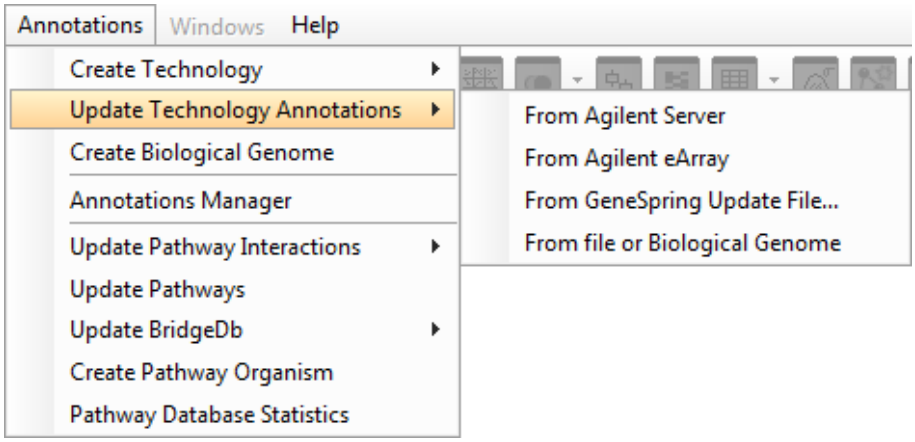


Figure 3.6: Update Technology Annotations

1. **From Agilent Server** - Updates are available on the server whenever new data libraries are made available by the chip manufacturers.
2. **From Agilent eArray** - Agilent technologies can be updated from eArray directly. You will require username and password of eArray to access.
3. **From GeneSpring Update File** - Standard technologies can be updated using GeneSpring Update file (provided by **GeneSpring** support on request)
4. **From file or Biological Genome** - It can be accessed from the menu ***Annotations***$\longrightarrow$ ***Update Technology Annotations***$\longrightarrow$ ***From file or Biological Genome***:
   (a) **Step 1 of 3** - Here the user specifies the technology as well as the source from which it has to be updated. The technology can be updated either from a file or from the **Biological Genome** of that organism. If the **Biological Genome** of that organism does not exist, then the user can create a genome from ***Annotations***$\longrightarrow$ ***Create Biological Genome***. For more details on the creation of a genome, refer to Biological Genome. If the user chooses to update from a file, then it should be chosen accordingly via the ***Choose file*** option. The file from which the update is to be performed has to be in a tabular format. This is seen in Figure 8.14.

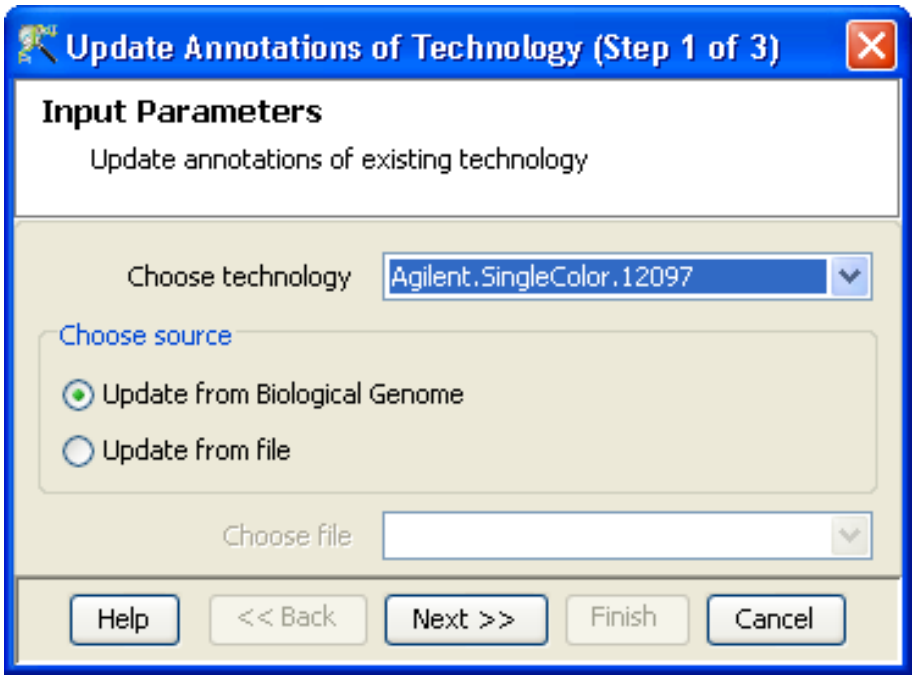


Figure 3.7: Input Parameters

(b) **Step 2 of 3** - This step appears only if the update source is a file. This step asks the user to input the file format of the annotations update file. This involves specifying format options, i.e., the Separator, Text qualifier, Missing value indicator and Comment Indicator of the file. This is seen in Figure 13.2.

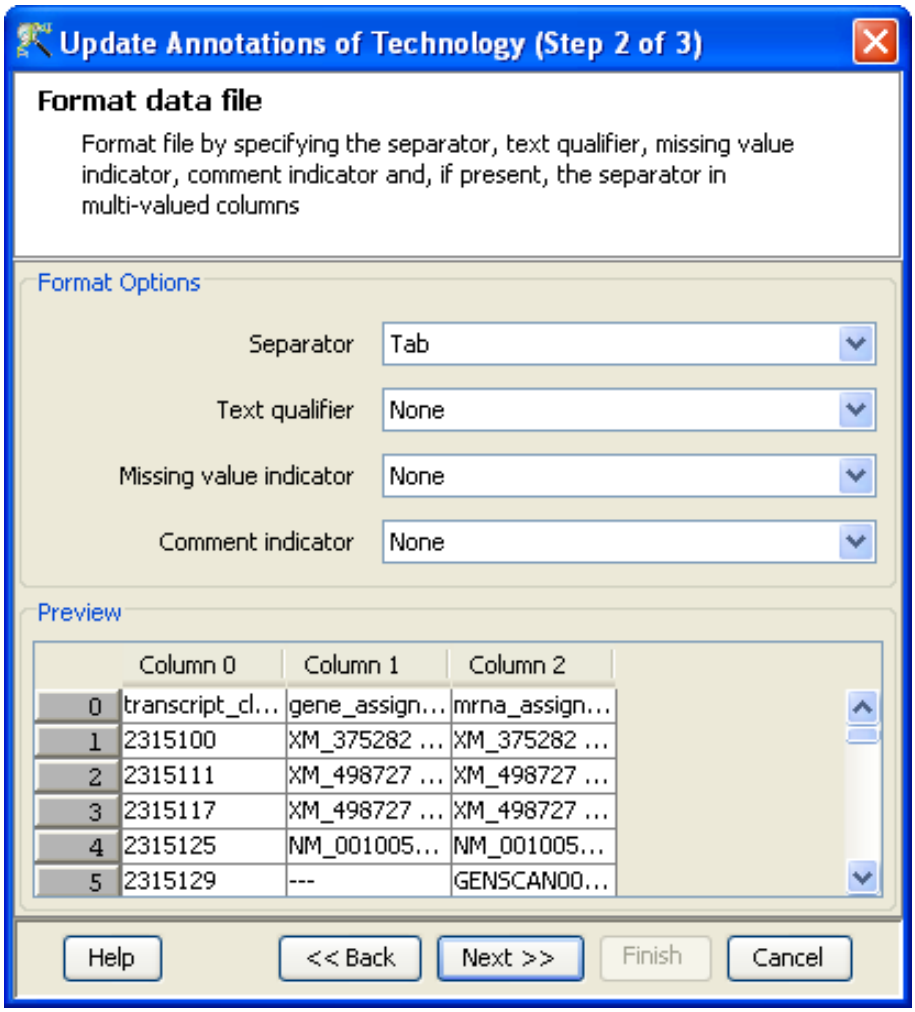


Figure 3.8: Format data file

(c) **Step 3 of 3** - The annotation columns are merged with the existing technology using a technology identifier. This step asks the user to specify the identifier and to choose the column to be updated from the annotation file/genome. While specifying the columns, column marks should be assigned (similar to how it was done while creating the Generic technology). Please note that the Entrez ID annotation column needs to be first specified as categorical and then as string in the dropdown. It is recommended that the user chooses a column with unique values as the identifier. Three kinds of updates are possible:

- Append to the existing information,
- Overwrite
- Fill in the missing values.

Appending the values will retain the original value as well as add the new value. Overwrite will replace the original value with the newer one, whereas fill in missing values will add values at places where previously there were none. This is seen in Figure 3.9.

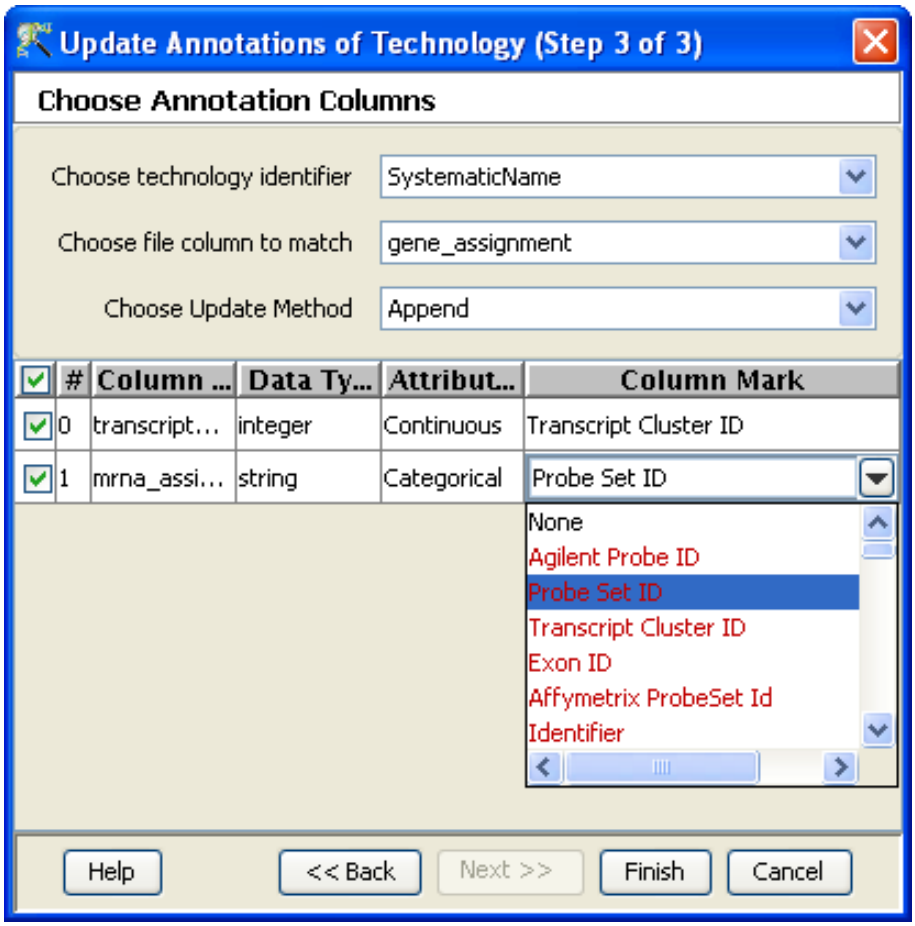


Figure 3.9: Choose Annotation Columns

The updated annotation values for existing columns can be seen by right click on *Experiment*⟶ *Inspect Technology*.

3.3 Annotations Manager

Annotations Manager, accessible from the **Annotations** menu is the central interface for updating, deleting or adding custom annotations for various experiment types. It allows the user to organize and view multiple builds of different organisms and their respective annotations for genes, transcripts, reference, and other annotation resources.

Updating annotations is required to compare experiment data and results of analyses with annotations on the GB for both imported Strand NGS experiments and core **GeneSpring** experiments like expression analysis, exon splicing analysis, copy number analysis, etc.

Note:

Reference annotations are specific to sequence based analyses, i.e. for imported Strand NGS experiments. However, the remaining annotation types are also used in experiments like expression, exon splicing, copy number and association experiments.

The Annotations Manager in **GeneSpring** comes prepackaged with GB annotations data for humans (hg18 and hg19 builds), mouse, rat, *C. elegans* and other organisms. See Figure 3.10. These data are obtained from various sources like UCSC (`http://hgdownload.cse.ucsc.edu/downloads.html`), NCBI (`ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/`) and ENSEMBL (`http://asia.ensembl.org/index.html`) and can be updated into **GeneSpring** via the Annotations Manager.

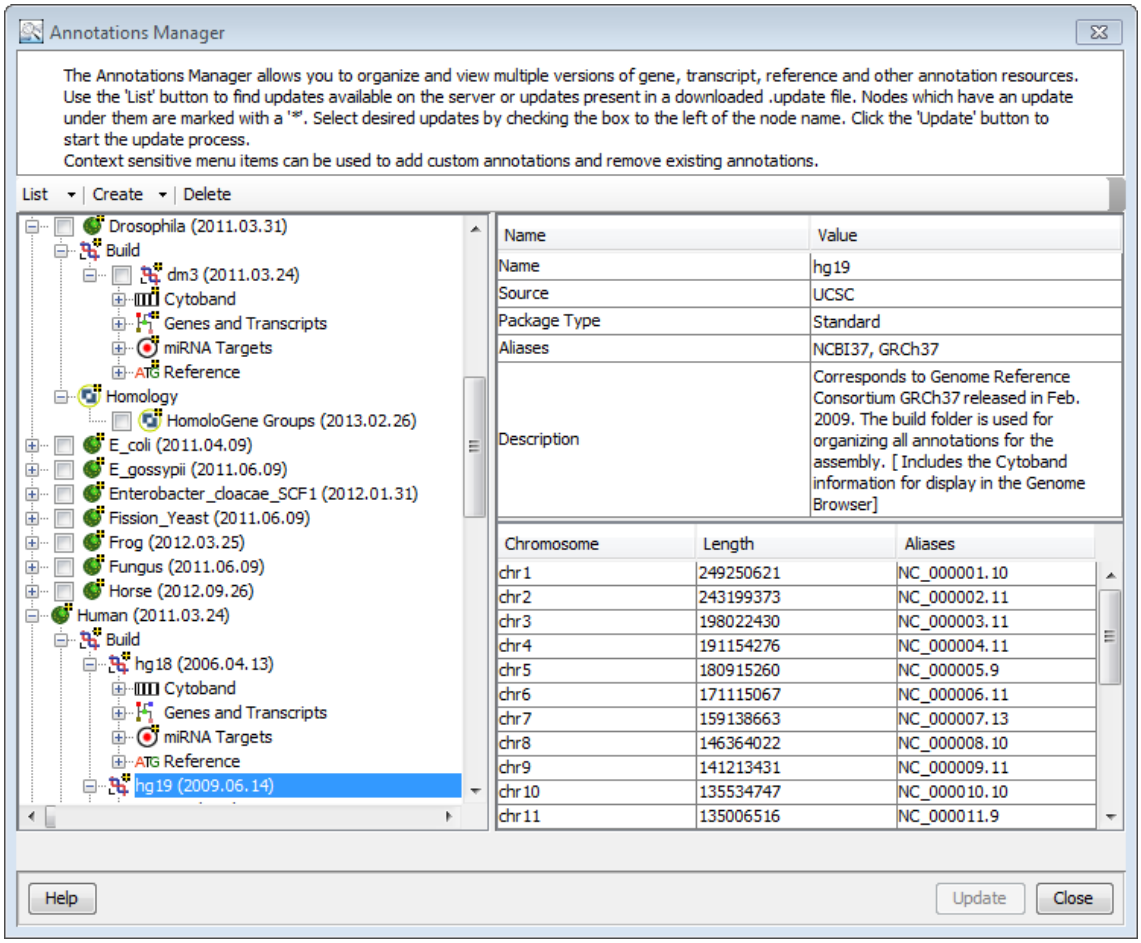


Figure 3.10: Annotations Manager

On the first launch, **GeneSpring** opens the Annotations Manager window which displays all the annotations that are available for download from the Agilent Server as annotations aren't packaged by default with the installer. From subsequent launches, Annotation Manager will show the annotations that have been downloaded. These annotations are organized in a tree hierarchy making it easy and quick to find all the information for a particular organism and build. The <**List**> option on the left can be used to find updates that haven't been downloaded and are available from the Agilent Server or in an '.update' file provided by the **GeneSpring** Sales or

Technical Support Team (informatics_support@agilent.com). Annotations which have updates available are marked with a '∗'. The desired updates can be selected by checking the box to the left of the node (annotation) name and by clicking on the < **Update**> button to download the data. After the update has been downloaded, the selected annotations get installed in the **GeneSpring** System. These annotations can be accessed from appropriate workflow links in the selected experiment type, e.g.: For all experiment types in the **GeneSpring** application, to view data on GB, the downloaded gene/transcript annotations are accessible from the ' **Annotations**' tab in the GB window.

If an organism, build and/or annotations are not listed in the Annotations Manager, they can be loaded from appropriate data files for that organism. The < **CREATE**> button can be used to create new *Organism*, *Build* and *Annotations*, thereby creating a custom genome by choosing appropriate annotation objects. The menu items show a list of appropriate annotation objects, based on the selected object in the navigator, i.e., when an ***Organism*** is selected (or clicked on), the Create menu shows '**New Homology**' and '**New Build**' as the only appropriate options, while '**New Transcript**' and '**New Gene Annotation**' (and others) will be shown when a Build is the selected object in the navigator.
A selected organism, build or annotations can be deleted using the <**Delete**> button in the menu item.

3.3.1 Create New Organism

If a desired organism is not listed in the Annotations Manager, a new organism can be created using the < **CREATE**> button in the Annotations Manager window. Selecting this and subsequently the 'Add Organism' option opens a window. See Figure 3.11. In this window it is essential to specify a Common name, Scientific name and Taxonomy ID for the organism. Taxonomy IDs can be obtained from http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Root. On selecting ***OK***, the Annotations Manager window gets refreshed to display the newly created organism in the tree hierarchy on the left. Using right click option on the created organism, homology and build information can be added. Any number of builds can be added to the organism with each build representing a distinct genomic assembly.

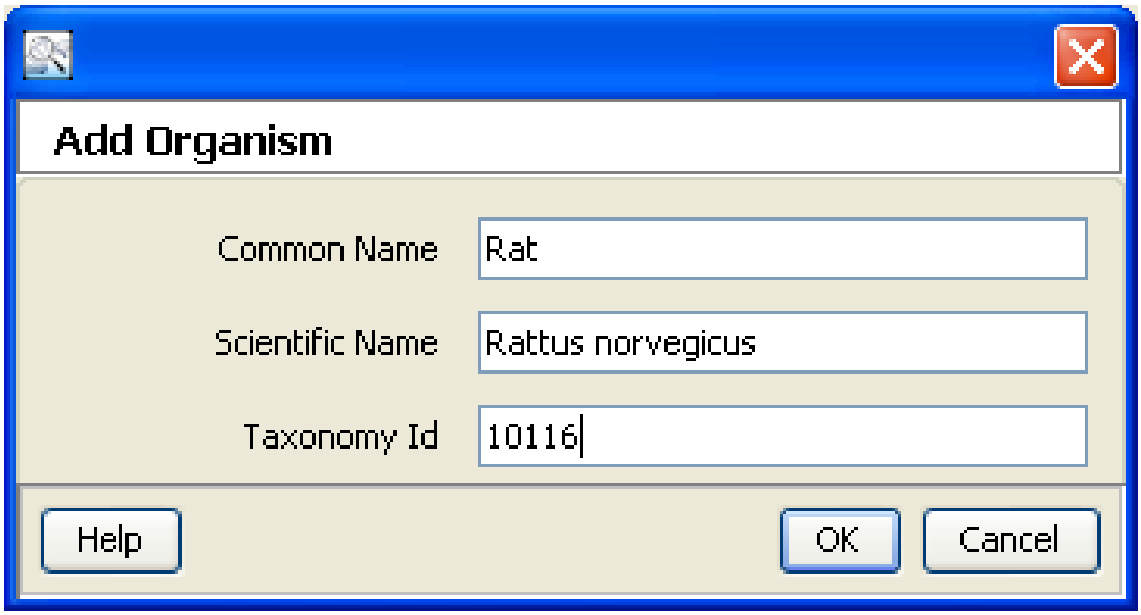


Figure 3.11: Create New Organism

3.3.2 Create New Build

A build is identified by a unique name and contains details of a specific genomic assembly of an organism. A build stores information about the number of chromosomes in a given organism and their lengths. Thus, a new build for an organism can be created by providing a two column file containing the names and lengths of each of the chromosomes. In the Annotations Manager, 'Add Build' option can be selected by right clicking on the desired organism. This opens a window where, a build name and source are to be entered, and a file uploaded with chromosome names and lengths in a tab delimited format. On selecting ***OK***, the build information gets added under the organism node. This created build can be examined in the Annotations Manager and Aliases can be added to the build itself or to individual chromosomes by editing the Aliases cell. See Figure 3.12.

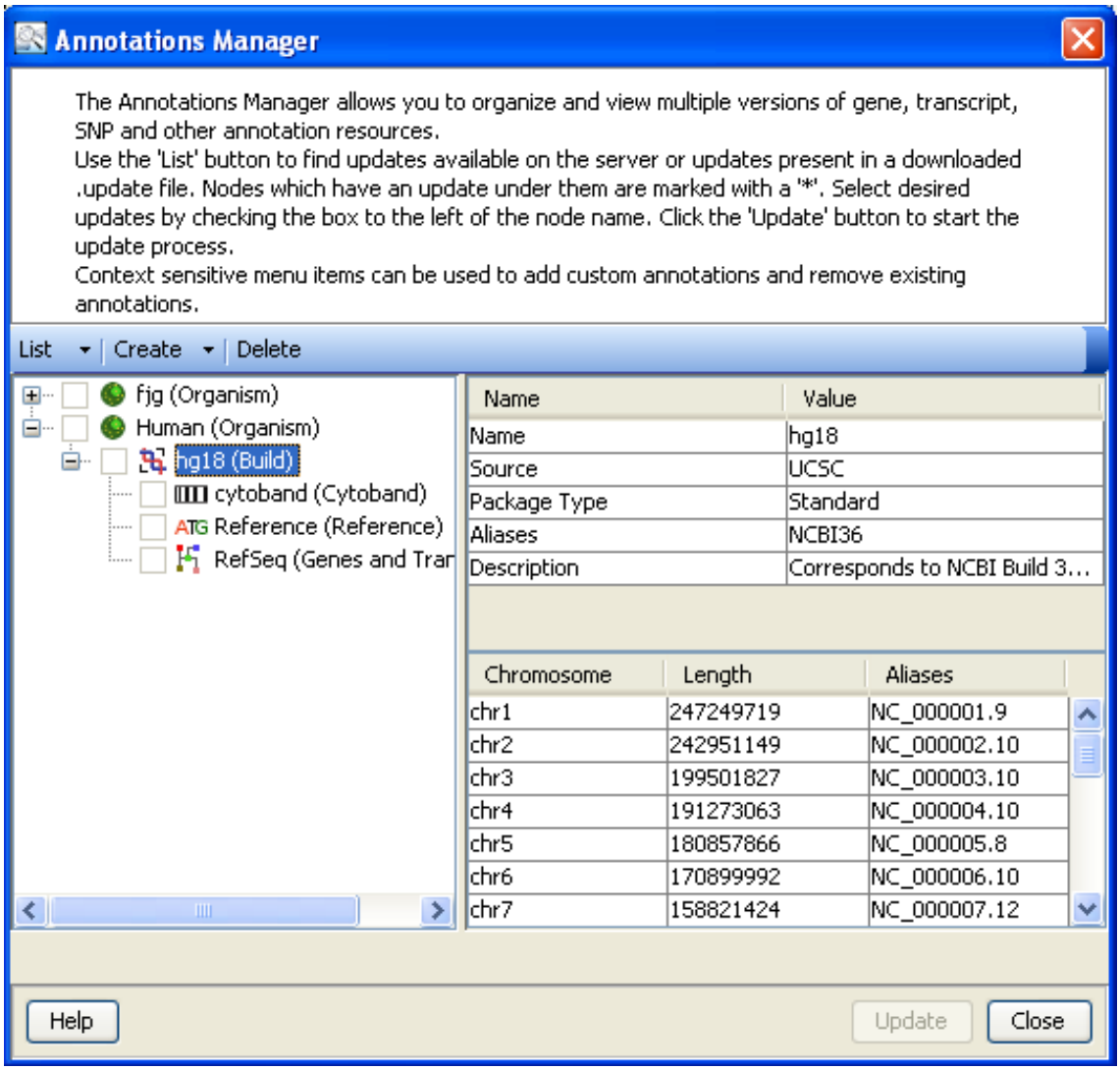


Figure 3.12: Create New Build

Further within a build, various annotations like Genes and Transcripts, Cytoband, and Reference can be added. These annotations can be selected by right clicking on the desired build. See Figure 3.13.

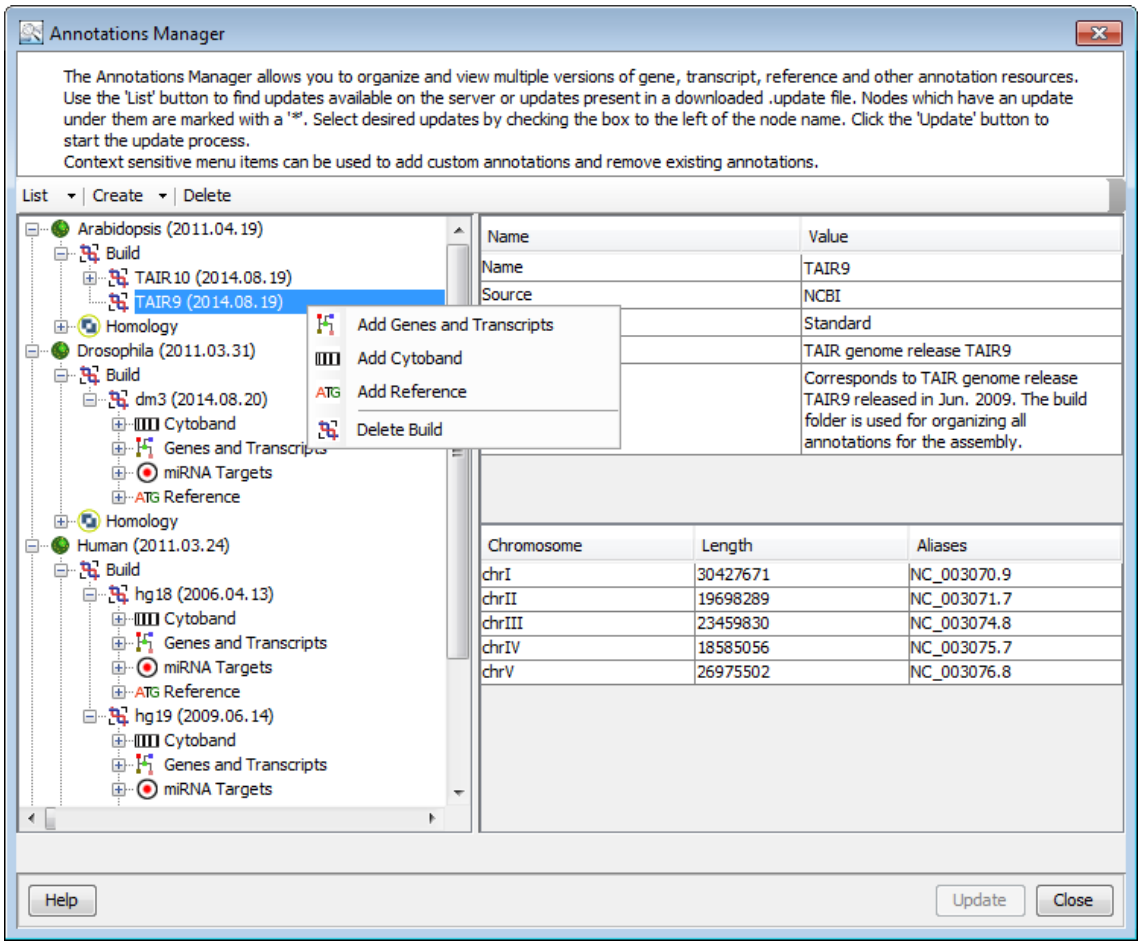


Figure 3.13: Add Annotations to a build

Gene and Transcript annotations can be used to compare experimental data/ results. Multiple gene and transcript annotations can be present in a single build. A list of standard gene and transcript annotations obtained from sources like UCSC, NCBI and ENSEMBL can be downloaded from the Annotations Manager.

For a given organism and a build, gene and transcript annotations from different sources can be selected to view on a single GB window. This can be used to compare the experimental data with annotations from different sources. Also for a given organism, multiple GB windows can be launched using different builds. Now, between multiple builds (i.e., between multiple GB windows), gene and transcript annotations from a single source can be compared.

New gene and transcript annotations can be added from the 'Add Gene and Transcript annotations' drop down by right clicking on the desired build. This opens a window as shown in Figure 3.14.

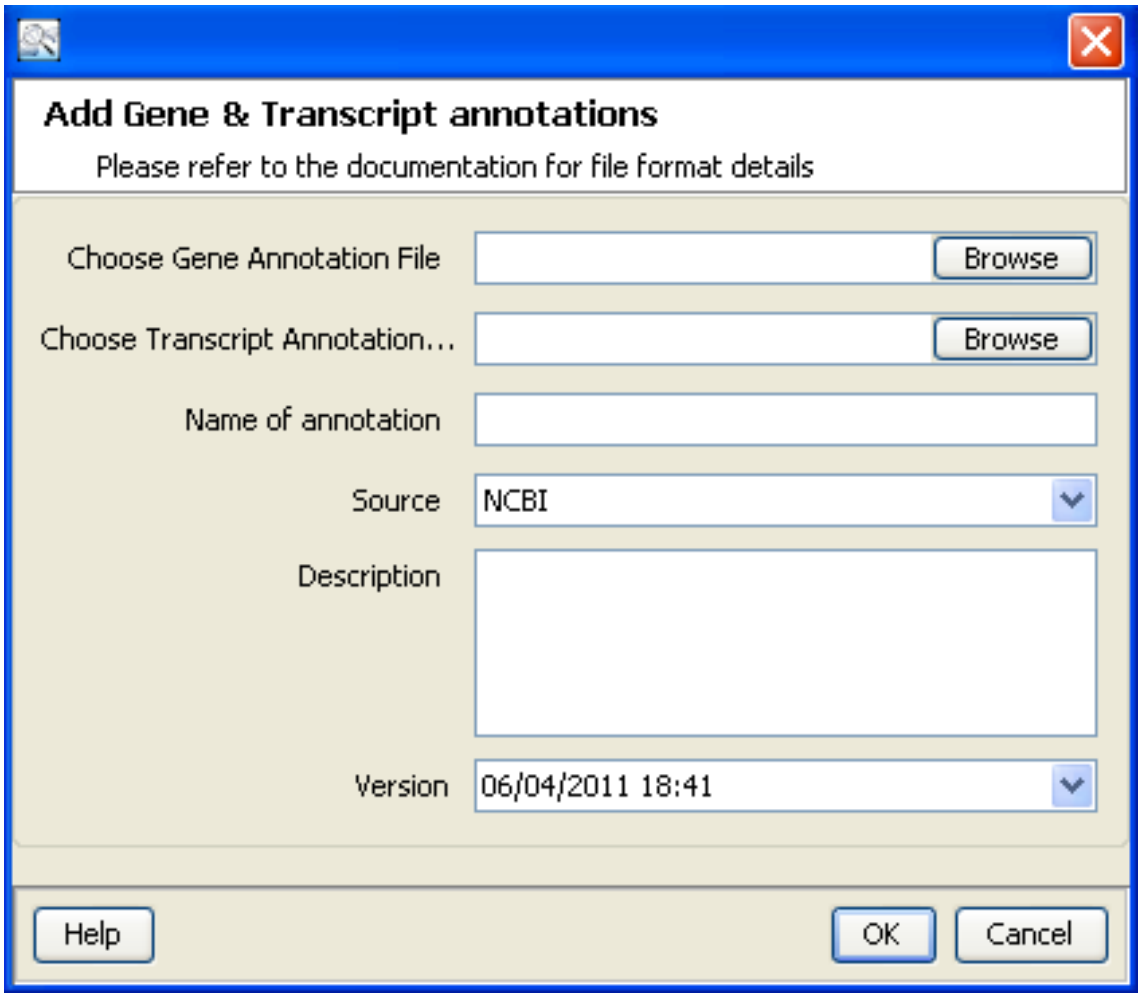


Figure 3.14: Add Gene and Transcript annotations

In this, the gene and transcript annotations should always be loaded together along with the annotation name, source and version attributes. If the transcript information is unavailable, an empty file (.txt file) can be selected along with the gene annotation file. Both the annotation files should have 1-based notation to represent chromosomal locations. The gene annotation file is a fixed format, i.e., a tab delimited file with a header row. See Table 3.1 and the sample gene annotation file below. Multiple values in columns 7-13 of the gene annotation should be separated by the '|' character. The transcript annotation file is also a tab delimited file but without any header row. See the Table 3.2 and sample transcript annotation file below.

Column	Name	Empty Allowed	Type	Constraints
1	Identifier	N	String	Unique in the file
2	Gene Symbol	N	String	
3	Chromosome	N	String	Match a chromosome in the build
4	Start	N	Integer	
5	End	N	Integer	
6	Strand	N	+ or -	
7	Aliases	Y	String	
8	Description	Y	String	
9	Gene Type	Y	String	
10	Entrez ID	Y	String	
11	Ensembl ID	Y	String	
12	GO Accession	Y	String	
13	DBXrefs	Y	String	

Table 3.1: Gene Annotation file format

Few lines from a sample gene annotation file are shown below:

```
Identifier Gene Symbol Chromosome Start End Strand Aliases Description Gene Type Entrez ID \
Ensembl ID GO Accession DBXrefs
373854 ODZ2 chr13 4581438 5056955 - TEN2 odz, odd Oz/ten-m homolog 2 (Drosophila) \
protein-coding 373854 ENSGALG00000001768 GO:0007165|GO:0016020|GO:0016021 CGNC:1242
373885 FZD10 chr15 3445651 3447893 - cFz-10 frizzled homolog 10 (Drosophila) protein-coding \
373885 ENSGALG00000002652 \
GO:0004926|GO:0004930|GO:0005515|GO:0007275|GO:0016020|GO:0016021|GO:0016055
373886 FZD9 chr19 254446 256193 -  frizzled homolog 9 (Drosophila) protein-coding 373886 \
 GO:0004926|GO:0004930|GO:0007275|GO:0016020|GO:0016021|GO:0016055 CGNC:538
373887 FZD4 chr1 193006476 193011404 + FZ-4 frizzled homolog 4 (Drosophila) protein-coding \
373887 ENSGALG00000017242 \
GO:0004926|GO:0004930|GO:0005515|GO:0007275|GO:0016020|GO:0016021|GO:0060070 CGNC:12957
373888 TFAM chr6 6292548 6297550 -  transcription factor A, mitochondrial protein-coding 373888 \
ENSGALG00000002703 GO:0003677|GO:0005515|GO:0005634|GO:0005739|GO:0033108|GO:0042645 CGNC:1948
373889 HK1 chr6 11979962 12004711 +  hexokinase 1 protein-coding 373889 ENSGALG00000004222 \
GO:0000166|GO:0004396|GO:0005524|GO:0005739|GO:0005829|GO:0005975|GO:0006096|GO:0016301|GO:0016740 \
CGNC:3101
373891 RECQL5 chr18 4769798 4805970 + RecQ5 RecQ protein-like 5 protein-coding 373891 \
ENSGALG00000002424 GO:0003676|GO:0004386|GO:0005524|GO:0006310|GO:0008026|GO:0016787
373892 ADD3 chr6 27211689 27268533 +  adducin 3 (gamma) protein-coding 373892 ENSGALG00000008542 \
 GO:0000794|GO:0005198|GO:0005737|GO:0005856|GO:0005911|GO:0005938|GO:0016020|GO:0046872
```

Column	Name	Empty Allowed	Type	Constraints
1	Gene Name	N	String	
2	Transcript Name	N	String	
3	Chromosome	N	String	Match a chromosome in the build
4	Strand	N	+ or -	
5	Start	N	Integer	
6	End	N	Integer	
7	Cds start	N	Integer	
8	Cds end	N	Integer	
9	Number exons	N	Integer	
10	Exon starts	N	Integer	Starts of the exons separated by comma
11	Exon ends	N	Integer	Ends of the exons separated by comma
12	Gene Id	N	String	Match an identifier in gene annotation file

Table 3.2: Transcript Annotation file format

Few lines from a sample transcript annotation file are shown below:

```
GHSR NM_004122 chr3 - 172165333 172166246 172165334 172166203 1 172165333 172166246, 2693
CCR8 NM_005201 chr3 + 39371197 39375171 39373823 39374890 2 39371197,39373809 39371320,\
39375171, 1237
OR5K2 NM_001004737 chr3 + 98216525 98217474 98216525 98217474 1 98216525 98217474, 402135
HRH1 NM_001098213 chr3 + 11178779 11304938 11300724 11302187 2 11178779,11300689 11179085,\
11304938, 3269
HTR1F NM_000866 chr3 + 88031726 88042919 88039900 88041000 2 88031726,88039858 88031841,\
88042919, 3355
HRH1 NM_001098212 chr3 + 11196214 11304938 11300724 11302187 2 11196214,11300689 11196240,\
11304938, 3269
HRH1 NM_001098211 chr3 + 11267669 11304938 11300724 11302187 2 11267669,11300689 11267745,\
11304938, 3269
CCDC71 NM_022903 chr3 - 49199969 49203785 49200238 49201641 2 49199969,49203700 49201693,\
49203785, 64925
IQCF5 NM_001145059 chr3 - 51907738 51909600 51907749 51909547 2 51907738,51909544 51908191,\
51909600, 389124
SST NM_001048 chr3 - 187386696 187388201 187386853 187388079 2 187386696,187387942 \
187387065,187388201, 6750
```

Note:

If a user selects a 0-based gene and a transcript annotation file, **GeneSpring** assumes it to be a 1-based file and proceeds with the processing.

Each build can only have a single cytoband annotation and a single reference sequence annotation. Cytoband information is used by the GB to display cytoband in Chromosome view and Genome view.

To create new reference sequence annotations for a build, 'Add Reference' should be chosen from the dropdown. In step 1 of the 'Create New Reference' wizard, apart from selecting the source of reference annotation, fasta files containing the chromosomal sequences should be uploaded. Each fasta file should have a single fasta sequence. The Annotations Manager proceeds with the import

of these files only if the number of selected fasta files match with the number of chromosomes in the build. Since chromosome naming conventions differ widely-chr1, c1, 1, NC_000001.10, etc., the Annotations Manager allows the user to verify the *fasta file* $\longrightarrow$ *chromosome name mapping* and make corrections if necessary. Now, the newly added reference sequence annotations can be seen in the navigator.

A selected organism, build and its corresponding annotations can be deleted either using the < **Delete**> option in the menu item or a right click on each of the organism, builds or annotation objects provides an option to delete the respective annotations.

Note:

- In the Annotations Manager, when the box on the left of the organism is selected, i.e., when an organism is chosen for update, all the nodes below (builds and annotations) also get selected. This way all the available annotations can be updated under an organism.Also, only relevant boxes can be left selected and the rest unselected to download only the desired annotations.
- When a build is selected for an update by selecting the box to the left of the build node, the corresponding organism and the annotations relevant to the build are also selected by default for update.
- When an annotation of a build is selected for update, then the build of the chosen annotation and the corresponding organism are also selected for update.

3.4 Translation

Translation is a feature that allows comparison of entity lists between experiments of different technologies. A standard use case of translation involves comparison of experiments done on a single organism but different technologies, e.g., Human samples on HG_U95Av2 and HG-U133_Plus_2. Another situation would be to identify the homologues, eg mapping Human genes to Mouse genome. The automated detection of homologs (similarity attributable to descent from a common ancestor) among the annotated genes of several completely sequenced eukaryotic genomes is performed using HomoloGene.

GeneSpring could have multiple experiments open at the same time. Exactly one of these experiments is active at any time. The desktop in the center shows views for the active experiment. You can switch active experiments by either clicking on the title bar of the experiment in the Navigator, or by clicking on the tab title of the experiment in the main Desktop. When the active experiment is changed, the active entity list of the project is also changed to the ''All Entities'' entity list of that experiment.

As mentioned before, if you click on another entity list of the active experiment, all views of that experiment are restricted to show only the entities in that entity list. In addition if you click on an entity list of an experiment other than the active one, the views are still constrained to show only that entity list.

Note that if the two experiments do not correspond to the same technology then entities in the entity list will need to be translated to entities in the active experiment. **GeneSpring** does this translation seamlessly for a whole range of organisms which are given in the table below.

Serial No.	Organism
1	Mus musculus
2	Rattus norvegicus
3	Magnaporthe grisea
4	Kluyveromyces lactis
5	Eremothecium gossypii
6	Arabidopsis thaliana
7	Oryza sativa
8	Schizosaccharomyces pombe
9	Saccharomyces cerevisiae
10	Neurospora crassa
11	Plasmodium falciparum
12	Caenorhabditis elegans
13	Anopheles gambiae
14	Drosophila melanogaster
15	Danio rerio
16	Pan troglodytes
17	Gallus gallus
18	Homo sapiens
19	Canis lupus familiaris
20	Bos taurus

Table 3.3: HomoloGene Table

This cross-organism translation is done via HomoloGene tables ftp://ftp.ncbi.nih.gov/pub/HomoloGene that map Entrez identifiers in one organism to Entrez identifiers in the other.

Consider a technology T1 from vendor V1 (Affymetrix, Illumina, Agilent,Generic(Entrez-ID must be present)) for organism O1 (Ex:Human) and another technology T2 from vendor V2 (Affymetrix, Illumina, Agilent, Generic(Entrez-ID must be present)) for organism O2 (Ex:Rat)

Translation compares the two cases:

T1V1O1=T2V2O2 via Entrez ID in the following situations:

- between same organism but different technologies

- between different organisms and different technologies.

3.4.1 Implementation

Translation is performed using Entrez Gene ID. The identifiers of the entity list to be translated are used to get the corressponding Entrez gene IDs say for technology T1. Using Homologene data, Entrez gene IDs are then retrieved for technology T2. These are then mapped to the identifiers of T2.

How is translation done?:

There are two ways to perform translation. The first method involves the following steps:

- Consider Entity list En1 from an experiment in T1 to be translated to T2.
- Keeping E2 as the active experiment, click on En1 in E1.
- This will restrict the view in E2 to the entity list selected in E1.
- Using this view (Spreadsheet, Box whisker, Profile plot), go to toolbar icon (create entity list)and create the entity list En2.

Alternative method to do translation involves Right clicking on En1 on E1 and selecting the option ***Translate list***.

- Step 1 of the Translation Inspector wizard appears. This is the **Input parameters** page and you can import values associated with your entity list such as p-value, fold change etc along with either the raw or normalized signal values. Also, Interpretation can be chosen here from the drop down. By default, 'All samples' is chosen. Click ***Next*** to proceed. See Figure 3.15.

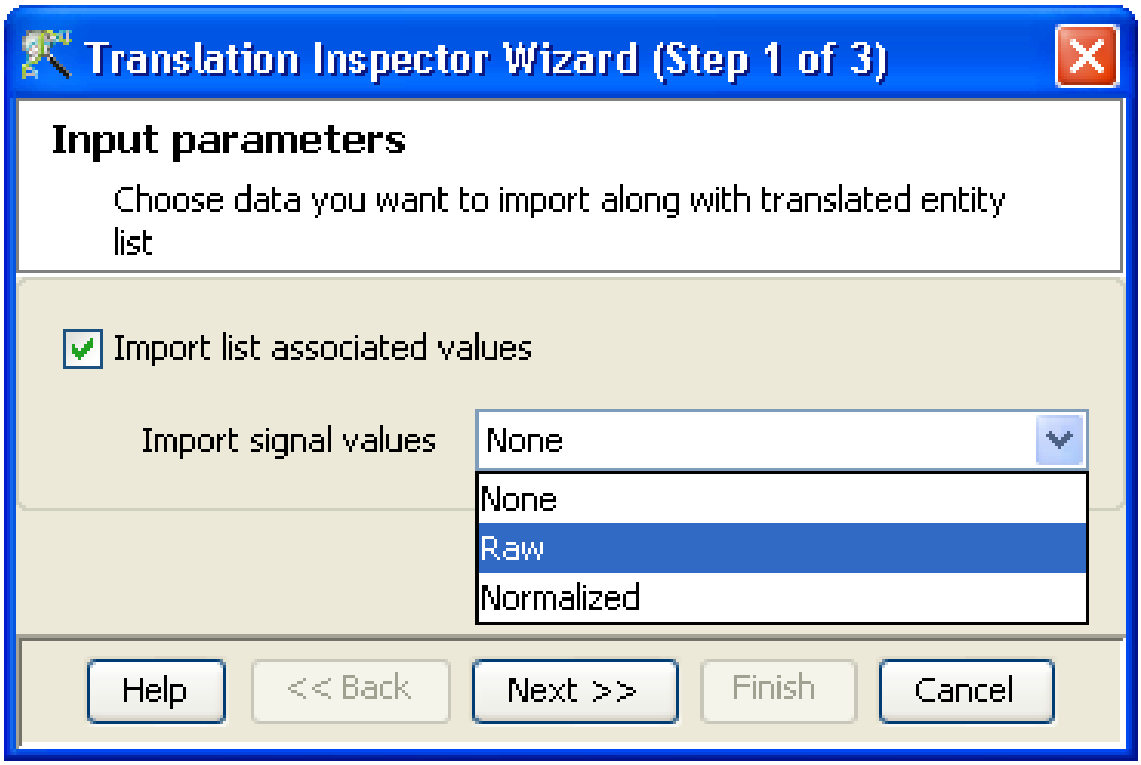


Figure 3.15: Input Parameters

- Step 2 shows the **Translation Table** page which has two tabs, the **Translated List** and **Translation mapping**. The Translation mapping table shows the mapping of the original entity list to the destination technology along with the annotations. The Translated list shows a list of probesets(destination technology identifiers) which represent the number of entities that have been translated along with the list associated values of the original entity list. The following rules are applied to the data associated with the entity lists while performing Translation:

  – The first scenario is when multiple entities of the entity list correspond to one entity of the active dataset.

  For example, when Translation is performed from Affymetrix HG_U95Av2 to Agilent Two-dye technology-12097, then values in the 'Translated List' would then correspond to the average of the 2 entities eg., the Agilent Probeset ID: A_23_P209059 corresponds to the Affymetrix probeset IDs: 38521_at and 38522_s_at. In the Translated List, values of the Affymetrix samples corresponding to the Probeset A_23_P209059 would be an average of 38521_at and 38522_s_at.

  – The other scenario is when one entry of the entity list corresponds to multiple entries in the active data set.

  If the above example is reversed, then the Probeset IDs A_23_P93015 and A_23_P85053 correspond to the Affymetrix 38523_f_at probeset. In this case, in the Translated List, the Probeset IDs A_23_P93015 and A_23_P85053 would report the same values, as that of 38523_f_at.

Annotations can be configured using *Configure Columns* button. See Figure 3.16.

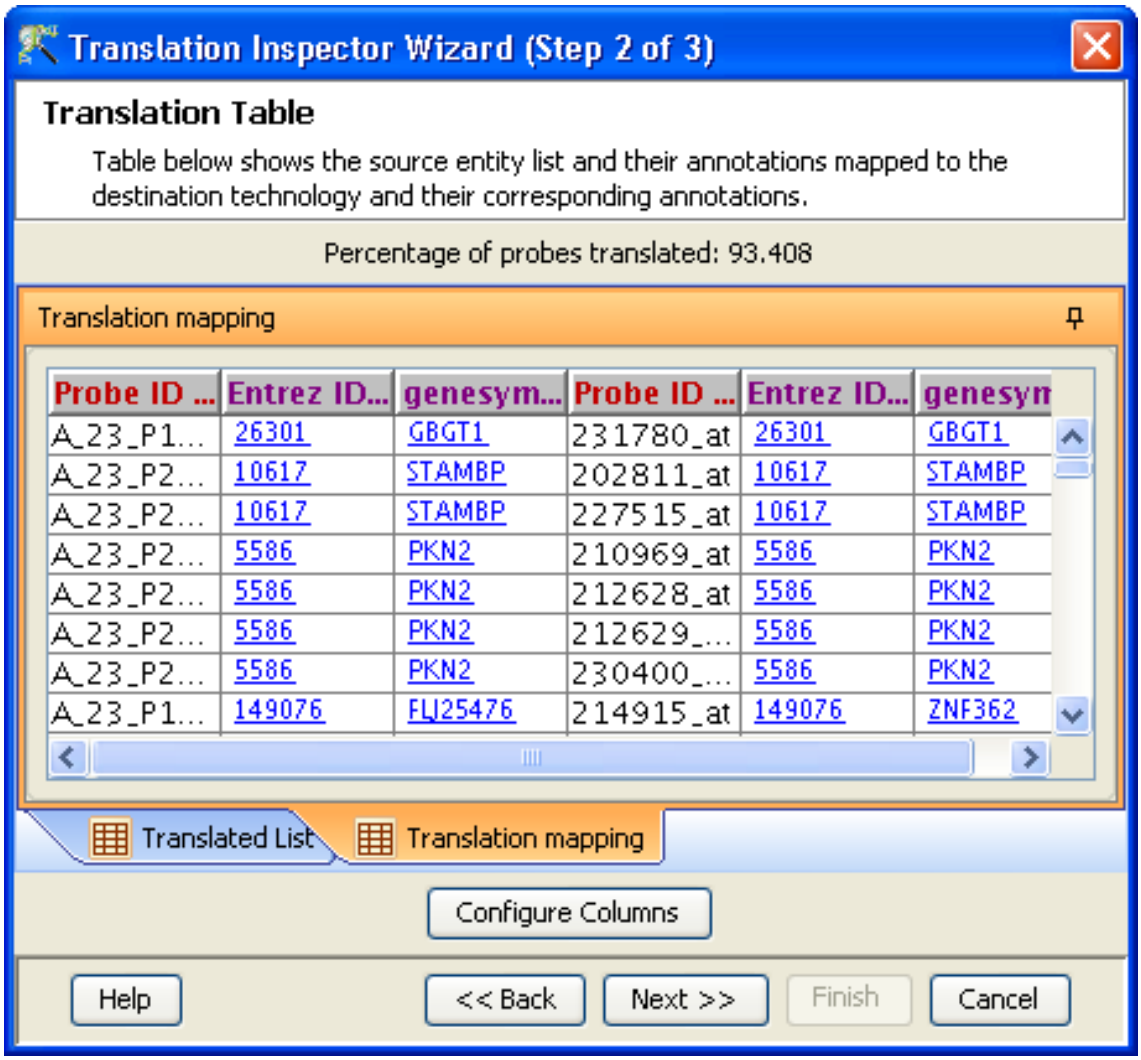


Figure 3.16: Translation Table

- Step 3 shows the **Save Entity List** window. This displays the details of the entity list created as a result of translation such as Creation date, modification date, owner, number of entities, notes etc. Click *Finish* and an entity list will be created and will be displayed in the experiment navigator of the destination experiment i.e., E2. Annotations can be configured using *Configure Columns* button. See Figure 3.17

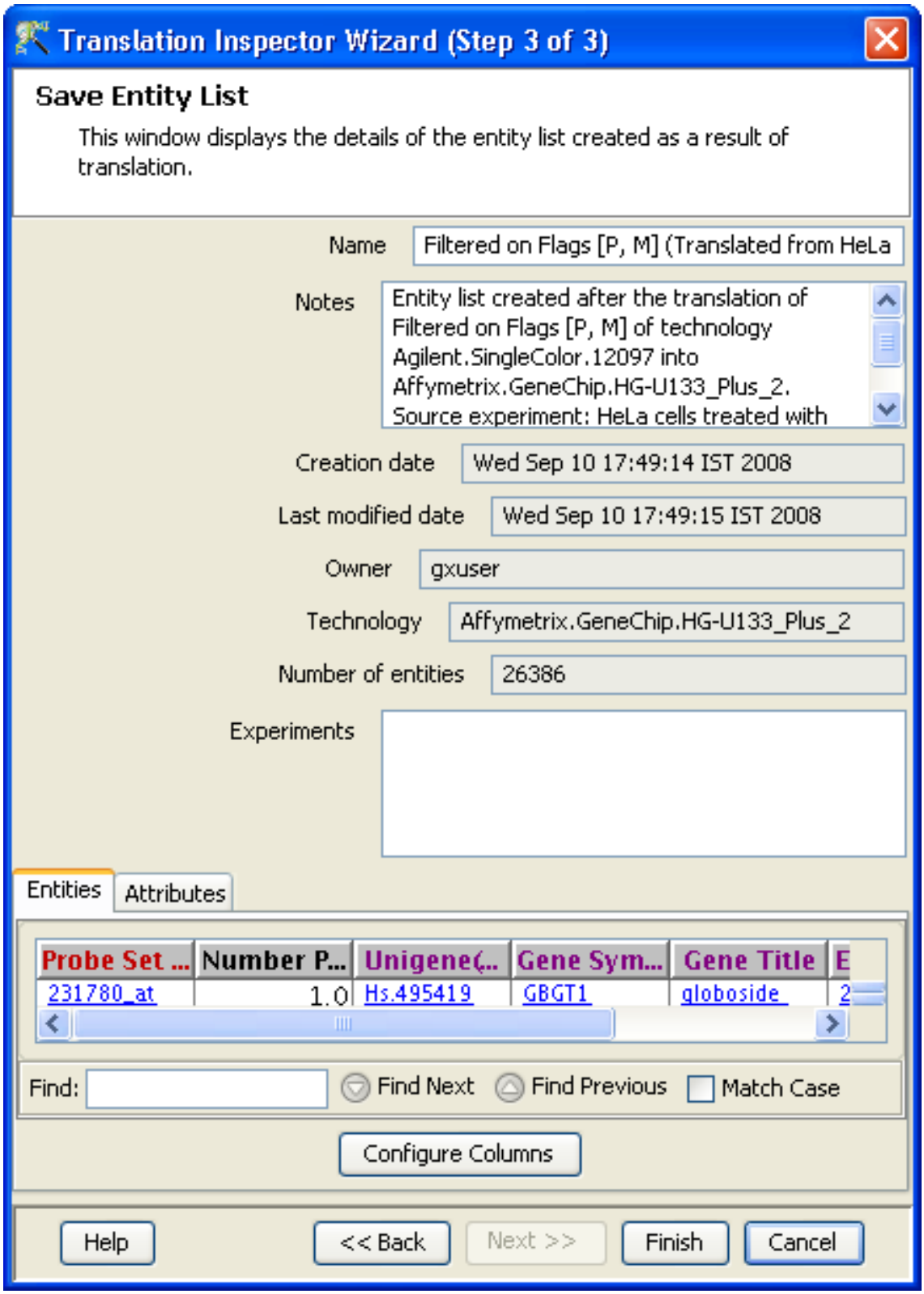


Figure 3.17: Save Entity List

Now any further analysis can be done and compared between En1 and En2.

3.4.2 Explicit Translation mapping

GeneSpring provides a way to explicitly define an annotation column for the source technology and an annotation column for the destination technology for translation, through the menu *Tool* → *Options* → *Miscellaneous* → *Translation Mapping*. Note that this explicit mapping will override the default EntrezID mapping. This feature is useful in translating data between a custom technology and a standard technology.

Go to *Tool* → *Options* → *Miscellaneous* → *Translation Mapping*. The window will allow the user to define the source and destination technologies, along with the name of those columns. There is a provision to add or remove technologies. An Error messages is shown if the source and destination technology are the same. If a mapping is already defined, duplicate mapping will not be allowed.

A typical use case is that of handling Affy text files during migration from GX 7.0 to GX 11.0. Migration tool cannot understand the text files as those of Affymetrix technology and an explicit mapping achieves the translation effortlessly.

Note: Explict translation mapping will override the default EntrezID mapping. Duplicate mapping will not be allowed.

3.4.3 Translation in Copy number and Association experiments

Translation in copy number and association experiments are slightly restricted in terms of the columns used as ID; see section Entity Lists and Translation rules in copy number for complete details. Note that explicit translation mapping does not work with copy number and association experiments.

3.5 Biological Genome

A Biological Genome refers to the collective set of all major annotations (Entrez-ID, GO IDs etc.) for any particular organism. It is created using the information available at NCBI and can be stored in **GeneSpring** . It is independent of any chip technology and once created can be used across multiple chip types and technologies. Biological Genome creation uses the following files from the NCBI site: All_Data.gene_info, gene2accession, gene2go, gene2refseq and gene2unigene. The NCBI site used for Biological Genome creation can be accessed from ***Tools***⟶ ***Options***⟶ ***Miscellaneous***⟶ ***NCBI ftp URL***. Since the Standard Technologies available from the update server usually contain all the annotations, Biological Genome is useful mainly in cases of custom technologies.

Biological Genome is essential in performing biological analyses in Generic experiments lacking

annotations. For eg., if a particular experiment does not have GO annotation columns, then the same can be obtained from Biological Genome and GO analysis can be performed. The Biological Genome can be created from ***Annotations*⟶ *Create Biological Genome*** using the following steps:

- On selecting ***Annotations*⟶ *Create Biological Genome***, a window appears with a list of organisms for which biological genomes can be created. This allows the user to select the species of interest.
- The user is also presented with an option to download the genomic data either from the NCBI ftp site or from a local folder. See figure 3.18. If the option to download from the NCBI site is chosen, then a confirmation window appears.

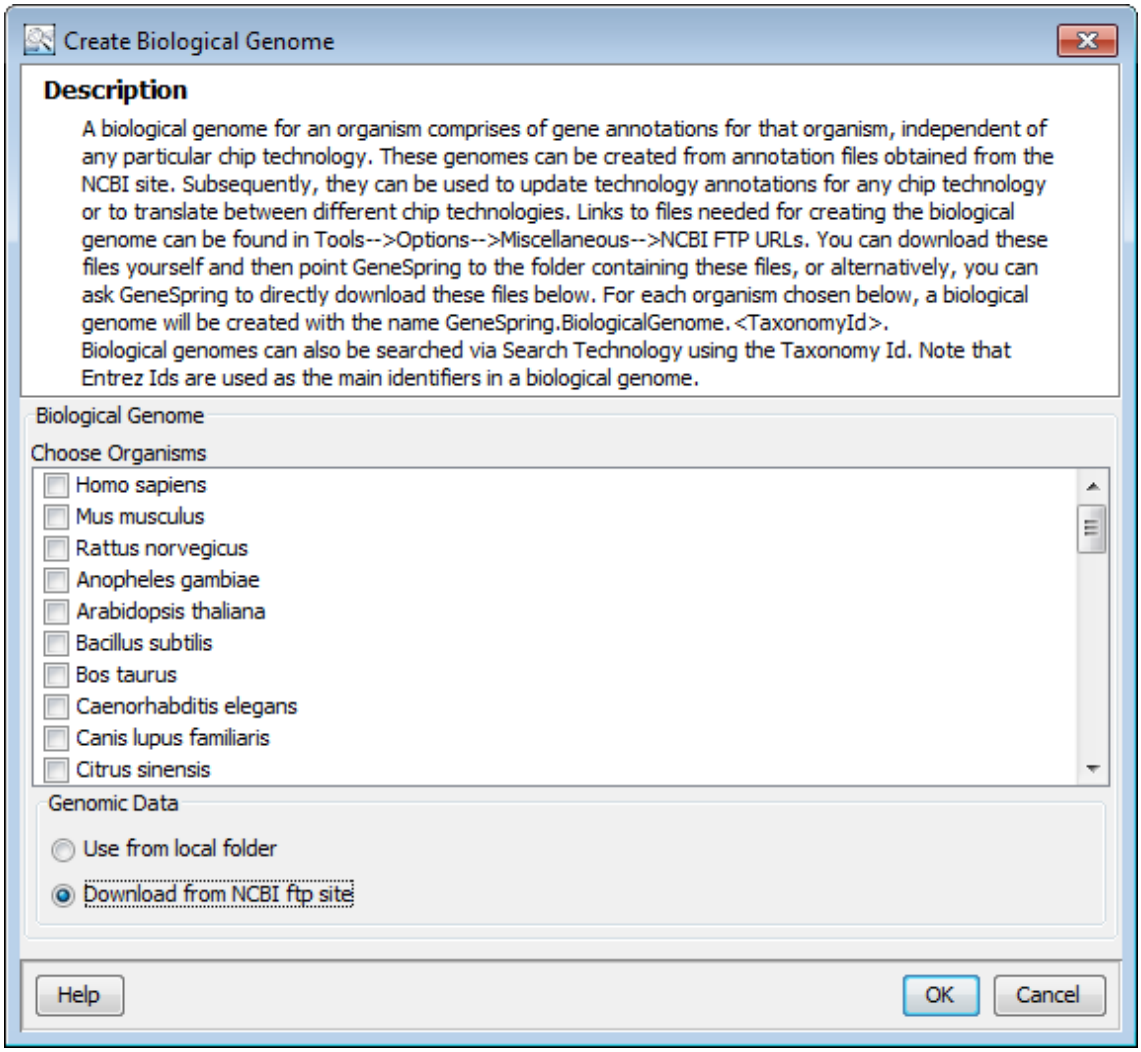


Figure 3.18: Create Biological Genome

- On choosing to go ahead, the user has to specify the folder in the system into which the files can be downloaded. This is a one time process as once the folder is created; subsequent creation of genomes for other organisms can be done from this folder by choosing the ***Use***

from local folder option. Alternatively the user can choose to download the files from the NCBI site directly into a local folder and utilize the option ***Use from local folder*** for the genomic data.

For using the Biological Genome created for an organism in an experiment, the user has to update the annotations for that particular technology from Tools-Update Technology Annotations-Update from Biological Genome. For more details on updating annotations, refer to Update Technology Annotations.

Chapter 4

Data Migration

This following sections explain about all the various data migration processes in **GeneSpring**.

4.1 GeneSpring Data Migration from GeneSpring GX 7

Migration in **GeneSpring** happens genome by genome. Migration of a genome involves migrating the corresponding samples, experiments, trees and genelists. The hierarchy of genelists and trees is also migrated. From here on, the phrase *migration of a genome* implies migration of all the above objects.

NOTE It is highly recommended that you upgrade to **GeneSpring GX 11.5.1** or later versions before migrating your **GX 7** data.

Migration of data from **GX 7** involves the following steps:

4.1.1 Migration Steps

Step 1 This step is needed only if **GX 7** and **GeneSpring** are installed on separate machines. In this case, copy the **Data** folder from **GX 7** to any location on (or accessible from) the machine where **GeneSpring** is installed. The Data folder for **GX 7** is located inside its installation folder.

Step 2 Launch **GeneSpring** and run *Tools⟶ Prepare for **GX 7** Migration*. Then provide the location of the **Data** folder described in Step 1 and click on the ***Start*** button. See Figure 4.1. This launches a procedure with the following properties:

- This procedure prepares the Data folder for migration to **GeneSpring**. Note that this procedure does not itself perform migration.
- This is a one-time procedure. Once completed, you can migrate genomes from **GX 7** to **GeneSpring** using the steps described further below. If any new experiment is added after the entire migration of the genome is finished, you have to run step 2 again. However only new experiments added will be migrated. Any changes to the existing experiments will not be reflected in the already migrated experiments.
- This procedure could be time consuming; a typical run comprising 28 experiments takes about 20 minutes. You can reduce the time needed by running Step 2 only on specific genomes of interest. To do this, create a new folder called XYZ (anywhere), then simply copy the relevant genome subfolder from the **Data** folder to within XYZ. Finally, in the dialog for Step 2, provide XYZ instead of the **Data** folder.
- This procedure will result in an increase in the size of the **GX 7** data directory. (A file *Exported_Experiments.zip* is created for every folder under the "Experiments" folder of the **GX 7** data directory. These files contain the normalized intensity values for all experiments in the **GX 7** genome and will be used by **GeneSpring** in the steps to create an experiment. Further, a *cache* directory is created inside the **Data** folder. However, this "cache" directory will be deleted at the end of the migration.)
- This procedure could give errors for two known reasons. The first situation is when it runs out of space in the system's temporary folders (on Windows systems this would typically be on the C: drive). If this happens then clear space and start Step 2 again. The second situation is when the **GX 7** cache file encounters an internal error; this could result in Step 2 hanging. In this situation, delete the cache files inside the **Data** folder and restart Step 2.

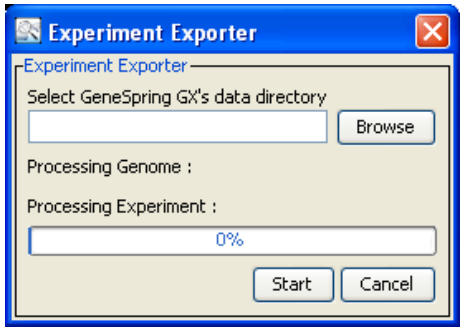


Figure 4.1: Experiment Exporter

Step 3 This step and subsequent steps focus on a particular genome of interest. To migrate this genome from **GX 7** to **GeneSpring**, run *Tools*⟶ *Migrate from* ***GX 7***. This will require confirmation whether Step 2 has been run on the genome (See Figure 4.2). Note that genomes on which Step 2 hasn't been run, will not be migrated. If Step 2 has been run on the genome of interest, click ***Ok*** and proceed further.

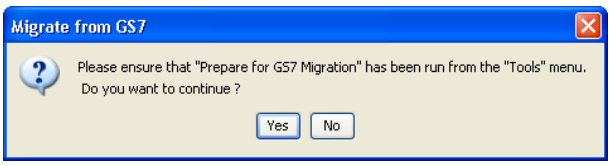


Figure 4.2: Confirmation Window

Step 4 The **GX 7 Data** folder needs to be provided at this step. Then **GeneSpring** will automatically detect all **GX 7** genomes within this **Data** folder. This will launch a window which shows the genome(s) selected for migration. By default all the genomes, which haven't been migrated before will be selected. Select only the genome(s) to be migrated and click ***OK***. See Figure 4.3. If the genome(s) was partially migrated before, it will launch another window showing the partially migrated genome(s). See Figure 4.4. Select the genome if a fresh migration has to be done or just click ***OK*** to resume migration of the genome from the point where it was left off. This step will eventually launch a progress bar showing the status of the migration. Migration can be a time consuming process depending on the amount of data to be migrated.

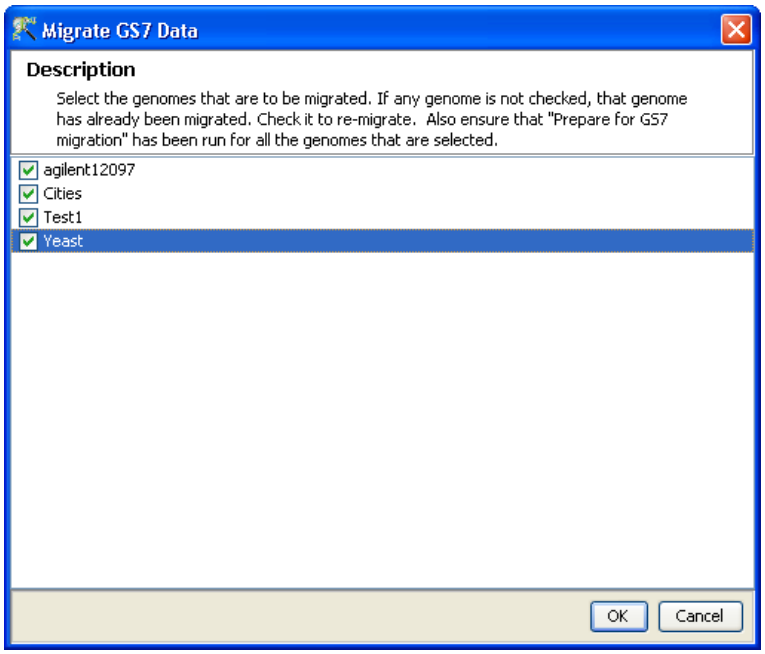


Figure 4.3: Migrate GS7 Data

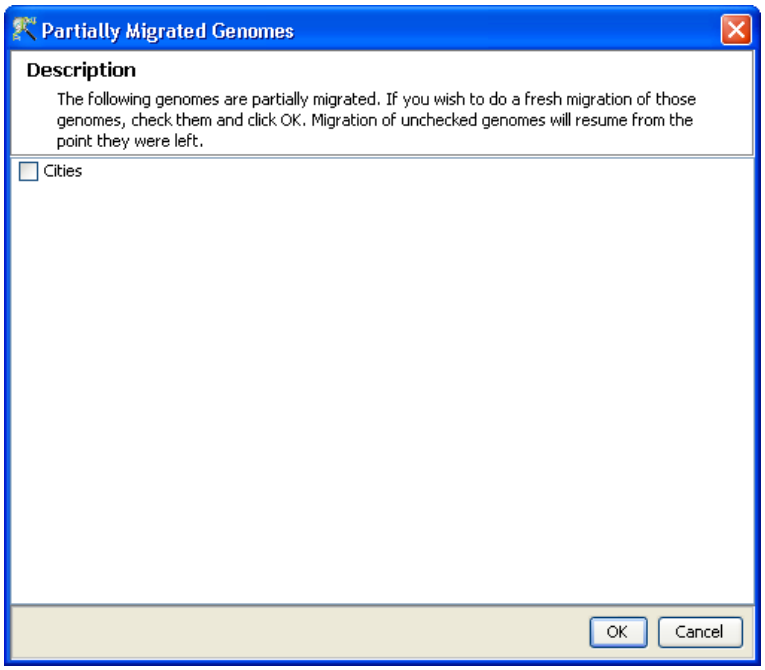


Figure 4.4: Partially Migrated Genomes

The General rate of migration is listed in the Table 4.1:

Task	Speed
Prepare for GS7 Migration	10 sample exp/1 min
Sample Migration	5MB/sec
Experiment Migration	10 sample exp/1 min
Other Objects	4 Objects/sec

Table 4.1: Migration Rate

The Migration timings for a HG_U133_Plus2 genome are listed in the tables below:
Machine: Windows XP Genuine Intel P4, 2.9Ghz, 1GB RAM, Xmx set to 1024m

Task	Speed
Prepare for GS7 Migration	10 sample exp/1 min
Sample Migration	5MB/sec
Experiment Migration	500 sample exp/ 90 min

Table 4.2: Migration Rate on Windows OS

Machine: Debian OS, Intel Xeon CPU X3220 2.4Ghz Quad Core 32 bit, 2 GB RAM, Xmx set to 1500m

Task	Speed
Prepare for GS7 Migration	10 sample exp/1 min
Sample Migration	5MB/sec
Experiment Migration	700 sample exp/ 65 min

Table 4.3: Migration Rate on Debian OS

To migrate experiments with around 1500 samples, the user needs to use a high end machine (64 bit, 8GB RAM)

The data that is brought in from **GX 7** will undergo the following transformations:

- **GeneSpring** works with data on the base 2 logarithmic scale while normalized values coming from **GX 7** are in linear scale; these are therefore converted to the log scale in **GeneSpring**.
- Prior to log transformation, **GeneSpring** will threshold the data so all values below 0.01 are thresholded to 0.01; this is consistent with **GX 7** as well.

4.1.2 Migrated Objects - Genomes, Projects, Experiments

Objects in **GX 7** were organized by genome, i.e., each object belonged to exactly one genome. In later versions of **GX 7**, an extra project tag was introduced; objects tagged with a particular project tag could be viewed as one collection. In contrast, the organization in **GeneSpring** is purely project based and not genome based. Hence the need for mapping from a genome based organization to a project based organization, which is done as follows.

For each genome in **GX 7**, a special project called the *Genome Project* is created in **GeneSpring**. This Genome Project contains all experiments associated with this genome. In addition, to reflect project tags on objects in **GX 7**, special *Project Projects* are created in **GeneSpring**; a Project Project contains only those experiments which have the corresponding project tag in **GX 7**.

Each of these experiments in turn contains other objects (Entity Lists, Gene Trees, Condition Trees and Classifications) associated with this genome. There are two cases here. If an experiment has a project tag then it contains only those objects which have the same project tag. And if an experiment has no project tag then it contains all objects in the genome which do not have any project tags. These objects appear classified into two groups, those which have no association with projects and those which do; the latter appear in appropriate folder structures which describe the project association. Within the above framework, the folder hierarchy for each object is preserved as in **GeneSpring**.

The data in an experiment comprises normalized values, raw values and flags for each entity (gene) and each associated sample, and experimental grouping information. These are migrated directly from **GX 7**, i.e., they are copied from **GX 7** and not recalculated in **GeneSpring**. Since algorithms and processing steps in **GeneSpring GX 7.3** and **GeneSpring GX 11.5** or a subsequent version are different, further operations on this data could give slightly different results in **GeneSpring GX 7.3** and **GeneSpring GX 11.5** or a subsequent version. For instance, if

this experiment has samples that are CEL files, then using these samples to re-summarize and create a new experiment could give slightly different results.

4.1.3 Migrated Objects - Experiments

When a **GX 7** experiment is migrated to **GeneSpring**, the following changes happen to objects contained therein.

All experiments other than Affymetrix and Agilent experiments with standard technologies will be migrated as what are called ‘‘custom” experiments. Each custom experiment will have Raw, Normalized and Control values exactly as derived from **GX 7**, with just the following change: normalized values will be displayed on the log scale while Raw and Control values will be displayed on the linear scale, in entity inspectors.

For Affymetrix experiments with standard technologies, Raw and Normalized values will be migrated from **GX 7**, with raw values kept in the linear scale and normalized values reported on the log scale. For Agilent single color experiments with standard technolgies, normalized values will be migrated from **GX 7** and raw values will be reread from the associated sample files (which may take some time). For Agilent two color experiments with standard technolgies, normalized values will be migrated from **GX 7** and raw Cy3 and Cy5 values will be reread from the associated sample files (which may take some time). In both cases, raw values are reported on the linear scale while normalized values are on the log scale.

Experimental Parameters and Interpretations: All experimental parameters, whether the parameter is numeric or not, parameter values for each such parameter, and the order of these values for each such parameter are migrated. The following interpretations are migrated:

- Interpretations having parameters with display modes as ‘‘Continuous’’ or ‘‘Non-continuous’’ are migrated and are preserved as ‘‘Numerical” and ‘‘Categorical” respectively.
- Interpretations having parameters with display mode ‘‘Color Code’’ are migrated without those parameters and with all conditions. These interpretations will be mentioned in the Summary Report that is launched after migration.

4.1.4 Migrated Objects - Samples

To describe this in more detail, we need to understand the various constituents of a sample in **GX 7**.

- Input Files: A typical sample was imported into **GX 7** starting with an input file, e.g., a CEL file or a .txt file. Typically, one input file contains one sample. However, there are

exceptions; one input file could contain multiple samples, as in the case of an Illumina input file. And there is the rare case of ImaGene generated two color raw files, where two input files together constitute one sample.

- Processed Sample: A processed sample is what is created from the above input files in **GX 7**. This processed sample contains the relevant segment of data from the input file with further transformations.
- Other Attachments: An attachment is an auxiliary file associated with a sample in **GX 7**. A sample could have one or more attachments, for instance, DAT files, ARR files etc.
- Sample Attributes: Attributes are other key value pairs associated with a sample and used typically for search.

Once migrated into **GeneSpring**, a sample from **GX 7** has the following possible outcomes.

Migration to Standard Samples: A standard sample is one for which **GeneSpring** understands the file format off-the-shelf and new experiments can be created with such samples directly via the Create New Experiment wizards. In addition, input files for these samples can be downloaded from the experiment navigator by right-clicking on the sample. Most Affymetrix samples (except those based on custom CDFs) and Agilent samples obtained from FE versions 8.5.x and 9.5.x will be converted to standard samples.

Migration to Raw Samples: A raw sample is one for which **GeneSpring GX 11.5** or a subsequent version does not understand the file format off-the-shelf. Except Affymetrix samples (except those based on custom CDFs) and Agilent samples obtained from FE versions 8.5.x and 9.5.x, all others will be converted to raw samples.

Migration Failure: This will only happen if the sample has no associated genome in **GX 7**, or an associated genome that is faulty for some reason.

Migration Scheme:

For migration into **GeneSpring** or a subsequent version, each processed sample in **GX 7** is considered in turn. For a particular processed sample, all its attachments in **GX 7** are scanned to see if any of these represents an input file which **GeneSpring GX 11.5** or a subsequent version can convert to a standard sample. If so, then that input file is migrated into **GeneSpring** as a standard sample. All other attachments with the processed sample in **GX 7** are added as attachments to this standard sample in **GeneSpring**, and all attributes of the processed sample are made attributes of this standard sample. Otherwise, if none of its attachments in **GX 7** represent an input file which **GeneSpring** can convert to a standard sample, the processed sample is itself migrated into **GX 7** as a raw sample (the associated technology name would be GS7.Custom.xxx), and the associated input files are migrated as raw files with the same corresponding sample names. All attachments/attributes with the processed sample are added as attachments/attributes to these corresponding raw files.

4.1.5 Entity Lists, Gene Trees, Condition Trees and Classifications

The organization of objects (Entity Lists, Gene Trees, Condition Trees and Classifications) within an experiment is different in **GeneSpring GX 11.5** or a subsequent version when compared to **GeneSpring GX 7.3**. All these objects appear within the Analysis subfolder in **GeneSpring GX 11.5** or a subsequent version but retaining the same hierarchy as in **GeneSpring GX 7.3**. Rules for whether or not an object appears within a particular experiment are as in the paragraphs above.

4.1.6 Migrated Objects - Other objects and miscellaneous rules

Other Objects: Other objects like bookmarks, pathways, etc. are not migrated

Translation mapping rules: A genome can contain a mixture of sample types; Affymetrix or Agilent samples that can be directly loaded in **GeneSpring** and those samples for which a custom technology needs to be created before they can be loaded into **GeneSpring** (Such technologies will be of the form : GS7.Custom.XXXX ; where XXXX is the Genome Name). In this scenario, the user need not explicitly define a translation mapping between the two technologies. (*Tools*$\longrightarrow$ *Options*$\longrightarrow$ *Miscellaneous*$\longrightarrow$ *Translation Mapping*). It will be automatically created during migration of such genomes.

Creation and Modification Dates: When **GX 7** objects are migrated to **GeneSpring**, the creation and modification dates of the new objects correspond to the date of migration rather than the date of creation/modification of the original **GX 7** objects. The latter dates are added as user attributes (namely, **GX 7** Creation Date and **GX 7** Last Modification Date) and are available for search.

4.1.7 Potential causes of Migration failure and Known Issues

Some causes of migration failure and some known issues are listed below.

- Running out of RAM is one key issue; 8GB of RAM will ensure that experiments with up to 1500 HG_U133_Plus2 samples can be migrated.
- During the "Preapre for Migration" step, a zip file "Exported_Experiments.zip" is created for every **GX 7** Experiment folder. If all the **GX 7** experiments are in a single folder, then this file can grow to be greater than 4 GB. If it does so, migration will fail. A work-around is to re-organize the **GX 7** Experiment folder into a few folders, thereby reducing the size of the created zip file.
- Unparseable characters in certain **GeneSpring GX 7.3** XML files, though rare, will cause the corresponding experiment to fail from being migrated.

- Condition trees that are malformed, possibly due to subsequent modification of conditions and/or deletion of the underlying experiment parameters, may not be migrated.
- Occasionally, there might be experiments for which the corresponding genome is empty (possibly on account of a deletion event); such experiments will not be migrated.
- If the **GX 7** genome has non-Affymetrix and non-Agilent samples, then all the genome annotations are migrated into the GS7.Custom.xxx technology (where xxx is the name of the genome). The default technology inspector view for that technology only displays a few technology annotation columns by default and thus, may not display important annotations like Gene Symbol or Entrez Id. As with any technology in **GeneSpring**, users can easily configure which annotations are to be shown by default by clicking on the *Configure Columns* button in the technology inspector.

Chapter 5

Data Visualization

5.1 View

Multiple graphical visualizations of data and analysis results are core features of **GeneSpring** that help discover patterns in the data. All views are interactive and can be queried, linked together, configured, and printed or exported into various formats. The data views provided in **GeneSpring** are the Spreadsheet, the Scatter Plot, the Profile Plot, the Heat Map, the Histogram, the Matrix Plot, the Summary Statistics, Bar Chart, BoxWhisker Plot, Violin Plot, MvA, Genome Browser, Plot List Associated Values and the Venn Diagram.

5.1.1 The View Framework in GeneSpring

In **GeneSpring** rich visualizations are used to present the results of algorithms. The user can interact with these views, change parameters and re-run the algorithm to get better results. The views also help in examining and inspecting the results and once the user is satisfied, these entity lists, condition trees, classification models, etc can be saved. The user can know the identity of a probe depicted by particular point on the view by pointing the mouse over it. You can also interact with the views and create custom lists from the results of algorithms. Details of the views associated with the Analysis: Biological Significance and the Data Import Wizard links will be detailed in the following sections.

In addition to presenting the results of algorithms as interactive views, views can also be launched on any entity list and interpretation available in the analysis from the view menu on the menu bar or from the tool bar. Also multiple views can be launched and each will show up as a separate tab in the central visualization panel. The Spreadsheet, the Scatter Plot, the Profile Plot, the Heat Map, the Histogram, the Matrix Plot, the Summary Statistics, BoxWhisker Plot, Violin Plot, MvA, Genome Browser, Plot List Associated Values and the Venn Diagram view can be launched from the View menu on the menu bar. The views will be launched with the current active entity list and interpretation in the experiment.

Note: The key driving force for all views derived from the view menu are the current active interpretation and the current active entity list in the experiment. The conditions in the interpretation provide the columns or the axes for the views and the current active entity list determines the entities that are displayed as rows or points in the view. While making another entity list in the same experiment, the active entity list will dynamically display those entities in the current view. Clicking on an entity list in another experiment will translate the entities in that experiment to the entities in the current experiment (based upon the technology and the homologies) and dynamically display those entities.

5.1.2 View Operations

All data views and algorithm results share a common menu and a common set of operations. There are two types of views, the plot derived views, like the Scatter Plot, the Profile Plot, the Histogram and the Matrix Plot; and the table derived views like the spreadsheet, the heat map view, and various algorithm result views. Plot views share a common set of menus and operations and table views share a common set of operations and commands.

In addition, some views like the heat map are provided with a tool bar with icons that are specific to that particular data view. The following section below gives details of the common view menus and their operations. The operations specific to each data view are explained in the following sections.

5.1.2.1 Common Operations on Plot Views

See Figure 5.1.

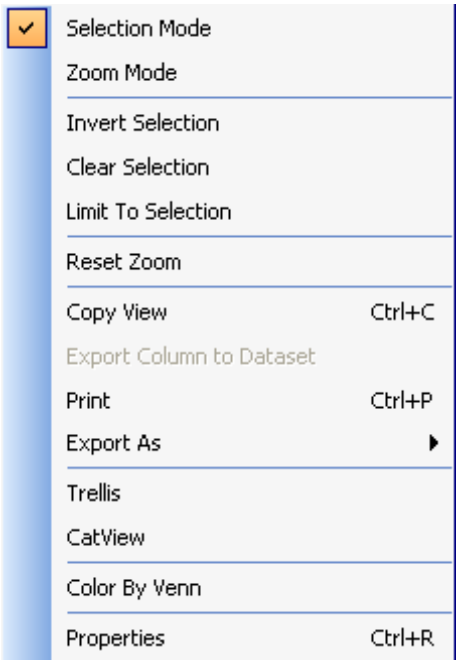


Figure 5.1: Menu accessible by Right-Click on the plot views

All data views and algorithm results that output a Plot share a common menu and a common set of operations. These operations are from Right-Click in the active canvas of the views. Views like the scatter plot, the 3D scatter plot, the profile plot, the histogram, the matrix plot, etc., share a common menu and common set of operations that are detailed below.

Selection Mode: All plots are by default launched in the Selection Mode. The selection toggles with the Zoom Mode where applicable. In the selection mode, left-clicking and dragging the mouse over the view draws a selection box and selects the elements in the box. Control + left-clicking and dragging the mouse over the view draws a selection box, toggles the elements in the box and adds to the selection. Thus if some elements in the selection box were selected, these would become selected and if some elements in the selection box were unselected, they would be added to the already present selection.

Selection in all the views are lassoed. Thus selection on any view will be propagated to all other views.

Zoom Mode: Certain plots like the Scatter Plot and the Profile Plot allow you to zoom into specific portions of the plot. The zoom mode toggles with the selection mode. In the zoom mode, left-clicking and dragging the mouse over the view draws a zoom window with dotted lines and expands the box to the canvas of the plot.

Invert Selection: This will invert the current selection. If no elements are selected, Invert Selection will select all the elements in the current view.

Clear Selection: This will clear the current selection.

Limit to Selection: Left-clicking on this check box will limit the view to the current selection. Thus only the selected elements will be shown in the current view. If there are no elements selected, there will be no elements shown in the current view. Also, when Limit to Selection

is applied to the view, there will is no selection color set and the the elements will be appear in the original color in the view. The status area in the tool will show the view as limited to selection along with the number of rows / columns displayed.

Reset Zoom: This will reset the zoom and show all elements on the canvas of the plot.

Copy View: This will copy the current view to the system clipboard. This can then be pasted into any appropriate application on the system, provided the other listens to the system clipboard.

Export Column to Dataset: Certain result views can export a column to the dataset. Whenever appropriate, the Export Column to dataset menu is activated. This will cause a column to be added to the current dataset.

Print: This will print the current active view to the system browser and will launch the default browser with the view along with the dataset name, the title of the view, with the legend and description. For certain views like the heat map, where the view is larger than the image shown, Print will pop up a dialog asking if you want to print the complete image. If you choose to print the complete image, the whole image will be printed to the default browser.

Export As: This will export the current view as a Report, Image, or HTML file or the values as a text, if appropriate. See Figure 5.2.

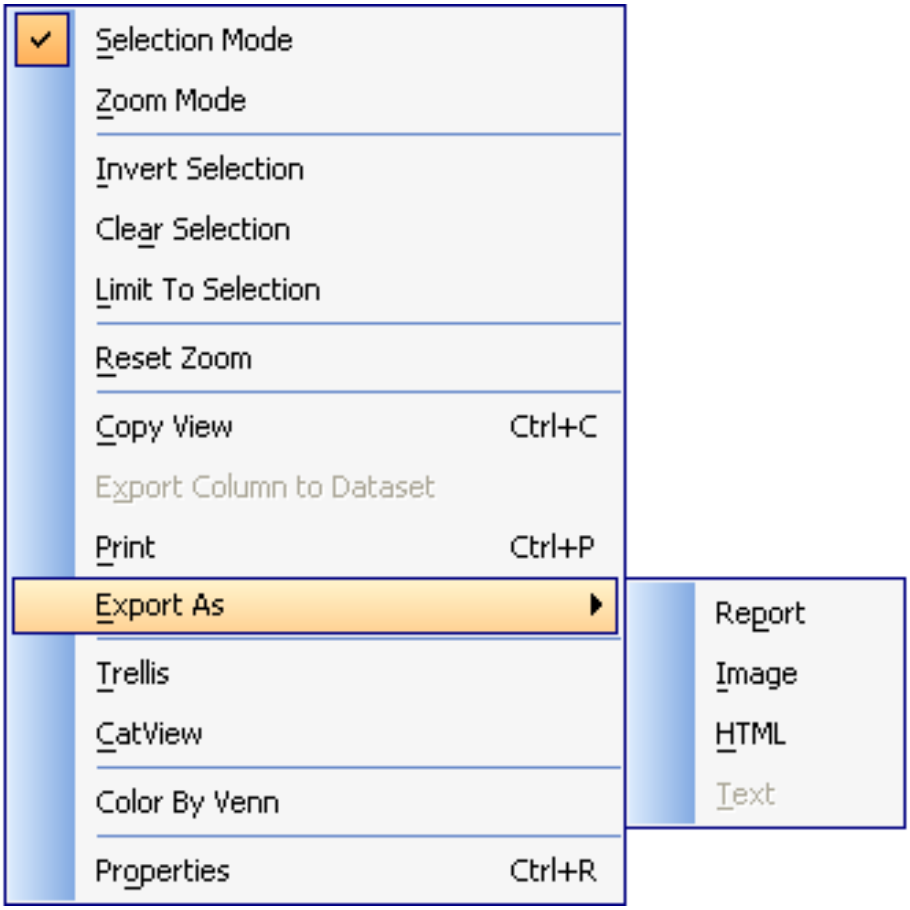


Figure 5.2: Export submenus

- Export As Report: Use this option to save the plot and its associated data as a report. See How to Create a Report.

- Export As Image: This will pop-up a dialog to export the view as an image. This functionality allows the user to export a very high quality image. You can specify any size of the image, as well as the resolution of the image by specifying the required dots per inch (dpi) for the image. Images can be exported in various formats. Currently supported formats include .tiff, .png, .jpg, .jpeg, .bmp, .svg, or .eps. Images of very large size and resolution can be printed in tiff, svg, or eps formats. Default parameters for exporting an image can be changed from **Tools** $\longrightarrow$ **Options** dialog under **Export as Image**. See Figure 5.3 and Figure 5.4.

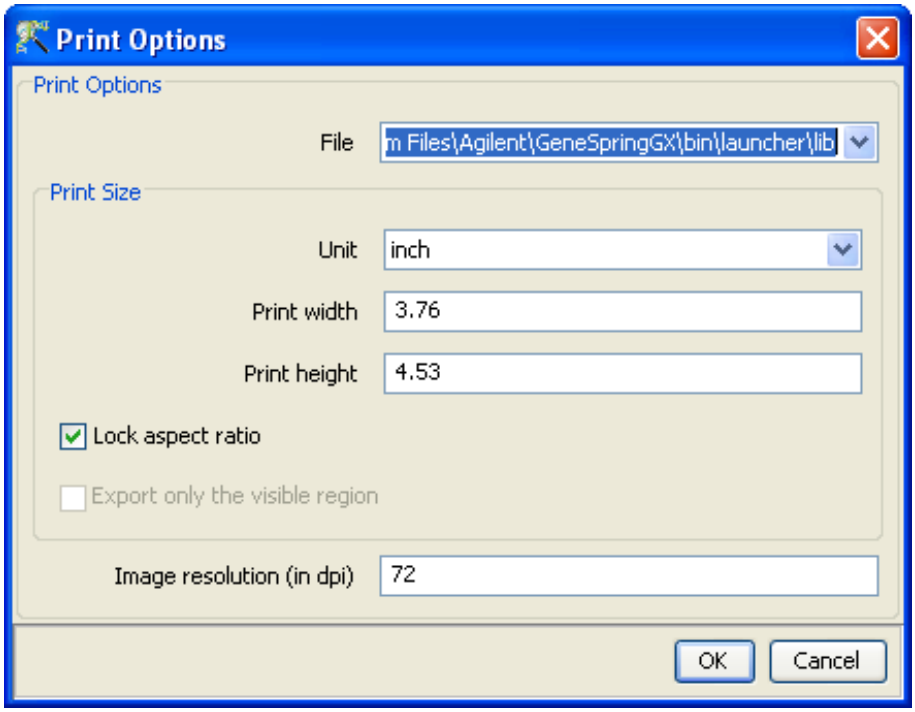


Figure 5.3: Export Image Dialog

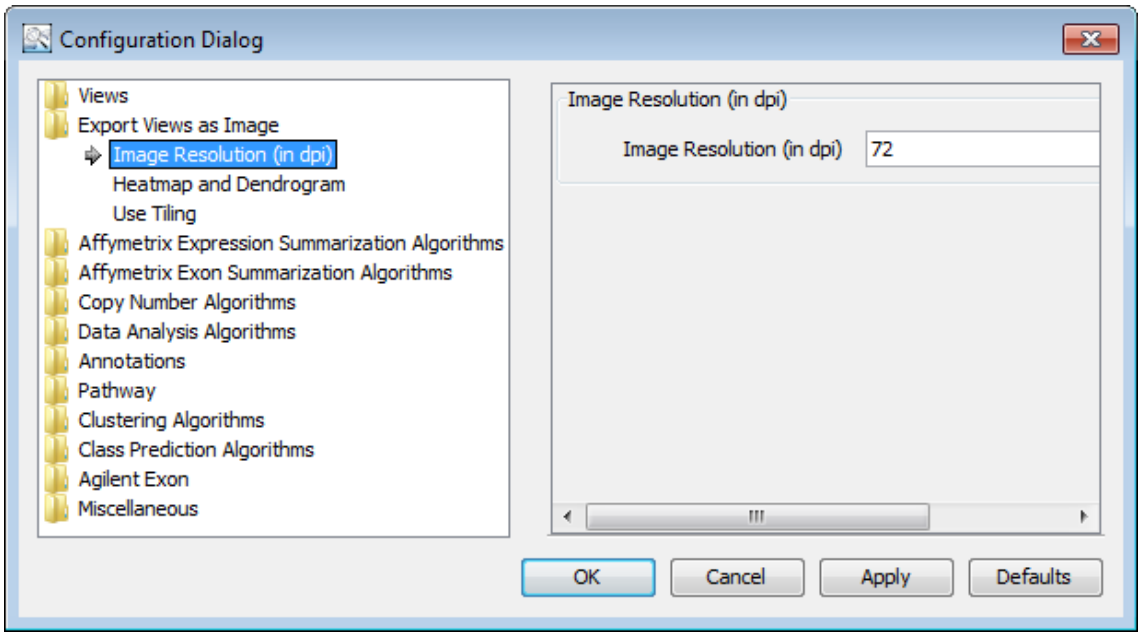


Figure 5.4: Tools - Options Dialog for Export as Image

Note: This functionality allows the user to create images of any size and with any resolution. This produces high-quality images and can be used for publications and posters. If you want to print very large images or images of very high-quality the size of the image will become very large and will require huge resources. If enough resources are not available, an error and resolution dialog will pop up, saying the image is too large to be printed and suggesting you to try the tiff option, reduce the size of image or resolution of image, or to increase the memory available to the tool by changing the XmxFraction option in `INSTALL_DIR/bin/packages/properties.txt` file. On **Mac OS X** the Java heap size parameters are set in the file `Info.plist` located in `INSTALL_DIR/GeneSpringGX.app/Contents/Info.plist`. Change the XmxFraction parameter appropriately. Note that in the Java heap size limit on Mac OS X is about 2048M. See Figure 5.5.

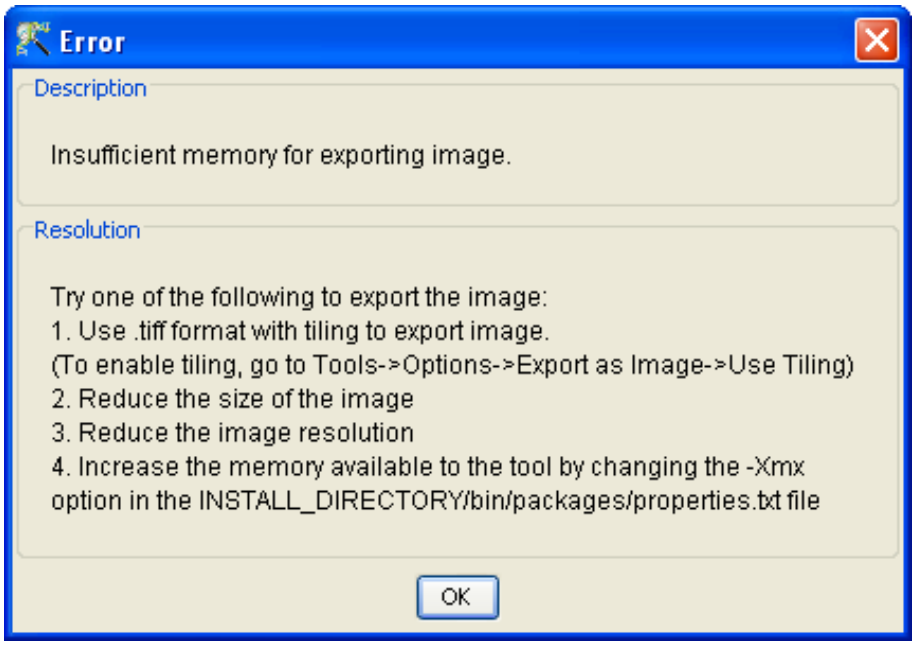


Figure 5.5: Error Dialog on Image Export

- Export As HTML: This will export the view as a html file. Specify the file name and the the view will be exported as a HTML file that can be viewed in a browser and deployed on the web.
- Export as Text: Not applicable for Plot views.

'Export As' will pop up a file chooser for the file name and export the view to the file. Images can be exported in .tiff, .png, .jpg, .jpeg, .bmp, .svg, or .eps formats. 'Export As Text' will enable export to a txt file.

Trellis: This option displays data with three or more dimensions effectively on a two-dimensional space.

CatView: This option displays the profile plot of an individual probe. Any probe can be selected from the drop-down.

Color By Venn: Use this option to visualize entities in a Profile plot, Scatter plot, or Cy3 *vs* Cy5 plot by differential colors based on the different regions identified in a previously created Venn diagram. The respective Venn Diagram has to be open for this feature to work.

Properties: This will launch the Properties dialog of the view as limited to selection along with the number of rows / columns displayed. the current active view. All Properties of the view can be configured from this dialog.

5.1.2.2 Common Operations on Table Views

See Figure 5.6.

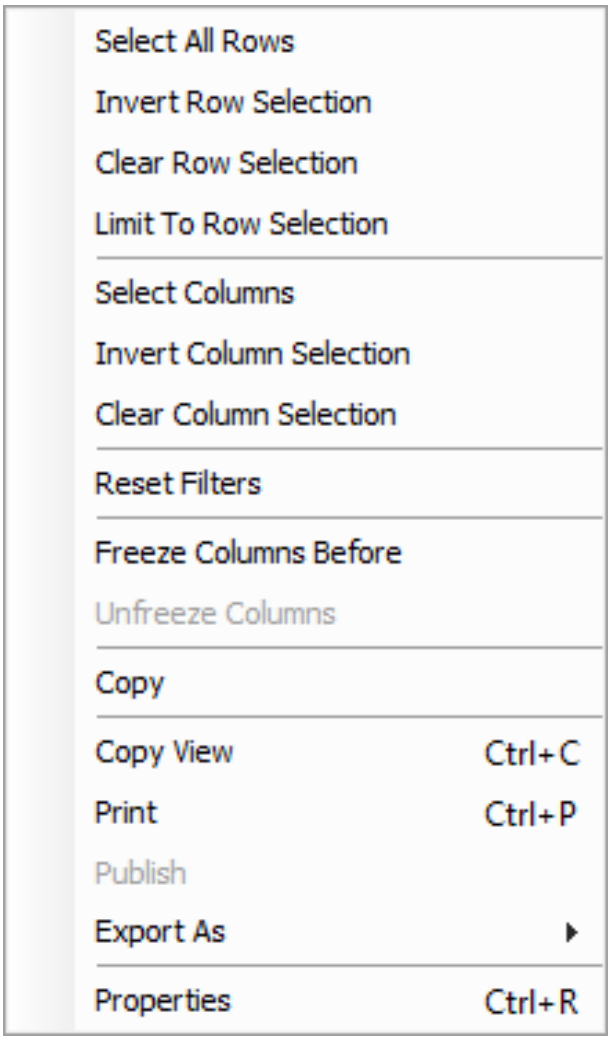


Figure 5.6: Menu accessible by Right-Click on the table views

All data views and algorithm results that output a table share a common menu and a common set of operations. Table views include the Spreadsheet View, the Heatmap, the Bar Chart, tables in inspectors, etc. These operations are accessed from Right-Click in the active canvas of the views. All these operations are accessible through the right-click menu (except for **Filter by Column Content**) and are described here:

Filter by Column Content: To filter a table by the content of one of its columns, Ctrl-Click the column header. The **Filter Dialog** dialog box opens.

If the column contains numerical values, you can determine the range of values for which you want to continue to display the corresponding rows in the table using the Minimum and Maximum sliders or type the desired value in the respective fields. See Figure 5.7.

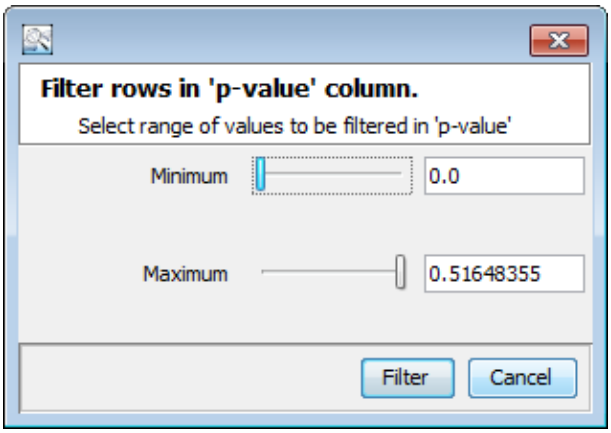


Figure 5.7: Filter Dialog for numerical values

If the column contains categorical values (for example identifiers, zygosity, or the '+' and '-' symbols in a Strand column), you can select or unselect which rows you want to display by moving items from **Available Categories** to **Selected Categories** and vice versa. See Figure 5.8.

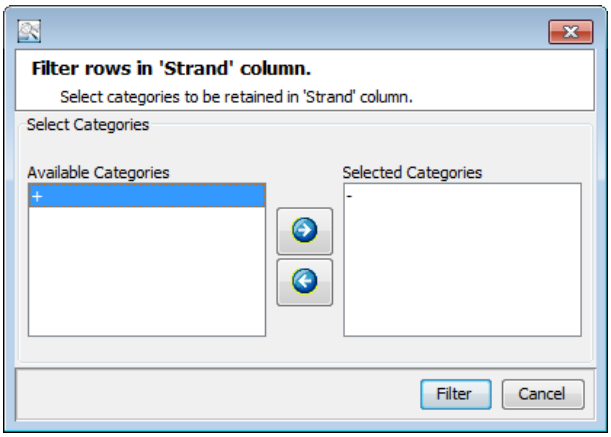


Figure 5.8: Filter Dialog for categorical values

A red border on the column header indicates an active filter.

To filter the table by a second column (equivalent to an 'AND' condition), Ctrl-Click another column and choose the filter settings in the **Filter Dialog** that opens. Both column headers are now highlighted by a red border. The second filter operation is only applied to the data displayed in the table after the first filter operation.

To reset the table to its original state (remove any filters), right-click over the table and select **Reset Filters**.

Selection: The table views are by default launched in the Selection Mode. Either columns or rows or both can be selected on the Table. Selection on all views is lassoed. Thus selection on the table will be propagated to all other views of the data (this does not include tables in inspectors). All Table views allow row and column selection.

Clicking on a cell in the table will select the column or row or both column and row of the table. If clicking on a cell selects rows, Left-Click and drag the mouse. This will select all

the rows. To select a large amount of continuous rows, Left-Click on the first row. Then scroll to the last row to be selected and Shift-Left-Click on the row. All rows between the first row and the last row will be selected and lassoed. Ctrl-Left-Click toggles the selection and adds to the current selection. Thus Ctrl-Left-Click on selected rows will unselect it, and Ctrl-Left-Click on unselected rows will add these rows to the selection.

Invert Row Selection: This will invert the current row selection. If no rows are selected, Invert Row Selection will select all the rows in the current table view.

Clear Row Selection: This will clear the current selection.

Limit to Selection: Left-Click on this check box will limit the table view to the current selection. Thus only the selected rows will be shown in the current table. If there are no selected rows, there will be no rows shown in the current table view. Also, when Limit to Selection is applied to the table view, there will is no selection color set and the the rows will be appear in the original color in the table view.

Select Column: This is a utility to select columns in any table view. Clicking on this will launch the Column Selector. To select columns in the table view, select the highlight the appropriate columns, move them to the Selected Items list box and click OK. This will select the columns in the table and lasso the columns in all the appropriate views.

Invert Column Selection: This will invert the current column selection. If no columns are selected, Invert Column Selection will select all the columns in the current table view.

Clear Column Selection: This will clear the current selection.

Copy Selected Columns: If there are any selected columns in the table, this will option will be enabled. Choosing this menu option will copy the selected column(s) on to the system clipboard. After copying to the clipboard, it will prompt an information messages saying it has Copied n column(s) to the clipboard. This can be later pasted into application on the user's desktop.

Copy View: This will copy the current view to the system clipboard. This can then be pasted into any appropriate application on the system, provided the other listens to the system clipboard.

Print: This will print the current active view to the system browser and will launch the default browser with the view along with the dataset name, the title of the view, with the legend and description. For certain views like the heat map, where the view is larger than the image shown, Print will pop up a dialog asking if you want to print the complete image. If you choose to print the complete image, the whole image will be printed to the default browser.

Export As: This option will enable to export the current view as an Image, HTML, or text. **Export As** will open up a file chooser for the file name and exports the view to a file. Images can be exported as a .tiff, .png, .jpg, .jpeg, .bmp, .png, .svg, or ,eps. 'Export As Text' will enable export to a txt file.

Properties: This will launch the Properties dialog of the current active view. All Properties of the view can be configured from this dialog.

5.2 The Spreadsheet View

A spreadsheet presents a tabular view of the data. The spreadsheet is launched from the view menu with the active interpretation and the active entity list. Alternately, Left-Click on the tool

bar icon to launch the spreadsheet. The spreadsheet will display the normalized signal values of the conditions in the current active interpretation as columns in the table. If the interpretation is averaged, it will show the normalized signal values averaged over the samples in the condition.

The rows of the table correspond to the entities in the current active interpretation. Clicking on another entity list in the analysis tree will make that entity list active and table will be dynamically updated with the corresponding entity list.

Thus if the current active interpretation in an experiment is a time averaged interpretation, where the normalized signal values for the samples are averaged for each time point, the columns in the table will correspond to these averaged normalized signal values at each time condition. The rows of the table will correspond to the active entity list. In addition, the identifier for the entity and the default set of entity annotation columns will be shown. The legend window shows the interpretation on which the scatter plot was launched.

Clicking on another entity list in the experiment will make that entity list active and the table will dynamically display the current active entity list. Clicking on an entity list in another experiment will translate the entities in that entity list to the current experiment and display those entities in the table. See Figure 5.9.

ProbeName	US22502705_251...	US22502705_251...	US22502705_2512...	US22502705_251...	US22502705_2512...	US22502705_251...
BrightCorner	0.5646615	-0.66926384	0.8790145	-0.5646615	-1.5560384	1.9043875
(-)3xSLv1	-0.04172516	-0.0068264008	0.006825447	-0.10069847	0.13784885	0.09785557
A_23_P146576	0.0812788	2.5978994	-0.21312332	4.437047	-0.08127785	-0.12073994
A_23_P125016	-0.025115013	-0.101350784	0.3015356	0.025115013	0.21356678	-0.08848953
A_23_P28555	0.05974579	0.043618202	0.017533302	-0.017533302	-0.045830727	-0.09685993
(+)E1A_r60_n9	0.0030794144	-0.025473595	0.09113979	-0.04561901	0.0198946	-0.0030794144
(+)Pro25G-02	-0.038568497	-0.2428379	0.08877087	-0.23419952	0.038568497	0.15631485
A_23_P23227	0.27430868	0.12372637	0.32570982	-0.36179018	-0.12372637	-0.5194464
A_23_P137543	0.13455391	0.14818573	0.1292963	-0.15054941	-0.2607293	-0.1292963
A_23_P501193	-0.6333108	-0.542676	-0.6671262	1.1745763	0.542676	0.61875343
A_23_P27247	0.19498634	0.43528652	-1.1547675	0.41011906	-0.6221757	-0.19498634
A_23_P323270	-1.6402297	1.3497367	1.6304193	4.0521417	-1.3497367	-1.3891988
A_23_P258433	-0.08195877	-0.07668781	0.07668877	-0.08682728	0.20853424	1.1098886
A_23_P28649	0.091856	0.059829712	0.10629368	-0.07480812	-0.059829712	-0.08853054
A_23_P96529	1.8983011	-0.21929932	-0.06592274	-0.22943878	0.06592274	5.149774
A_23_P2322	0.11109567	-0.17040038	0.21625638	0.0057222843	-0.31952405	-0.0057222843
A_23_P372910	0.45046568	0.03280258	0.25379562	-0.35731792	-0.33713055	-0.03280258
A_23_P343357	2.056336	-0.21929932	-0.06592274	-0.22943878	0.06592274	2.6238565
A_23_P501634	-0.10973024	0.028988361	0.855937	-0.028988361	-0.553936	1.0748968
A_23_P203419	-0.18279457	-0.24509144	-0.11606693	0.20826244	0.24380493	0.11606693
A_23_P330490	0.34325314	0.27876472	0.3412056	-0.4059124	-0.4662056	-0.27876472
A_23_P122674	0.25163913	0.23183322	0.36654544	-0.24110532	-0.23183322	-0.3746693
A_23_P250800	-0.068247795	-0.11196613	0.20245552	-0.22711277	0.06824875	0.23601913
A_23_P392384	-0.33146	-0.35491467	-0.2805872	0.3379135	0.2805872	0.3394251
A_23_P23611	-0.6574874	-1.0203571	0.69216347	-0.024581432	0.02458191	0.6013808
A_23_P68884	0.1021862	0.05596161	0.04508686	-0.04508686	-0.08346939	-0.13256931
A_23_P111402	1.0591068	0.8690548	1.267076	-2.5859656	-3.1925607	-0.8690548
A_23_P136095	-0.12815094	-0.12287998	3.122302	-0.13301945	0.16234207	0.12287998
A_23_P81867	-0.18833542	-0.24621487	-0.12946606	0.1642189	0.1672821	0.12946606
A_23_P382705	0.019473553	-0.019473553	0.27335787	-0.3655157	-0.5355873	0.119407654
A_23_P165457	-0.08195877	-0.07668781	0.07668877	-0.08682728	0.20853424	1.195066
A_23_P72372	-0.22457027	0.9433999	-0.06592274	-0.22943878	0.06592274	2.9713688
A_23_P85503	0.6817112	-0.92747116	3.5792794	-0.9376106	0.77732897	-0.6817112
A_23_P24784	0.39361048	-0.39361	4.5427938	-0.69071865	-0.39535713	2.5880451

Spreadsheet | Scatter Plot

Figure 5.9: Spreadsheet

5.2.1 Spreadsheet Operations

Spreadsheet operations are available by Right-Click on the canvas of the spreadsheet. Operations that are common to all views are detailed in the section Common Operations on Table Views above. In addition, some of the spreadsheet specific operations and the spreadsheet properties are explained below:

Sort: The Spreadsheet can be used to view the sorted order of data with respect to a chosen column. Click on the column header to sort the data based on values in that column. Mouse clicks on the column header of the spreadsheet will cycle though an ascending values sort, a descending values sort and a reset sort. The column header of the sorted column will also be marked with the appropriate icon.

Thus to sort a column in the ascending, click on the column header. This will sort all rows of the spreadsheet based on the values in the chosen column. Also an icon on the column header will denote that this is the sorted column. To sort in the descending order, click again on the same column header. This will sort all the rows of the spreadsheet based on the decreasing values in this column. To reset the sort, click again on the same column. This will reset the sort and the sort icon will disappear from the column header.

Selection: The spreadsheet can be used to select entities, and conditions Entities can be selected by clicking on any cell in the table. Conditions can be selected from the properties dialog of the spreadsheet as detailed below. The selection will be shown by the default selection color on the spreadsheet.

Entity Selection: Entities can be selected by left-clicking on any cell and dragging along the rows. Ctrl-Left-Click selects subsequent entities and Shift-Left-Click selects a consecutive set of entities. The selected entities can be used to create a new entity list by a Left-Click on the 'Create Entity List from Selection' icon. This will launch an entity list inspector where you can provide a name for the entity list, add notes and choose the columns for the entity list. This newly created entity list from the selection will be added to the analysis tree in the navigator.

5.2.2 Spreadsheet Properties

The Spreadsheet Properties Dialog is accessible by right-clicking on the spreadsheet and choosing **Properties** from the menu. The spreadsheet view can be customized and configured from the spreadsheet properties. See Figure 5.10.

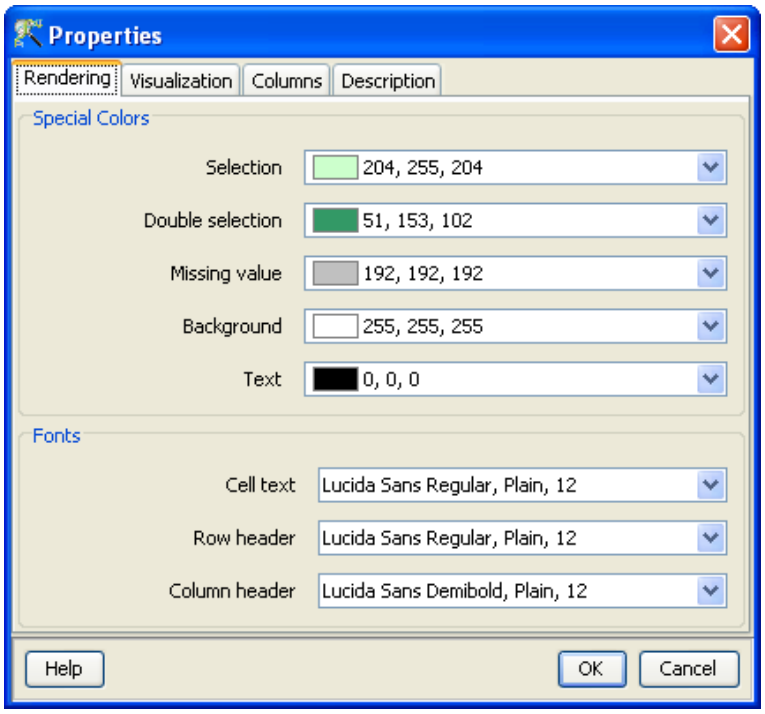


Figure 5.10: Spreadsheet Properties Dialog

Rendering: The rendering tab of the spreadsheet dialog allows you to configure and customize the fonts and colors that appear in the spreadsheet view.

Special Colors: All the colors in the Table can be modified and configured. You can change the Selection color, the Double Selection color, Missing Value cell color and the Background color in the table view. To change the default colors in the view, Right-Click on the view and open the Properties dialog. Click on the Rendering tab of the properties dialog. To change a color, click on the appropriate color bar. This will pop-up a Color Chooser. Select the desired color and click OK. This will change the corresponding color in the Table.

Fonts: Fonts that occur in the table can be formatted and configured. You can set the fonts for Cell text, row Header and Column Header. To change the font in the view, Right-Click on the view and open the Properties dialog. Click on the Rendering tab of the Properties dialog. To change a Font, click on the appropriate drop-down box and choose the required font. To customize the font, click on the customize button. This will pop-up a dialog where you can set the font size and choose the font type as bold or italic.

Visualization: The display precision of decimal values in columns, the row height and the missing value text, and the facility to enable and disable sort are configured and customized by options in this tab.

The visualization of the display precision of the numeric data in the table, the table cell size and the text for missing value can be configured. To change these, Right-Click on the table view and open the Properties dialog. Click on the visualization tab. This will open the Visualization panel.

To change the numeric precision. Click on the drop-down box and choose the desired precision. For decimal data columns, you can choose between full precision and one to four decimal places, or representation in scientific notation. By default, full precision is displayed.

You can set the row height of the table, by entering a integer value in the text box and pressing Enter. This will change the row height in the table. By default the row height is set to 16.

You can enter any a text to show missing values. All missing values in the table will be represented by the entered value and missing values can be easily identified. By default all the missing value text is set to an empty string.

You can also enable and disable sorting on any column of the table by checking or unchecking the check box provided. By default, sort is enabled in the table. To sort the table on any column, click on the column header. This will sort the all rows of the table based on the values in the sort column. This will also mark the sorted column with an icon to denote the sorted column. The first click on the column header will sort the column in the ascending order, the second click on the column header will sort the column in the descending order, and clicking the sorted column the third time will reset the sort.

Columns: The order of the columns in the spreadsheet can be changed by changing the order in the Columns tab in the Properties Dialog.

The columns for visualization and the order in which the columns are visualized can be chosen and configured from the List Selector in the **Column** tab of the **Properties** dialog. Right-Click the view and select **Properties** to open the dialog. Click the **Columns** tab. The List Selector panel shows the **Available items** on the left and the **Selected items** on the right. The items under **Selected items** are the columns that are displayed in the view in the same order in which they appear here.

To move columns from the **Available list** box to the **Selected list** box, select the required items in the **Available items** box and click the **Move right** button between the list boxes. This adds the selected columns from the **Available items** box to the items already in the *Selected items* box. To move columns from the **Selected items** box to the **Available items** box, select the required items in the **Selected items** box and click the **Move left** button. This moves the selected columns from the **Selected items** box to the *Available items* box in the exact position or order in which the column appears in the experiment.

You can also change the order of the columns in which they appear in the view by selecting items in the **Selected items** box and using the **Move Up** and **Move Down** buttons to arrange the columns in the desired order. If multiple items are selected, the first click consolidates the selected items (bring all the selected items together) with the first item in the specified direction. Subsequent clicks on the **Move Up** and **Move Down** buttons will move the selected items as a block in the specified direction, one step at a time until it reaches the end of the list. If only one item or contiguous items are selected in the *Selected items* box, then these will be moved in the specified direction, one step at a time until the end of the list is reached. To reset the order of the columns to the original order in which they currently appear in the plot view, click on the **Reset order** button located between the **Move Up** and **Move Down** buttons.

To select items, Left-Click on the required item. To select multiple items in any of the list boxes, Left-Click and Shift-Left-Click will highlight all contiguous items, and Ctrl-Left-Click will add the clicked-on item to the previously selected elements.

The lower portion of the Columns tab provides a utility to select items in the **List Selector** based on a search by column name or mark. You can either match by **Name** or **Column Mark** whichever is appropriate. By default, the Match by *Name* option is active.

- To match by **Name**, select **Name** from the **Match by item** drop-down list, enter a string in the **Name** text box and hit Enter. **GeneSpring** performs a substring match with the columns listed in the **Available items** and the **Selected items** boxes and selects any matching entries.

- To match by **Mark**, select **Mark** from the **Match by item** drop-down list. The column marks (for example, Affymetrix ProbeSet Id, raw signal, etc.) available in the experiment are shown in the **Mark** drop-down list. Choose a **Mark**, and **GeneSpring** selects the corresponding columns of the experiment.

Description: The title for the view and description or annotation for the view can be configured and modified from the **Description** tab in the **Properties** dialog. Right-Click the view and select **Properties** to open the dialog. Click the **Description** tab to display the the current **Title** and **Description** of the view. Both **Title** and **Description** can be modified by changing the text in the corresponding text boxes. If you modify the default title, the new title is displayed in the Legend and used when the view is exported as **Report**. The change does not reflect in the title tab of the view in the **GeneSpring** desktop. The **Description** (default and modified) appears in the **Legend** window on the right. By default, if the view is derived from running an algorithm, the description will contain the algorithm and the parameters used.

5.3 MvA plot

The MvA plot is a scatter plot of the difference vs. the average of probe measurements between two samples. This plot is specifically used to assess quality and relation between samples. The MvA plot is used more in the two-color spotted arrays to assess the relation between the Cy3 and the Cy5 channels of each hybridization.

The MvA plot is launched from the view menu on the main menu bar with the active entity list in the experiment. Launching the plot from the menu in a two color experiment asks for the channel which can either be a sample or a condition depending on the interpretation chosen. It then shows the relation between the Cy3 and Cy5 channels of individual samples if the interpretation chosen is **All Samples**. In the case of other interpretations, it takes the average of Cy3 and Cy5 for all samples of a condition of the chosen interpretation to show the relation. See figure 5.11 In single color experiments, the plot asks for two inputs for calculating M and A. Depending on the interpretation chosen, the inputs could either be individual samples or conditions. The points in the MvA plot correspond to the entities in the active entity list.

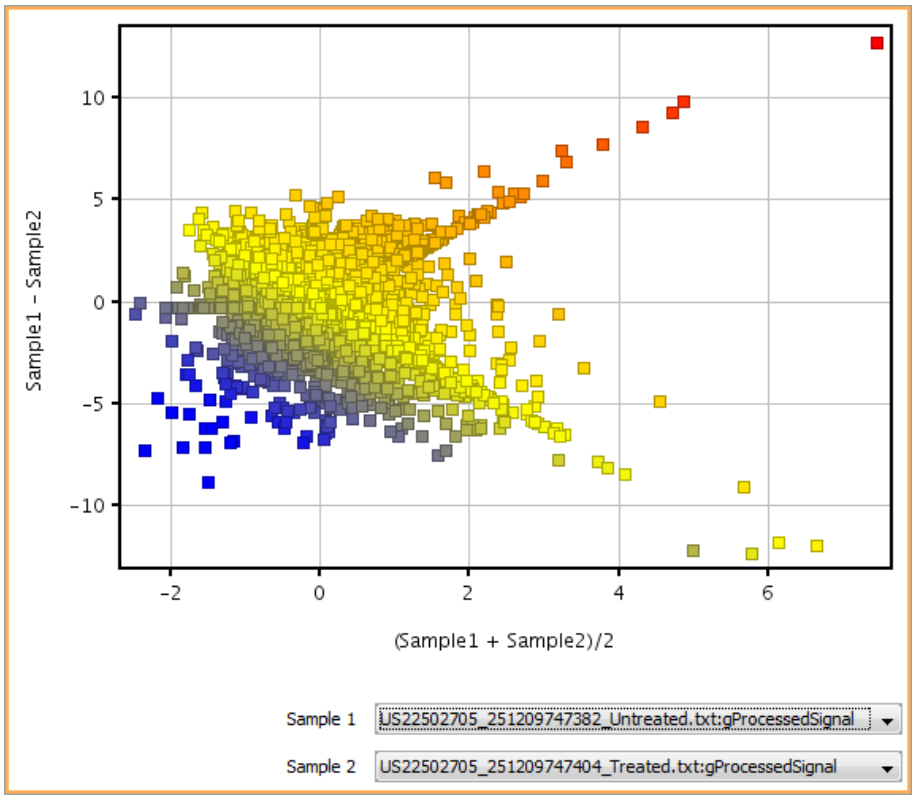


Figure 5.11: MvA Plot

Clicking on another entity list in the experiment will make that entity list active and the MvA plot will dynamically display the current active entity list. Clicking on an entity list in another experiment will translate the entities in that entity list to the current experiment and display those entities in the scatter plot.

The MvA Plot is a lassoed view, and supports both selection and zoom modes. Most elements of the MvA Plot, like color, shape, size of points etc. are configurable from the properties menu described in the properties section of scatter plot.

5.4 The Scatter Plot

The Scatter Plot is launched from view menu on the main menu bar with the active interpretation and the active entity list in the experiment. Alternately, Left-Click on the tool bar icon to launch the scatter plot.The Scatter Plot shows a 2-D scatter of all entities of the active entity list along the first two conditions of the active interpretation by default. If the active interpretation is a unaveraged interpretation, the axes of the scatter plot will be the normalized signal values of the first two samples. If the interpretation is averaged, the axes of the scatter plot will be the averaged normalized signal values of the samples in each condition. The axes of the scatter plot can be changed from the axes chooser on the view. The points in the scatter plot are colored by the normalized signal values of the first sample (or the averaged normalized signal values of the

first condition) and are shown in the scatter plot legend window. The legend window also display the interpretation on which the scatter plot was launched.

Clicking on another entity list in the experiment will make that entity list active and the scatter plot will dynamically display the current active entity list. Clicking on an entity list in another experiment will translate the entities in that entity list to the current experiment and display those entities in the scatter plot.

The Scatter Plot is a lassoed view, and supports both selection and zoom modes. Most elements of the Scatter Plot, like color, shape, size of points etc. are configurable from the properties menu described below. See Figure 5.12.

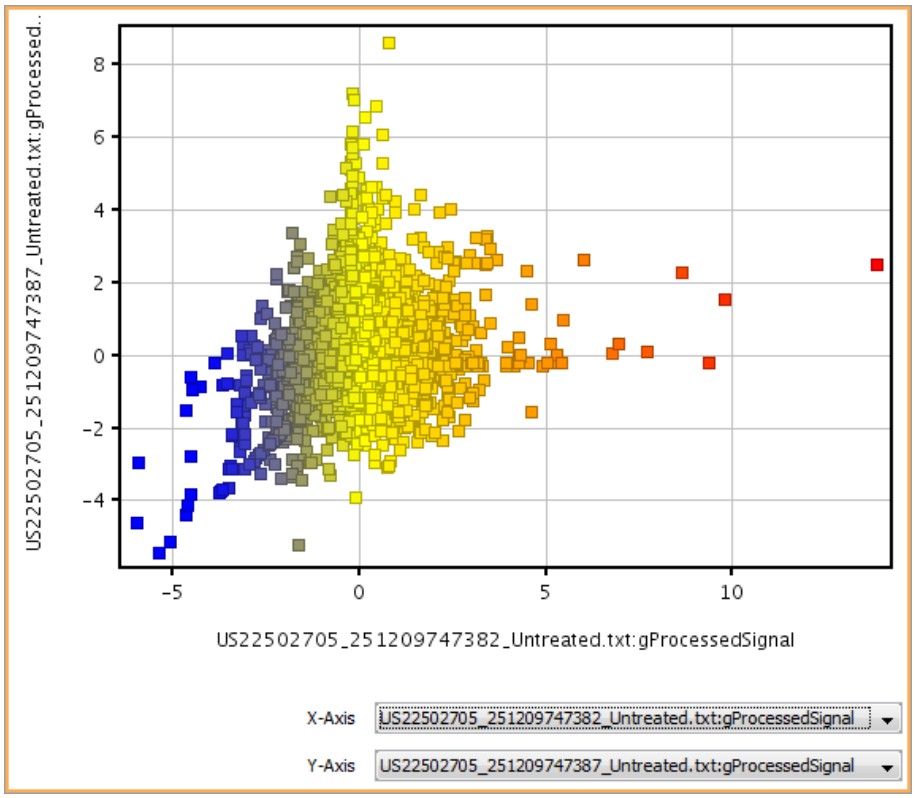


Figure 5.12: Scatter Plot

5.4.1 Scatter Plot for 2 color experiments

For 2 color experiments, a Cy3 vs Cy5 scatter plot can also be launched. This plot shows the raw signal values after log2 transformation. The menu and the tool bar icon give options to plot a regular scatter plot or a Cy3 vs Cy5 plot for 2 color experiments.

5.4.2 Scatter Plot Operations

Scatter Plot operations are accessed by right-clicking on the canvas of the Scatter Plot. Operations that are common to all views are detailed in the section Common Operations on Plot Views. Scatter Plot specific operations and properties are discussed below.

Selection Mode: The Scatter Plot is launched in the selection mode by default. In selection mode, Left-Click and dragging the mouse over the Scatter Plot draws a selection box and all entities within the selection box will be selected. To select additional entities, Ctrl-Left-Click and drag the mouse over desired region. You can also draw and select regions within arbitrary shapes using Shift-Left-Click and then dragging the mouse to get the desired shape.

Selections can be inverted from the pop-up menu on Right-Click inside the Scatter Plot. This selects all unselected points and unselect the selected entities on the scatter plot. To clear the selection, use the Clear selection option from the Right-Click pop-up menu.

The selected entities can be used to create a new entity list by a Left-Click on the 'Create Entity List from Selection' icon. This will launch an entity list inspector where you can provide a name for the entity list, add notes and choose the columns for the entity list. This newly created entity list from the selection will be added to the analysis tree in the navigator.

Zoom Mode: The Scatter Plot can be toggled from the Selection Mode to the Zoom Mode from the right-click drop-down menu on the scatter plot. While in the zoom mode, left-clicking and dragging the mouse over the selected region draws a zoom box and will zoom into the region. Reset zoom from the right-click menu on the scatter plot, to revert back to the default, showing all the points in the dataset.

Save Entities: This allows the user to save entities with respect to the fold change. On selecting this option, a window appears with 4 boxes, each representing a set of entities satisfying a particular condition of fold change.

Scatter plot in Log10/Linear Scale: In **GeneSpring**, the data is usually in log 2 scale and the plots are launched with this data. In Scatter plot, there is an option to launch with data in log10 or linear scale, from menu *View* → *Plot Log10/Linear Values*. Refer section Plot Log10/Linear Values for details.

5.4.3 Scatter Plot Properties

The Scatter Plot view offers a wide variety of customization with log and linear scale, colors, shapes, sizes, drawing orders, error bars, line connections, titles and descriptions from the Properties dialog. These customizations appear in three different tabs on the Properties window, labelled Axis, Visualization, Rendering, Description. See Figure 5.13.

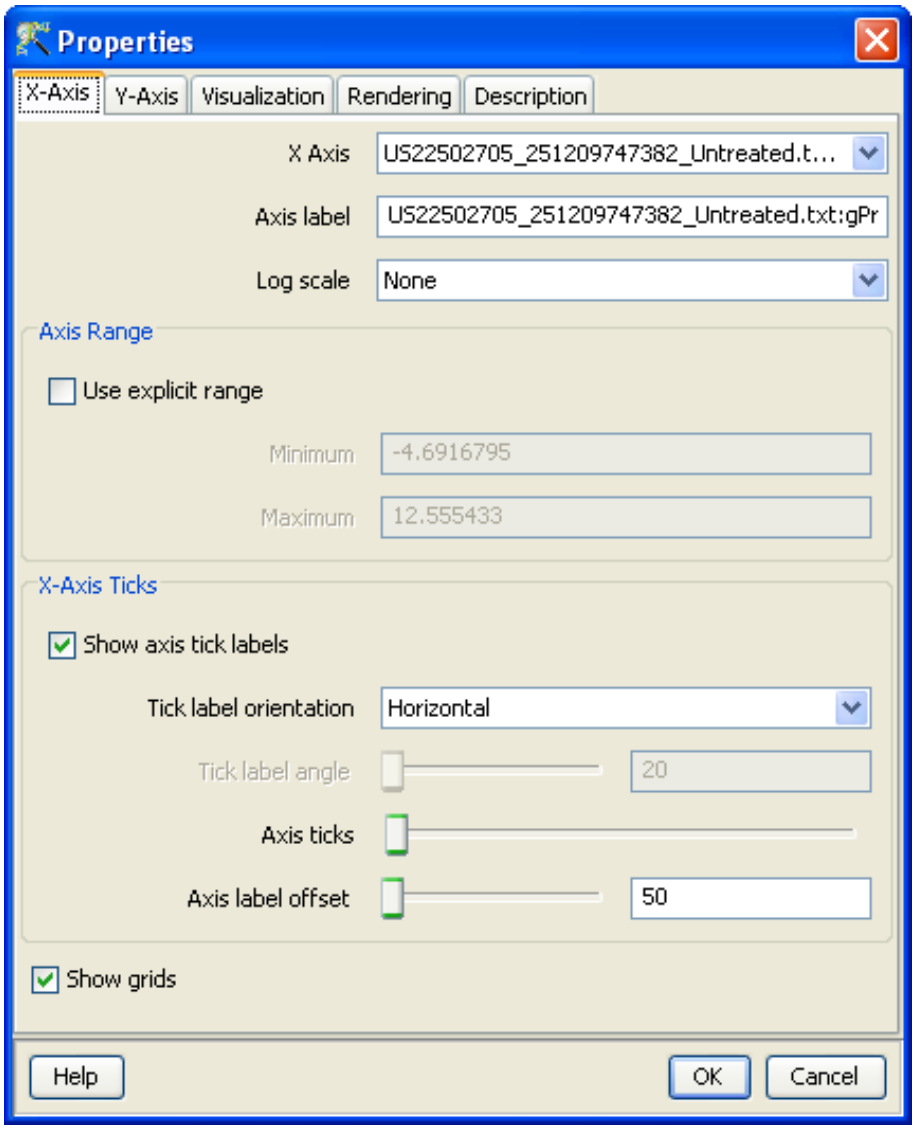


Figure 5.13: Scatter Plot Properties

Axis: The axes of the Scatter Plot can be set from the Properties Dialog or from the Scatter Plot itself. When the Scatter Plot is launched, it is drawn with the first two conditions of the interpretation. These axes can be changed from the Axis selector in the drop down box in this dialog or in the Scatter Plot itself.

The axis for the plot, axis titles, the axis scale, the axis range, the axis ticks, tick labels, orientation and offset, and the grid options of the plot can be changed and modified from the axis tabs of the scatter plot properties dialog.

To change the scale of the plot to the log scale, click on the log scale option for each axis. This will provide a drop-down of the log scale options.

None: If None is chosen, the points on the chosen axis is drawn on the linear scale

Log:, If Log Scale is chosen, the points on the chosen axis is drawn on the log scale, with log of negative values if any being marked at missing values and dropped from the plot.

$(if x > 0), x = log(x)$
$(if x <= 0), x =$ missing value

Symmetric Log: If Symmetric Log is chosen, the points along the chosen axis are transformed such that for negative values, the log of the 1− absolute value is taken and plotted on the negative scale and for positive values the log of 1+ absolute value is taken and plotted on the positive scale.
$(if x >= 0), x = log(1 + x)$
$(if x < 0), x = -log(1 - x)$

To use an explicit range for the scatter plot, check this option and set the minimum and maximum range. By default, the minimum and maximum will be set to the minimum and maximum of the corresponding axis or column of the dataset. If explicit range is explicitly set in the properties dialog, this will be maintained even if the axis columns are changed.

The grids, axes labels, and axis ticks of the plots can be configured and modified from the Properties dialog. To modify these, Right-Click the view, select **Properties** and click on the **Axis** tab.

The plot can be drawn with or without the grid lines by clicking on the 'Show grids' option.

The ticks and axis labels are automatically computed and shown on the plot. You can show or remove the axis labels by clicking on the Show Axis Labels check box. Further, the orientation of the tick labels for the X-Axis can be changed from the default horizontal position to a slanted position or vertical position by using the drop down option and by moving the slider for the desired angle.

The number of ticks on the axis are automatically computed to show equal intervals between the minimum and maximum and displayed. You can increase the number of ticks displayed on the plot by moving the Axis Ticks slider. For Numerical data columns, you can double the number of ticks shown by moving the slider to the maximum. For categorical columns, if the number of categories are less than ten, all the categories are shown and moving the slider does not increase the number of ticks.

Visualization: The colors, shapes and sizes of points in the Scatter Plot are configurable.

Color By: The points in the Scatter Plot can be plotted in a fixed color by clicking on the Fixed radio button. The color can also be determined by values in one of the columns by clicking the 'By Columns' radio button and choosing the column to color by, as one of the columns in the dataset. This colors the points based on the values in the chosen columns. The color range can be modified by clicking the Customize button.

Shape By: The shape of the points on the scatter plot can be drawn with a fixed shape or be based on values in any categorical column of the active dataset. To change the 'Shape By' column, click on the drop down list provided and choose any column. Note that only categorical columns in the active dataset will be shown list. To customize the shapes, click on the customize button next to the drop down list and choose appropriate shapes.

Size By: The size of points in the scatter plot can be drawn with a fixed shape, or can be drawn based upon the values in any column of the active dataset. To change the 'Size By' column, click on the drop down box and choose an appropriate column. This will change the plot sizes depending on the values in the particular column. You can also customize the sizes of points in the plot, by clicking on the customize button. This will pop up a dialog where the sizes can be set.

Drawing Order: In a Scatter Plot with several points, multiple points may overlap causing only the last in the drawing order to be fully visible. You can control the drawing order of points by specifying a column name. Points will be sorted in increasing order of value

in this column and drawn in that order. This column can be categorical or numerical. If this column is numeric and you wish to draw in decreasing order instead of increasing, simply scale this column by -1 using the scale operation and use this column for the drawing order.

Error Bars: When visualizing profiles using the scatter plot, you can also add upper and lower error bars to each point. The length of the upper error bar for a point is determined by its value in a specified column, and likewise for the lower error bar.

If error columns are available in the current dataset,this can enable viewing Standard Error of Means via error bars on the scatter plot.

Jitter: If the points on the scatter plot are too close to each other, or are actually on top of each other, then it is not possible to view the density of points in any portion of the plot. To enable visualizing the density of plots, the jitter function is helpful. The jitter function will perturb all points on the scatter plot within a specified range, randomly, and the draw the points. the Add jitter slider specifies the range for the jitter. By default there is no jitter in the plots and the jitter range is set to zero. the jitter range can be increased by moving the slider to the right. This will increase the jitter range and the points will now be randomly perturbed from their original values, within this range.

Connect Points: Points with the same value in a specified column can be connected together by lines in the Scatter Plot. This helps identify groups of points and also visualize profiles using the scatter plot. The column specified must be a categorical column. This column will be used to group the points together. The order in which these will be connected by lines is given by another column, namely the 'Order By' column. This 'Order By' column can be categorical or numerical. See Figure 5.14.

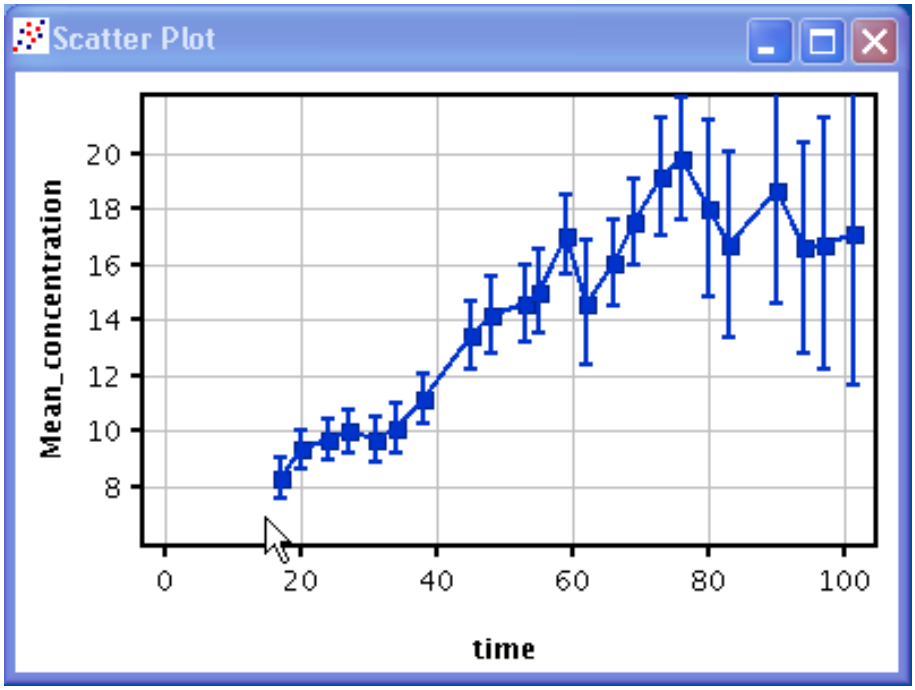


Figure 5.14: Viewing Profiles and Error Bars using Scatter Plot

Labels: You can label each point in the plot by its value in a particular column; this column can be chosen in the Label Column drop-down list. Alternatively, you can choose to label only the selected points.

Fold Change Lines: This option allows the user to draw fold change lines on the scatter plot based on the following equations:

- y=x+log(FC), y = x , y = x-log(FC) ------(1)
- y=(FC) x , y = x, y = (1/FC) x ----------(2)

The lines that are drawn on the Scatter Plot depends on the following two parameters:

1. Datatype
   - **Normalized Data**: If the scales chosen for the x-axis and y-axis is not same then no lines are drawn. If the scales chosen are same and are 'None' then lines given by (1) above are drawn. If the scales chosen are 'Log' then no lines are drawn. If the scales chosen are 'Symmetric Log' then no lines are drawn.
   - **Raw Data**: If the scales chosen for the x-axis and y-axis is not same then no lines are drawn.If the scales chosen are same and are 'None' then lines given by (2) above are drawn. If the scales chosen are 'Log' then lines given by (1) above are drawn. If the scales chosen are 'Symmetric Log' then no lines are drawn.
2. Axis Scale

The default fold change value given is 2.0. The user can change the default value either by moving the slider or entering the value in the appropriate box. When the default value is changed, the view gets dynamically altered to reflect the new fold change lines. See Figure 5.15.

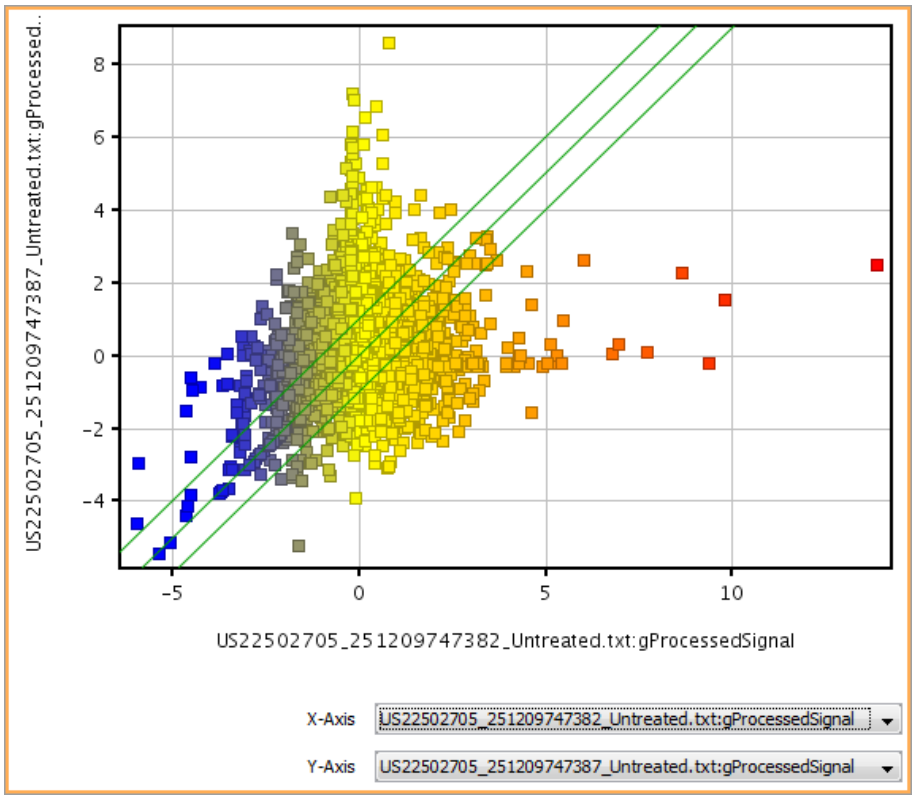


Figure 5.15: Scatter plot with Fold Change lines

Rendering: The Scatter plot allows all aspects of the view to be customized. Fonts, colors, offsets, etcetera can all be configured.

Fonts: All fonts on the plot can be formatted and configured. To change the font in the view, Right-Click on the view and open the Properties dialog. Click on the **Rendering**

tab of the **Properties** dialog. To change a **Font**, click on the appropriate drop-down box and choose the required font. To customize the font, click on the customize button. This will pop-up a dialog where you can set the font size and choose the font type as bold or italic.

Special Colors: All the colors that occur in the plot can be modified and configured. The plot Background color, the Axis color, the Grid color, the Selection color, as well as plot specific colors can be set. To change the default colors in the view, Right-Click on the view and open the Properties dialog. Click on the Rendering tab of the Properties dialog. To change a color, click on the appropriate arrow. This will pop-up a **Color Chooser**. Select the desired color and click **OK**. This will change the corresponding color in the View.

Offsets: The bottom, top, left, and right offsets of a plot can be modified and configured. These offsets can be changed if the axis labels or axis titles are not completely visible in the plot, or if only the graph portion of a plot is required. To change the offsets, Right-Click on the view and open the Properties dialog. Click on the Rendering tab. To change plot offsets, move the corresponding slider, or enter an appropriate value in the text box provided. This will change the particular offset in the plot.

Miscellaneous: The quality of the plot can be enhanced by anti aliasing all the points in the plot. this is done to ensure better print quality. To enhance the plot quality, click on the High Quality Plot option.

Column Chooser: The column chooser can be disable and removed from the scatter plot if required. The plot area will be increased and the column chooser will not be available on the scatter plot. To remove the column chooser from the plot, uncheck the Show Column Chooser option.

Description: The title for the view and description or annotation for the view can be configured and modified from the **Description** tab in the **Properties** dialog. Right-Click the view and select **Properties** to open the dialog. Click the **Description** tab to display the the current **Title** and **Description** of the view. Both **Title** and **Description** can be modified by changing the text in the corresponding text boxes. If you modify the default title, the new title is displayed in the Legend and used when the view is exported as **Report**. The change does not reflect in the title tab of the view in the **GeneSpring** desktop. The **Description** (default and modified) appears in the **Legend** window on the right. By default, if the view is derived from running an algorithm, the description will contain the algorithm and the parameters used.

5.5 The Profile Plot View

The Profile Plot is launched from the view menu on the main menu bar. Alternately, Left-Click on the tool bar icon to launch the profile plot. The profile plot (referred to as 'Graph View' in earlier versions of **GeneSpring**) is one of the important visualizations of normalized expression value data against the chosen interpretation. In fact, the default view of visualizing interpretations is the profile plot launched by clicking on the interpretation in the experiment and making it the active interpretation. See Figure 5.16.

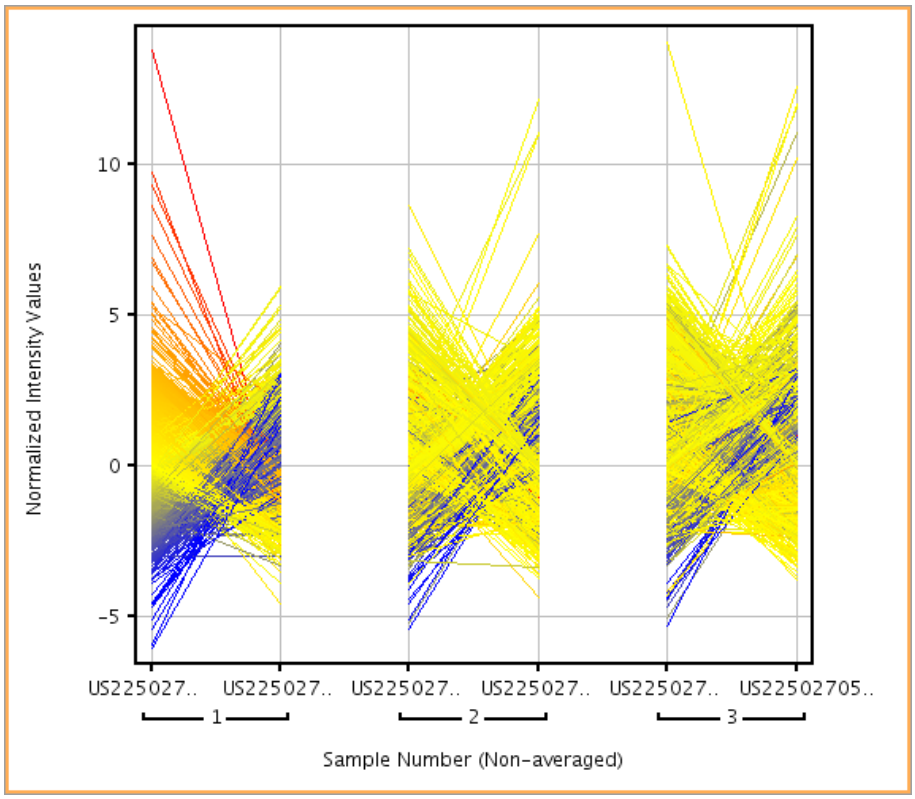


Figure 5.16: Profile Plot

When the profile plot is launched from the view menu, it is launched with the active interpretation and the active entity list in the experiment. The profile plot shows the conditions in the active interpretation along the x-axis and the normalized expression values in the y-axis. Each entity in the active entity list is shown as a profile in the plot. Depending upon the interpretation, whether averaged or non-averaged, the profile of the entity in each group is split and displayed along the conditions in the interpretation.

5.5.1 Time Series Plot

Starting from **GeneSpring** version 11.5, Time Series Plots are also supported in the product. A Time Series Plot is a special type of Profile Plot where the interpretation on which it is launched contains only one 'Numerical' parameter with numeric values only. In a Time Series Plot, the X axis values are scaled according to the numerical parameter values

The Figure 5.17 shows the numerical/categorical parameters of an interpretation, Figure 5.18 shows the conditions for that interpretation and the Figure 5.19 shows the Time Series plot launched for this interpretation.

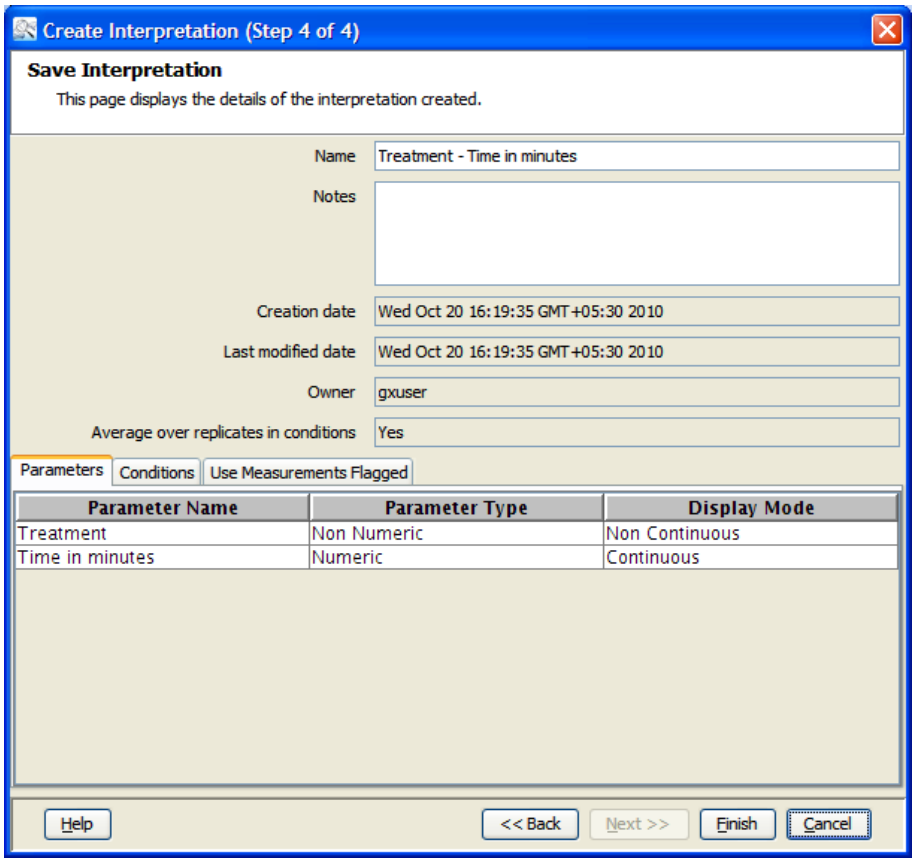


Figure 5.17: Interpretation for which a Time Series Plot will be launched (Parameters)

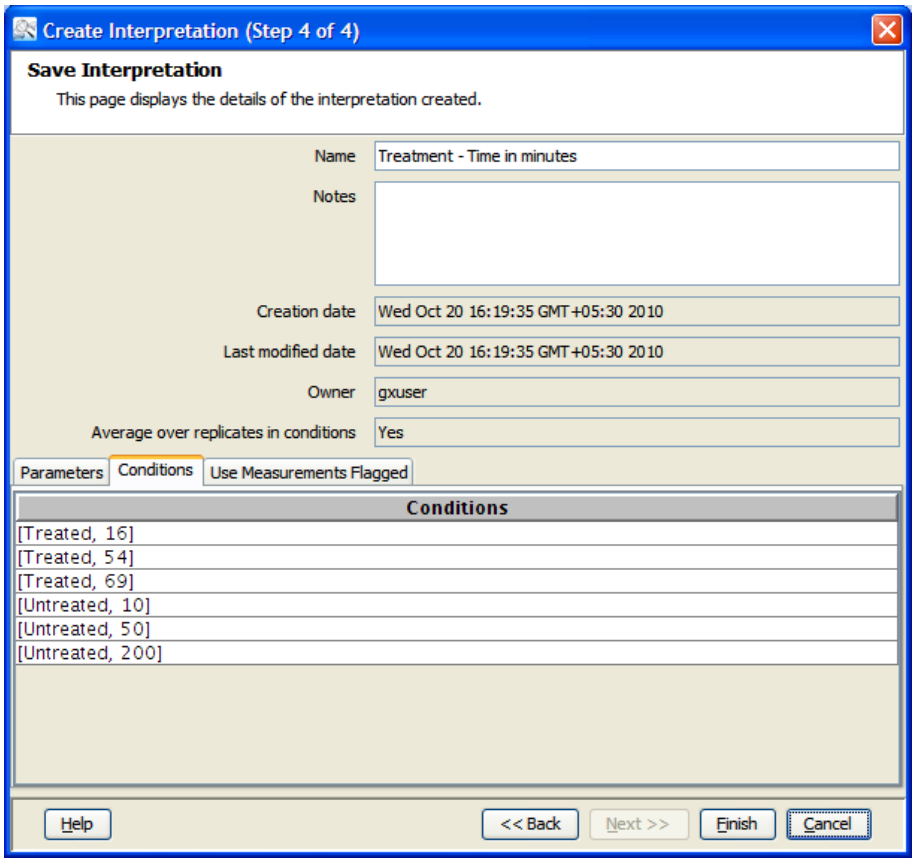


Figure 5.18: Interpretation for which a Time Series Plot will be launched (Conditions)

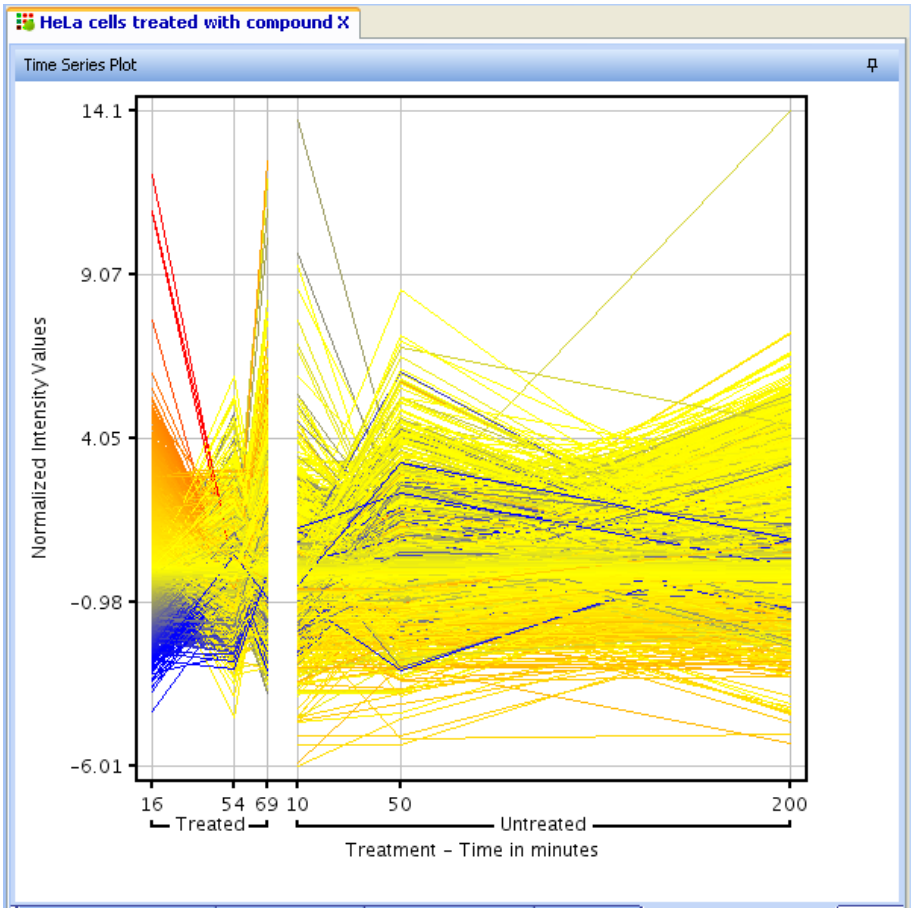


Figure 5.19: Time Series Plot

Profile Plot for All Samples: If the active interpretation is the default *All Samples* interpretation, then each sample is shown in the x-axis and the normalized expression values for each entity in the active entity list is connected across all the samples.

Profile Plot of Non-Averaged Interpretation: If the active interpretation is *Non-Averaged* over the replicates, then the samples in each condition are grouped together along the x-axis, and the profile plot of the entities in the active interpretation is numerical within the samples in a condition and split across the conditions.

Profile Plot of Averaged Interpretation: If the active interpretation is averaged, over the replicates, then the conditions in the interpretation are plotted on the x-axis. The profile plot of the entities in the active entity list is displayed in numerical mode with the averaged condition. And if there are multiple parameters in the interpretation, the profile plot will be shown according to the display mode specified in the interpretation. Say, there are two parameters 'dosage' (which is 'Numerical') and 'Gender' (which is 'Categorical'). These two parameters combine to make conditions, and the profile will be numerical with dosage and split along Gender.

Clicking on another entity list in the experiment will make that entity list active and the profile plot will dynamically display the current active entity list. Clicking on an entity list in another

experiment will translate the entities in that entity list to the current experiment and display those entities in the profile plot.

The Profile Plot supports both the Selection Mode and the Zoom Modes. The profile plot is launched with the selection mode as default and colored by the values in the first condition. The interpretation of the profile plot and the color band are displayed in the legend window.

5.5.2 Profile Plot Operations

The Profile Plot operations are accessed by right-clicking on the canvas of the Profile Plot. Operations that are common to all views are detailed in the section Common Operations on Plot Views. Profile Plot specific operations and properties are discussed below.

Selection Mode: The Profile Plot is launched, by default, in the selection mode. While in the selection mode, left-clicking and dragging the mouse over the Profile Plot will draw a selection box and all profiles that intersect the selection box are selected. To select additional profiles, Ctrl-Left-Click and drag the mouse over desired region. Individual profiles can be selected by clicking on the profile of interest.

Zoom Mode: While in the zoom mode, left-clicking and dragging the mouse over the selected region draws a zoom box and will zoom into the region. Reset Zoom will revert back to the default, showing the plot for all the entities in the active entity list.

5.5.3 Profile Plot Properties

The following properties are configurable in the Profile Plot. See Figure 5.20.

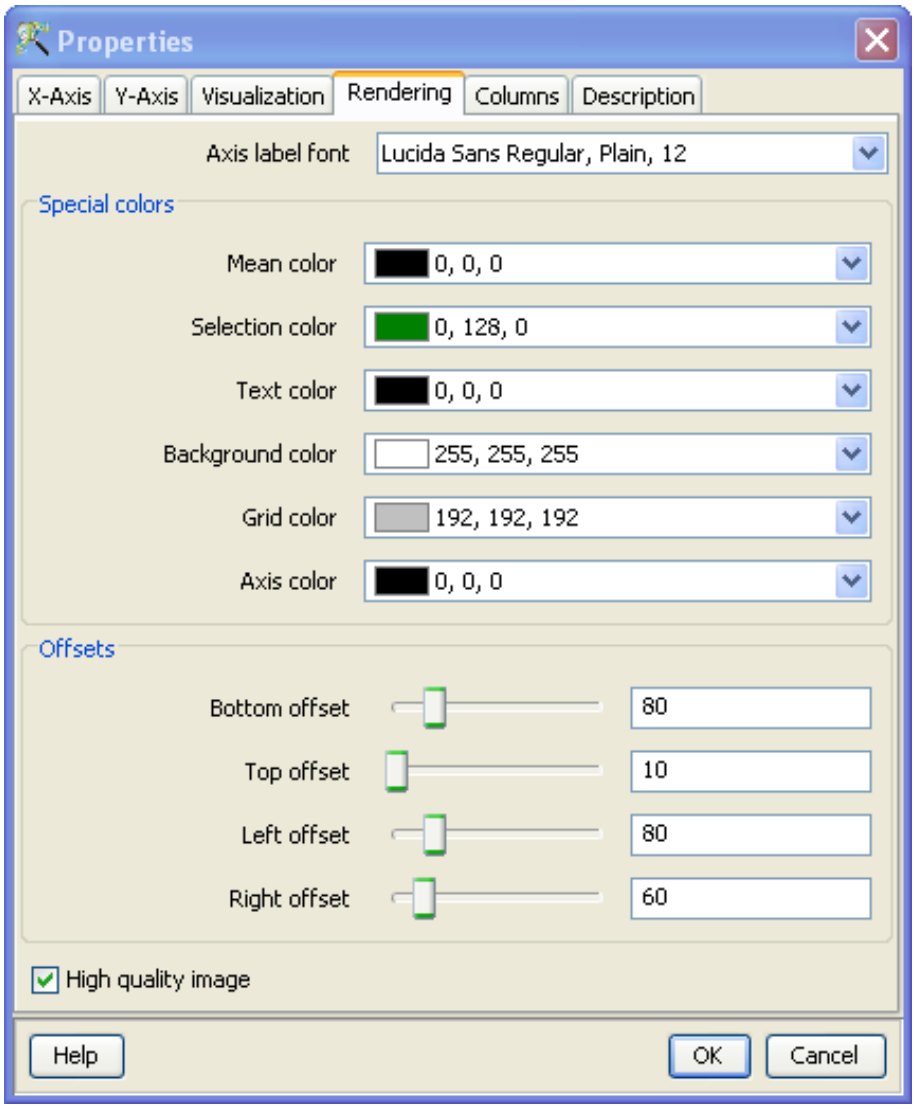


Figure 5.20: Time Series Plot

Axis: The grids, axes labels, and axis ticks of the plots can be configured and modified from the Properties dialog. To modify these, Right-Click the view, select **Properties** and click on the **Axis** tab.

The plot can be drawn with or without the grid lines by clicking on the 'Show grids' option.

The ticks and axis labels are automatically computed and shown on the plot. You can show or remove the axis labels by clicking on the Show Axis Labels check box. Further, the orientation of the tick labels for the X-Axis can be changed from the default horizontal position to a slanted position or vertical position by using the drop down option and by moving the slider for the desired angle.

The number of ticks on the axis are automatically computed to show equal intervals between the minimum and maximum and displayed. You can increase the number of ticks displayed on the plot by moving the Axis Ticks slider. For Numerical data columns, you can double the number of ticks shown by moving the slider to the maximum. For categorical columns, if the number of categories are less than ten, all the categories are shown and moving the

slider does not increase the number of ticks.

Visualization: The Profile Plot displays the mean profile over all rows by default. This can be hidden by unchecking the Display Mean Profile check box.

The colors of the Profile Plot can be changed from the properties dialog. The colors of the profile plot can be changed from this dialog. You can choose a fixed color or use one of the data columns to color the profile plot by choosing a column from the drop-down list. The colors range of the profile plot and the middle color can be customized by clicking on the *Customize* button and choosing the minimum color, the middle color and the maximum color. By default, the minimum color is set to the median value of the data column.

Rendering: The rendering of the fonts, colors and offsets on the Profile Plot can be customized and configured.

Fonts: All fonts on the plot can be formatted and configured. To change the font in the view, Right-Click on the view and open the Properties dialog. Click on the **Rendering** tab of the **Properties** dialog. To change a **Font**, click on the appropriate drop-down box and choose the required font. To customize the font, click on the customize button. This will pop-up a dialog where you can set the font size and choose the font type as bold or italic.

Special Colors: All the colors that occur in the plot can be modified and configured. The plot Background color, the Axis color, the Grid color, the Selection color, as well as plot specific colors can be set. To change the default colors in the view, Right-Click on the view and open the Properties dialog. Click on the Rendering tab of the Properties dialog. To change a color, click on the appropriate arrow. This will pop-up a **Color Chooser**. Select the desired color and click **OK**. This will change the corresponding color in the View.

Offsets: The bottom, top, left, and right offsets of a plot can be modified and configured. These offsets can be changed if the axis labels or axis titles are not completely visible in the plot, or if only the graph portion of a plot is required. To change the offsets, Right-Click on the view and open the Properties dialog. Click on the Rendering tab. To change plot offsets, move the corresponding slider, or enter an appropriate value in the text box provided. This will change the particular offset in the plot.

Quality Image: The Profile Plot image quality can be increased by checking the High-Quality anti-aliasing option. This is slow however and should be used only while printing or exporting the Profile Plot.

Column: The Profile Plot is launched with a default set of columns. The set of visible columns can be changed from the Columns tab. The columns for visualization and the order in which the columns are visualized can be chosen and configured from the List Selector in the **Column** tab of the **Properties** dialog. Right-Click the view and select **Properties** to open the dialog. Click the **Columns** tab. The List Selector panel shows the **Available items** on the left and the **Selected items** on the right. The items under **Selected items** are the columns that are displayed in the view in the same order in which they appear here.

To move columns from the **Available list** box to the **Selected list** box, select the required items in the **Available items** box and click the **Move right** button between the list boxes. This adds the selected columns from the **Available items** box to the items already in the *Selected items* box. To move columns from the **Selected items** box to the **Available items** box, select the required items in the **Selected items** box and click the **Move left** button. This moves the selected columns from the **Selected items** box to the *Available items* box in the exact position or order in which the column appears in the experiment.

You can also change the order of the columns in which they appear in the view by selecting items in the **Selected items** box and using the **Move Up** and **Move Down** buttons

to arrange the columns in the desired order. If multiple items are selected, the first click consolidates the selected items (bring all the selected items together) with the first item in the specified direction. Subsequent clicks on the **Move Up** and **Move Down** buttons will move the selected items as a block in the specified direction, one step at a time until it reaches the end of the list. If only one item or contiguous items are selected in the *Selected items* box, then these will be moved in the specified direction, one step at a time until the end of the list is reached. To reset the order of the columns to the original order in which they currently appear in the plot view, click on the **Reset order** button located between the **Move Up** and **Move Down** buttons.

To select items, Left-Click on the required item. To select multiple items in any of the list boxes, Left-Click and Shift-Left-Click will highlight all contiguous items, and Ctrl-Left-Click will add the clicked-on item to the previously selected elements.

The lower portion of the Columns tab provides a utility to select items in the **List Selector** based on a search by column name or mark. You can either match by **Name** or **Column Mark** whichever is appropriate. By default, the Match by *Name* option is active.

- To match by **Name**, select **Name** from the **Match by item** drop-down list, enter a string in the **Name** text box and hit Enter. **GeneSpring** performs a substring match with the columns listed in the **Available items** and the **Selected items** boxes and selects any matching entries.
- To match by **Mark**, select **Mark** from the **Match by item** drop-down list. The column marks (for example, Affymetrix ProbeSet Id, raw signal, etc.) available in the experiment are shown in the **Mark** drop-down list. Choose a **Mark**, and **GeneSpring** selects the corresponding columns of the experiment.

Description: The title for the view and description or annotation for the view can be configured and modified from the **Description** tab in the **Properties** dialog. Right-Click the view and select **Properties** to open the dialog. Click the **Description** tab to display the the current **Title** and **Description** of the view. Both **Title** and **Description** can be modified by changing the text in the corresponding text boxes. If you modify the default title, the new title is displayed in the Legend and used when the view is exported as **Report**. The change does not reflect in the title tab of the view in the **GeneSpring** desktop. The **Description** (default and modified) appears in the **Legend** window on the right. By default, if the view is derived from running an algorithm, the description will contain the algorithm and the parameters used.

5.6 The Heatmap View

The heatmap is launched from View Menu on the main menu bar with the active interpretation and the active entity list in the experiment. Alternately, Left-Click on the tool bar icon to launch the heat map view. The Heatmap displays the normalized signal values of the conditions in the active interpretation for all the entities in the active entity list. The legend window displays the interpretation on which the heat map was launched.

Clicking on another entity list in the experiment will make that entity list active and the heatmap will dynamically display the current active entity list. Clicking on an entity list in another experiment will translate the entities in that entity list to the current experiment and display those entities in the heat map.

The expression value of each gene is mapped to a color-intensity value. The mapping of expression values to intensities is depicted by a color-bar created by the range of values in the conditions of the interpretation. This provides a birds-eye view of the values in the dataset. The tool tip on a cell in the heat map shows the normalized expression value of the entity.

The heat map allows selecting the entities (rows) and selecting the conditions (columns) and these are lassoed in all the views. To select contiguous cells in the heat map, click and drag to draw a rectangular box on the canvas of the heat map. The corresponding entities and conditions will be selected and lassoed across all views. See Figure 31.53.

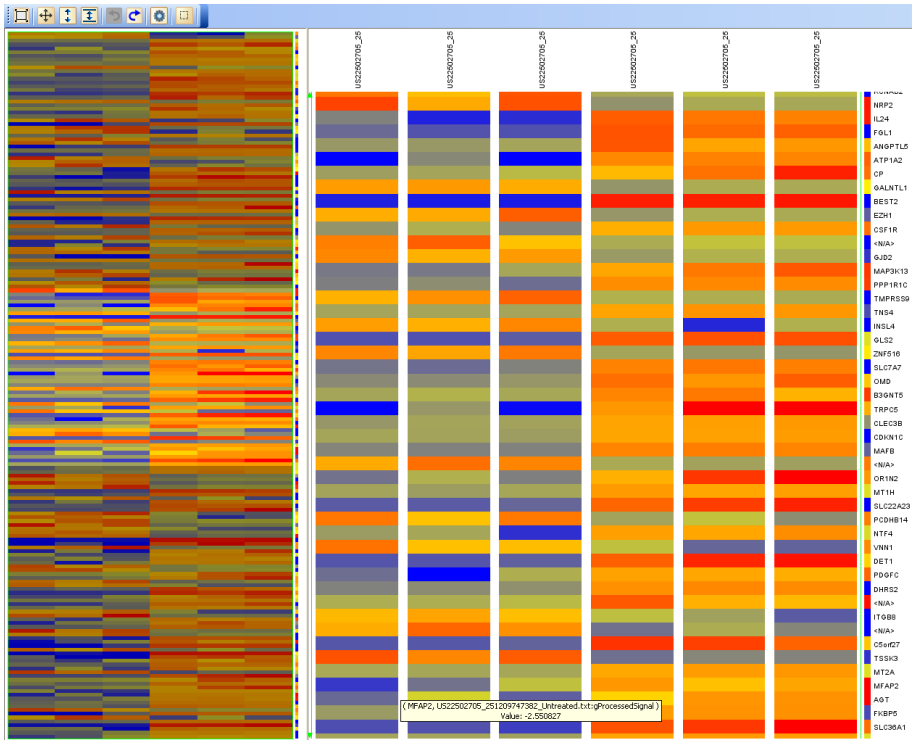


Figure 5.21: Heatmap

5.6.1 Heatmap Operations

Operations on heat map are accessible on the tool bar and Right-Click menu on the canvas of the heat map. Operations that are common to all views are detailed in the section Common Operations on Table Views above. In addition, some of the heat specific operations and the HeatMap properties are explained below: See Figure 5.22.

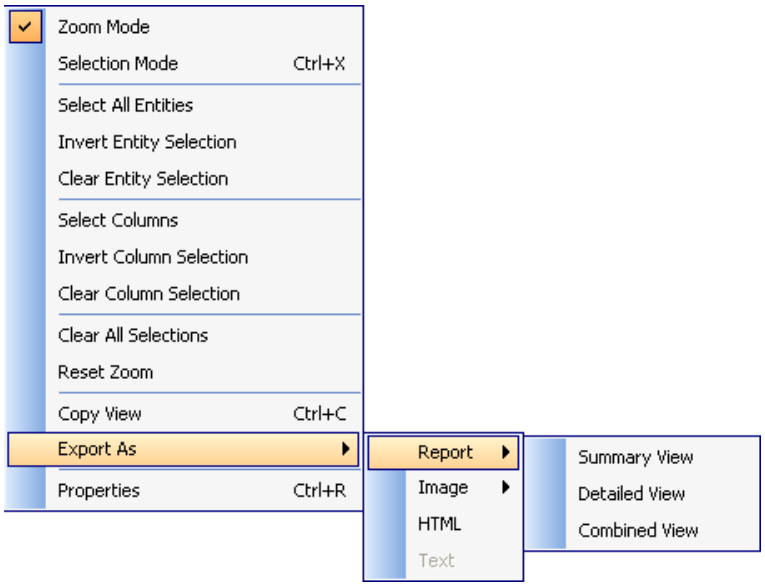


Figure 5.22: Heatmap Export submenu

Select All Entities: Selects all entities in the heatmap.

Invert Entity Selection: Inverts the entity selection.

Clear Entity Selection: Deselects any selected entities.

Select Columns: Allows you to select individual samples/columns, which consequently appear highlighted in the cluster tree. See "Select Columns" on page 652.

Invert Column Selection: Inverts the column selection.

Clear Column Selection: Deselects any selected columns.

Export As Image: You can export only the visible region or the whole heatmap image. Images of any size can be exported with high quality. For png, jpg, jpeg, and bmp format exports, large size imags will be split into multiple images. In such cases, the individual images will be written separately and reported. However, if a .tiff, .svg, or .eps image format is chosen, large size images will also be exported as a single image.

> Note: You can export the whole heatmap as a single image with any size and desired resolution. Images of any size can be exported as a .tiff, .svg, or .eps file. This image can be opened on any machine with enough resources for handling large image files.

Export as HTML: If a whole heatmap image export is chosen, multiple images will be exported and can be opened in a browser.

5.6.2 Heatmap Toolbar

The icons on the heatmap and their operations are listed below: See Figure 5.23.

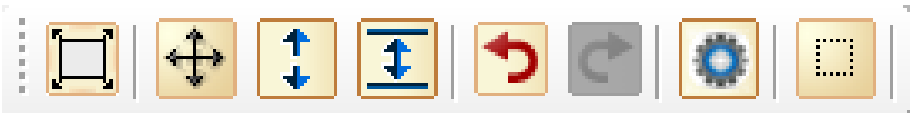


Figure 5.23: Heatmap Toolbar

Toggle fullscreen mode icon: Expands the Heatmap view to appear in fullscreen mode. Click the icon again to return to the original view size.

Switch heatmap orientation: Changes the orientation of the heatmap.

Show all entity labels: Zooms into the heatmpa within the selected area to a level where the maximum number of entity labels can be displayed.

Fit all entities to a page: Zooms out of the heatmap to display all the entities (rows) in the selected area.
Note: Entity labels are not visible when the spacing becomes too small to display labels. Zooming in or resetting will restore the labels.

Go to previous zoom rectangle: Returns the view to the previously selected or zoomed area.

Go to next zoom rectangle: Displays the next selected or zoomed area.

Properties: Opens the Heatmap Properties dialog box.

Select all in zoomed view: Selects all the entities in the selected area.

5.6.3 Heatmap Properties

The heatmap views supports the following configurable properties. See Figure 5.24.

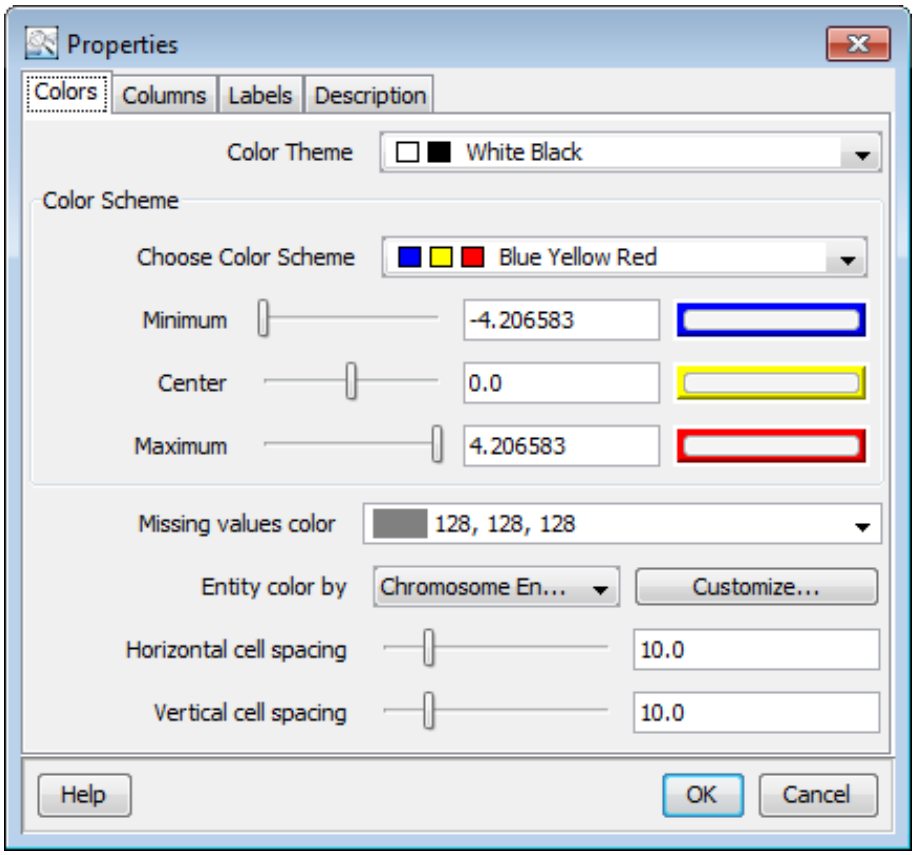


Figure 5.24: Heatmap Properties

Colors

- **Color Theme**: Choose a color theme from the drop-down list or select **Custom Color Theme** to create your own theme.
- **Color Scheme**: Choose a color scheme from the drop-down list or select **Custom Color Scheme** to create your own color scheme. Use the available Minimum, Center, Maximum sliders to modify the coloring values of the selected color scheme.
- **Other options**: Determine a Missing values color, select the annotation to use for coloring the entities from the **Entity color by** drop-down list and use the **Customize** option to change the default color range. You can also define the **Horizontal and Vertical cell spacing**.

Columns

- **Sort entities by**: Select the annotation to use for determining the order in which the entities are displayed in the heatmap from this drop-down list.
- **List Selector**: Choose the columns (samples) to be displayed in the heatmap and determine the order in which they appear using this selection functionality.
- **Highlight Column**: Use this functionality to identify columns (samples) from the columns listed in the **List Selector** with a common characteristic in their Name or that contain the same Mark.

Labels Choose whether to display Entity and Condition labels, the annotation to display for entity labels, the label width and location as well as the font.

Description Edit the **Title** and **Description** for the heatmap.

5.6.4 Heatmap for viewing Copy Number Analysis Results

Heatmap view is supported for visualizing the results of Copy Number Anaysis in **GeneSpring** . Copy Number and LOH values can be visualized in the heatmap for a chosen entity list and interpretation. See Section Heatmap View for Copy Number.

5.7 The Histogram View

The Histogram is launched from View menu on the main menu bar with the active interpretation and the active entity list in the experiment. Alternately, Left-Click on the tool bar icon to launch the histogram. This toolbar provides the option to view either a single histogram or multiple histograms in one view. Multiple histograms will be launched with different samples in the experiment and there is an option in 'Properties → Rendering' to set the number of samples to be shown in the view.

The view shows a histogram of one condition in the active interpretation as a bar chart of the frequency or number of entities in each interval of the condition. This is done by binning the normalized signal value of the condition into equal interval bins and plotting the number of entities in each bin. If the default *All Samples* interpretation is chosen, the histogram will correspond to the normalized signal values of the first sample. If an averaged interpretation is active interpretation, then the histogram will correspond to the averaged normalized signal values of the samples in the first condition. You can change the condition on which the histogram is drawn from the drop-down list on the view. The legend window displays the interpretation on which the histogram was launched. See Figure 5.25.

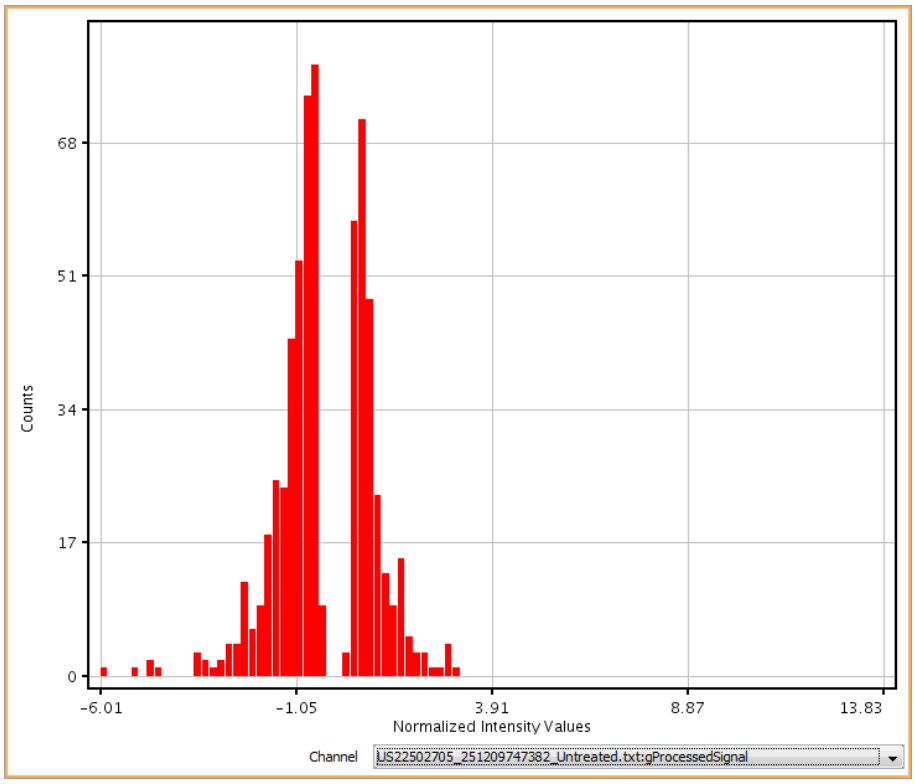


Figure 5.25: Histogram

Clicking on another entity list in the experiment will make that entity list active and the histogram will dynamically display the frequency of this entity list on the condition. Clicking on an entity list in another experiment will translate the entities in that entity list to the current experiment and display the frequency of those entities in the histogram.

The frequency in each bin of the histogram is dependent upon the lower and upper limits of binning, and the size of each bin. These can be configured and changed from the **Properties** dialog.

When 'multiple histogram' is launched, the view shows the histogram of multiple samples (as set in the Number of samples options in 'Rendering' section of **Properties** dialog. In this case, the setting in **Properties** dialog apply to all the samples and hence all the histograms.

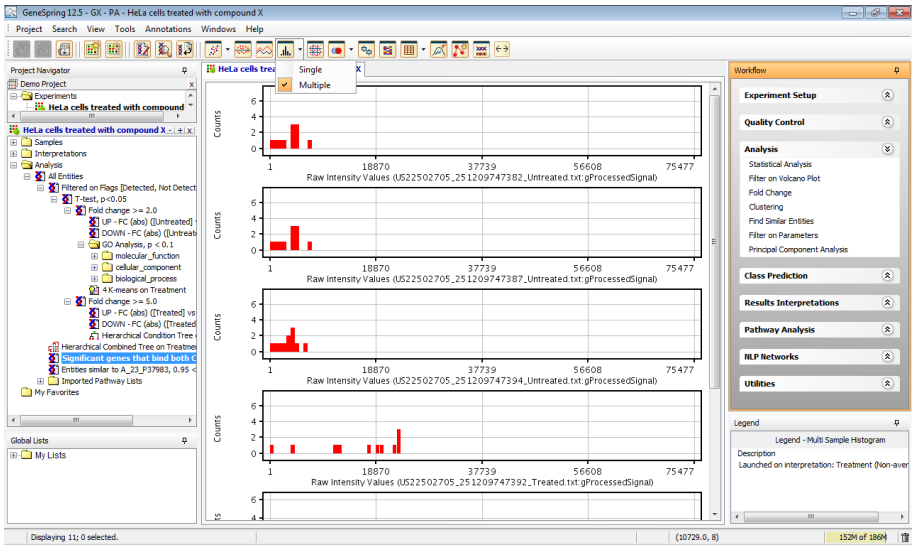


Figure 5.26: Histogram Viewing Options

5.7.1 Histogram Operations

The Histogram operations are accessed by Right-Click on the canvas of the Histogram Operations that are common to all views are detailed in the section Common Operations on Plot Views. Histogram-specific operations and properties are discussed below.

Selection Mode: The Histogram supports only the Selection mode. Left-Click and dragging the mouse over the Histogram draws a selection box and all bars that intersect the selection box are selected and lassoed. Clicking on a bar also selects the elements in that bar. To select additional elements, Ctrl-Left-Click and drag the mouse over the desired region.

5.7.2 Histogram Properties

The Histogram can be viewed with different channels, user-defined binning, different colors, and titles and descriptions from the Histogram Properties Dialog. See Figure 5.27.

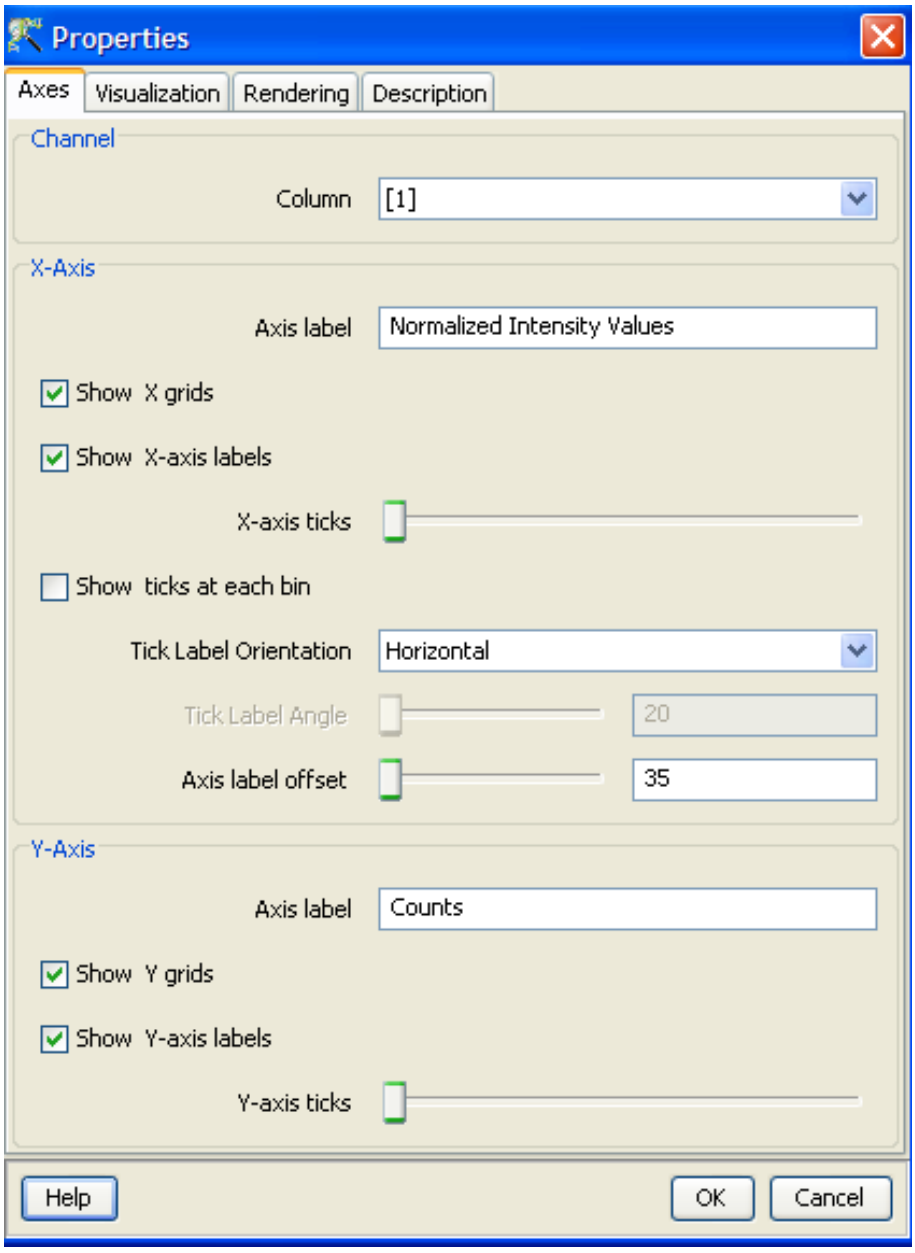


Figure 5.27: Histogram Properties

The Histogram Properties Dialog is accessible by right-clicking on the histogram and choosing **Properties** from the menu. The histogram view can be customized and configured from the histogram properties.

Axis: The histogram channel can be changed from the Properties menu. Any column in the dataset can be selected here.

The grids, axes labels, and axis ticks of the plots can be configured and modified from the Properties dialog. To modify these, Right-Click the view, select **Properties** and click on the **Axis** tab.

The plot can be drawn with or without the grid lines by clicking on the 'Show grids' option.

The ticks and axis labels are automatically computed and shown on the plot. You can show or remove the axis labels by clicking on the Show Axis Labels check box. Further, the orientation of the tick labels for the X-Axis can be changed from the default horizontal position to a slanted position or vertical position by using the drop down option and by moving the slider for the desired angle.

The number of ticks on the axis are automatically computed to show equal intervals between the minimum and maximum and displayed. You can increase the number of ticks displayed on the plot by moving the Axis Ticks slider. For Numerical data columns, you can double the number of ticks shown by moving the slider to the maximum. For categorical columns, if the number of categories are less than ten, all the categories are shown and moving the slider does not increase the number of ticks.

Visualization: Color By: You can specify a Color By column for the histogram. The Color By should be a categorical column in the active dataset. This will color each bar of the histogram with different color bars for the frequency of each category in the particular bin.

Explicit Binning: The Histogram is launched with a default set of equal interval bins for the chosen column. This default is computed by dividing the interquartile range of the column values into three bins and expanding these equal interval bins for the whole range of data in the chosen column. The Histogram view is dependent upon binning and the default number of bins may not be appropriate for the data. The data can be explicitly re-binned by checking the Use Explicit Binning check box and specifying the minimum value, the maximum value and the number of bins using the sliders. The maximum - minimum values and the number of bins can also be specified in the text box next to the sliders. Please note that if you type values into the text box, you will have to hit Enter for the values to be accepted.

Bar Width: the bar width of the histogram can be increased or decreased by moving the slider. The default is set to 0.9 times the area allocated to each histogram bar. This can be reduced if desired.

Channel chooser: The Channel Chooser on the histogram view can be disabled by unchecking the check box. This will afford a larger area to view the histogram.

Rendering: This tab provides the interface to customize and configure the fonts, the colors and the offsets of the plot.

Fonts: All fonts on the plot can be formatted and configured. To change the font in the view, Right-Click on the view and open the Properties dialog. Click on the **Rendering** tab of the **Properties** dialog. To change a **Font**, click on the appropriate drop-down box and choose the required font. To customize the font, click on the customize button. This will pop-up a dialog where you can set the font size and choose the font type as bold or italic.

Special Colors: All the colors that occur in the plot can be modified and configured. The plot Background color, the Axis color, the Grid color, the Selection color, as well as plot specific colors can be set. To change the default colors in the view, Right-Click on the view and open the Properties dialog. Click on the Rendering tab of the Properties dialog. To change a color, click on the appropriate arrow. This will pop-up a **Color Chooser**. Select the desired color and click **OK**. This will change the corresponding color in the View.

Offsets: The bottom, top, left, and right offsets of a plot can be modified and configured. These offsets can be changed if the axis labels or axis titles are not completely visible in the plot, or if only the graph portion of a plot is required. To change the offsets,

Right-Click on the view and open the Properties dialog. Click on the Rendering tab. To change plot offsets, move the corresponding slider, or enter an appropriate value in the text box provided. This will change the particular offset in the plot.

There is also a provision to set the number of samples that can be seen in the view, if the histogram tool bar was launched with the 'Multiple' option.

Description: The title for the view and description or annotation for the view can be configured and modified from the **Description** tab in the **Properties** dialog. Right-Click the view and select **Properties** to open the dialog. Click the **Description** tab to display the the current **Title** and **Description** of the view. Both **Title** and **Description** can be modified by changing the text in the corresponding text boxes. If you modify the default title, the new title is displayed in the Legend and used when the view is exported as **Report**. The change does not reflect in the title tab of the view in the **GeneSpring** desktop. The **Description** (default and modified) appears in the **Legend** window on the right. By default, if the view is derived from running an algorithm, the description will contain the algorithm and the parameters used.

5.8 The Bar Chart

The Bar Chart is launched on the default interpretation from the following script:
`script.view.BarChart().show()`

Launch the Script Editor from the main menu via Tools → Script Editor and copy-paste this script into the upper pane of the editor. Click the **Run the script** icon. By default, the Bar Chart is launched with all numerical columns in the active dataset. The Bar Chart provides a view of the range and distribution of values in the selected column. The Bar Chart is a tabular view and thus all operations that are possible on a table are possible here. The Bar Chart can be customized and configured from the **Properties** dialog accessed from the Right-Click menu on the canvas of the Chart. See Figure 5.28.

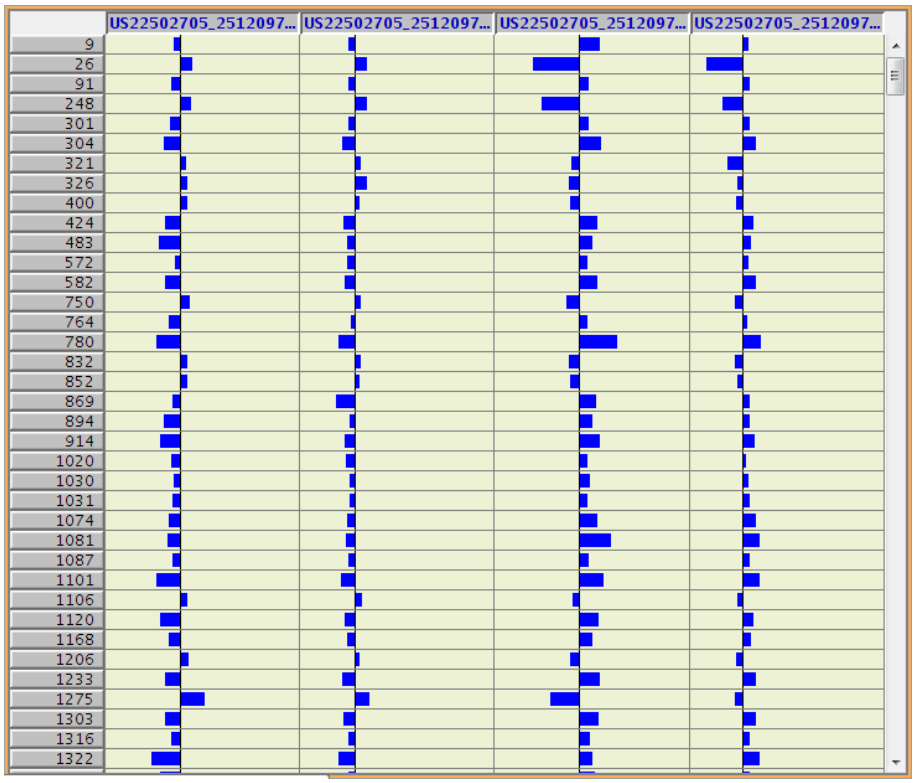


Figure 5.28: Bar Chart

Note that the Bar Chart will show only the numerical columns in the current dataset.

5.8.1 Bar Chart Operations

The Operations on the Bar Chart is accessible from the menu on Right-Click on the canvas of the Bar Chart. Operations that are common to all views are detailed in the section Common Operations on Table Views above. In addition, some of operations and the bar chart properties are explained below:

Sort: The Bar Chart can be used to view the sorted order of data with respect to a chosen column as bars. Sort is performed by clicking on the column header. Mouse clicks on the column header of the bar chart will cycle though an ascending values sort, a descending values sort and a reset sort. The column header of the sorted column will also be marked with the appropriate icon.

Thus to sort a column in the ascending order, click on the column header. This will sort all rows of the bar chart based on the values in the chosen column. Also an icon on the column header will denote that this is the sorted column. To sort in the descending order, click again on the same column header. This will sort all the rows of the bar chart based on the decreasing values in this column. To reset the sort, click again on the same column. This will reset the sort and the sort icon will disappear from the column header.

Selection: The bar chart can be used to select rows, columns, or any contiguous part of the dataset. The selected elements can be used to create a subset dataset by a Left-Click on the 'Create dataset from Selection' icon.

Row Selection: Rows are selected by left-clicking on the row headers and dragging along the rows. Ctrl-Left-Click selects subsequent items and Shift-Left-Click selects a consecutive set of items. The selected rows will be shown in the lasso window and will be highlighted in all other views.

Column Selection: Columns can be selected by left-clicking in the column of interest. Ctrl-Left-Click selects subsequent columns and Shift-Left-Click selects consecutive set of columns. The current column selection on the bar chart usually determines the default set of selected columns used when launching any new view, executing commands or running algorithm. The selected columns will be lassoed in all relevant views and will be show selected in the lasso view.

5.8.2 Bar Chart Properties

The Bar Chart Properties Dialog is accessible by Right-Click on the bar chart and choosing **Properties** from the menu. The bar chart view can be customized and configured from the bar chart properties.

Rendering: The rendering tab of the bar chart dialog allows you to configure and customize the fonts and colors that appear in the bar chart view.

Special Colors: All the colors in the Table can be modified and configured. You can change the Selection color, the Double Selection color, Missing Value cell color and the Background color in the table view. To change the default colors in the view, Right-Click on the view and open the Properties dialog. Click on the Rendering tab of the properties dialog. To change a color, click on the appropriate color bar. This will pop-up a Color Chooser. Select the desired color and click OK. This will change the corresponding color in the Table.

Fonts: Fonts that occur in the table can be formatted and configured. You can set the fonts for Cell text, row Header and Column Header. To change the font in the view, Right-Click on the view and open the Properties dialog. Click on the Rendering tab of the Properties dialog. To change a Font, click on the appropriate drop-down box and choose the required font. To customize the font, click on the customize button. This will pop-up a dialog where you can set the font size and choose the font type as bold or italic.

Visualization: The display precision of decimal values in columns, the row height, the missing value text, and the facility to enable and disable sort are configured and customized by options in this tab.

The visualization of the display precision of the numeric data in the table, the table cell size and the text for missing value can be configured. To change these, Right-Click on the table view and open the Properties dialog. Click on the visualization tab. This will open the Visualization panel.

To change the numeric precision. Click on the drop-down box and choose the desired precision. For decimal data columns, you can choose between full precision and one to four decimal places, or representation in scientific notation. By default, full precision is displayed.

You can set the row height of the table, by entering a integer value in the text box and pressing Enter. This will change the row height in the table. By default the row height is set to 16.

You can enter any a text to show missing values. All missing values in the table will be represented by the entered value and missing values can be easily identified. By default all the missing value text is set to an empty string.

You can also enable and disable sorting on any column of the table by checking or unchecking the check box provided. By default, sort is enabled in the table. To sort the table on any column, click on the column header. This will sort the all rows of the table based on the values in the sort column. This will also mark the sorted column with an icon to denote the sorted column. The first click on the column header will sort the column in the ascending order, the second click on the column header will sort the column in the descending order, and clicking the sorted column the third time will reset the sort.

Columns: The order of the columns in the bar chart can be changed by changing the order in the Columns tab in the Properties Dialog.

The columns for visualization and the order in which the columns are visualized can be chosen and configured from the List Selector in the **Column** tab of the **Properties** dialog. Right-Click the view and select **Properties** to open the dialog. Click the **Columns** tab. The List Selector panel shows the **Available items** on the left and the **Selected items** on the right. The items under **Selected items** are the columns that are displayed in the view in the same order in which they appear here.

To move columns from the **Available list** box to the **Selected list** box, select the required items in the **Available items** box and click the **Move right** button between the list boxes. This adds the selected columns from the **Available items** box to the items already in the *Selected items* box. To move columns from the **Selected items** box to the **Available items** box, select the required items in the **Selected items** box and click the **Move left** button. This moves the selected columns from the **Selected items** box to the *Available items* box in the exact position or order in which the column appears in the experiment.

You can also change the order of the columns in which they appear in the view by selecting items in the **Selected items** box and using the **Move Up** and **Move Down** buttons to arrange the columns in the desired order. If multiple items are selected, the first click consolidates the selected items (bring all the selected items together) with the first item in the specified direction. Subsequent clicks on the **Move Up** and **Move Down** buttons will move the selected items as a block in the specified direction, one step at a time until it reaches the end of the list. If only one item or contiguous items are selected in the *Selected items* box, then these will be moved in the specified direction, one step at a time until the end of the list is reached. To reset the order of the columns to the original order in which they currently appear in the plot view, click on the **Reset order** button located between the **Move Up** and **Move Down** buttons.

To select items, Left-Click on the required item. To select multiple items in any of the list boxes, Left-Click and Shift-Left-Click will highlight all contiguous items, and Ctrl-Left-Click will add the clicked-on item to the previously selected elements.

The lower portion of the Columns tab provides a utility to select items in the **List Selector** based on a search by column name or mark. You can either match by **Name** or **Column Mark** whichever is appropriate. By default, the Match by *Name* option is active.

- To match by **Name**, select **Name** from the **Match by item** drop-down list, enter a string in the **Name** text box and hit Enter. **GeneSpring** performs a substring match with the columns listed in the **Available items** and the **Selected items** boxes and selects any matching entries.

- To match by **Mark**, select **Mark** from the **Match by item** drop-down list. The column marks (for example, Affymetrix ProbeSet Id, raw signal, etc.) available in the experiment are shown in the **Mark** drop-down list. Choose a **Mark**, and **GeneSpring** selects the corresponding columns of the experiment.

Description: The title for the view and description or annotation for the view can be configured and modified from the **Description** tab in the **Properties** dialog. Right-Click the view and select **Properties** to open the dialog. Click the **Description** tab to display the the current **Title** and **Description** of the view. Both **Title** and **Description** can be modified by changing the text in the corresponding text boxes. If you modify the default title, the new title is displayed in the Legend and used when the view is exported as **Report**. The change does not reflect in the title tab of the view in the **GeneSpring** desktop. The **Description** (default and modified) appears in the **Legend** window on the right. By default, if the view is derived from running an algorithm, the description will contain the algorithm and the parameters used.

5.9 The Matrix Plot View

The Matrix Plot is launched from the View menu on the main menu bar with the active interpretation and the active entity list. Alternately, Left-Click on the tool bar icon will bring up the Matrix plot. The Matrix Plot shows a matrix of pairwise 2D scatter plots for conditions in the active interpretation. The X-Axis and Y-Axis of each scatter plot corresponding to the conditions in the active interpretation are shown in the corresponding row and column of the matrix plot. See Figure 5.29.

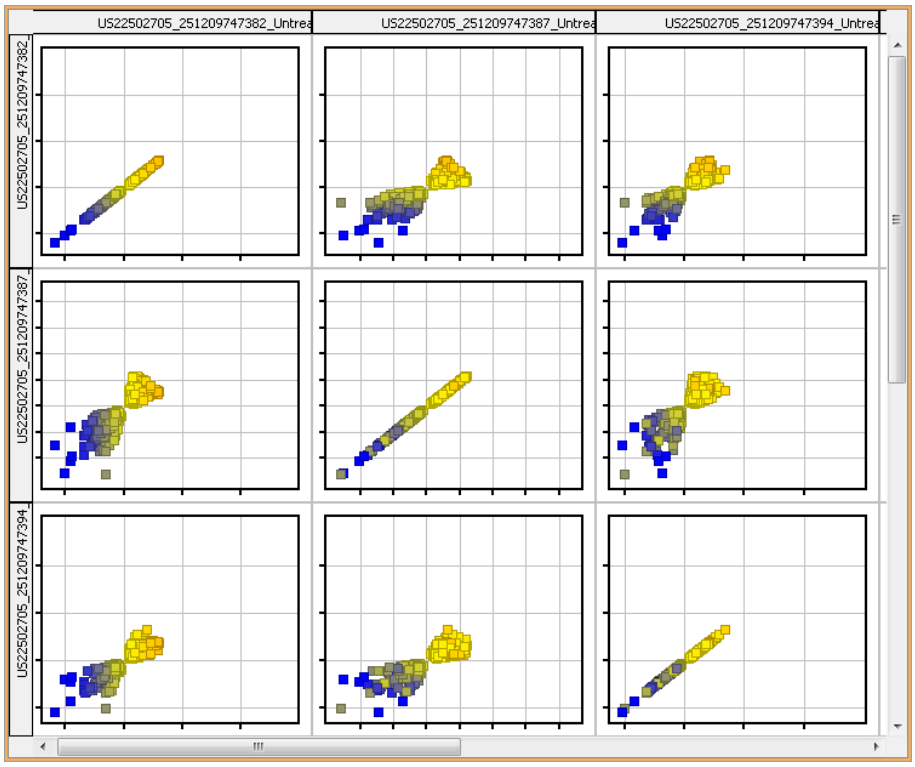


Figure 5.29: Matrix Plot

If the active interpretation is the default *All Samples* interpretation, the matrix plot shows the normalized expression values of each sample against the other. If an averaged interpretation is the active interpretation, then the matrix plot will show the averaged normalized signal values of the samples in each condition against the other. The points in the matrix plot correspond to the entities in the active entity list. The legend window displays the interpretation on which the matrix plot was launched.

Clicking on another entity list in the experiment will make that entity list active and the matrix plot will dynamically display the current active entity list. Clicking on an entity list in another experiment will translate the entities in that entity list to the current experiment and display those entities in the matrix plot.

The main purpose of the matrix plot is to get an overview of the correlation between conditions in the dataset, and detect conditions that separate the data into different groups.

By default, a maximum of 10 conditions can be shown in the matrix plot. If more than 10 conditions are present in the active interpretation, only ten conditions are projected into the matrix plot and other columns are ignored with a warning message. The matrix plot is interactive and can be lassoed. Elements of the matrix plot can be configured and altered from the properties menu described below.

5.9.1 Matrix Plot Operations

The Matrix Plot operations are accessed from the main menu bar when the plot is the active windows. These operations are also available by right-clicking on the canvas of the Matrix Plot. Operations that are common to all views are detailed in the section Common Operations on Plot Views. Matrix Plot specific operations and properties are discussed below.

Selection Mode: The Matrix Plot supports only the Selection mode. Left-Click and dragging the mouse over the Matrix Plot draws a selection box and all points that intersect the selection box are selected and lassoed. To select additional elements, Ctrl-Left-Click and drag the mouse over the desired region. Ctrl-Left-Click toggles selection. This selected points will be unselected and unselected points will be added to the selection and lassoed.

5.9.2 Matrix Plot Properties

The matrix plot can be customized and configured from the properties dialog accessible from the Right-Click menu on the canvas of the Matrix plot. The important properties of the scatter plot are all available for the Matrix plot. These are available in the Axis tab, the Visualization tab, the Rendering tab, the Columns tab and the description tab of the properties dialog and are detailed below. See Figure 5.30.

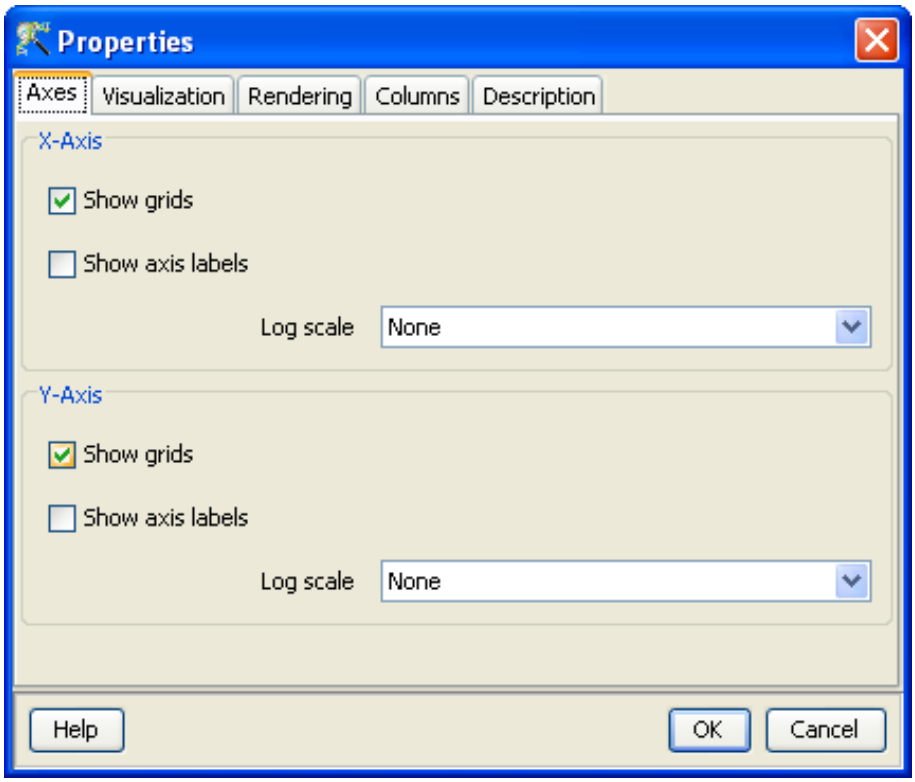


Figure 5.30: Matrix Plot Properties

Axis: The Axes on the Matrix Plot can be toggled to show or hide the grids, or show and hide the axis labels.

Visualization: The scatter plots can be configured to Color By any column of the active dataset, Shape By any categorical column of the dataset, and Size by any column of the dataset.

Rendering: The fonts on the Matrix Plot, the colors that occur on the Matrix Plot, the Offsets, the Page size of the view and the quality of the Matrix Plot can be be altered from the Rendering tab of the Properties dialog.

Fonts: All fonts on the plot can be formatted and configured. To change the font in the view, Right-Click on the view and open the Properties dialog. Click on the **Rendering** tab of the **Properties** dialog. To change a **Font**, click on the appropriate drop-down box and choose the required font. To customize the font, click on the customize button. This will pop-up a dialog where you can set the font size and choose the font type as bold or italic.

Special Colors: All the colors that occur in the plot can be modified and configured. The plot Background color, the Axis color, the Grid color, the Selection color, as well as plot specific colors can be set. To change the default colors in the view, Right-Click on the view and open the Properties dialog. Click on the Rendering tab of the Properties dialog. To change a color, click on the appropriate arrow. This will pop-up a **Color Chooser**. Select the desired color and click **OK**. This will change the corresponding color in the View.

Offsets: The bottom, top, left, and right offsets of a plot can be modified and configured. These offsets can be changed if the axis labels or axis titles are not completely visible in the plot, or if only the graph portion of a plot is required. To change the offsets, Right-Click on the view and open the Properties dialog. Click on the Rendering tab. To change plot offsets, move the corresponding slider, or enter an appropriate value in the text box provided. This will change the particular offset in the plot.

Page: The visualization page of the Matrix Plot can be configured to view a specific number of scatter plots in the Matrix Plot. If there are more scatter plots in the Matrix plot than in the page, scroll bars appear and you can scroll to the other plot of the Matrix Plot.

Plot Quality: The quality of the plot can be enhanced to be anti-aliased. This will produce better points and will produce better prints of the Matrix Plot.

Columns: The Columns for the Matrix Plot can be chosen from the Columns tab of the Properties dialog.

The columns for visualization and the order in which the columns are visualized can be chosen and configured from the List Selector in the **Column** tab of the **Properties** dialog. Right-Click the view and select **Properties** to open the dialog. Click the **Columns** tab. The List Selector panel shows the **Available items** on the left and the **Selected items** on the right. The items under **Selected items** are the columns that are displayed in the view in the same order in which they appear here.

To move columns from the **Available list** box to the **Selected list** box, select the required items in the **Available items** box and click the **Move right** button between the list boxes. This adds the selected columns from the **Available items** box to the items already in the *Selected items* box. To move columns from the **Selected items** box to the **Available items** box, select the required items in the **Selected items** box and click the **Move left** button. This moves the selected columns from the **Selected items** box to the *Available items* box in the exact position or order in which the column appears in the experiment.

You can also change the order of the columns in which they appear in the view by selecting items in the **Selected items** box and using the **Move Up** and **Move Down** buttons

to arrange the columns in the desired order. If multiple items are selected, the first click consolidates the selected items (bring all the selected items together) with the first item in the specified direction. Subsequent clicks on the **Move Up** and **Move Down** buttons will move the selected items as a block in the specified direction, one step at a time until it reaches the end of the list. If only one item or contiguous items are selected in the *Selected items* box, then these will be moved in the specified direction, one step at a time until the end of the list is reached. To reset the order of the columns to the original order in which they currently appear in the plot view, click on the **Reset order** button located between the **Move Up** and **Move Down** buttons.

To select items, Left-Click on the required item. To select multiple items in any of the list boxes, Left-Click and Shift-Left-Click will highlight all contiguous items, and Ctrl-Left-Click will add the clicked-on item to the previously selected elements.

The lower portion of the Columns tab provides a utility to select items in the **List Selector** based on a search by column name or mark. You can either match by **Name** or **Column Mark** whichever is appropriate. By default, the Match by *Name* option is active.

- To match by **Name**, select **Name** from the **Match by item** drop-down list, enter a string in the **Name** text box and hit Enter. **GeneSpring** performs a substring match with the columns listed in the **Available items** and the **Selected items** boxes and selects any matching entries.
- To match by **Mark**, select **Mark** from the **Match by item** drop-down list. The column marks (for example, Affymetrix ProbeSet Id, raw signal, etc.) available in the experiment are shown in the **Mark** drop-down list. Choose a **Mark**, and **GeneSpring** selects the corresponding columns of the experiment.

Description: The title for the view and description or annotation for the view can be configured and modified from the **Description** tab in the **Properties** dialog. Right-Click the view and select **Properties** to open the dialog. Click the **Description** tab to display the the current **Title** and **Description** of the view. Both **Title** and **Description** can be modified by changing the text in the corresponding text boxes. If you modify the default title, the new title is displayed in the Legend and used when the view is exported as **Report**. The change does not reflect in the title tab of the view in the **GeneSpring** desktop. The **Description** (default and modified) appears in the **Legend** window on the right. By default, if the view is derived from running an algorithm, the description will contain the algorithm and the parameters used.

5.10 Summary Statistics View

The Summary Statistics View is launched from view menu on the main menu bar with the active interpretation and the active entity list in the experiment. Alternately, Left-Click on the tool bar icon to display the summary statistics. This view shows the summary statistics of the conditions in the active interpretation with respect to the active entity list. Thus, each column of the summary statistics shows the mean, standard deviation, median, percentiles and outliers of the conditions in the active interpretation with active entity list. In **GeneSpring**, points that lie outside the quartiles i.e., 25th percentile value- 1.5 * (interquartile range) and 75th percentile value + 1.5 * (interquartile range) are considered outliers. The interquartile range is between 75th percentile and 25th percentile.

Let's say you have 100 values in your dataset. If you sort them in ascending order, the 25th value is

4, and the 75th value is 7. Therefore, the interquartile range is 7 - 4 = 3. 1.5 * interquartile range = 1.5 * 3 = 4.5. Therefore, all values in the dataset which are less than or equal to $4 - 4.5 = -0.5$ and all values which are more than or equal to $7 + 4.5 = 11.5$ are considered as outliers.

If the active interpretation is the default *All Samples* interpretation, the table shows the summary statistics of each sample with respect to the active entity list. If an averaged interpretation is the active interpretation, the table shows the summary statistics of the conditions in the averaged interpretation with respect to the active entity list. The legend window displays the interpretation on which the summary statistics was launched.

Clicking on another entity list in the experiment will make that entity list active and the summary statistics table will dynamically display the current active entity list. Clicking on an entity list in another experiment will translate the entities in that entity list to the current experiment and display those entities in the summary statistics table.

This Summary Statistics View is a tabular view and thus all operations that are possible on a table are possible here. The summary statistics table can be customized and configured from the **Properties** dialog accessed from the Right-Click menu on the canvas of the Chart. See Figure 5.31.

This view presents descriptive statistics information on the active interpretation, and is useful to compare the distributions of different conditions in the interpretation.

	US22502705_...	US22502705_...	US22502705_...
No. of Observations	630	630	630
No. of Missing Values	0	0	0
Minimum	-5.945851	-5.195895	-5.3121476
Maximum	2.9420753	2.3684192	3.3038101
Mean	-0.3261476	-0.30182755	-0.21017464
Trimmed Mean	-0.31360033	-0.28923494	-0.20104119
Median	-0.5929078	-0.5630168	-0.5549085
Std. Deviation	1.2126718	1.1539652	1.2245678
Trimmed Std. Deviation	1.0999714	1.0666007	1.1378244
No. Of Outliers	5	5	4
Percentile 1.0	-3.5079732	-3.4143982	-3.1954958
Percentile 5.0	-2.2516694	-2.184887	-2.1405253
Percentile 10.0	-1.6517274	-1.6230108	-1.6094166
Percentile 25.0	-1.0678427	-1.0102725	-0.94769526
Percentile 50.0	-0.5929078	-0.5630168	-0.5549085
Percentile 75.0	0.63223314	0.6369066	0.7316594
Percentile 90.0	1.0590799	1.0828253	1.4146017
Percentile 95.0	1.5467687	1.4110379	1.7696872
Percentile 99.0	2.5480194	2.221198	2.2974193

Figure 5.31: Summary Statistics View

5.10.1 Summary Statistics Operations

The Operations on the Summary Statistics View are accessible from the menu on Right-Click on the canvas of the Summary Statistics View. Operations that are common to all views are detailed in the section Common Operations on Table Views above. In addition, some of the Summary Statistics View specific operations and the bar chart properties are explained below:

Column Selection: The Summary Statistics View can be used to select conditions or columns. The selected columns are lassoed in all the appropriate views.

Columns can be selected by left-clicking in the column of interest. Ctrl-Left-Click selects subsequent columns and Shift-Left-Click consecutive set of columns. The current column selection on the bar chart usually determines the default set of selected columns used when launching any new view, executing commands or running algorithms. The selected columns will be lassoed in all relevant views and will be shown selected in the lasso view.

Export As Text: The *Export* → *Text* option saves the tabular output to a tab-delimited file that can be opened in **GeneSpring**.

5.10.2 Summary Statistics Properties

The Summary Statistics View Properties Dialog is accessible by right-clicking on the Summary Statistics View and choosing **Properties** from the menu. The Summary Statistics View can be customized and configured from the Summary Statistics View properties. See Figure 5.32.

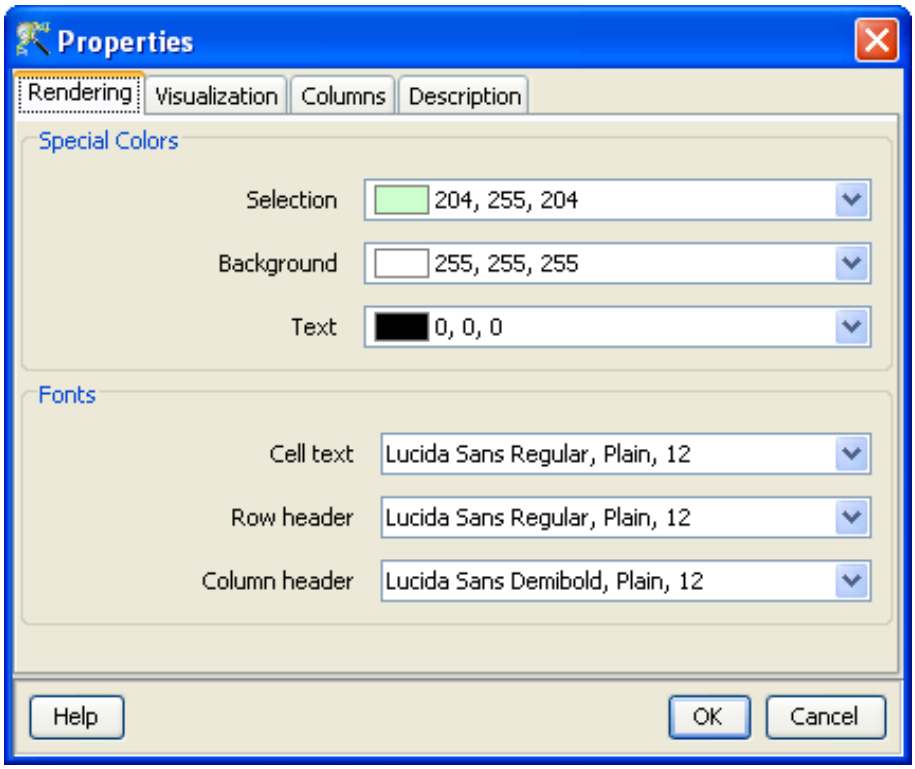


Figure 5.32: Summary Statistics Properties

Rendering: The rendering tab of the Summary Statistics View dialog allows you to configure and customize the fonts and colors that appear in the Summary Statistics View view.

Special Colors: All the colors in the Table can be modified and configured. You can change the Selection color, the Double Selection color, Missing Value cell color and the Background color in the table view. To change the default colors in the view, Right-Click on the view and open the Properties dialog. Click on the Rendering tab of the properties dialog. To change a color, click on the appropriate color bar. This will pop-up a Color Chooser. Select the desired color and click OK. This will change the corresponding color in the Table.

Fonts: Fonts that occur in the table can be formatted and configured. You can set the fonts for Cell text, row Header and Column Header. To change the font in the view, Right-Click on the view and open the Properties dialog. Click on the Rendering tab of the Properties dialog. To change a Font, click on the appropriate drop-down box and choose the required font. To customize the font, click on the customize button. This will pop-up a dialog where you can set the font size and choose the font type as bold or italic.

Visualization: The display precision of decimal values in columns, the row height and the missing value text, and the facility to enable and disable sort are configured and customized by options in this tab.

The visualization of the display precision of the numeric data in the table, the table cell size and the text for missing value can be configured. To change these, Right-Click on the table view and open the Properties dialog. Click on the visualization tab. This will open the Visualization panel.

To change the numeric precision. Click on the drop-down box and choose the desired precision. For decimal data columns, you can choose between full precision and one to four decimal places, or representation in scientific notation. By default, full precision is displayed.

You can set the row height of the table, by entering a integer value in the text box and pressing Enter. This will change the row height in the table. By default the row height is set to 16.

You can enter any a text to show missing values. All missing values in the table will be represented by the entered value and missing values can be easily identified. By default all the missing value text is set to an empty string.

You can also enable and disable sorting on any column of the table by checking or unchecking the check box provided. By default, sort is enabled in the table. To sort the table on any column, click on the column header. This will sort the all rows of the table based on the values in the sort column. This will also mark the sorted column with an icon to denote the sorted column. The first click on the column header will sort the column in the ascending order, the second click on the column header will sort the column in the descending order, and clicking the sorted column the third time will reset the sort.

Columns: The order of the columns in the Summary Statistics View can be changed by changing the order in the Columns tab in the Properties Dialog.

The columns for visualization and the order in which the columns are visualized can be chosen and configured from the List Selector in the **Column** tab of the **Properties** dialog. Right-Click the view and select **Properties** to open the dialog. Click the **Columns** tab. The List Selector panel shows the **Available items** on the left and the **Selected items** on the right. The items under **Selected items** are the columns that are displayed in the view in the same order in which they appear here.

To move columns from the **Available list** box to the **Selected list** box, select the required items in the **Available items** box and click the **Move right** button between the list boxes. This adds the selected columns from the **Available items** box to the items already in the *Selected items* box. To move columns from the **Selected items** box to the **Available items** box, select the required items in the **Selected items** box and click the **Move left** button. This moves the selected columns from the **Selected items** box to the *Available items* box in the exact position or order in which the column appears in the experiment.

You can also change the order of the columns in which they appear in the view by selecting items in the **Selected items** box and using the **Move Up** and **Move Down** buttons to arrange the columns in the desired order. If multiple items are selected, the first click consolidates the selected items (bring all the selected items together) with the first item in the specified direction. Subsequent clicks on the **Move Up** and **Move Down** buttons will move the selected items as a block in the specified direction, one step at a time until it reaches the end of the list. If only one item or contiguous items are selected in the *Selected items* box, then these will be moved in the specified direction, one step at a time until the end of the list is reached. To reset the order of the columns to the original order in which they currently appear in the plot view, click on the **Reset order** button located between the **Move Up** and **Move Down** buttons.

To select items, Left-Click on the required item. To select multiple items in any of the list boxes, Left-Click and Shift-Left-Click will highlight all contiguous items, and Ctrl-Left-Click will add the clicked-on item to the previously selected elements.

The lower portion of the Columns tab provides a utility to select items in the **List Selector** based on a search by column name or mark. You can either match by **Name** or **Column Mark** whichever is appropriate. By default, the Match by *Name* option is active.

- To match by **Name**, select **Name** from the **Match by item** drop-down list, enter a string in the **Name** text box and hit Enter. **GeneSpring** performs a substring match

with the columns listed in the **Available items** and the **Selected items** boxes and selects any matching entries.

- To match by **Mark**, select **Mark** from the **Match by item** drop-down list. The column marks (for example, Affymetrix ProbeSet Id, raw signal, etc.) available in the experiment are shown in the **Mark** drop-down list. Choose a **Mark**, and **GeneSpring** selects the corresponding columns of the experiment.

Description: The title for the view and description or annotation for the view can be configured and modified from the **Description** tab in the **Properties** dialog. Right-Click the view and select **Properties** to open the dialog. Click the **Description** tab to display the the current **Title** and **Description** of the view. Both **Title** and **Description** can be modified by changing the text in the corresponding text boxes. If you modify the default title, the new title is displayed in the Legend and used when the view is exported as **Report**. The change does not reflect in the title tab of the view in the **GeneSpring** desktop. The **Description** (default and modified) appears in the **Legend** window on the right. By default, if the view is derived from running an algorithm, the description will contain the algorithm and the parameters used.

5.11 The Box Whisker Plot

The Box Whisker Plot is launched from View menu on the main menu bar with the active interpretation and the active entity list in the experiment. Alternately, Left-Click on the tool bar icon to launch the boxwhisker plot.The Box Whisker Plot presents the distribution of the of the conditions in the active interpretation with respect to the active entity list in the experiment. The box whisker shows the median in the middle of the box, the 25th percentile and the 75th percentile, or the 1st and 3rd quartile. The whiskers are extensions of the box, snapped to the point within 1.5 times the interquartile. The points outside the whiskers are plotted as they are, but in a different color and could normally be considered the outliers. See Figure 5.33.

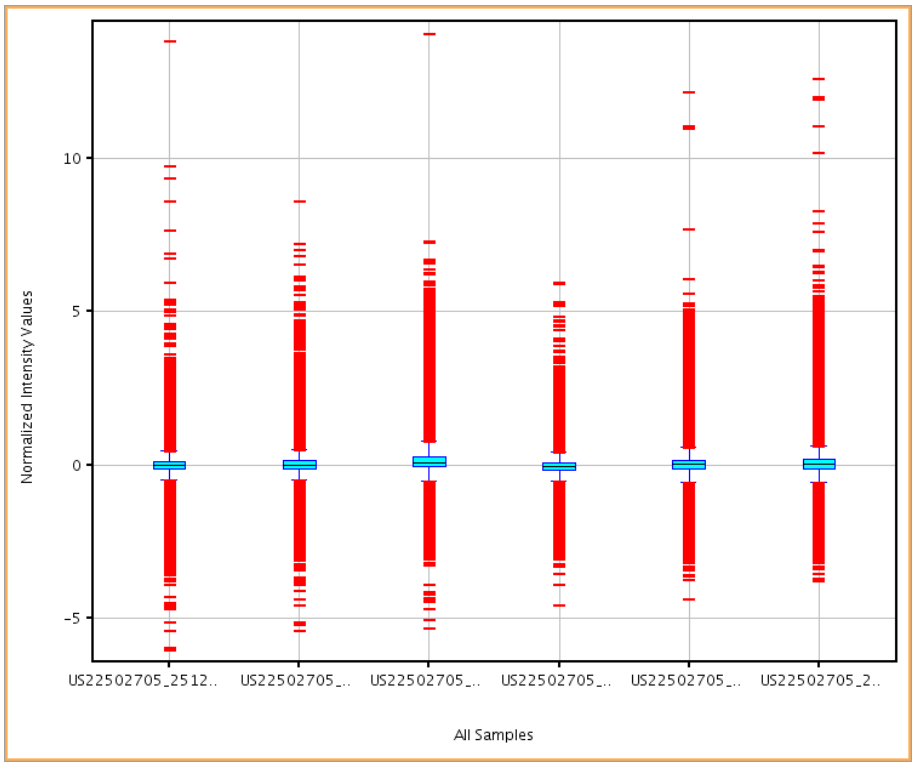


Figure 5.33: Box Whisker Plot

If the active interpretation is the default *All Samples* interpretation, the box whisker plot the distribution of each sample with respect to the active entity list. If an averaged interpretation is the active interpretation, the box whisker plot shows the distribution of the conditions in the averaged interpretation with respect to the active entity list. The legend window displays the interpretation on which the box whisker plot was launched.

Clicking on another entity list in the experiment will make that entity list active and the box whisker plot will dynamically display the current active entity list. Clicking on an entity list in another experiment will translate the entities in that entity list to the current experiment and display those entities in the box whisker plot.

The operations on the box whisker plot are similar to operations on all plots and are discussed below. The box whisker plot can be customized and configured from the **Properties** dialog.

5.11.1 Box Whisker Operations

The Box Whisker operations are accessed from the toolbar menu when the plot is the active window. These operations are also available by right-clicking on the canvas of the Box Whisker. Operations that are common to all views are detailed in the section Common Operations on Plot Views. Box Whisker specific operations and properties are discussed below.

Selection Mode: The Selection on the Box Whisker plot is confined to only one column of plot. This is so because the box whisker plot contains box whiskers for many columns and each of them contain all the rows in the active dataset. Thus selection has to be confined to only to one column in the plot. The Box Whisker only supports the selection mode. Thus, left-clicking and dragging the mouse over the box whisker plot confines the selection box to only one column. The points in this selection box are highlighted in the density plot of that particular column and are also lassoed highlighted in the density plot of all other columns. Left-clicking and dragging, and shift-left-clicking and dragging selects elements and Ctrl-Left-Click toggles selection like in any other plot and appends to the selected set of elements.

5.11.2 Box Whisker Properties

The Box Whisker Plot offers a wide variety of customization and configuration of the plot from the Properties dialog. These customizations appear in four different tabs on the Properties window, labeled Axis, Rendering, Columns, and Description. See Figure 5.34.

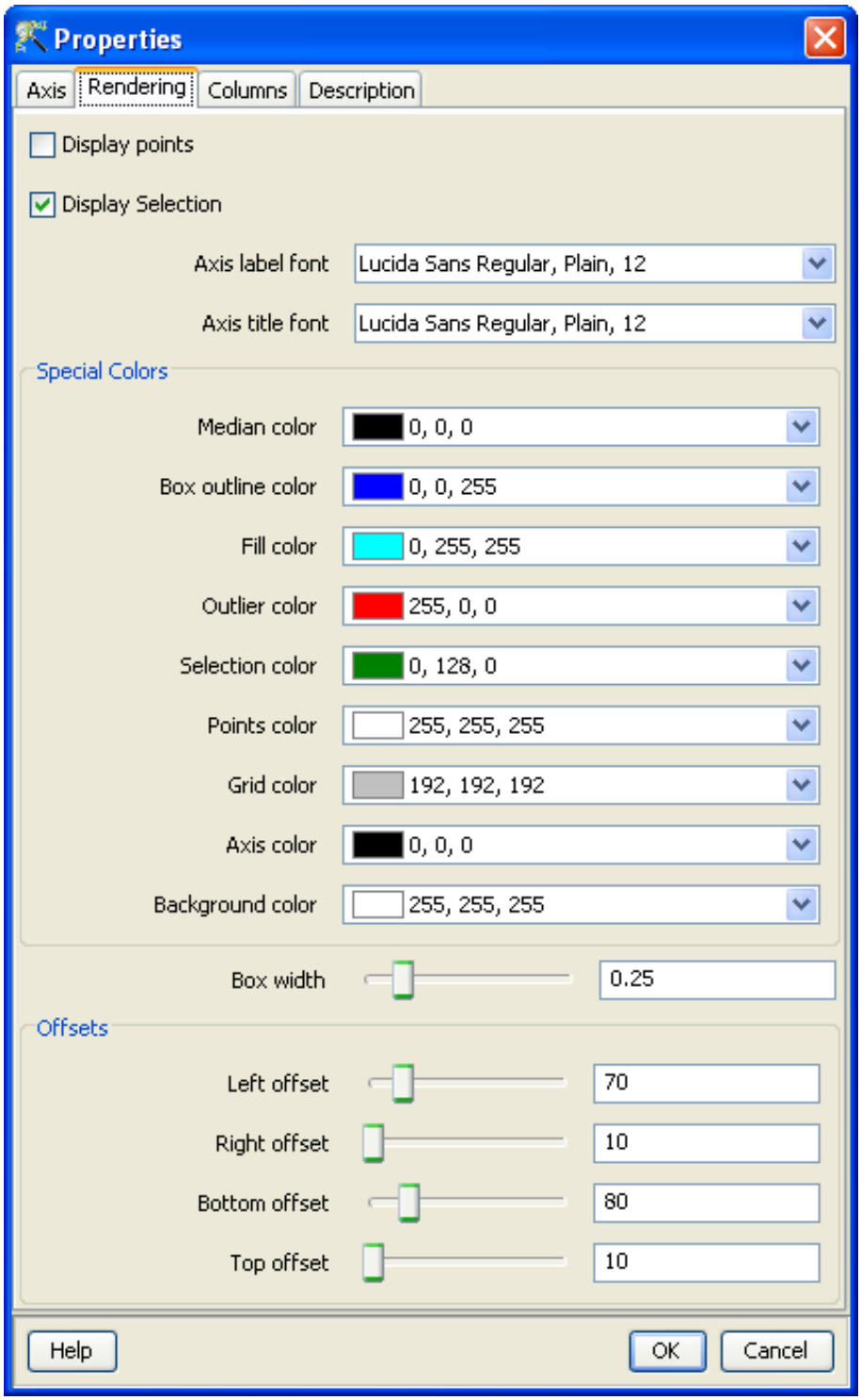


Figure 5.34: Box Whisker Properties

Axis: The grids, axes labels, and axis ticks of the plots can be configured and modified from the Properties dialog. To modify these, Right-Click the view, select **Properties** and click on the **Axis** tab.

The plot can be drawn with or without the grid lines by clicking on the 'Show grids' option.

The ticks and axis labels are automatically computed and shown on the plot. You can show or remove the axis labels by clicking on the Show Axis Labels check box. Further, the orientation of the tick labels for the X-Axis can be changed from the default horizontal position to a slanted position or vertical position by using the drop down option and by moving the slider for the desired angle.

The number of ticks on the axis are automatically computed to show equal intervals between the minimum and maximum and displayed. You can increase the number of ticks displayed on the plot by moving the Axis Ticks slider. For Numerical data columns, you can double the number of ticks shown by moving the slider to the maximum. For categorical columns, if the number of categories are less than ten, all the categories are shown and moving the slider does not increase the number of ticks.

Rendering: The Box Whisker Plot allows all aspects of the view to be configured including fonts, the colors, the offsets, etc.

Show Selection Image: The Show Selection Image, shows the density of points for each column of the box whisker plot. This is used for selection of points. For large datasets and for many columns this may take a lot of resources. You can choose to remove the density plot next to each box whisker by unchecking the check box provided.

Fonts: All fonts on the plot can be formatted and configured. To change the font in the view, Right-Click on the view and open the Properties dialog. Click on the **Rendering** tab of the **Properties** dialog. To change a **Font**, click on the appropriate drop-down box and choose the required font. To customize the font, click on the customize button. This will pop-up a dialog where you can set the font size and choose the font type as bold or italic.

Special Colors: All the colors on the box whisker can be configured and customized.

All the colors that occur in the plot can be modified and configured. The plot Background color, the Axis color, the Grid color, the Selection color, as well as plot specific colors can be set. To change the default colors in the view, Right-Click on the view and open the Properties dialog. Click on the Rendering tab of the Properties dialog. To change a color, click on the appropriate arrow. This will pop-up a **Color Chooser**. Select the desired color and click **OK**. This will change the corresponding color in the View.

Box Width: The box width of the box whisker plots can be changed by moving the slider provided. The default is set to 0.25 of the width provided to each column of the box whisker plot.

Offsets: The bottom, top, left, and right offsets of a plot can be modified and configured. These offsets can be changed if the axis labels or axis titles are not completely visible in the plot, or if only the graph portion of a plot is required. To change the offsets, Right-Click on the view and open the Properties dialog. Click on the Rendering tab. To change plot offsets, move the corresponding slider, or enter an appropriate value in the text box provided. This will change the particular offset in the plot.

Columns: The columns drawn in the Box Whisker Plot and the order of columns in the Box whisker Plot can be changed from the Columns tab in the Properties Dialog.

The columns for visualization and the order in which the columns are visualized can be chosen and configured from the List Selector in the **Column** tab of the **Properties** dialog. Right-Click the view and select **Properties** to open the dialog. Click the **Columns** tab. The List Selector panel shows the **Available items** on the left and the **Selected items** on the right. The items under **Selected items** are the columns that are displayed in the view in the same order in which they appear here.

To move columns from the **Available list** box to the **Selected list** box, select the required items in the **Available items** box and click the **Move right** button between the list boxes. This adds the selected columns from the **Available items** box to the items already in the *Selected items* box. To move columns from the **Selected items** box to the **Available items** box, select the required items in the **Selected items** box and click the **Move left** button. This moves the selected columns from the **Selected items** box to the *Available items* box in the exact position or order in which the column appears in the experiment.

You can also change the order of the columns in which they appear in the view by selecting items in the **Selected items** box and using the **Move Up** and **Move Down** buttons to arrange the columns in the desired order. If multiple items are selected, the first click consolidates the selected items (bring all the selected items together) with the first item in the specified direction. Subsequent clicks on the **Move Up** and **Move Down** buttons will move the selected items as a block in the specified direction, one step at a time until it reaches the end of the list. If only one item or contiguous items are selected in the *Selected items* box, then these will be moved in the specified direction, one step at a time until the end of the list is reached. To reset the order of the columns to the original order in which they currently appear in the plot view, click on the **Reset order** button located between the **Move Up** and **Move Down** buttons.

To select items, Left-Click on the required item. To select multiple items in any of the list boxes, Left-Click and Shift-Left-Click will highlight all contiguous items, and Ctrl-Left-Click will add the clicked-on item to the previously selected elements.

The lower portion of the Columns tab provides a utility to select items in the **List Selector** based on a search by column name or mark. You can either match by **Name** or **Column Mark** whichever is appropriate. By default, the Match by *Name* option is active.

- To match by **Name**, select **Name** from the **Match by item** drop-down list, enter a string in the **Name** text box and hit Enter. **GeneSpring** performs a substring match with the columns listed in the **Available items** and the **Selected items** boxes and selects any matching entries.
- To match by **Mark**, select **Mark** from the **Match by item** drop-down list. The column marks (for example, Affymetrix ProbeSet Id, raw signal, etc.) available in the experiment are shown in the **Mark** drop-down list. Choose a **Mark**, and **GeneSpring** selects the corresponding columns of the experiment.

Description: The title for the view and description or annotation for the view can be configured and modified from the **Description** tab in the **Properties** dialog. Right-Click the view and select **Properties** to open the dialog. Click the **Description** tab to display the the current **Title** and **Description** of the view. Both **Title** and **Description** can be modified by changing the text in the corresponding text boxes. If you modify the default title, the new title is displayed in the Legend and used when the view is exported as **Report**. The change does not reflect in the title tab of the view in the **GeneSpring** desktop. The **Description** (default and modified) appears in the **Legend** window on the right. By default, if the view is derived from running an algorithm, the description will contain the algorithm and the parameters used.

5.12 The Venn Diagram

The *Venn diagram* is a special view that is used for capturing commonalities between entity lists or pathway lists, even across experiments. In **GeneSpring**, the user can choose entity lists or

pathway lists from not only the active experiment but also from other experiments in a project. The experiments need to be open in the project for enabling selection of entity lists/pathway lists across them. For entity lists, this is enabled by performing translation of entities across experiments. For more details, kindly refer to the section on Translation.

The Venn diagram is launched from the *View* menu on the main menu bar or from the icon in the tool bar. Launching the Venn diagram from either place, will bring up a window to choose entity lists for this view. Choose multiple entity lists or pathway lists from one experiment using the Ctrl+left click operation and click on **OK**. This will launch the Venn diagram with the chosen entity lists or pathway lists as circles of the Venn diagram after performing translation across experiments, if required. Please note that when choosing entity lists across experiments, the Venn diagram would only show translated entities for lists from other experiments. For example, if the translated entity list originally contains hundred entities out of which only fifty entities get translated into the active experiment, then only these entities would show up in the intersecting and non intersecting regions of the view. The remaining fifty will not be shown as these cannot be viewed in the current experiment due to their absence from the technology of the experiment. This exception does not apply to pathway lists.

NOTE The functionality to create a Venn diagram on entity lists is not enabled in a MOA experiment. However, if you launch a Venn Diagram on Entity Lists from another single open experiment in the same project as the MOA experiment, you can drag-and-drop entity lists from the open MOA experiment to add to the open Venn diagram (see Drag and drop operations on Venn diagram).

A Venn diagram can be launched with a minimum of two and a maximum of four entity lists or pathway lists. See Figure 5.35. Also multiple Venn diagrams can be launched, which will appear in separate tabs in the central visualization panel. You can also create new entity lists or pathway lists from a selection of regions in the Venn diagram by using the 'Create entity list' or 'Create pathway list' icon present in the tool bar of the Venn diagram view. Using these icons also allows you to carry over list associated values from the input entity lists/pathway lists. This is described in detail in the next section.

Please note, that the 'Create entity list' icon is also present in the main toolbar and can also be used to create a new entity list from a selection in a Venn diagram. However, creating an entity list from the option in the main tool bar does not allow you to carry over any list associated values. This is described in detail in the next section.

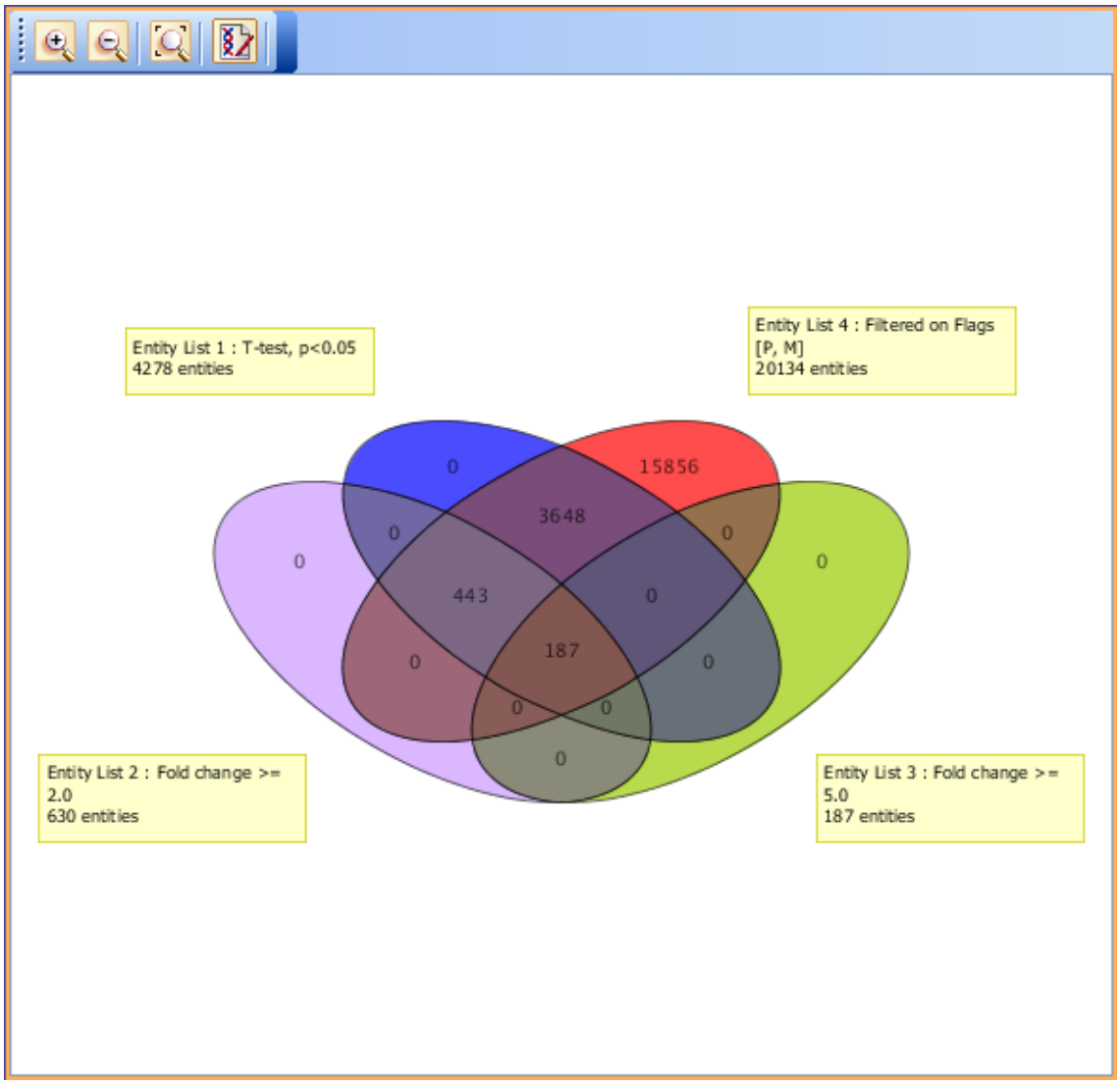


Figure 5.35: The Venn Diagram

5.12.1 Venn Diagram Operations

5.12.1.1 Drag and drop operations on Venn diagram

After launching the Venn diagram, it is possible to add or replace an entity list, provided translation is possible between the chosen entity lists. From the navigator, choose an entity list and drag it into the Venn diagram view. Drop outside the circles to add this entity list into the Venn diagram. Note that addition is possible only if the original Venn diagram has less than 4 entity lists. Drop inside the non-overlapping part of a circle in the Venn diagram to replace that entity list with the chosen one.

While dragging an entity list into the Venn diagram, the tool-tip indicates if addition/replacement of entity lists is possible or not.

5.12.1.2 Create Area-proportional representation of a Venn diagram

In **GeneSpring**, Venn diagrams can be visualized as area-proprtional diagrams using the area-proportional diagram icon. However in these diagrams, regions with no data are not displayed. Right-click $\longrightarrow$ Properties $\longrightarrow$ Rendering options can be altered to view percentages instead of absolute values. Entity lists can be added to the area-proptional digram view by dragging and dropping. A 2-way, 3-way Venn diagram, or a maximum of three entity lists can be utilized to visualize area-proportional diagrams.

5.12.1.3 Create new entity or pathway lists from overlapping and unique regions

From the Venn diagram/ Area-proportional diagram, select any region (overlapping or unique/non-overlappping part of the circles) with a left-click; click on the Create entity list icon to create a new entity list, or the Create pathway list icon to create a new pathway list. The ‘‘Create New Entity List” or ‘‘Create New Pathway List” workflow opens.

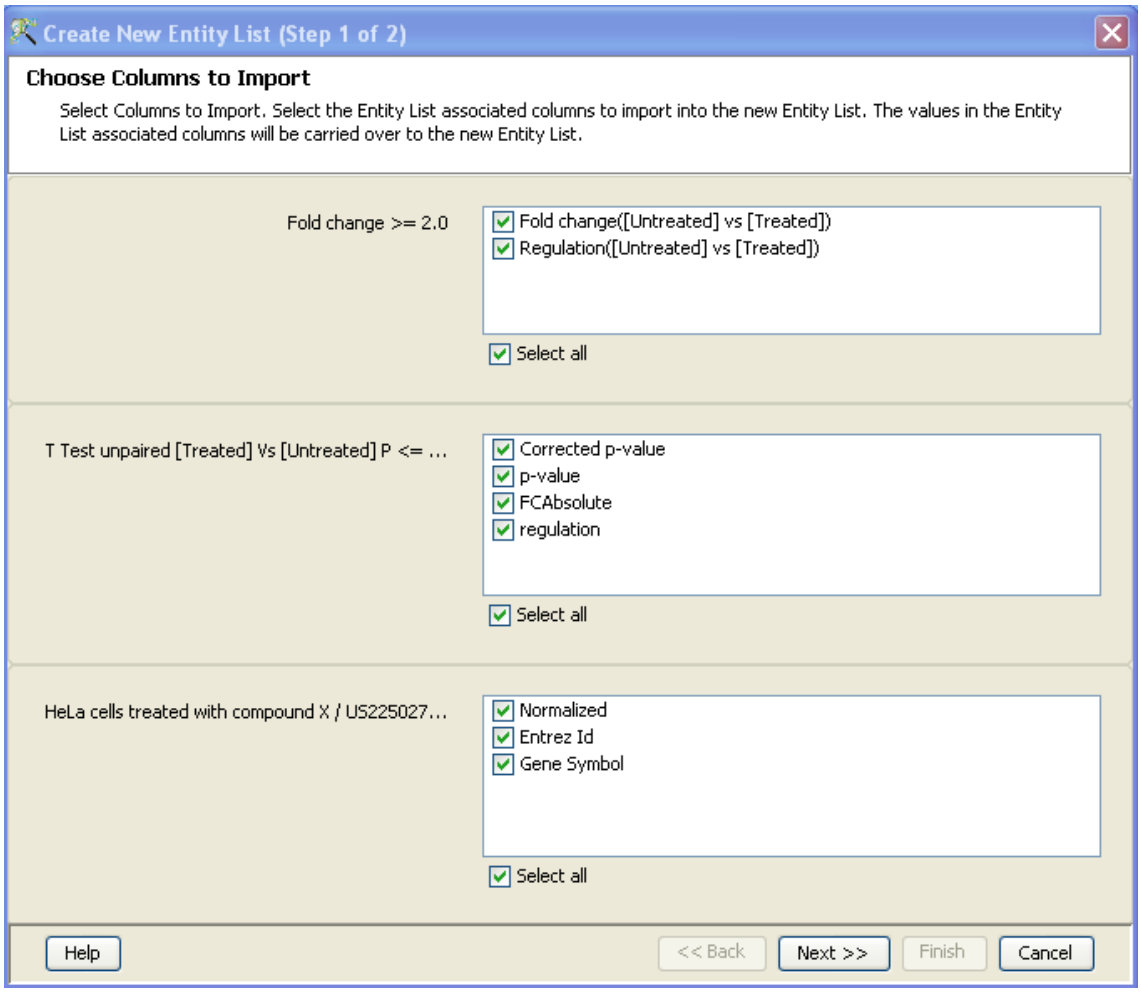


Figure 5.36: Create New Entity List from Venn Diagram

- Step 1 of 2: The entity lists or pathway lists corresponding to the selected region are shown here along with their list associated values/columns. For each of the entity lists/pathway lists, all or a subset of the list associated values/columns can be chosen. Some of the entity lists/pathway lists may be greyed out depending on whether they contribute to the selected region from the Venn diagram or not.

- Step 2 of 2: The results including the selected entity lists/pathway lists and their list associated values/columns and corresponding annotations are displayed here. A **Find** functionality allows you to locate entities/pathways containing certain list associated values or annotations. It is possible to configure the columns by clicking the **Configure Columns** button. **Finish** exits the workflow.

  A newly created entity list along with the chosen columns is saved in the Experiment Navigator of the active experiment as a child node. A newly created pathway list is saved in the **Imported Pathway Lists** folder of the active experiment. If this is the first pathway list to be added to the experiment, the folder is newly created at this time.

NOTE: Create New Entity List from the Venn diagram allows creation of new entity lists with the union of associated values/columns from the participating entity lists.

5.12.1.4 Right-click operations

The operations on Venn diagram are accessible from the Right-Click menu on the Venn diagram. These operations are similar to the menu available on any plot. The Venn diagram is a lassoed view. Thus you can select any area within the Venn diagram. This will be shown with a yellow border and the genes in any in this area will be lassoed all across the project. Multiple regions in the Venn diagram can be selected using Ctrl and left click. Further, if the *Show selection/total for each set* option is selected in the Rendering tab of of the Properties dialog (see below), the Venn diagram displays the number of selected genes in any other view, for example a Profile Plot, compared to the total number of genes in each area.

5.12.2 Venn Diagram Properties

The Properties dialog of the Venn diagram is accessible by Right-Click on the Venn diagram. See Figure 5.37.

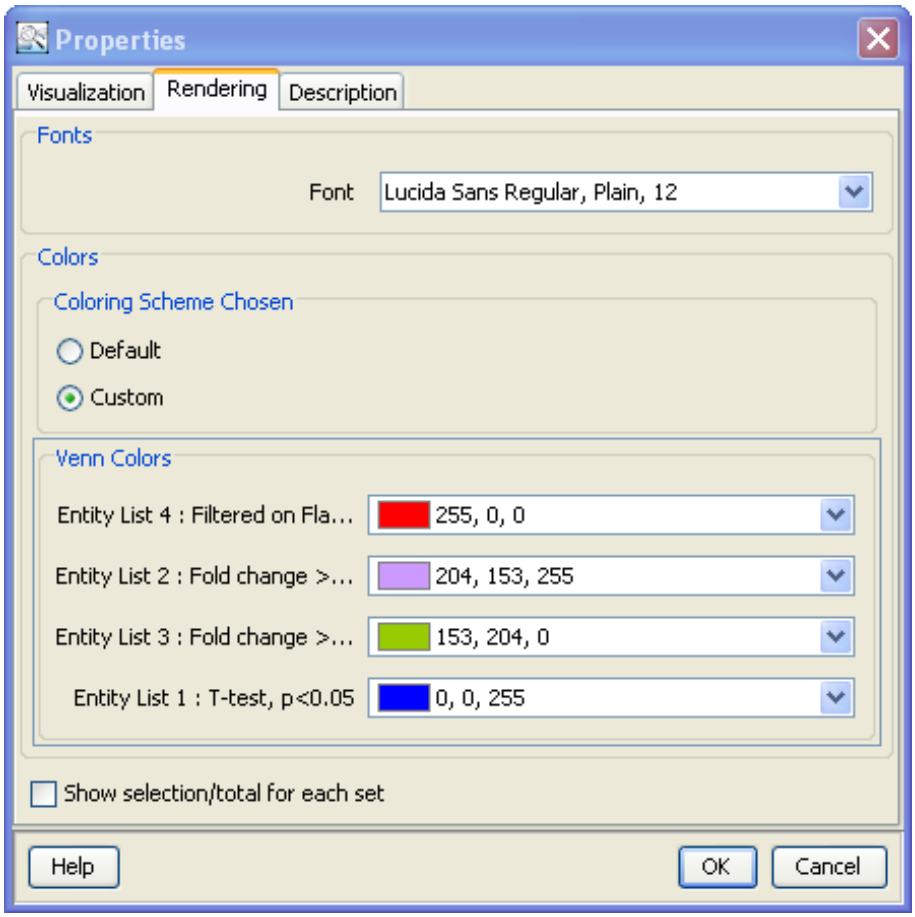


Figure 5.37: The Venn Diagram Properties

Visualization: The Venn diagram is drawn with upto four chosen entity lists. The visualization tab allows you to view the Venn diagram with all the permutations and combinations of the chosen entity lists. For instance, if the Venn diagram was launched with entity lists E1, E2, E3 and E4 from the visualization tab, user can choose to view the venn diagram with only E1 and E2 or E1 and E3 E2 and E4.

Rendering: The rendering tab in the properties dialog allows you to configure and customize the font and colors of the different entity lists shown in the Venn diagram. Also, selecting the *Show selection/total for each set* option will show you how many entities selected in other views form a part of the total entities in that region of the Venn diagram.

Description: The description for the view can be configured and modified from the Description tab in the Properties dialog. By default, if the view is derived from running an algorithm, the description will contain the algorithm and the parameters used. The description, if any, appears in the Legend situated at the bottom right of the main **GeneSpring** window.

5.13 The Violin Plot

You can launch the violin plot from the View menu on the main menu bar with the active interpretation and the active entity list of the experiment. Alternately, Left-Click the Violin Plot icon in the Toolbar to launch the violin plot.

A Violin plot is a combination of a box plot and a kernel density plot which illustrates the probability density function of normalized intensity values. Kernel density is a non-parametric way to estimate the probability density function derived from Gaussian distribution. Kernels of a violin plot illustrate density of data at different levels and are rotated to surround each side of the box plots. Box plots group normalized intensity values through their quartiles. The box plot of a violin plot displays the median in the middle box, the 25th percentile and the 75th percentile, or the 1st and 3rd quartile along with lower and upper limits. Violin plots (see Figure 5.38) are good visual aids to compare multiple groups within large datasets.

Refer "Formula for drawing a Violin plot in **GeneSpring** " on page 184 for details on the mathematical formula used for violin plot rendering in **GeneSpring**.

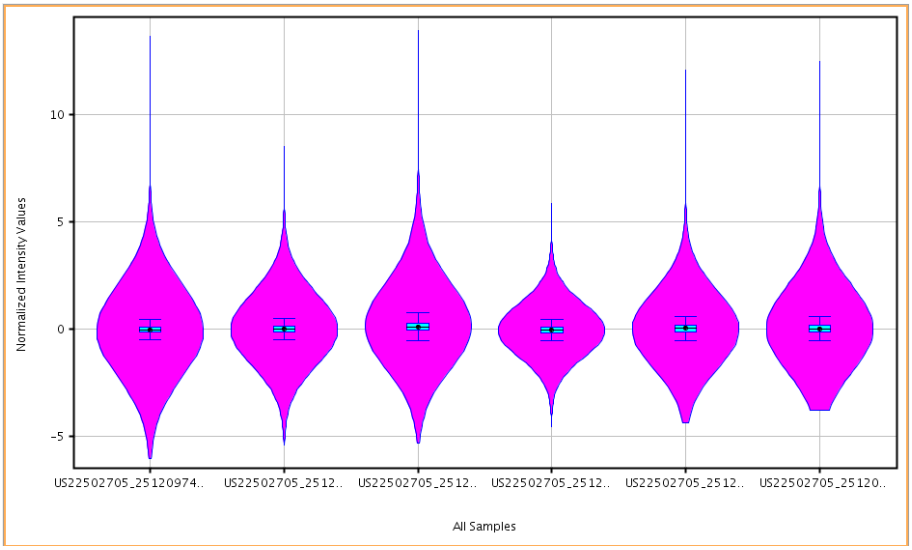


Figure 5.38: Violin Plot

If the active interpretation is the default All Samples interpretation, the violin plot displays the distribution of each normalized intensity value with respect to the active entity list. If an averaged interpretation is the active interpretation, the violin plot displays the distribution of the normalized intensity values of conditions in the averaged interpretation with respect to the active entity list. The legend window displays the interpretation on which the violin plot is launched.

Select a different entity list to make it active and the violin plot will dynamically illustrate the normalized intensity values from the active entity list. Clicking on an entity list in another experiment will translate the entities in that entity list to the current experiment and display those entities in the violin plot.

You can customize and configure the violin plot from the **Properties** dialog (see "Violin Plot Properties" on page 181).

5.13.1 Violin Plot Operations

Right-click options on a violin plot are common to all views and are described in the section Common Operations on Plot Views.

5.13.2 Violin Plot Properties

You can customize and configure the violin plot from the Properties dialog. There are four tabs in the Properties window: Axis, Rendering, Columns, and Description. See Figure 5.39.

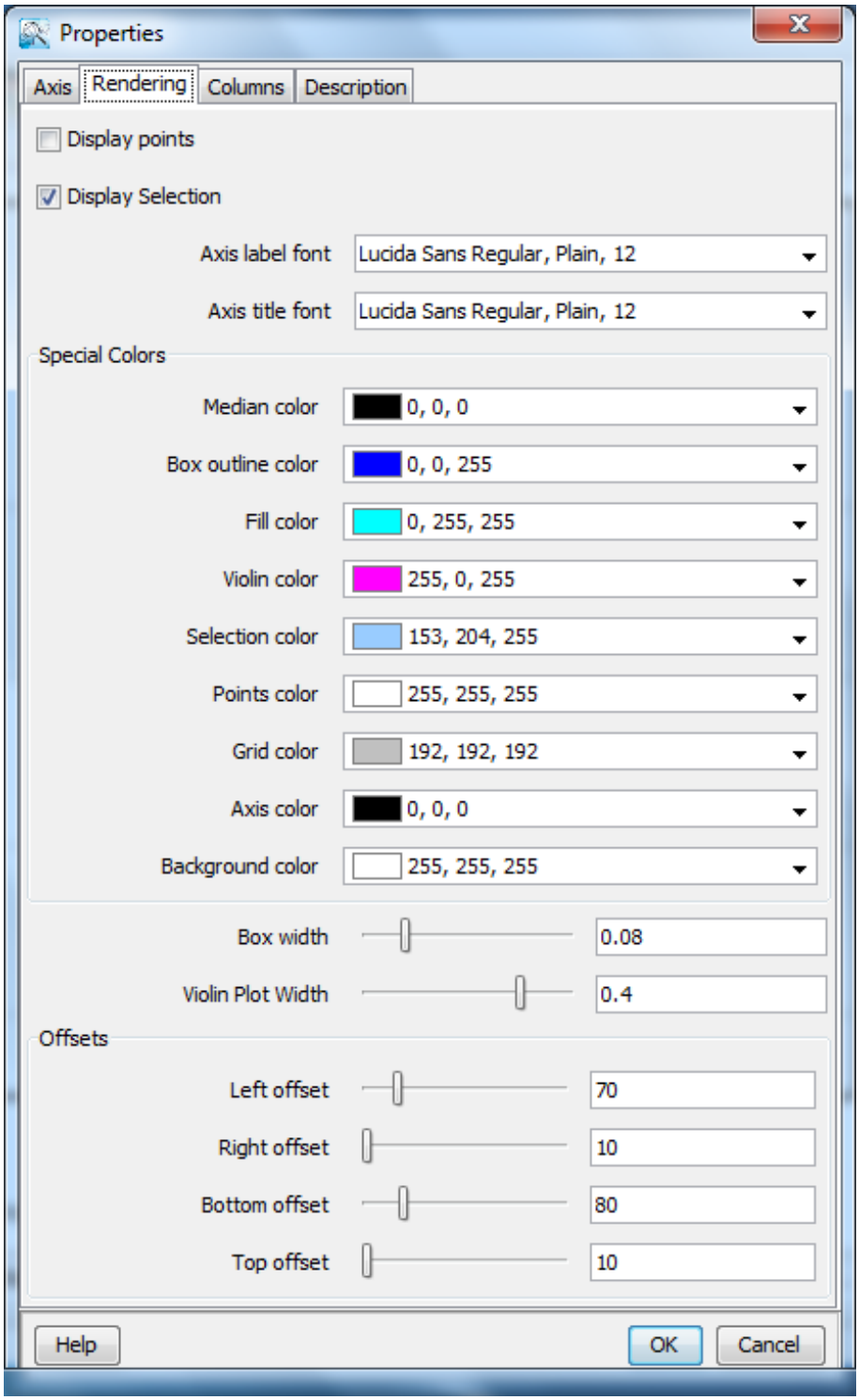


Figure 5.39: Violin Plot Properties

Axis: The grids, axes labels, and axis ticks of the plots can be configured and modified from the Properties dialog. To modify these, Right-Click the view, select **Properties** and click on the **Axis** tab.

The plot can be drawn with or without the grid lines by clicking on the 'Show grids' option.

The ticks and axis labels are automatically computed and shown on the plot. You can show or remove the axis labels by clicking on the Show Axis Labels check box. Further, the orientation of the tick labels for the X-Axis can be changed from the default horizontal position to a slanted position or vertical position by using the drop down option and by moving the slider for the desired angle.

The number of ticks on the axis are automatically computed to show equal intervals between the minimum and maximum and displayed. You can increase the number of ticks displayed on the plot by moving the Axis Ticks slider. For Numerical data columns, you can double the number of ticks shown by moving the slider to the maximum. For categorical columns, if the number of categories are less than ten, all the categories are shown and moving the slider does not increase the number of ticks.

Rendering: You can configure all aspects of the violin plot view including fonts, colors, and offsets.

Display Points: Display points display the density of points for each column of the violin plot. This is used for the selection of points.

Fonts: All fonts on the plot can be formatted and configured. To change the font in the view, Right-Click on the view and open the Properties dialog. Click on the **Rendering** tab of the **Properties** dialog. To change a **Font**, click on the appropriate drop-down box and choose the required font. To customize the font, click on the customize button. This will pop-up a dialog where you can set the font size and choose the font type as bold or italic.

Special Colors: You can configure and customize all the colors of the violin plot.

All the colors that occur in the plot can be modified and configured. The plot Background color, the Axis color, the Grid color, the Selection color, as well as plot specific colors can be set. To change the default colors in the view, Right-Click on the view and open the Properties dialog. Click on the Rendering tab of the Properties dialog. To change a color, click on the appropriate arrow. This will pop-up a **Color Chooser**. Select the desired color and click **OK**. This will change the corresponding color in the View.

Box Width: Customize the box width of the violin plot using the slider. The default is set to 0.08 of the width provided to each column of the box whisker in the violin plot.

Violin Plot Width: Customize the width of the violin plot using the slider. The default is set to 0.4.

Offsets: The bottom, top, left, and right offsets of a plot can be modified and configured. These offsets can be changed if the axis labels or axis titles are not completely visible in the plot, or if only the graph portion of a plot is required. To change the offsets, Right-Click on the view and open the Properties dialog. Click on the Rendering tab. To change plot offsets, move the corresponding slider, or enter an appropriate value in the text box provided. This will change the particular offset in the plot.

Columns: You can customize the columns drawn in the violin plot and the order of columns from the Columns tab in the Properties Dialog.

The columns for visualization and the order in which the columns are visualized can be chosen and configured from the List Selector in the **Column** tab of the **Properties** dialog. Right-Click the view and select **Properties** to open the dialog. Click the **Columns** tab. The List Selector panel shows the **Available items** on the left and the **Selected items** on the right. The items under **Selected items** are the columns that are displayed in the view in the same order in which they appear here.

To move columns from the **Available list** box to the **Selected list** box, select the required items in the **Available items** box and click the **Move right** button between the list boxes.

This adds the selected columns from the **Available items** box to the items already in the *Selected items* box. To move columns from the **Selected items** box to the **Available items** box, select the required items in the **Selected items** box and click the **Move left** button. This moves the selected columns from the **Selected items** box to the *Available items* box in the exact position or order in which the column appears in the experiment.

You can also change the order of the columns in which they appear in the view by selecting items in the **Selected items** box and using the **Move Up** and **Move Down** buttons to arrange the columns in the desired order. If multiple items are selected, the first click consolidates the selected items (bring all the selected items together) with the first item in the specified direction. Subsequent clicks on the **Move Up** and **Move Down** buttons will move the selected items as a block in the specified direction, one step at a time until it reaches the end of the list. If only one item or contiguous items are selected in the *Selected items* box, then these will be moved in the specified direction, one step at a time until the end of the list is reached. To reset the order of the columns to the original order in which they currently appear in the plot view, click on the **Reset order** button located between the **Move Up** and **Move Down** buttons.

To select items, Left-Click on the required item. To select multiple items in any of the list boxes, Left-Click and Shift-Left-Click will highlight all contiguous items, and Ctrl-Left-Click will add the clicked-on item to the previously selected elements.

The lower portion of the Columns tab provides a utility to select items in the **List Selector** based on a search by column name or mark. You can either match by **Name** or **Column Mark** whichever is appropriate. By default, the Match by *Name* option is active.

- To match by **Name**, select **Name** from the **Match by item** drop-down list, enter a string in the **Name** text box and hit Enter. **GeneSpring** performs a substring match with the columns listed in the **Available items** and the **Selected items** boxes and selects any matching entries.
- To match by **Mark**, select **Mark** from the **Match by item** drop-down list. The column marks (for example, Affymetrix ProbeSet Id, raw signal, etc.) available in the experiment are shown in the **Mark** drop-down list. Choose a **Mark**, and **GeneSpring** selects the corresponding columns of the experiment.

Description: The title for the view and description or annotation for the view can be configured and modified from the **Description** tab in the **Properties** dialog. Right-Click the view and select **Properties** to open the dialog. Click the **Description** tab to display the the current **Title** and **Description** of the view. Both **Title** and **Description** can be modified by changing the text in the corresponding text boxes. If you modify the default title, the new title is displayed in the Legend and used when the view is exported as **Report**. The change does not reflect in the title tab of the view in the **GeneSpring** desktop. The **Description** (default and modified) appears in the **Legend** window on the right. By default, if the view is derived from running an algorithm, the description will contain the algorithm and the parameters used.

5.13.3 Formula for drawing a Violin plot in GeneSpring

In **GeneSpring**, the violin plot is rendered depending on the number of available data points (probe intensity values) in the active experiment.

In an experimental data set:

- If the data points are < 500, the violin plot is rendered by the formula:

$$Est_i = (\frac{1}{nh\sqrt{2\pi}})\sum_{j=1}^{n} e^{-}\frac{1}{2}(\frac{Xeval_i - X_j}{h})^2$$

  where:

  - Est_i - Estimated kernel density value of $Xeval_i$.
  - **n** - Maximum number of data points.
  - **h** - Kernel density curve smoothing factor

$$= \sigma * (\frac{4}{3 * n})^{\frac{1}{5}}$$

  - $Xeval_i$ - A given evaluation point i.
  - X_j - j^{th} sample element or each data point in the data set.

- If the data points are ≥ 500, the violin plot is rendered by the formula:

$$Est_i = (\frac{1}{nh\sqrt{2\pi}})\sum_{j=1}^{n} f_j * e^{-}\frac{1}{2}(\frac{Xeval_i - X_j}{h})^2$$

  where:

  - Est_i - Estimated kernel density value of $Xeval_i$.
  - **n** - Number of bins with frequency > 0

$$= 8 * log_{10}(Totalnumberofdatapoints)$$

  - **h** - Kernel density curve smoothing factor

$$= \sigma * (\frac{4}{3 * n})^{\frac{1}{5}}$$

  - $Xeval_i$ - A given evaluation point i.
  - X_j - mid point of j^{th} bin.
  - f_j - frequency of j^{th} bin.

5.14 LD Plot

You can launch the LD Plots of individual Chromosomes (results of LD Analysis).

The following steps guide you to launch an LD Plot:

- Click on an LD Plot node in the experiment navigator to launch the LD Plot.
- Select an option from the LD Measure menu: r^2 or $D - prime(D')$ (default option).
- Drag the mouse pointer over the plot to select blocks of interest (Figure 5.40).

- Click on the "Create entity list from selection" icon in the toolbar, and save the blocks as entity lists.
  - Select "Single Entitylist" to save the entities in all the selected blocks as a single entity list.
  - Select "Blockwise Entitylist" to save the entities in each block as a separate entity list.

  Refer to LD Analysis section for information on LD Analysis.

5.14.1 LD Plot Toolbar

You can select `Selection` (default), `Clear Selection`, `Zoom`, and `Reset Zoom` modes from the toolbar. The icon corresponding to the mode that you select is grayed out in the toolbar.

Figure 5.40: LD Plot Toolbar

Zoom Mode:
Select a block of SNPs to launch an LD Plot of the block.

Selection Mode (default):

- Select the blocks of SNPs.
- Click on the "Create entity list from selection" icon to save each selected block as a separate entity list or all the blocks as a single entity list.

5.14.2 LD Measure Options

GeneSpring provides two LD measure visualizations: r^2 and D-prime.

r^2:
Plots the raw r^2 score for a given marker pair. The r^2 is a measure of linkage disequilibrium

between two genetic markers. For SNPs that have not been separated by recombination or have the same allele frequencies (perfect LD), $r^2 = 1$. In such cases, the SNPs are said to be redundant. Lower r^2 values indicate less degree of LD.

D-prime:
This is the default plot and is displayed when the user turns on the LD plot track. This track plots the raw D-prime score for a given marker pair. D-prime is a measure of linkage disequilibrium between two genetic markers. Absolute value of D-prime = 1 (complete LD) indicates that two SNPs have not been separated by recombination, while values of `D-prime < 1` (incomplete LD) indicate that the ancestral LD was disrupted during the history of the population [38].

> **Note:** Only D-prime values near one are a reliable measure of LD extent; lower D-prime values are usually difficult to interpret as the reliability of D-prime strongly depends on sample size.

Source: http://hapmap.ncbi.nlm.nih.gov/gbrowse_help.html#genotypes

5.14.3 LD Plot Properties

You can open the Properties dialog from the context (right-click) menu.

The Properties dialog has three tabs, viz., Visualization, Rendering, and Description.

Visualization:

You can select the plot label from the drop-down box, and adjust the label offset, label length, and the margins (left, right, bottom, and top) using the respective sliders.

Rendering:

You can configure the color range adjusting the sliders for the Minimum, Center, and Maximum values, and the respective Color Chooser box.

> **Note:** D-Prime: ranges from -1 to +1 R-Square: ranges from 0 to +1

Description
You can add an appropriate Title and Description for the plot and click OK.

Export As:

- You can export the plot as an Image or HTML page from the context (right-click) menu option.
- You save the Image as a .tiff, .png, .jpeg, .jpg, .bmp, .svg, or .eps file.

5.15 Genome Browser

The **GeneSpring** genome browser allows the viewing of expression data imposed against the genomic features. For more details on the same, refer to the chapter on Genome Browser

5.16 Plot Options

5.16.1 Plot Log10/Linear Values

In **GeneSpring**, the data in the experiments are in log2 scale and the views are launched with the data in log2 scale. This option *'Plot Log10/Linear Values'* allows the user to view the scatter plot and the profile plot in log10 scale or in linear scale. The signal values for this plot can be chosen to be in raw or normalized form and the plot will be launched with the chosen interpretation.

On clicking *View* → *Plot Log10/Linear Values*, a window comes up with options to choose the interpretation, the type of signal values (raw or normalized), the scale (log 10 or linear), and the plot (scatter or profile plot).

The legend accompanying the resultant plot will show the chosen parameters with which the plot was launched. The plot operations and properties remain the same as with the regular scatter plot and profile plot and are described in sections Scatter Plot Operations, Scatter Plot Properties,Profile Plot Operations and Profile Plot Properties.

5.16.2 Plot List Associated Values

This option available from the View menu allows the user to visually inspect the data associated with two entity lists either as a scatter plot, a histogram or as a profile plot. The list associated values include the columns obtained during analysis such as Fold Change, Significance Analysis etc. It does not include the annotation columns associated with the entity list.

On selecting this option, a window appears in which the user needs to provide the necessary inputs. The option to choose entity lists as well as an interpretation is provided. The user can also select the type of visualization from the options (Scatter Plot, Histogram and Profile Plot) provided in the drop down. See Figures 5.42, 5.43 and 5.44. There is also an option to see either the raw or normalized signal values for the entities in the entity list in the view. See 5.41.

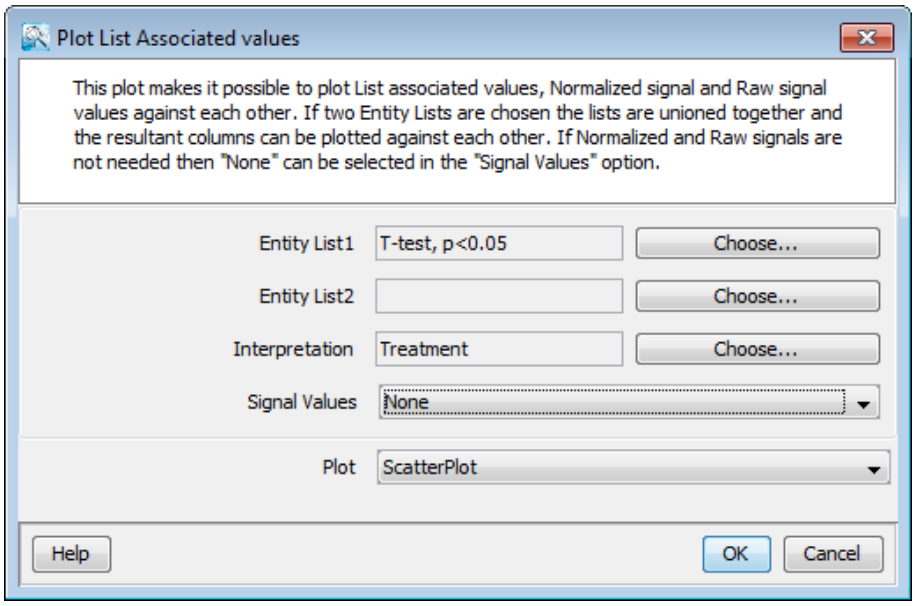


Figure 5.41: Plot List Associated Values

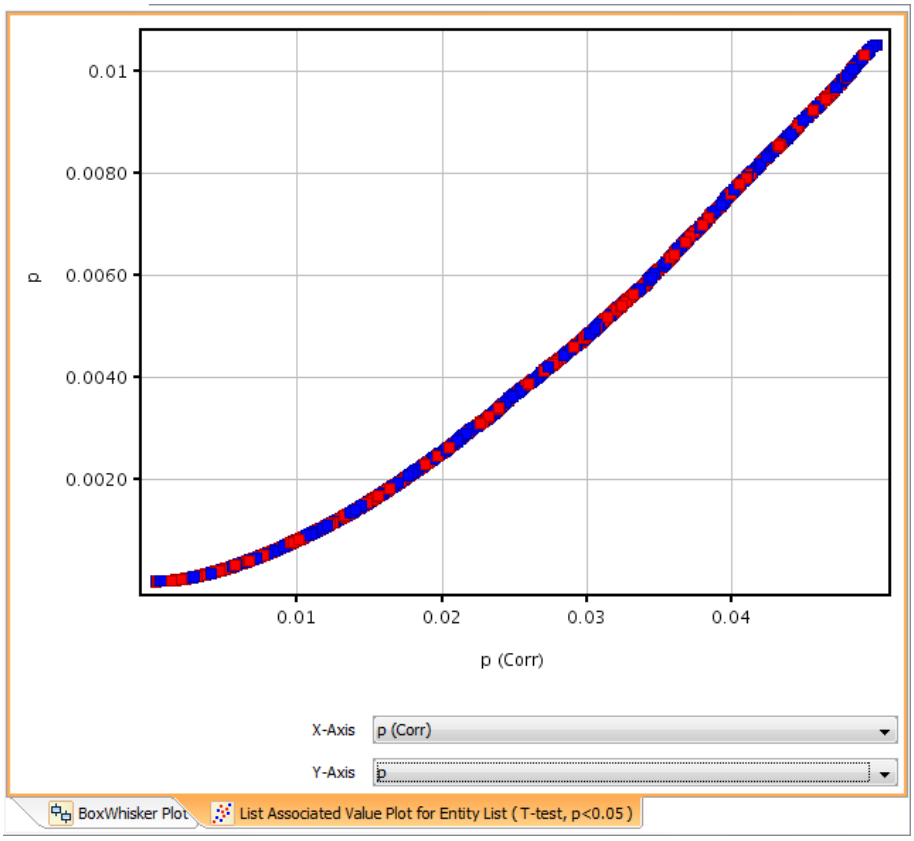


Figure 5.42: Plot List Associated Values-Scatter plot

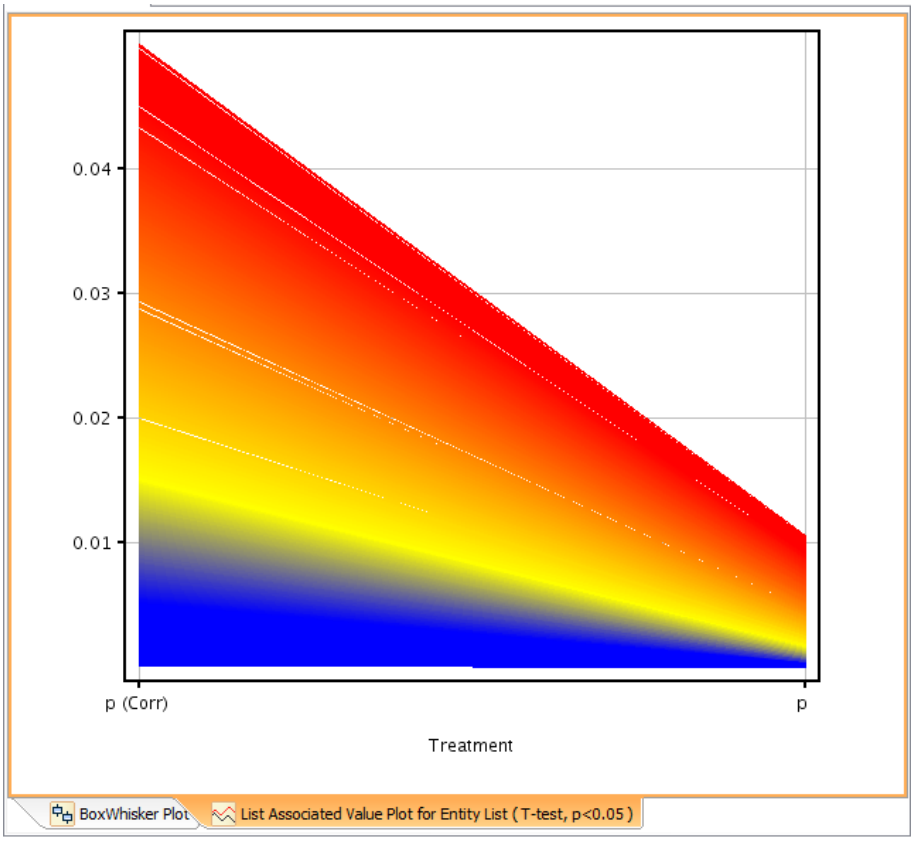


Figure 5.43: Plot List Associated Values-Profile plot

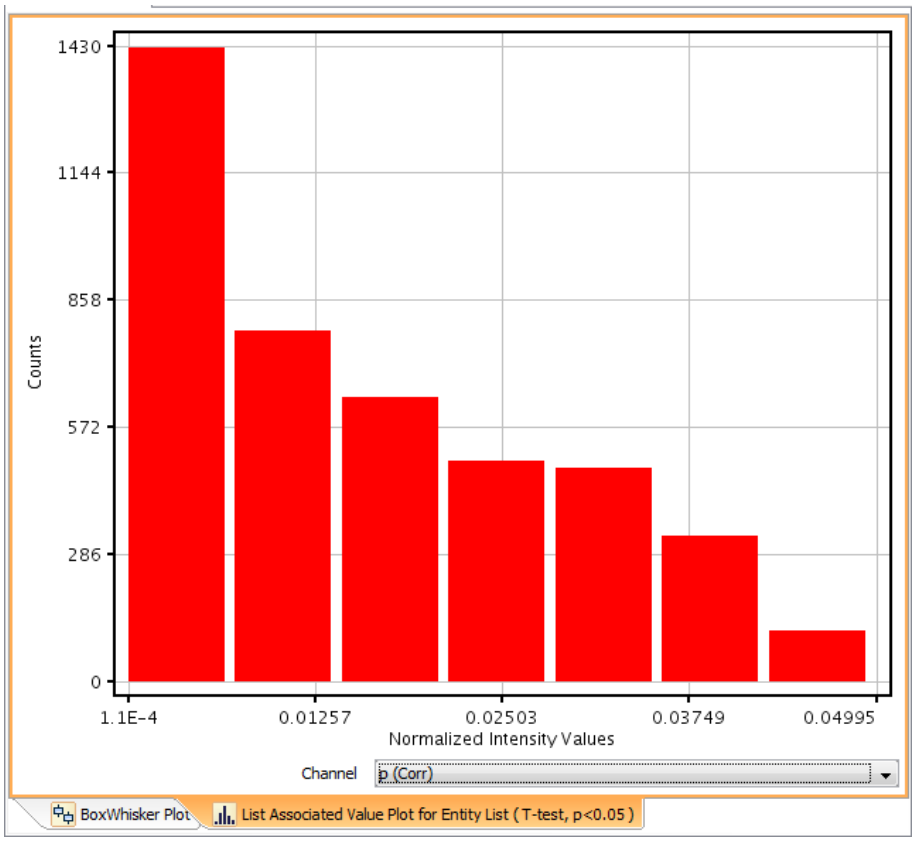


Figure 5.44: Plot List Associated Values-Histogram

On clicking on ***OK***, the tool shows the view specified. By default, the X-axis is the second column of the first entity list and the Y-axis is the third column of the second entity list (the first column in both being the identifier). The user is provided with an option of all the list associated values in the dropdown for the axes and can choose as required.

5.17 Miscellaneous operations

5.17.1 Save Current view

In **GeneSpring**, open views (Heatmaps, Classifications, Scatter Plot etc) are not saved in the experiment by default; so reopening or refreshing the experiment will not automatically bring up these views. Clicking on this option saves all currently open views in the experiment so these can be restored when the experiment is reopened.

Note: This option is disabled when a Venn diagram or a Pathway list are open in an active experiment.

There is one caveat though. Technology updates will not reflect in these saved views; so after a technology update, it is advisable to regenerate views which need updated annotation information.

5.17.2 Find Entity

Find Entity is a search functionality available through the *View* menu or with the key binding Ctrl-F. This brings up a window listing all entities. The *Find* tag at the bottom allows the user to input a string/value for the search. T he other tags, *Find Next* and *Find Previous* select and highlight the next/previous entity that matches the search condition. *Select All* will select all entities that matches the search string entered in the *Find* text box. Configuration of columns can also be done through this step. Any selection here will reflect throughout the tool in all views.

5.17.3 Inspect Entities

Inspect Entities brings up the Entity inspector with the selected entities. Can also be called by the key binding Ctrl-I.

5.17.4 Properties

Properties or Ctrl-R brings up the properties windows relevant to the view on focus.

5.17.5 Exporting Views

Views can be explicitly copied and pasted on any application and or exported as image or html. For more details, see section Exporting Views.

5.17.6 Configuring Views

Many of the properties of the views can be configured from the menu **Tools** → **Options** → **Views**. Different views are listed here along with the handles for configuring them. These configuration options are remembered by the tool and shown by default whenever a view is launched. They are also saved during the BackUp / Restoration Process.

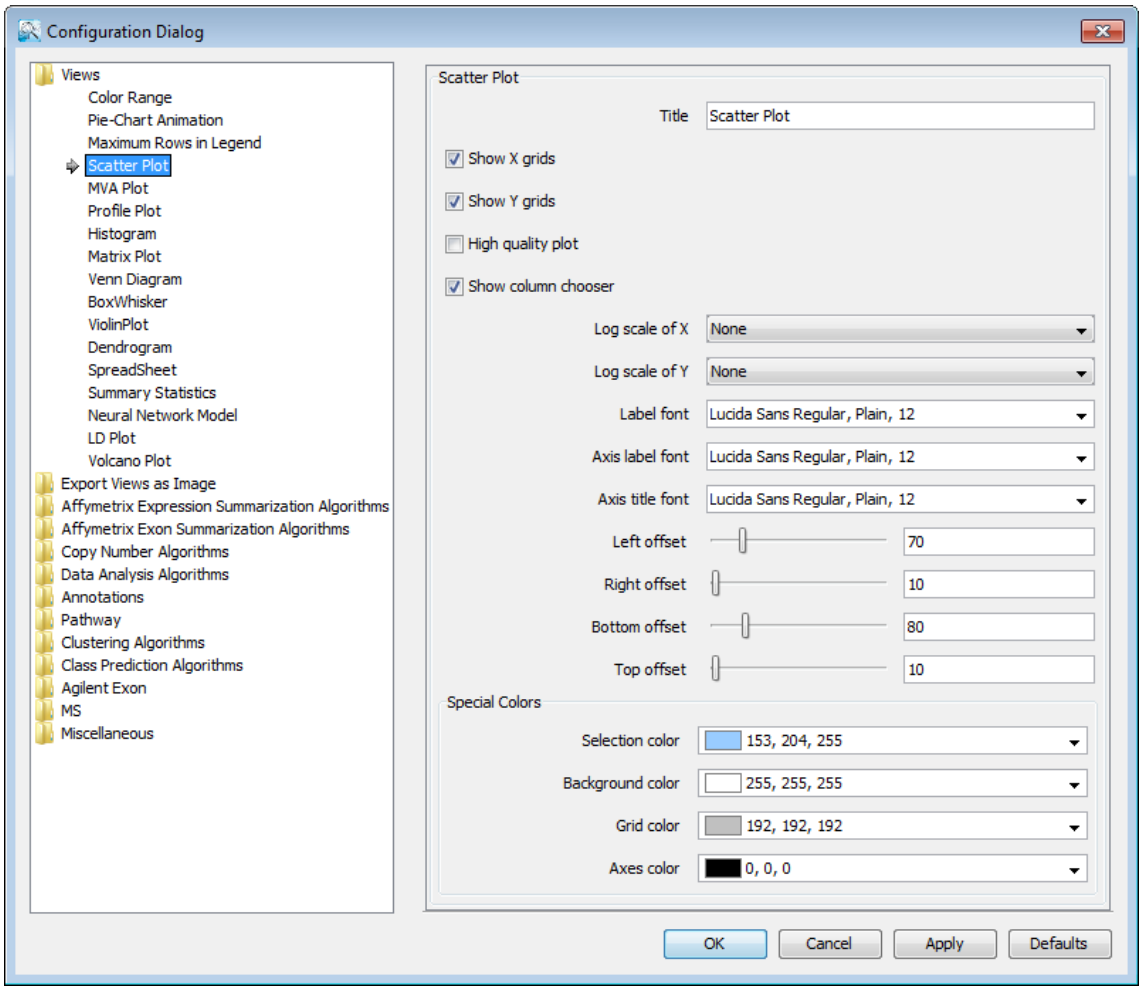


Figure 5.45: Configuration options for Views

Chapter 6

Genome Browser

6.1 Introduction

Genome Browser (GB) in **GeneSpring** is a powerful visualization tool to view data, results, and annotations associated with the genome in one place.

The Genome Browser can be used to:

- Visualize data from different experiment types of any technology in one view; however all the data must be from the same organism.
- Visualize data from entity lists.
- Merge different data tracks and analyze data across technologies/experiments/entities.
- Pan the entire chromosome and drill down (zoom into) to the gene level or transcript level for viewing data.
- Import annotation data for any of the listed organisms or create annotations data for a custom organism using Annotations Manager, and view it in the Genome Browser along with the experimental data.
- Export entity lists from the experimental data tracks.
- Publish and share images.
- Export plots as images for presentations/publications.
- Drag and drop files containing chromosome information from outside **GeneSpring** for viewing.
- View associated data of selected region in the chromosome in a spreadsheet view; export selected region for later analysis.
- Search for option to find specific genes, transcripts and SNPs (only rsIDs downloaded from UCSC are searchable) using the 'Search Result' option.

Any experiment chosen for visualization in the GB should contain annotations describing the chromosomal location of the entities (genes/transcripts/SNPs/CN probesets, etc.) in the experiment. This includes:

1. The chromosome number.
2. Start location and stop location.For SNPs, the start and stop locations would be identical .
3. Strand information. This is optional. However, it is essential for displaying the gene or transcript structure (if 'gene' or 'transcript' is chosen as an annotation track). Strand information is also required to see the upstream region of a gene.

For standard technologies supported by **GeneSpring**, all the above information is available within the experiment and the GB will pick this up automatically. For custom technologies, the user needs to ensure that the above information is present.

6.2 Getting started with the Genome Browser

6.2.1 Launching the Genome Browser

In order to launch and view data on the GB, GB annotations are to be updated first. While the organism and build information are a must to launch the GB and view data, other annotations like genes, transcripts, and SNPs obtained from standard sources like UCSC, NCBI, and ENSEMBL can be used to compare the experimental data/results.

Upon the first launch of **GeneSpring**, the user will be prompt to update the GB annotations via the Annotations Manager window. However, if the GB annotations are not updated on the first launch of the tool, these can be brought in (updated) via *Annotations menu* $\longrightarrow$ *Annotations*. For more details on Annotations Manager refer to the Section 3.3.

Once the Genome Browser annotations are updated, there are 3 ways to launch GB in **GeneSpring** . They are;

- From the menu bar by selecting *View* $\longrightarrow$ *Genome Browser*.
- From the tool bar by selecting 'Launch Genome Browser' icon.
- From the Workflow link; *Views* $\longrightarrow$ *Genome Browser* (Only for Copy Number and Association experiments).

Genome Browser can be launched only on an active experiment (an open experiment). By default, the GB identifies the organism information from the active experiment. Thus, when the GB is launched by any of the above mentioned options, a window titled 'Select Build' will come up with

a dropdown to select a build for that organism. See Figure 6.1. Upon selecting the build and selecting ***OK***, the GB appears on the display window.

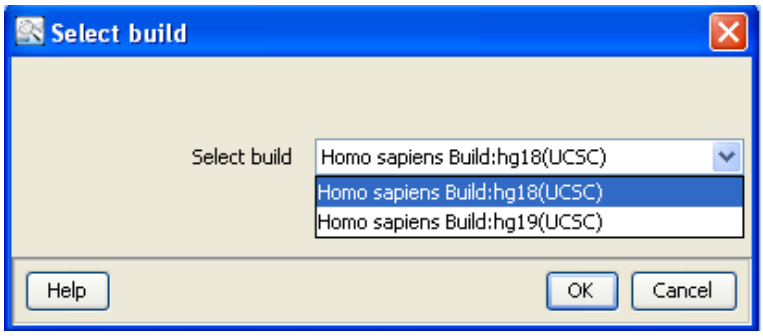


Figure 6.1: Genome Browser- Select Build

Note:

If the Genome Browser is launched without updating the annotations, an error dialog is displayed informing the user to update GB annotations from the Annotations Manager. See Figure 6.2.

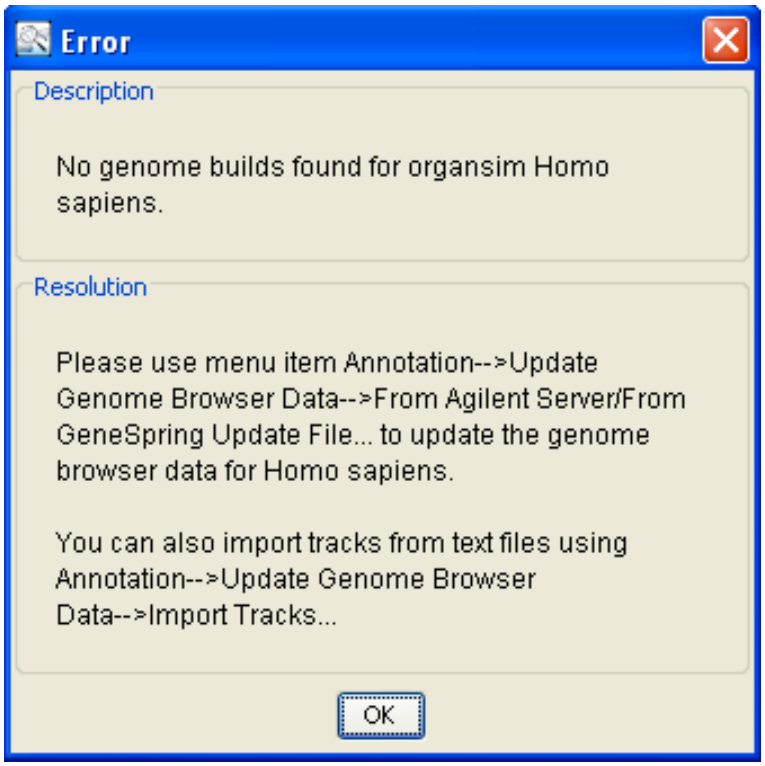


Figure 6.2: Update Genome Browser Annotations

6.3 Visualization in Genome Browser

On launch, GB shows annotation information pertaining to Chromosome 1 of the chosen organism and build. See Figure 6.3.

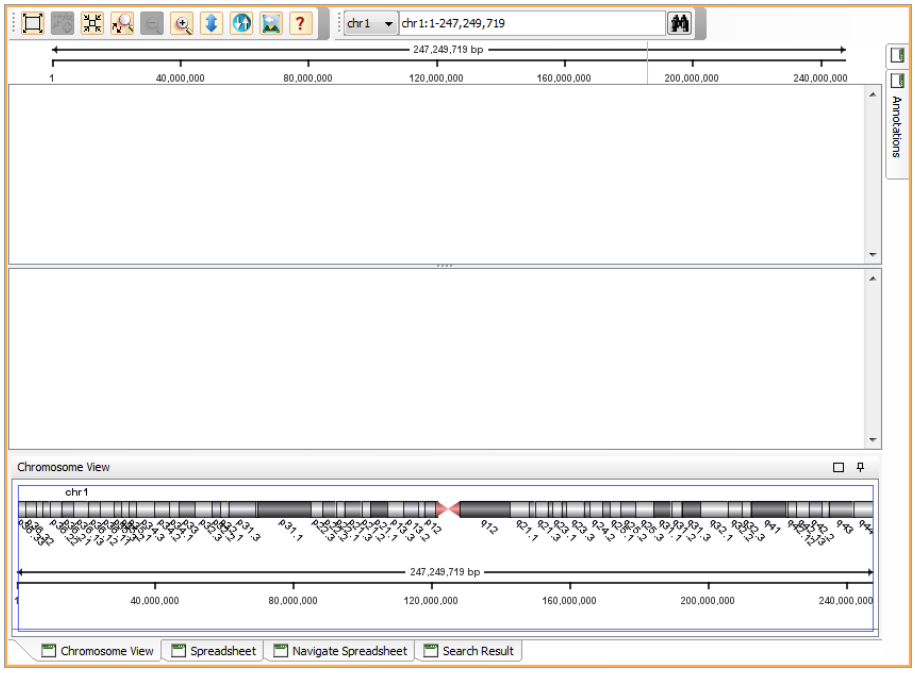


Figure 6.3: Genome Browser- on launch

The GB UI has been organized into different visualization sections. See Figure 6.4. The main section of the UI is called the 'Region View', where, the data (experimental data, results, and annotations data) is displayed. Apart from the Region View, there are other views like Chromosome View, Spreadsheet, Search Result, Genome View and Annotations. All these are present as tabs below and on the right side of the Region View. Mouse over any of these tabs displays the view corresponding to that tab. The windows corresponding to these tabs can be shifted to different locations within the GB window simply by clicking on the tab, dragging and dropping it to a different location. All these windows are dockable frames and can be docked/undocked by clicking on the Toggle auto-hide option shown at the top right corner of each frame. Details about each of the above mentioned views are discussed in the following sections.

Figure 6.4: Genome Browser Organization

6.3.1 Region View

Data is displayed as 'Tracks' in the Region View of the GB window. A 'Track' is a plot of the chromosomal location in one axis (x-axis) against the chosen data in the other axis (y-axis). The data could be annotation data of the chromosomes, raw or normalized signal values from expression experiments, data related to copy number or association experiments.

The Region View is divided into two panels called, 'Main' and 'Footer'. By default, the data that comes from the experiments, and are specific to the samples, (i.e. data tracks) are displayed in the Main panel. Annotations data of the organism that is independent of the sample information and is downloaded from the Annotations Manager to compare with the experimental data (i.e. Annotations tracks) is displayed in the Footer panel. However, data tracks can be loaded in the Footer panel and Annotation tracks can be loaded in the Main panel as well. In each of these panels, any number of tracks can be loaded/ shown simultaneously, one stacked above the other. An inner scroll appears within the track when a plot (data) loaded is larger than the track size. The tracks as well as the panels can be resized by scrolling over the track or panel separators and dragging up and down. A track can be selected by clicking on it (anywhere on the area of that track). Multiple tracks can be selected using Shift or the Ctrl key. By default, when multiple tracks are loaded, the recently added track (last track in the order of loading) gets automatically selected. A selected track displays a set of icons on the top right corner of the track. See Figure 6.5.

These icons can be used to perform different operations on a selected track. It is possible to move

the track from one panel to the other by clicking on the triangle shaped icon. Also, if there are more than one tracks in the panel, a track can be minimized using the icon and maximized using icon. Right icon and the left icon arrows can be utilized to find data on either sides of a selected track. Close icon can be used to remove a track from the GB view.

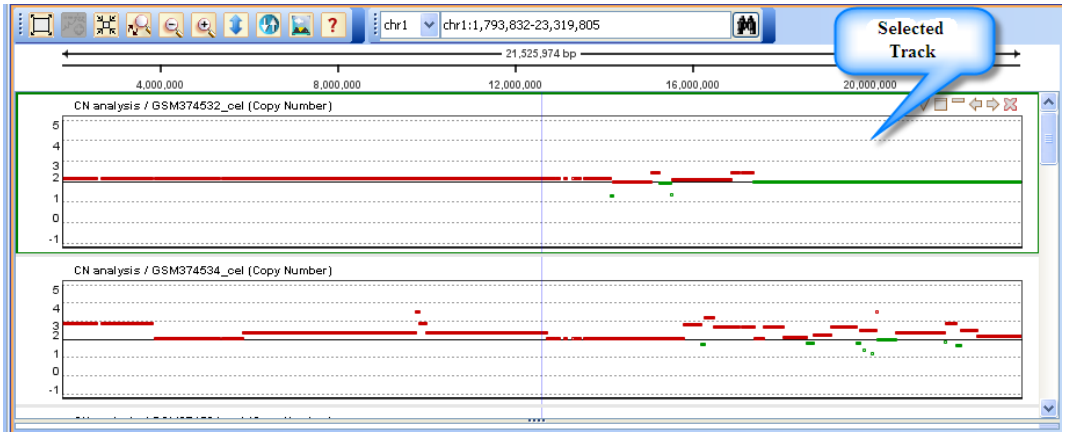


Figure 6.5: Selecting a track

A cross-hair appears inside the Region view on mouse over at any location, indicating the chromosomal position at that location (value along the x-axis). It also shows the value along the y-axis, provided the track is an experimental data track. Also a vertical line is displayed at the centre of all the tracks which turns into a double line at high zoom and indicates the width of a base.

Annotation tracks for genes, transcripts, dbSNPs, etc. that have been downloaded from the Annotations Manager can be selected and added from the Annotations tab on the right. The experimental data can be added to the GB in any of the following ways:

- Drag and Drop experiments
- Drag and Drop entity lists
- Drag and Drop files from anywhere

6.3.1.1 Drag and Drop experiments

Experiments from the navigator panel (which lists all the experiments in the project) can be dragged and dropped into the GB to populate it. A window titled 'Select Data' appears on dragging and dropping an experiment. See Figure 6.6. On this window, user can choose the following:

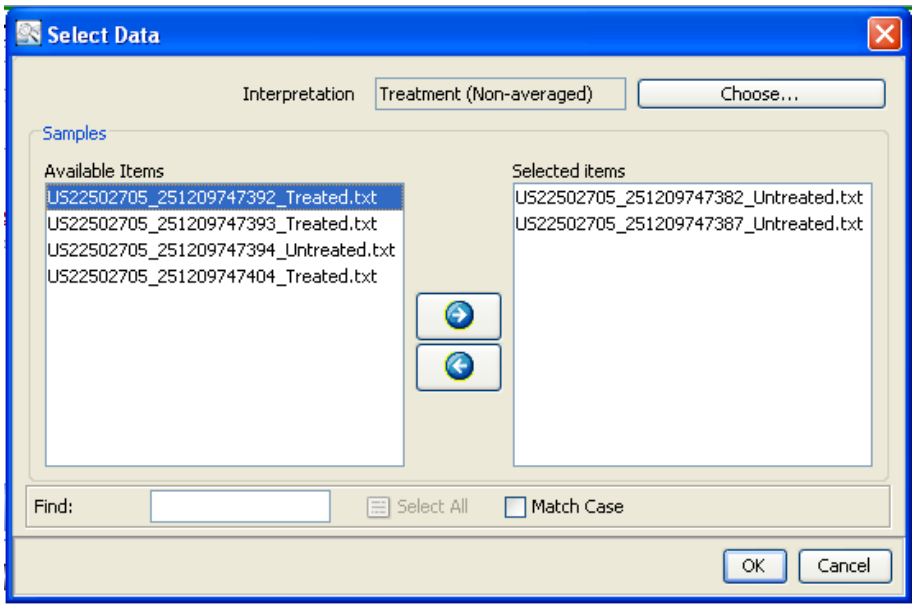


Figure 6.6: Genome Browser- Select Data

1. **An interpretation**: For expression and exon experiments, if a Non-averaged interpretation is chosen, the individual samples are considered and when an averaged interpretation is chosen, only the conditions are taken. However, for Copy number experiments only samples are considered irrespective of whether an averaged or Non-averaged interpretation is chosen. For Association experiments, there is an option to view genotype calls or allele frequencies; for the former, samples are taken and for the latter, the conditions are taken.

2. **Samples:** Samples can be chosen from the 'Select Data' window displayed in GB. (Samples listed in the 'Available Items' field can be selected and dropped in the 'Selected Items' field and vice versa using the Right and Left arrow buttons). Utilities like, 'Find', 'Select All' and 'Match case' can be used to locate the samples easily.

3. **Type of data:** In experiments other than Copy number and Association, for a given sample the Raw and Normalized signal values can be viewed in a single data track of the Region View. By default, when an experiment is dragged and dropped, only 'Normalized' signal values are launched on the track. To view 'Raw' signal values, right click on the track and choose 'Edit track properties (For more details on 'Edit track properties' and viewing options, see Section 6.4). This opens a new window where the user can select (click on) the 'Raw' check box listed under the Region View node and click OK at the bottom. See Figure 6.7. This launches the Raw signal values in the same track as that of the sample's Normalized signal values. Both Normalized and Raw signal values have different scales and are separated by a divider. A scroll over on the plot displays the Normalized and Raw labels. See Figure 6.8.

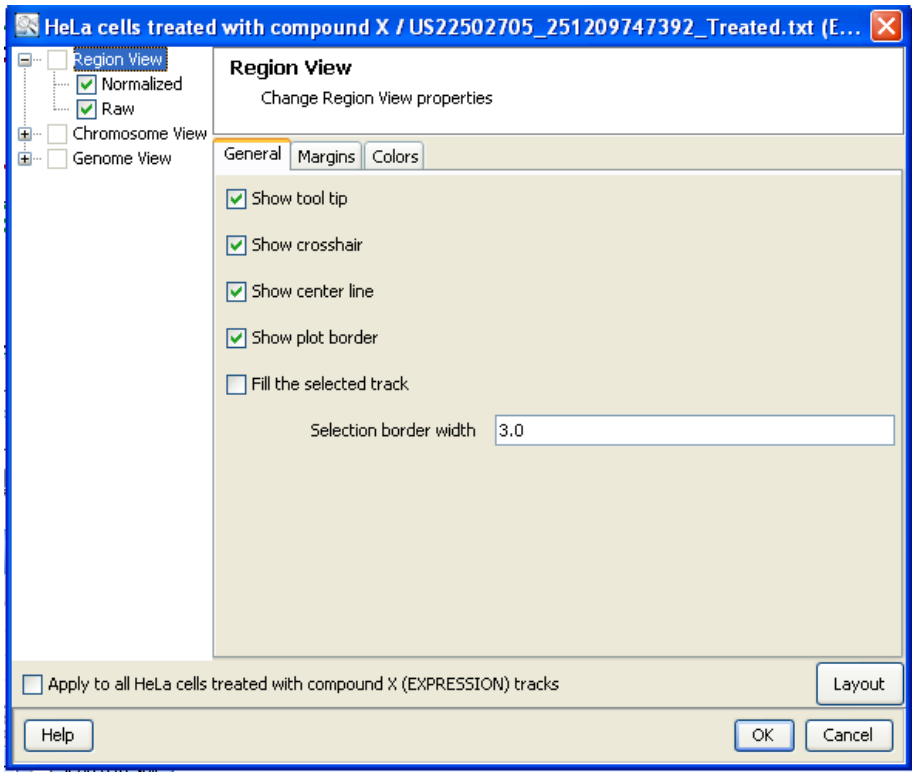


Figure 6.7: View expression data

Figure 6.8: View expression data- Normalized and Raw signal values

For a given interpretation in a Copy Number experiment, from the 'Select Data' window, Copy Number, Copy Number Confidence, Log Ratios, Mean Log Ratios, LOH, Allele Specific Copy Number and Parent Specific Copy Number can be selected fro viewing in the GB. See Figure 6.9.

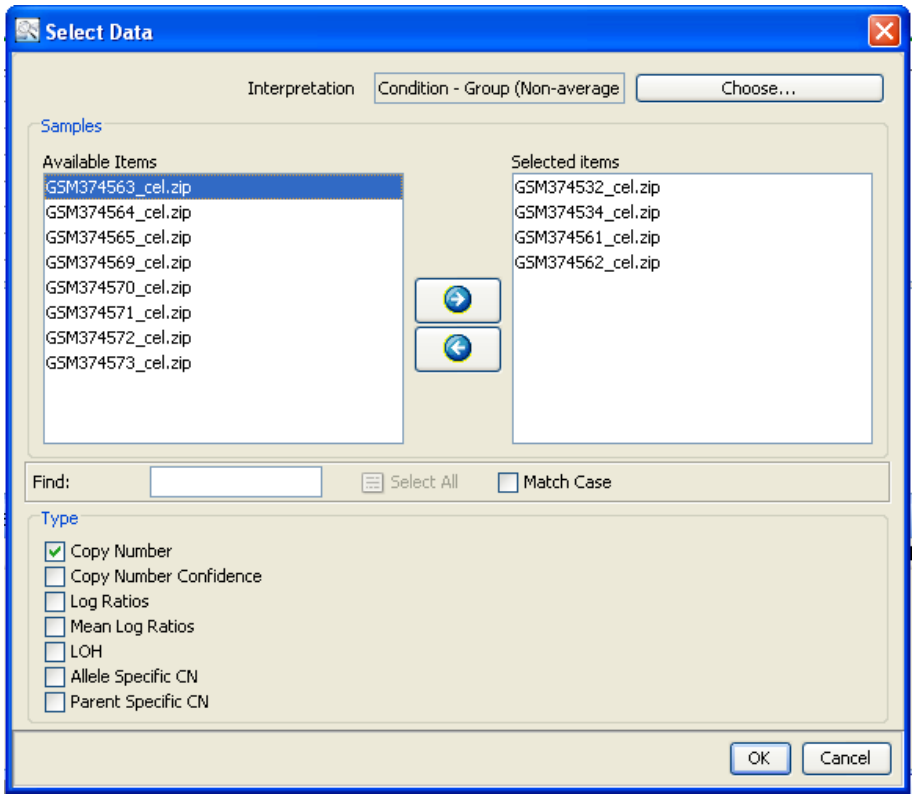


Figure 6.9: Copy Number Analysis- Select Data

For a given sample, each type of data selected from the above mentioned list is launched as a separate track. Also data values specific to each track are displayed in the Spreadsheet view. (For more details on Spreadsheet view, refer Section 6.3.5).

However, when a user selects Allele Specific Copy Number (ASCN) and Parent Specific Copy Number (PSCN) option in the 'Select Data' window, multiple types of data associated with ASCN and PSCN can be viewed in a single track. For example, ASCN is associated with its P value, Allele A and Allele B copy numbers. Thus, for a given sample, from the 'Select Data' window, when ASCN is chosen, by default ASCN P Value is launched in GB. In order to view Allele A and/or Allele B copy numbers also, right click on the track and select 'Edit track properties' and then select the required data from the Region View node. On clicking OK, the ASCN track gets refreshed with the chosen data. By default, these three values/plots are overlaid on top of each other and can be split into individual plots within the same track by the 'Edit track properties' option. For more details on 'Overlay' and 'Split', refer to Section 6.4.1. Similarly, when PSCN is selected, all its associated data i.e. PSCN difference, PSCN Max and PSCN Min values can be viewed in the same GB track.

For Association analysis, there is an option to view Genotype calls and Allele frequencies in GB. These options are available in the form of tabs on dragging and dropping an experiment into GB. Genotype calls can be launched only on samples as independent tracks. Allele frequencies can be launched on conditions (parameters) (Eg: Cases and Controls). The plot can be viewed only for interpretations on which the 'Calculate Allele Frequencies' workflow step has been run. When Allele frequencies are launched for a condition, both Allele A and Allele B frequency plots are

displayed in the same track. These plots are overlaid on top of the other and can be split into individual plots.

6.3.1.2 Drag and Drop entity lists

For open experiments, analyses results i.e., entity lists (created in the Analysis folder in the experiment navigator) can be dragged and dropped into the GB of an active experiment. Multiple entity lists can be viewed by dragging and dropping the lists one at a time. Each of the entity lists appears as a separate track.

By default, the track is shown with the location of the entity, i.e., 'Regions' against the location on the chromosome. However, along with the entity list any associated data (i.e. list associated data) can be viewed in the same track in GB. For example, in an expression analysis experiment, when an entity list created using the analysis 'Find Similar Entities' is dragged and dropped in GB, both the location of the entities (Regions) and the similarity values can be viewed. For this, right click on the desired entity list track, choose the Region view in the 'Edit track properties' utility, and check the data nodes to view in GB.

Note:

In general, while multiple data types can be viewed in a single track, by default only one of the data types is shown in GB. To view other associated data types in the track, utilize right click 'Edit track properties' option.

6.3.1.3 Drag and Drop File from anywhere

Files containing minimal information on chromosome name, start and end location can be dragged from anywhere in the file system and dropped into the GB. These files are always displayed in the Main panel. Even if there is a build conflict with the active experiment, available information from the files can be viewed. These files need not be generated from **GeneSpring** but the GB should have been launched with an active experiment prior to the dragging and dropping.

When an external file is dragged and dropped into GB, a two step wizard titled 'Drag & Drop Import Track' appears;

- Step1: **Format File**

  A name of the feature and a template are to be chosen (default is Generic Segments). The user can choose to format the file using Format Options. A preview of the file is shown at the bottom. If the first row of the file is chosen for column names (check box at the bottom), these are chosen accordingly in the step 2.

- Step2: **Choose Columns**

In this step, the data type and the column types for each of the columns can be defined. Drop down options are available for each cell of the Data type and Column type. Default values will be based on the 'template' option selected in Step 1.'. 'Save template as' is provided to save and use the created templates again. Subsequently, saved templates will be available in the 'Choose template' field of Step 1.

6.3.2 Genome View

By default, Genome View tab appears at the top right beside the Region view. See Figure 6.10. Genome View displays the cytoband representation (when the cytoband information is available for the build else the visual length of the each chromosome is represented with a band) for all the chromosomes in the organism in one window and a chromosome can be selected for viewing. By default, the chromosome that is selected in the Region View is also highlighted in the Genome View. Clicking on a different chromosome in the Genome View will change the Chromosome View and the tracks in the Region View to display information for that chromosome. A region can also be selected by left clicking and dragging a selection on that selected chromosome and the Region View responds to focus all its tracks at that selected region. The chromosome to be viewed can also be selected from the drop-down menu available on the GB toolbar. The Genome View can also display data tracks to the right of the cytoband representation for every chromosome. However, the number of data tracks that can be viewed in the Genome View is limited to one. If multiple tracks are selected to be viewed in Genome View, only the latest track is displayed. Displaying data track in Genome View allows to navigate quickly to the chromosome of interest and is particularly useful to visualize sparsely populated results. To visualize a track on the Genome View, right click on the track of interest in the Region view, select the option 'Edit track Properties' and select the appropriate checkboxes under the Genome View node of the left side panel.

Figure 6.10: Genome Browser Organization

Note:

- Data tracks in Genome View can be loaded only from the data tracks already loaded in the Region View.
- If multiple data are present in a single data track, they can be visualized in Genome View. However, multiple data tracks themselves cannot be viewed simultaneously in the Genome View.

6.3.3 Chromosome View

The Chromosome View shows a single chromosome (Chromosome 1 by default) with cytoband graphics (when the cytoband information is available for the build) and allowsthe selection of a region within the chromosome by a left click and drag. The selected region is indicated by a rectangular box and will become the new focus of the Region View. See Figure 6.11. Chromosome View can display one or more tracks of the data. This can be done by selecting the information that needs to be displayed in the 'Edit track properties' right click menu similar to the description for the Genome View.

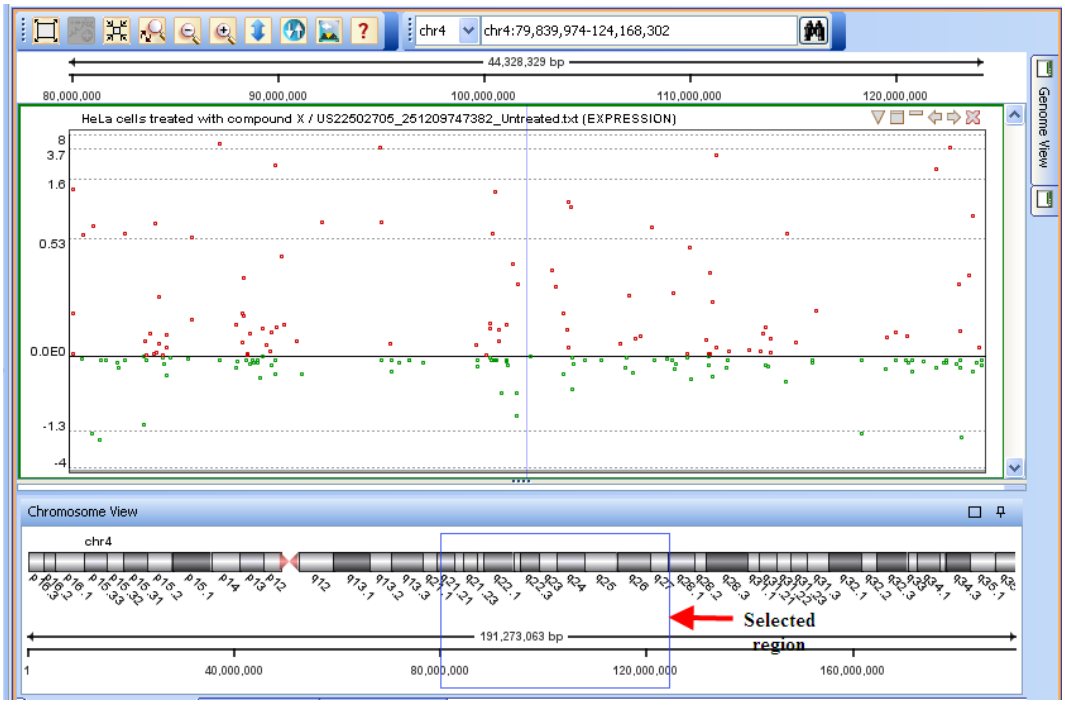


Figure 6.11: Chromosome View-Selecting the region

Data tracks in chromosome View can be loaded only from the data tracks already loaded in the Region View.

6.3.4 Annotations Tab

For all experiment types in the **GeneSpring** application, the downloaded gene/transcript/SNP annotations from Annotations Manager are accessible from the 'Annotations tab' in GB window. See Figure 6.10. Annotation tracks can be added to the GB by selecting the annotations of interest from the Annotations tab. Mouse over the Annotations tab displays a list of available annotations and double clicking on any of the entries will display a track for that annotation in the Footer panel of the Region View. Alternatively, tracks can also be added for annotations by selecting them from the list and clicking on the 'Add' button provided at the bottom of the frame. Annotations being displayed will be shown in bold in the Annotations tab.

6.3.5 Spreadsheet

Spreadsheet displays the data corresponding to the selected (active in case of single selection and first in case of multiple selections) track of the Region View in a tabular form. For data tracks where more than one plot (data type) can be loaded in the same track,(E.g.: ASCN, PSCN data in Copy Number experiments and Raw and Normalized signal values in Expression experiments) irrespective of the plots selected for viewing, a single Spreadsheet view displays data

values associated with all the plots in different columns. In Figure 6.12, though Normalized signal values are displayed in the Region View, the spreadsheet displays the Raw signal values aswell.

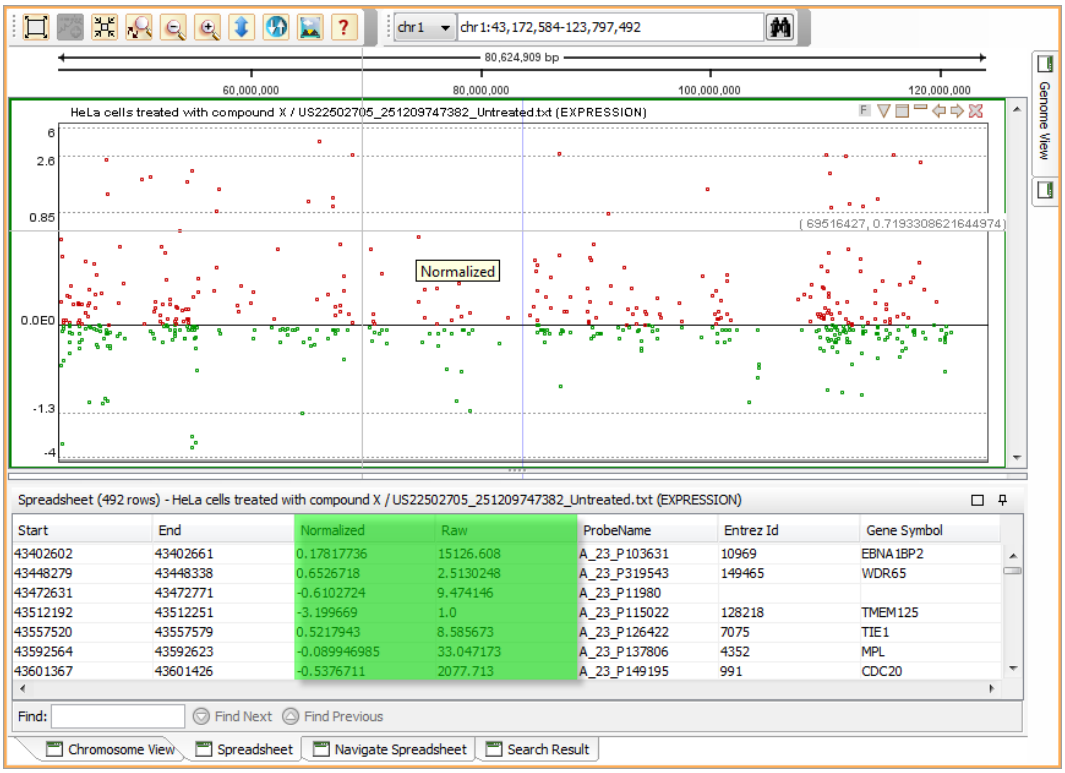


Figure 6.12: Spreadsheet View

A right click on any row displays the following options:

- **Export as Text** allows to export the rows as a text file.
- **Clear Highlight** clears the highlighting of the selected region corresponding to the selected row of the spreadsheet in all the tracks.

  'Highlight' refers to the highlighting of a certain region in the tracks displayed. A region can be highlighted by selecting a specific row or a set of rows (Shift or Ctrl key can be used to select multiple rows) in the Spreadsheet View. Single click on a row highlights the region in the track (Appears as a grey band to represent the region of selection. It will look like a thin line if the region is very small with respect to the current zoom that one is looking at.), while a double click on the row zooms in the region of interest (the region appears highlighted by a double line as the zoom is high). By default, all the tracks that are loaded are highlighted in the selected region. Highlighting can be cleared by a right click on the spreadsheet data and selecting the 'Clear Highlight' option.
- **Configure Columns** option can be used to select the data columns to be viewed in a Spreadsheet.

6.3.6 Search Result

Search Result displays the results of the search query provided in the GB tool bar. All the annotations updated using Annotations Manager i.e. gene, transcript, SNPs annotations, etc. are searchable. An appropriate search string should be specified in the text box provided on the tool bar. See Figure 6.13. The search is case insensitive and also allows the wildcard character '∗'. For e.g. the search string 'brca∗' can be used to find both 'brca1' and 'brca2'. Also, one could restrict the search to a specific region by giving the chromosome name, start and end locations in the format; chromosome name; start-end. E.g.: chr1; 1-10000. The results of the search are shown in the 'Search Result' spreadsheet. Each row in the table refers to a unique search result. While asingle click on a row highlights the corresponding region if the chromosome and region are in the view, double clicking on a row brings the selected chromosome and region into view in all the tracks.

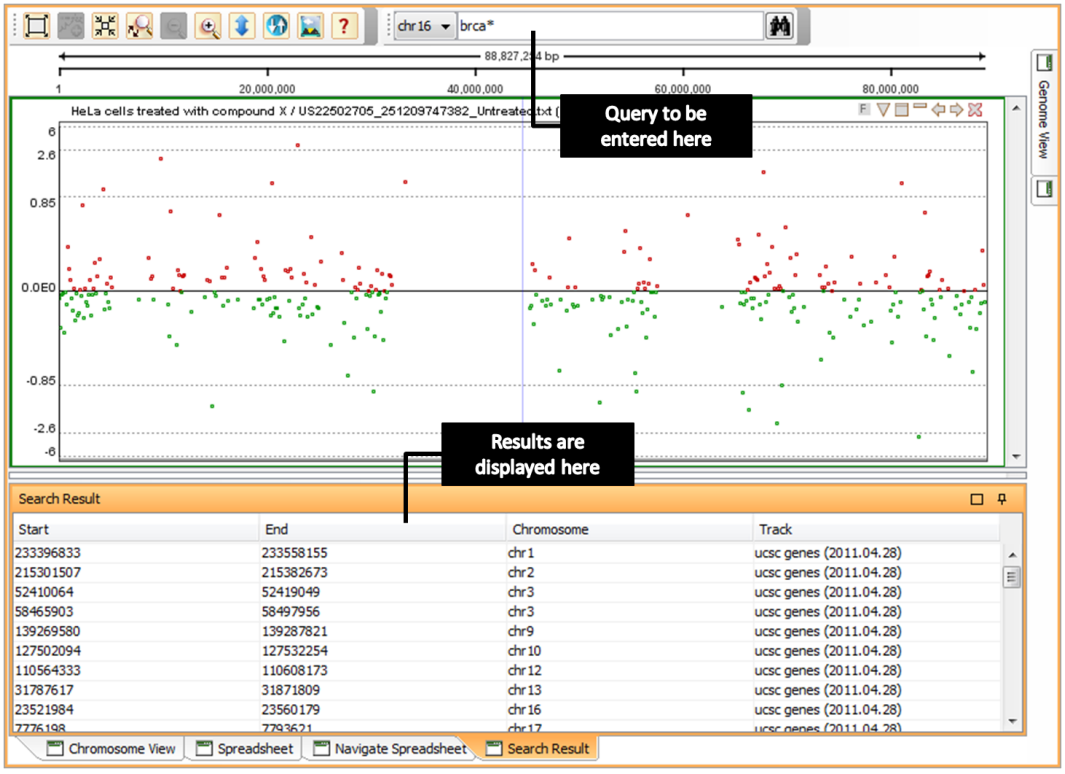


Figure 6.13: Search Result

6.4 View Customizations

Various options are provided in GB to customize the data visualization. All these options are discussed below.

6.4.1 Edit Track Properties

The properties of a track can be edited by right clicking on the desired track in the Region View and selecting the option 'Edit track properties'. This option opens a window with a panel on the left side showing three sets of nodes, one for each of the three views i.e. Region View, Genome View and the Chromosome View. See Figure 6.14. Each of these nodes has the selected data track item listed under them. By default, the data that is loaded in GB is selected (checked) under the Region View node in this window. The selected track can also be loaded (displayed) in the Chromosome View and/or Genome View in GB by selecting the appropriate options in this window (i.e. against the options listed under Chromosome View and Genome View nodes).

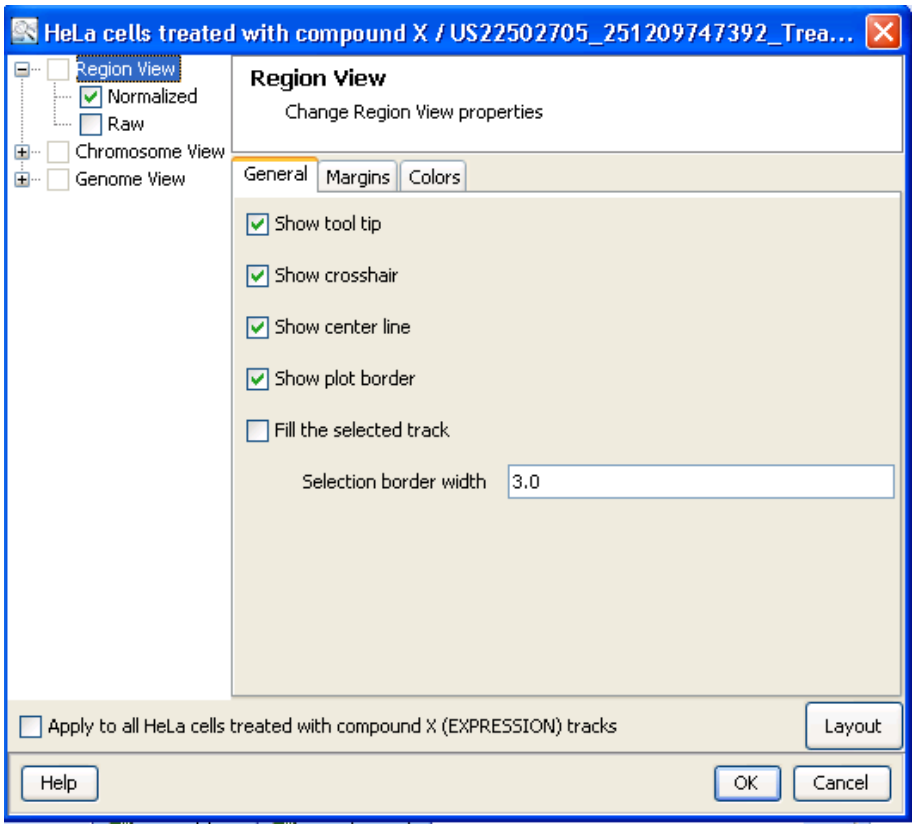


Figure 6.14: Edit Track Properties Window

For tracks which have multiple data types associated with it, by default one of the data plots is selected for viewing. To enable the other values, appropriate items (options) should be selected from the Region View node in the Edit track properties window.

Clicking on the headings Region View / Chromosome View / Genome View in the left side panel will show the edit options available for that view on the right side under three tabs: General, Margins and Colors.

- Under the 'General' tab, options to turn on/off the display of tool tip (which displays the data labels on scrolling over the plot), crosshair, and center line and plot border are available.

User can also specify if the selected track needs to be filled in (colored). There is a text box option also to specify the width of the selection border (which gets displayed when a track is selected by clicking at any point on the track).

- Under the 'Margins' tab, the borders of the plot area within the track can be specified by specifying the offsets on all the four sides.
- Under the 'Colors' tab, colors can be specified for crosshair, background, zoom band, track selection and highlight. Zoom band refers to the band that is shown when a region is selected on the track for zooming in.

Note:

All these view level properties are applied to all the tracks present in that specific view.

Selecting the name of an item (data plot) listed under the Region View / Chromosome View / Genome View node shows the edit properties available for the plot on the right side. Depending on the data plot displayed (whether the data is from experiment, entity list from analysis or an annotation data), the properties are broadly divided into three tabs: General, Ruler and Rendering. The various properties that can be edited under each of these tabs are detailed below:

- **General**
  - **Show Labels:** The display of labels for each data point represented in the track can be turned on/off. Labels can be actual values or the names of the data points. For e.g.: the labels can be the actual signal values for Raw or Normalized expression data, for an entity list, labels can be the name of the entity or any list associated data. The data type can also be labeled using 'Label by' option present under 'Show Labels'.
  - **Apply Sampling:** - When Sampling is not applied, each data point along the visible region of the chromosome is shown in the GB view. Except in very high zoom, this will result in plotting multiple values at a pixel and can be confusing. To avoid these issues, instead of plotting the multiple values falling at a pixel, Sampling derives a single value from these multiple values and displays only that. The drop down to choose the 'Sampling type' indicates the way to arrive at the single value.
  - **Apply Smoothing** When Smoothing is applied, the plot is smoothened by displaying at every pixel the average of the values in a window around that pixel. The size of this window can be specified in the option 'Smoothing window size'.
  - **Collapse regions of track:** By default, if there are any overlapping regions, GB displays them one below the other. Applying 'Collapse regions of track' clubs the overlapping regions and displays them as a single row.
- **Ruler** (Only for non-region based plots)
  - **Show ruler:** Check box to indicate if the ruler (y-axis) should be shown or not.
  - **Show ruler lines:** Check box to indicate if the ruler lines indicating the y-axis divisions should be shown or not.
  - **Linear ruler:** A linear ruler will divide the region above the reference and below the reference into equally spaced divisions. If 'Linear ruler' option is unchecked, the plot will be non-linear; the regions will be divided so as to show a maximum zoom in the area just above and below the reference and this area will progressively decrease as you move away from the reference line.

- **Maximum:** The value on the y-axis above which the view is cut-off.
- **Minimum:** The value on the y-axis below which the view is cut-off.
- **Reference lines:** The horizontal line used as a reference to plot the data. Data values above and below reference lines are colored differently for easy visual depiction.
- **Divisions above reference:** Number of divisions into which the view above the reference line is divided.
- **Divisions below reference:** Number of divisions into which the view below the reference line is divided.
- **Auto calibrate:** Auto calibrate option sets the maximum and minimum values in the ruler based on all the data values for the track.
- **Color for data above reference:** The color with which the plot above the reference line needs to be drawn.
- **Color for data below reference:** The color with which the plot below the reference line needs to be drawn.

- **Rendering**
  - **Plot:** From the drop down,a desired plot type can be chosen for viewing the data in GB.
  - **Show profile points:** Check this box to indicate if the points representing the data also need to be shown along with the line connecting them in the profile plots.
  - **Point Size and Point shape:** The size and shape of the data points in scatter/profile plots can be specified.
  - **Profile plot fill color:** Color to be filled between the plot and the reference line.
  - **Enable Density Plot:** This displays the density distribution of any data across a chromosome in the selected track. In general, 'blue' color indicates less density, 'yellow' indicates medium and 'red' indicates maximum density of data in a given chromosome.
  - **Draw region borders:** This option can be used to display the border for a region.
  - **Fill regions:** This option can be used to fill color inside a region.
  - **Color by:** This option can be used to color entities using the associated data values.

The 'Layout' button at the bottom of the Edit track properties window can be used to reorder, 'Overlay' or 'Split' multiple data types (features) that are launched in the same track. Selecting the Layout button opens up a window titled, 'Customize Features Layout'. See Figure 6.15. Selecting the features (using CTRL or Shift keys) enables the 'up' and 'down' arrows that can be used to reorder the tracks loaded in the Region View. Overlay button can be used to overlay the multiple data types (Normalized and Raw signal values in the above figure) on top of each other in the track. Clicking OK refreshes the GB with the overlaid view. Selecting the Layout button in an already overlaid track gives an option to the user to split the data types using 'Split' option.

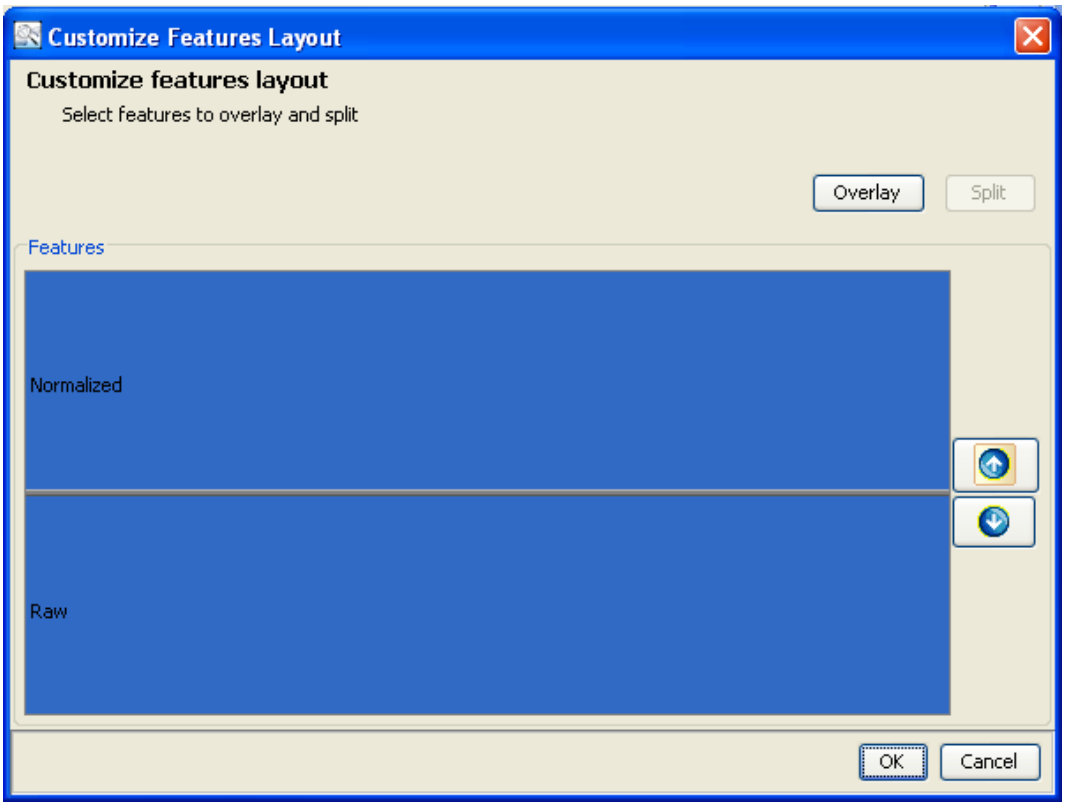


Figure 6.15: Customize Features Layout

6.4.2 Other Right Click Options of a Selected Track

- **Change track size:** Track size can be altered by right clicking on the selected track and choosing the option 'Change track size'. This allows specifying the track size using a slider/text box. The tracks to which the specified size needs to be applied can be chosen from the options displayed. The option 'Fit all tracks to frame' ignores the track size and fits all the tracks to the GB frame with equal size (minimum size of 50) for all the tracks.
- **Export as:** This gives an option to export as text and image. File chooser window comes up for text and image export.
- **Remove track:** A track can be removed by right clicking on a track in the GB view and selecting the option.
- **Filter options:** Data in a track can be filtered to retain only the desired data points in a track. Right clicking on a track and selecting 'Filter Options' opens up a window wherein the conditions for the filter can be specified. 'Column name' and the 'Condition' can be selected from the drop down menus provided. The list of conditions chosen depends on the type of the data track selected. 'Value' can be entered as text in the box provided. Conditions can be added or removed using the 'Add/Remove' buttons provided at the top. The result of the filter will be the set of rows that satisfy the specified conditions and those are displayed in GB.

6.5 Track Operations

Toolbar operations at the top of the GB window allows carrying out various track operations as listed below:

- **Full Screen:** This icon can be selected to view a full screen mode of all the tracks launched in the Region View.
- **Merge:** This icon can be used to merge multiple data tracks selected. Using the CTRL or the SHIFT option, multiple data tracks can be selected. This enables the Merge icon only when all the selected data tracks have only one kind of data type in them. Right clicking on a merged track provides an option to split the merged tracks.

  Note:

  - While Overlay is used for multiple data types of the same track, Merge option is useful for a single data type across multiple tracks.
  - Edit track properties option is not available for the merged tracks. However, its available for an overlaid track.
- **Fit to View:** This icon adjusts the size of the loaded tracks such that all are visible on the Region View with a minimum size of 50.
- **Full View:** Full View icon and Zoom out icon.
- **Orientation:** By clicking Orientation icon, orientation view can be changed to make it horizontal or vertical.
- **Reorder Tracks:** The order of tracks shown in the view can be altered by selecting the Reorder Tracks icon.
- **Export Selected Track as Image:** One or more tracks can be exported as an image, maintaining the order of tracks. Select the tracks for exporting and click on the 'Export Track as image' icon.
- **Help:** The Help icon directs the user to the appropriate sections of the **GeneSpring** documentation.

6.6 Elastic Genome Browser

This feature is available for all **GeneSpring** experiments, but particularly useful to imported **Strand NGS** experiments.

Many NGS data analyses software generate predictions of novel splice junctions, gene fusions, and genome structural variants. However, visually verifying these results remains a challenge since the predictions involve large genomic regions or even multiple genomic regions present on different chromosomes. Traditional NGS genome browsers can display detailed read information when

zoomed in, but display only summary information in the form of coverage plots when zoomed out. Moreover, NGS genome browsers can typically display only one genomic region in detail.

The new elastic genome browser allows a user to simultaneously view multiple regions of the genome with detailed read level information. This is not possible with common genome browsers which have only controls such as jump (allows you to jump to a specified region of interest), pan (shifts the region of interest) and zoom (shows a selected portion of the current view in greater detail). However, these controls are not sufficient to verify results involving multiple genomic regions.

Figure 6.16 shows a novel splice junction spanning a 42 Kbp region. The genome browser in this figure only displays the coverage track since the region under investigation is large. Even if the reads were to be rendered, the size of the exons on the screen would make it difficult to verify that the spanning reads agree perfectly with the exon boundaries.

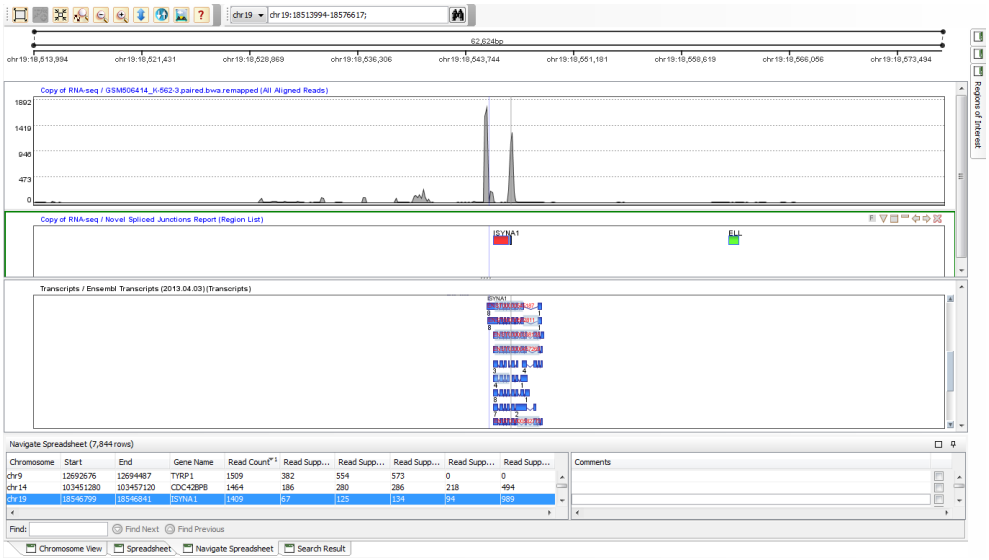


Figure 6.16: Novel splice junction discovered by **Strand NGS** in the ISYNA1 gene

Figure 6.17 displays the exact same genomic region in the Elastic Genome Browser (EGB); the two distant regions containing the exons 4, 5 have been stretched to occupy a larger screen width and uninteresting upstream and downstream regions have been compressed. The EGB clearly shows that the novel splice junction prediction is indeed accurate due to a part of exon 5 being retained which is shown in the novel detection report track.

Figure 6.17: An elastic view of the ISYNA1 gene in which exons 4,5 alone are shown in great detail. The elastic band located at the top of the browser contains the pin controls that govern the resolution at which different genomic regions are displayed in the viewing area

The EGB view is controlled by pins which define a transformation between the logical space of genome co-ordinates and the screen space. Some of the operations that are possible in EGB are

- Pins can be added, moved and removed.
- They can be added at the desired genomic region using **Right/left click⟶Add Pin** at the top of the browser showing the displayed genomic size.
- This would create a pin pair and the line connecting the pins shows the genomic region size it represents from the left hand side.
- Many pins can be added to a single view and all these would also be visible in the **Regions of Interest** tab in the right side of the EGB. This tab shows the chromosomal coordinates of each pin as well the space it is being assigned in the entire frame.
- A pin can be moved by left click and then dragging the mouse. This would expand this desired region at the cost of the other displayed genomic regions.
- Pins can be removed by clicking on the specific pin and selecting **Delete Pin** or from the **Regions of Interest** tab.
- Multiple pins can also be merged from the **Regions of Interest** tab.
- The move and remove operations can result in a change in the mapping function. This may in-turn trigger repainting the view.
- The first and last pin-pairs are immovable and cannot be removed.
- The user can also launch the EGB by giving the necessary genomic coordinates for the desired regions in the GB toolbar like **chr1:45241617-45242351:0.5;chr10:65241620-65242359:0.5;**.

- The EGB view is launched automatically when the **Navigate in Genome Browser** functionality is launched for the region lists resulting from *Gene Fusion Detection* and **Find Novel Spliced Junctions**. These algorithms output bookmarks customized to each result region as shown in figure 6.18.
- The user can exit the EGB by deleting all the pins or by merging all the regions from **Regions of Interest** tab or by giving just one genomic region in the search bar(e.g.,chr10:65241620-65242359;)

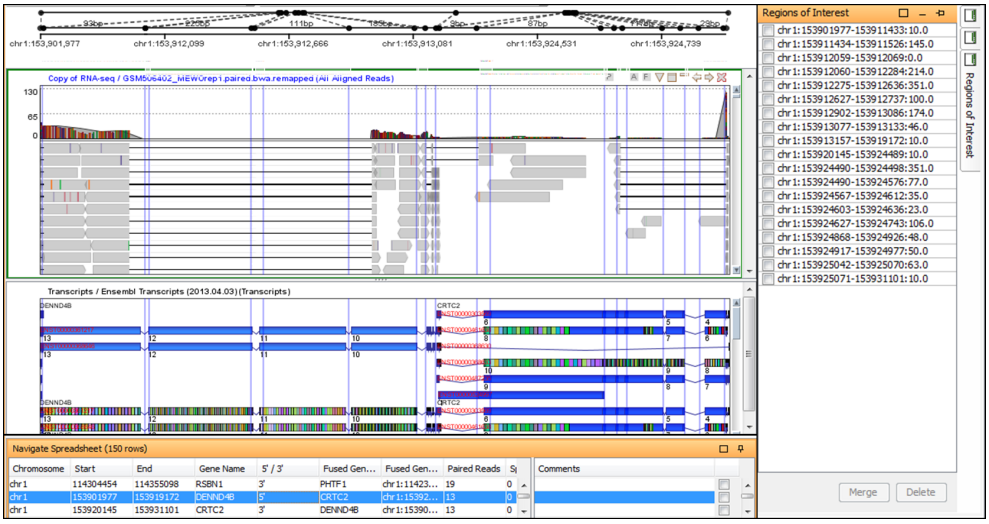


Figure 6.18: Gene Fusion discovered by **Strand NGS** between DENND4B and CRTC2 genes

An example is given below on how to compress intronic regions to visualize the reads spanning exonic boundaries. Figure 6.19 shows a transcript with two exons separated by a 100 bp intron. One possible sequence of steps that would help shorten the intron width is:

1. Add two pin-pairs demarcating the genomic region to be compressed,
2. Move the first pin right, thereby increasing the space occupied by the first exon,
3. Move the second pin left to complete the intron-compression.

A sequence of such addition and movement of pins can help compress/expand any number of sections of the genome.

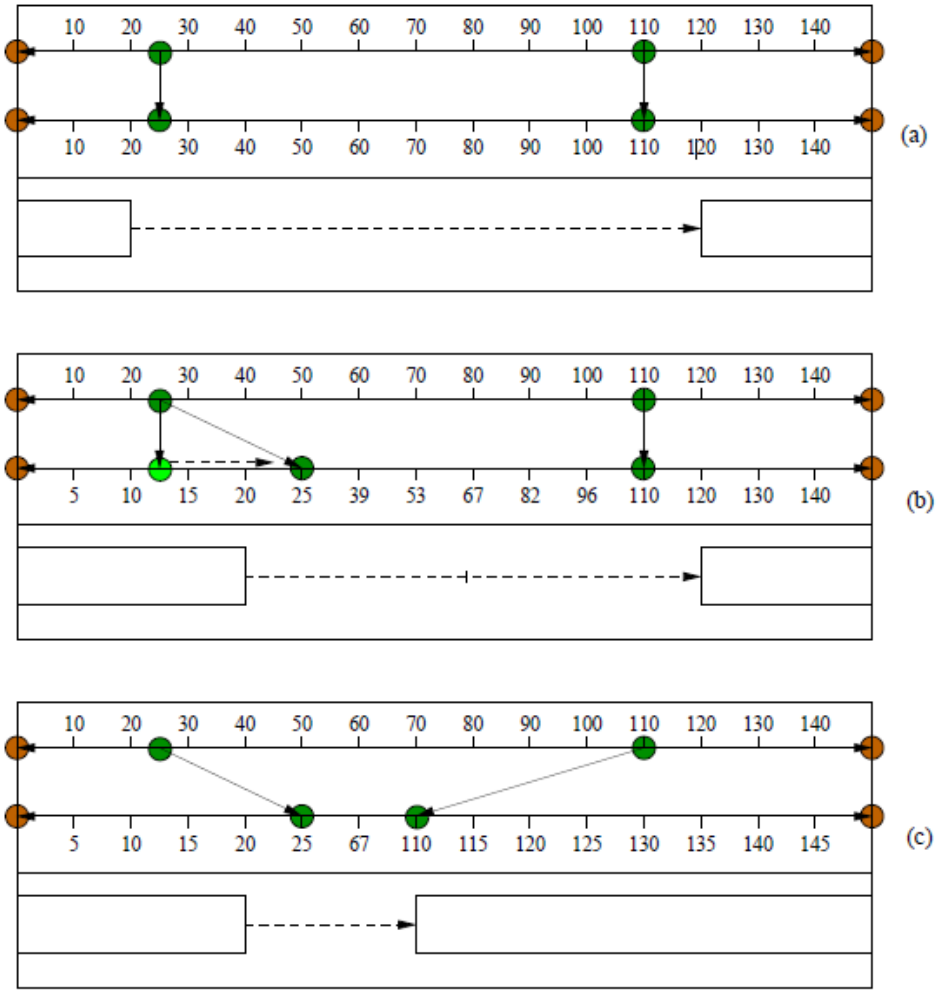


Figure 6.19: Panel (a) shows an EGB view in which two pin-pairs have been added. Panel (b) shows the view after one pin has been moved to a new screen location. Panel (c) shows the view after the second pin has also been moved

6.7 A few utilities in Copy Number Experiments

- LOH Score cannot be viewed in GB for Illumina Copy Number experiments.
- The chromosome start and end position are same for SNP and Copy Number probes. For Copy Number probes, it is taken as (Start+End)/2.
- For Chromosome Y, Birdseed algorithm does not generate calls and clusters for more than 60% of the SNP probes; and so Copy Number Analysis does not run for SNP probes for chromosome Y. Hence SNP probes are not seen when Chromosome Y is viewed in the GB.
- For Affymetrix SNP V6.0 Array experiments, the copy number is computed for all SNP and CN probes while the ASCN is computed only for SNP probes and hence there are fewer points in ASCN track. PSCN track will also show only SNP probes.

Chapter 7

Analyzing Expression Data

7.1 Overview

This chapter explains details and **GeneSpring** specific workflows 'common' to all the expression experiments.

7.1.1 Summary of experiment types

- Analyzing Agilent Single Color Expression Data
- Analyzing Agilent Two Color Expression Data
- Analyzing Affymetrix Expression Data
- Analyzing Illumina Data
- Analyzing Generic Single Color Expression Data
- Analyzing Generic Two Color Expression Data
- Analyzing Agilent Exon Expression Data
- Analyzing Affymetrix Exon Expression Data
- Analyzing Affymetrix Exon Splicing Data

7.2 Running the Expression Experiment Workflow

7.2.1 Creating the experiment

Upon launching **GeneSpring**, the startup is displayed with 3 options.

- **Create new project**
- **Open existing project**
- **Open recent project**

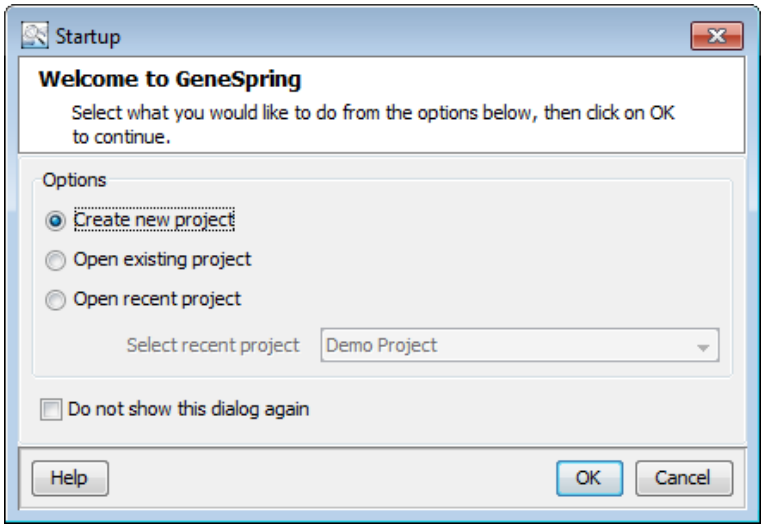


Figure 7.1: Welcome Screen

Either a new project can be created or a previously generated project can be opened and re-analyzed.

On selecting **Create new project**, a window appears in which details (Name of the project and Notes) can be recorded. **Open recent project** lists all the projects that were recently worked on and allows the user to select a project. After selecting any of the above 3 options, click on ***OK*** to proceed.

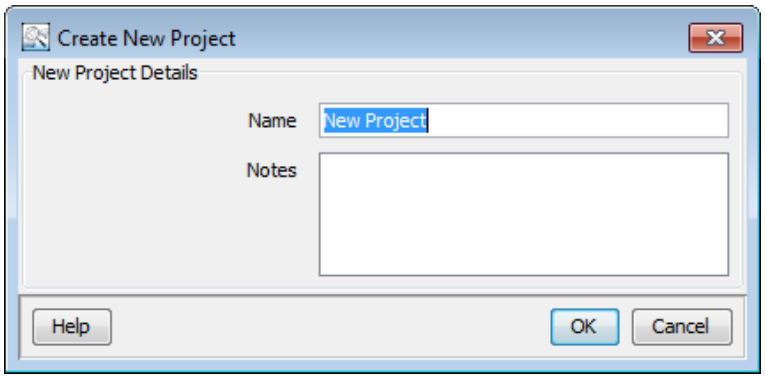


Figure 7.2: Create New Project

If **Create new project** is chosen, then an Experiment Selection dialog window appears with two options

1. **Create new experiment**: This allows the user to create a new experiment. (steps described below).
2. **Open existing experiment**: This allows the user to use existing experiments from previous projects for further analysis.

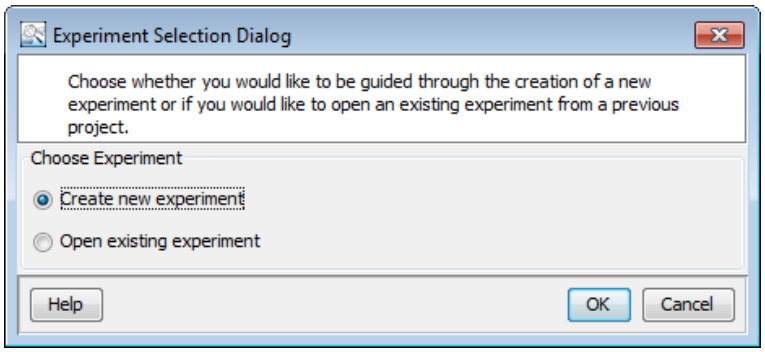


Figure 7.3: Experiment Selection

Clicking on **Create new experiment** opens up a New Experiment dialog in which **Experiment name** can be assigned. The drop-down menu for the analysis type gives the user the option to choose between multiple analysis types namely Association, Copy Number, Exon, Expression, Mass Profiler, Pathway, RTPCR and miRNA. After selecting the desired analysis type, the user needs to choose the experiment type of interest.

Next, the workflow type needs to be selected from the options provided below, based on the user convenience.

1. **Analysis: Biological Significance**

2. **Data Import Wizard**

Analysis: Biological Significance is primarily meant for a new user and is designed to assist the user through the creation and basic analysis of an experiment. Analysis involves default parameters which are not user configurable. However in **Data Import Wizard**, the parameters can be changed to suit individual requirements.

Upon selecting the workflow, a window opens with the following options:

1. Choose Files(s)
2. Choose Samples
3. Reorder
4. Remove

An experiment can be created using either the data files or else using samples. **GeneSpring** differentiates between a data file and a sample. A data file refers to the hybridization data obtained from a scanner. On the other hand, a sample is created within **GeneSpring**, when it associates the data files with its appropriate technology (See the section on Technology). Thus a sample created with one technology cannot be used in an experiment of another technology. These samples are stored in the system and can be used to create another experiment of the same technology via the ***Choose Samples*** option. For selecting data files and creating an experiment, click on the ***Choose File(s)*** button, navigate to the appropriate folder and select the files of interest. Click on ***OK*** to proceed.

The technology specific for any chip type needs to be created or downloaded only once. Thus, upon creating an experiment of a specific chip type for the first time, **GeneSpring** prompts the user to download the technology from the update server. If an experiment has been created previously with the same technology, **GeneSpring** then directly proceeds with experiment creation. Clicking on the ***Choose Samples*** button, opens a sample search wizard, with the following search conditions:

1. **Search field**: Requires one of the 6 following parameters- Creation date, Modified date, Name, Owner, Technology, Type can be used to perform the search.
2. **Condition**: Requires one of the 4 parameters- Equals, Starts with, Ends with and Includes Search value.
3. **Search Value**

Multiple search queries can be executed and combined using either *AND* or *OR*.

Samples obtained from the search wizard can be selected and added to the experiment by clicking on ***Add*** button, or can be removed from the list using *Remove* button. Files can either be removed or reordered during the data loading step using the ***Remove*** or ***Reorder*** button.

7.2.2 Data Processing

The data processing varies depending on the technology and is explained in individual chapters:

- Analyzing Agilent Single Color Expression Data
- Analyzing Agilent Two Color Expression Data
- Analyzing Affymetrix Expression Data
- Analyzing Illumina Data
- Analyzing Generic Single Color Expression Data
- Analyzing Generic Two Color Expression Data
- Analyzing Agilent Exon Expression Data
- Analyzing Affymetrix Exon Expression Data
- Analyzing Affymetrix Exon Splicing Data

7.2.3 Analysis: Biological Significance

The *Analysis: Biological Significance* wizard appears with the sequence of steps on the left hand side with the current step being highlighted. The workflow allows the user to proceed in schematic fashion and does not allow the user to skip steps.

NOTE: Analysis: Biological Significance analysis is not supported for Generic experiments, Agilent Exon experiments and Affymetrix Exon Splicing experiments.

Summary report (Step 1 of 8): The Summary report displays the summary view of the created experiment. It shows a Box Whisker plot, with the samples on the X-axis and the Log Normalized Expression values on the Y axis. An information message on the top of the wizard shows the sample processing details. By default, *Analysis: Biological Significance* workflow does RMA and baseline transformation to median of all samples. If the number of samples are more than 30, they are represented in a tabular column. On clicking the *Next* button it will proceed to the next step and on clicking *Finish*, an entity list will be created on which analysis can be done. By placing the cursor on the screen and selecting by dragging on a particular probe, the probe in the selected sample as well as those present in the other samples are displayed in green. On doing a right click, the options of invert selection is displayed and on clicking the same the selection is inverted i.e., all the probes except the selected ones are highlighted in green. Figure 7.4 shows the Summary report with box-whisker plot.

Note: In *Analysis: Biological Significance* workflow, these default parameters cannot be changed. To choose different parameters, use *Data Import Wizard*.

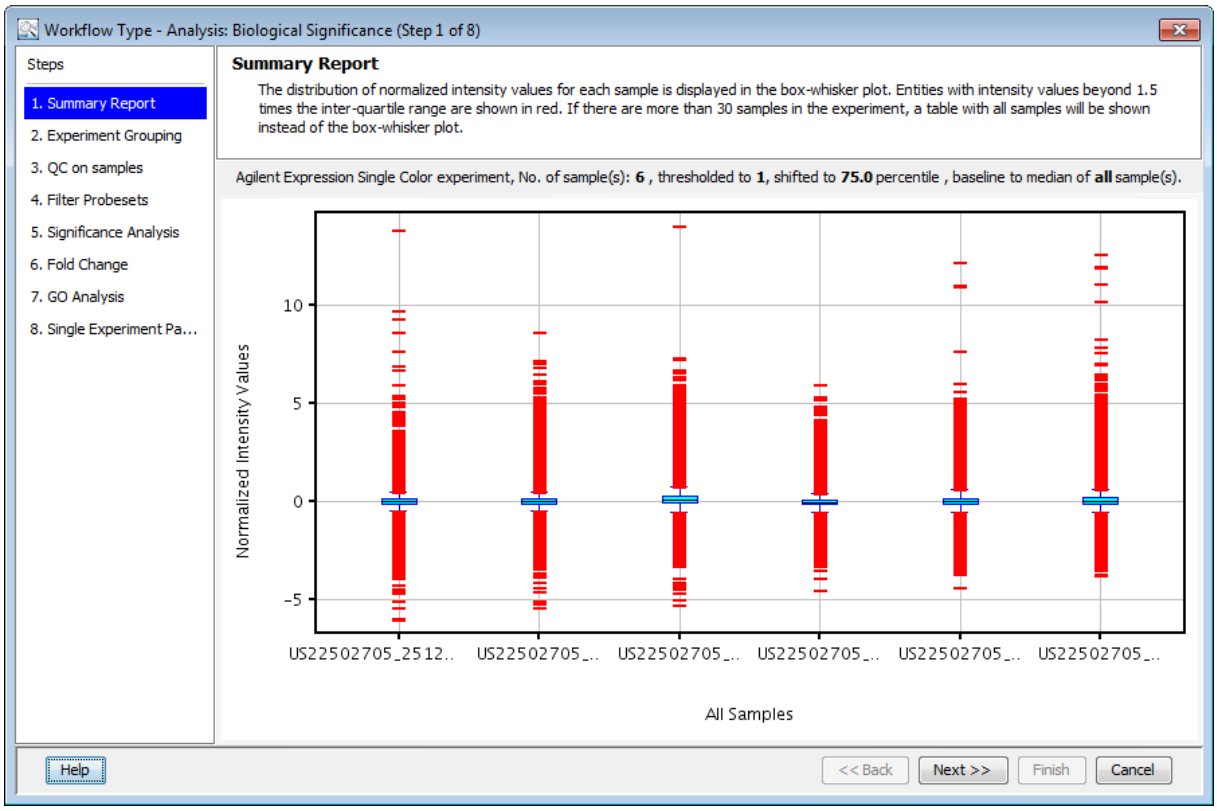


Figure 7.4: Summary Report

Experiment Grouping (Step 2 of 8): On clicking next, the 2nd step in the Analysis: Biological Significance workflow appears which is **Experiment Grouping**. It requires the adding of parameters to help define the grouping and replicate structure of the experiment. Parameters can be created by clicking on the **Add parameter** button. Sample values can be assigned by first selecting the desired samples and inputing the value. For removing a particular value, select the sample and click on **Clear**. Press OK to proceed. Although any number of parameters can be added, only the first two will be used for analysis in the Analysis: Biological Significance workflow. The other parameters can be used in the Data Import Wizard.

Note: The Analysis: Biological Significance workflow cannot proceed without grouping information.

To load experiment parameters form file, click on icon1 and navigate to the required file. The information concerning **Experiment Grouping** information should be present in a tab or comma separated text file. The file should contain a column containing sample names; in addition, it should have one column per factor containing the grouping information for that factor. Here is an example of a tab separated file.

Sample genotype dosage

A1.txt NT 20
A2.txt T 0
A3.txt NT 20
A4.txt T 20
A5.txt NT 50

Reading this tab file generates new columns corresponding to each factor.

On clicking on the icon2 the entered experiment parameters are saved to file in a tab separated text file and the same parameters can be used again in subsequent experiments by clicking on the icon1 button. On clicking on the icon3 the experiment parameters can be imported from previous samples. After selecting a parameter column, On clicking the icon4 button the selected parameter column can be moved left and on clicking the icon5 they can be moved right. On clicking the icon6 the parameter values can be reordered in a selected parameter column. Also the samples and the parameters can be grouped by a left click on the heading in the tabular column.

Unwanted parameter columns can be removed by using the **Properties** option from the right-click menu. Alternatively the **Delete parameter** button allows the deletion of the selected column. Multiple parameters can also be deleted at the same time. Similarly, by clicking on the **Edit parameter** button the parameter name as well as the values assigned to it can be edited.

Note: The Analysis: Biological Significance workflow default creates averaged and unaveraged interpretations based on parameters and conditions. It takes average interpretation for analysis in the Analysis: Biological Significance wizard.

Windows for Experiment Grouping and Parameter Editing are shown in Figures 7.5 and 7.6 respectively.

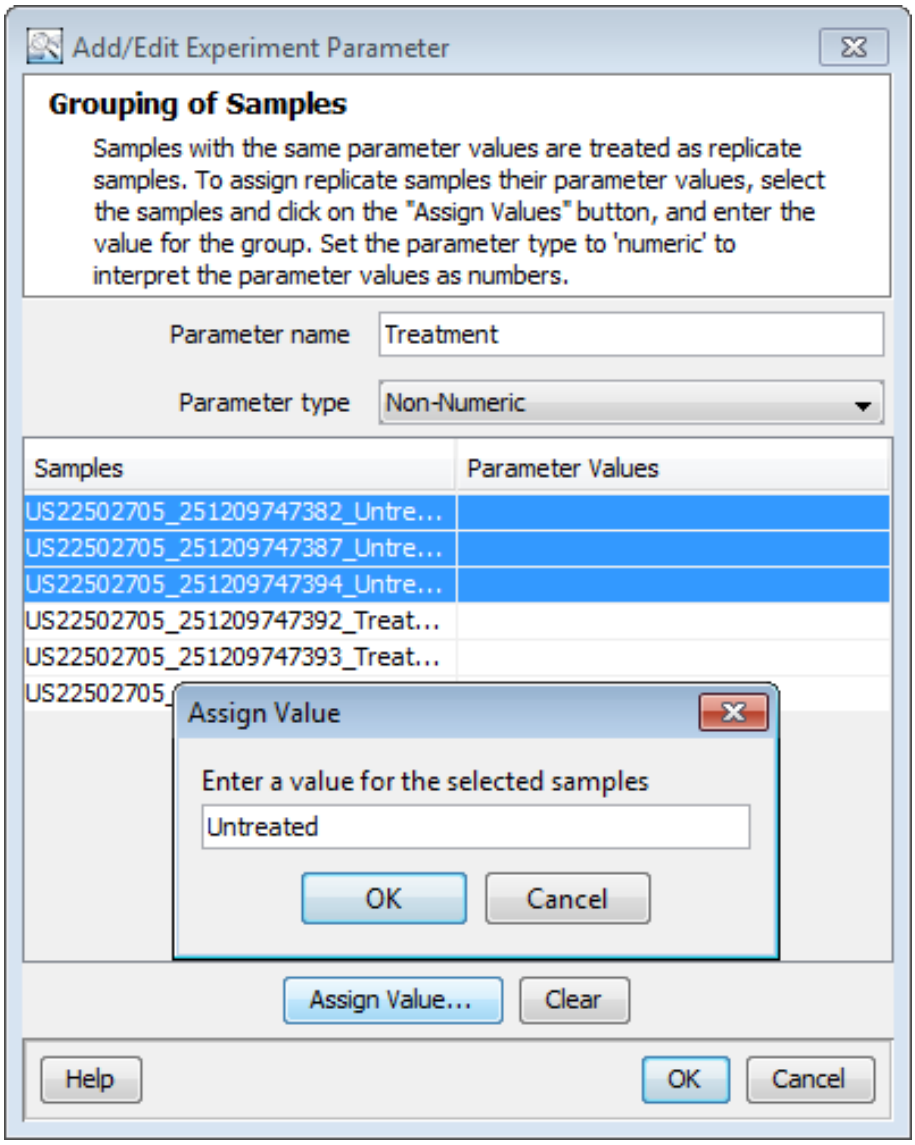


Figure 7.5: Experiment Grouping

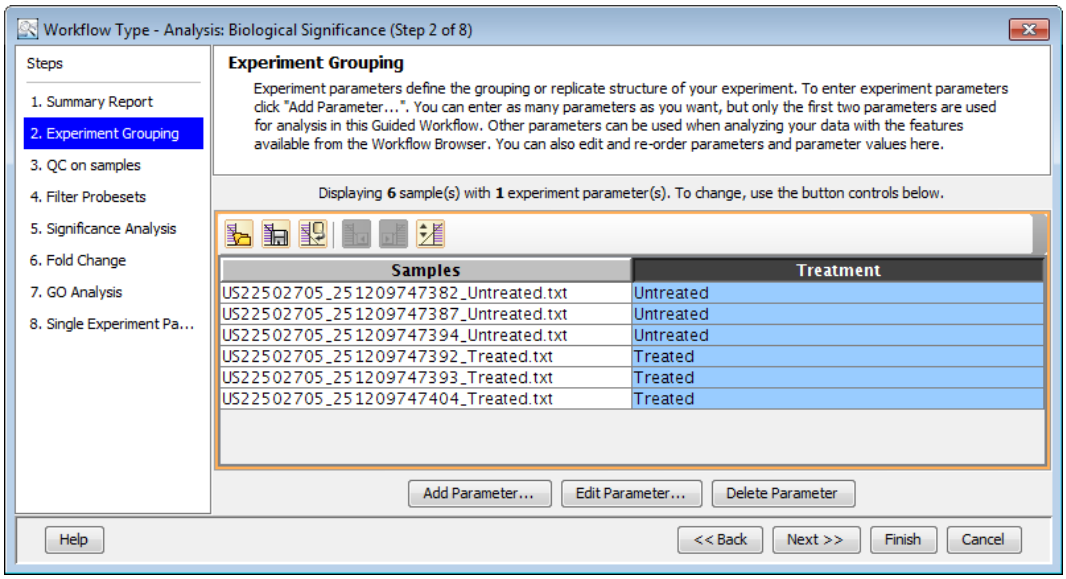


Figure 7.6: Edit or Delete Parameters

QC on Samples (Step 3 of 8): The Quality Control (QC) on samples varies with technology; see individual chapters for details.

- Analyzing Agilent Single Color Expression Data
- Analyzing Agilent Two Color Expression Data
- Analyzing Affymetrix Expression Data
- Analyzing Illumina Data
- Analyzing Affymetrix Exon Expression Data

Filter probesets (Step 4 of 8): This operation removes by default, the lowest **20** *percentile* of all the intensity values or filters the data set on existing flag values and generates a profile plot of the filtered entities. This operation is performed on the raw signal values. The plot is generated using the normalized (not raw) signal values and samples grouped by the active interpretation. The plot can be customized via the right-click menu. This filtered Entity List will be saved in the Navigator window. The Navigator window can be viewed after exiting from *Analysis: Biological Significance* workflow. Double clicking on an entity in the Profile Plot opens up an *Entity Inspector* giving the annotations corresponding to the selected profile. Annotations can be removed or added using the *Configure Columns* button on the Entity Inspector. Additional tabs in the *Entity Inspector* give the raw and the normalized values for that entity. The cut-off for filtering is set at 20 percentile, or 'Detected' and 'Not Detected' by default and can be changed using the *Rerun Filter* option (see Figure 7.8). Additional entity lists are generated with each run of the filter and saved in the Navigator. Figure 7.7 displays the profile plot obtained from an experiment grouping with one parameter.

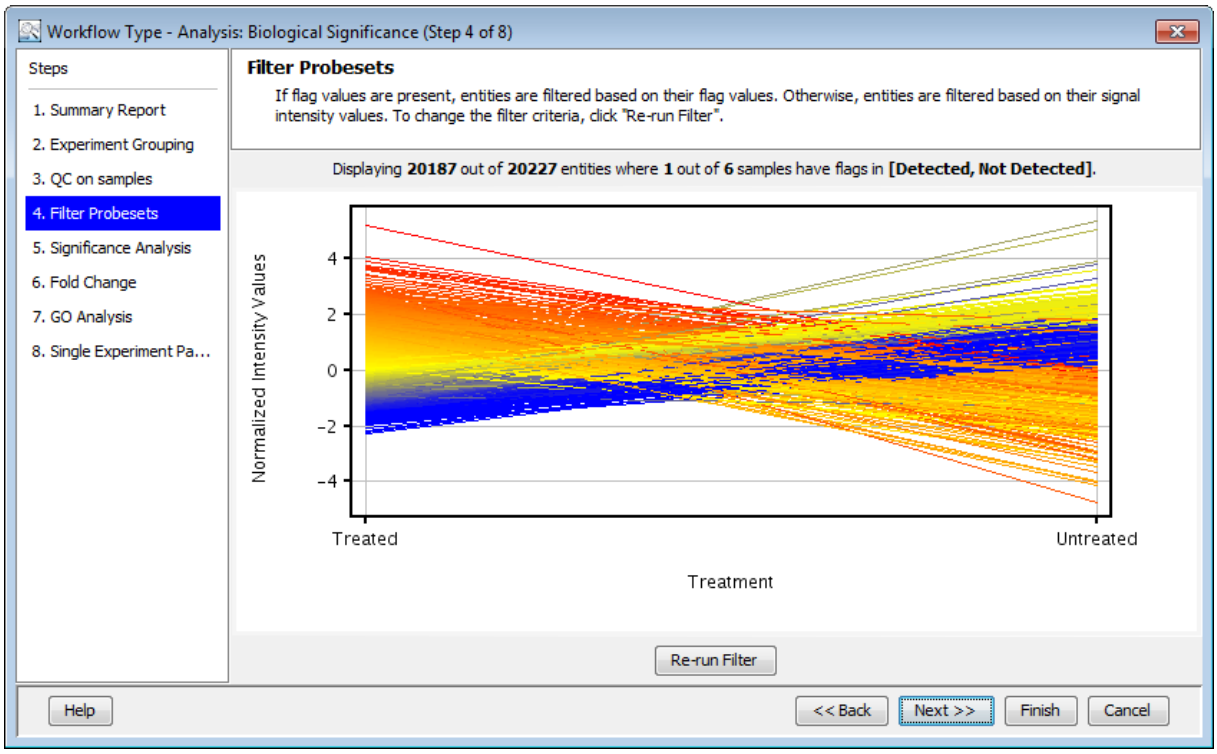


Figure 7.7: Filter Probesets-Single Parameter

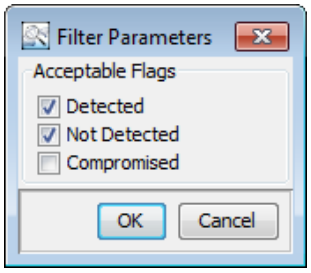


Figure 7.8: Re-run Filter

Significance Analysis (Step 5 of 8): Entities are filtered based on their p-values calculated from statistical analysis. To apply a new p-value cut-off, use the **p-value cut-off** slider at the bottom of the window. If none of the entities pass the p-value cut-off, the workflow is not able to proceed. Depending upon the sample grouping of the experiment, **GeneSpring** performs either t-test or Analysis of Variance (ANOVA). The tables below illustrate the types of statistical test performed given a specific experimental grouping with examples:

- **Example Sample Grouping I:** The example outlined in Table 15.4 has two groups, Normal and Tumor, each with replicates. In such a situation, **GeneSpring** performs a Filter on Volcano Plot and an unpaired t-test.

Samples	Grouping
S1	Normal
S2	Normal
S3	Normal
S4	Tumor
S5	Tumor
S6	Tumor

Table 7.1: Sample Grouping and Significance Tests I

- **Example Sample Grouping II:** In this example (see Table 15.5), only one group, Tumor, is present. The Filter on Volcano Plot and *t*-test against zero is performed here.

Samples	Grouping
S1	Tumor
S2	Tumor
S3	Tumor
S4	Tumor
S5	Tumor
S6	Tumor

Table 7.2: Sample Grouping and Significance Tests II

- **Example Sample Grouping III:** When three groups are present (in this example Normal, Tumor1 and Tumor2) and one of the groups (Tumor2 in this case) does not have replicates, statistical analysis cannot be performed.

NOTE In cases where some conditions do not have replicates, statistical analysis can only be performed on conditions with replicates if the samples representing the conditions with out replicates (Sample 6 in this example) are removed from the **Interpretation** used for analysis. Therefore, we recommend that you create a new **Interpretation** (excluding the condition without replicate samples) once experiment creation is complete and re-run statistical analysis using the option available from the **Workflow Browser**.

Samples	Grouping
S1	Normal
S2	Normal
S3	Normal
S4	Tumor1
S5	Tumor1
S6	Tumor2

Table 7.3: Sample Grouping and Significance Tests III

- **Example Sample Grouping IV:** When there are three groups within an Interpretation and each group has replicates (see Table 15.7), **GeneSpring** performs a one-way ANOVA.

Samples	Grouping
S1	Normal
S2	Normal
S3	Tumor1
S4	Tumor1
S5	Tumor2
S6	Tumor2

Table 7.4: Sample Grouping and Significance Tests IV

- **Example Sample Grouping V:** The Table 15.8 shows an example of an experiment where two parameters are present. The presence of two parameters means that **GeneSpring** looks for samples in all the possible parameter value pairings (in this example: Normal/10 min, Normal/50 min, Tumor/10 min, Tumor/50 min). However, there are no samples for the conditions Normal/50 min and Tumor/10 min. Due to the absence of such samples, no statistical significance test is performed.

Samples	Grouping A	Grouping B
S1	Normal	10 min
S2	Normal	10 min
S3	Normal	10 min
S4	Tumor	50 min
S5	Tumor	50 min
S6	Tumor	50 min

Table 7.5: Sample Grouping and Significance Tests V

- **Example Sample Grouping VI:** For experiments with the kind of experiment grouping illustrated in Table 15.9, a two-way ANOVA is performed.

Samples	Grouping A	Grouping B
S1	Normal	10 min
S2	Normal	10 min
S3	Normal	50 min
S4	Normal	50 min
S5	Tumor	50 min
S6	Tumor	50 min
S7	Tumor	10 min
S8	Tumor	10 min

Table 7.6: Sample Grouping and Significance Tests VI

- **Example Sample Grouping VII:** In the example shown in Table 15.10, a two-way ANOVA is performed. However, the result only contains a p-value for each parameter, *i.e.* for Grouping A and Grouping B. The p-value for the combined parameters, Grouping A- Grouping B is not computed. In this particular example, there are six conditions (Normal/10min, Normal/30min, Normal/50min, Tumor/10min, Tumor/30min, Tumor/50min), which is the same as the number of samples. The p-value for the combined parameters can only be computed when the number of samples exceeds the number of possible groupings.

Samples	Grouping A	Grouping B
S1	Normal	10 min
S2	Normal	30 min
S3	Normal	50 min
S4	Tumor	10 min
S5	Tumor	30 min
S6	Tumor	50 min

Table 7.7: Sample Grouping and Significance Tests VII

Results are displayed in the following ways:

- ***t*-test**: *t*-test unpaired is chosen as a test of choice for experimental samples as shown in Table 15.4. Upon completion of the *t*-test the results are displayed as tiled windows.
  - A ***p*-value table** consisting of Probe Names, *p*-values, corrected *p*-values, and Fold change. If an entity is 2-fold up or 2-fold down, it will be called +2.0 or -2.0 respectively. The fold change value is colored in Red or Blue by default for up-regulated and down-regulated entities, respectively.
  - The **Differential expression analysis report** containing the Test Description, *i.e.* the test that has been used for computing *p*-values, as well as the selected

type of p-value computation (**Asymptotic** or **Permutative**) and multiple testing correction. This report also provides a result summary that indicates the number of entities satisfying different p-value cut-offs.

– A **Volcano plot** is only present if there are two groups provided in Experiment Grouping. In this plot, the entities are colored depending on their passing the default cut-offs:

  * **Red**: Entities which pass both **Corrected p-value** and **Fold change** cut-offs and are up-regulated.
  * **Blue**: Entities which pass both **Corrected p-value** and **Fold change** cut-offs and are down-regulated.
  * **Grey**: Entities which neither pass the **Corrected p-value** cut-off nor the **Fold change** cut-off.

  The Volcano plot shows the negative log_{10} of p-value vs log_2 of fold change. Entities with large fold-change and low p-value are easily identifiable on this view.

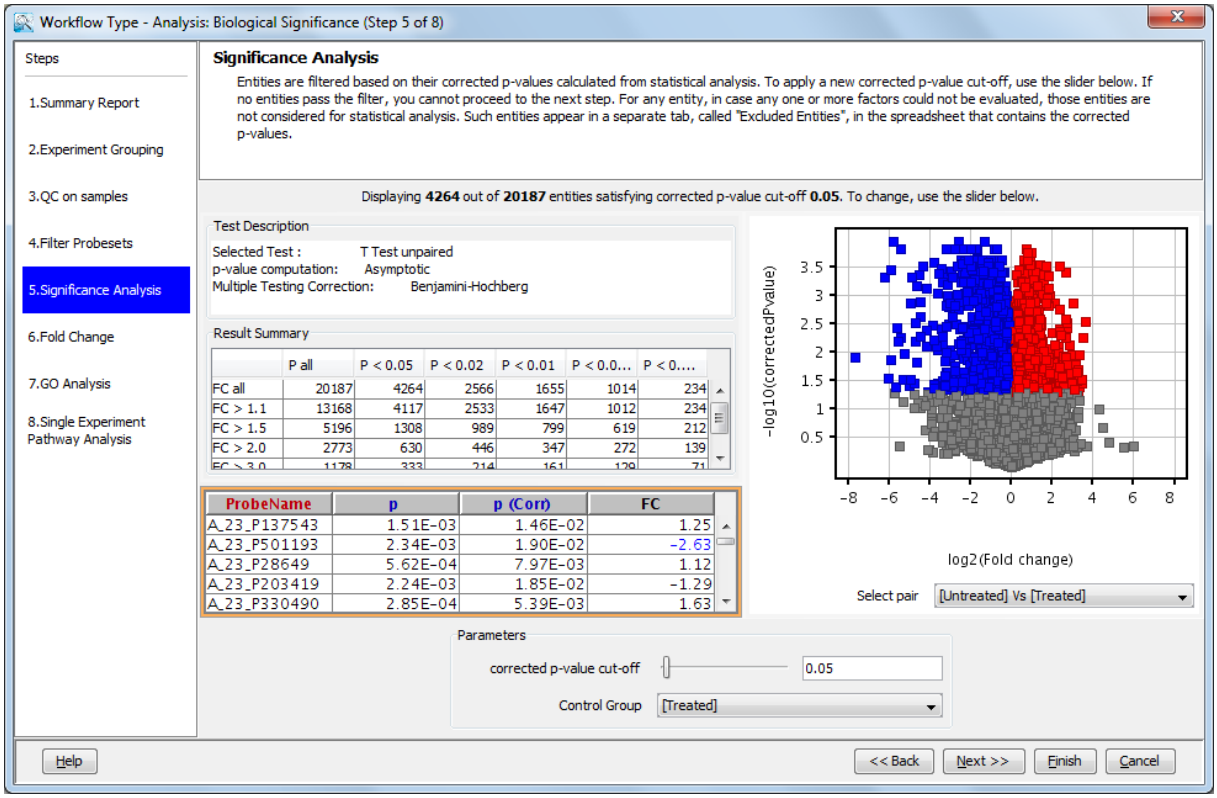


Figure 7.9: Significance Analysis

- **ANOVA**: For 2-way ANOVAs, the p-value table contains SS ratio columns in addition to the Fold change columns and the Volcano plot is replaced by a Venn Diagram:

  – A **p-value table** consisting of Probe Names, p-values, corrected p-values, SS ratio and Fold change. The SS ratio is the mean of the sum of squared deviates (SSD) as an aggregate measure of variability between and within groups.

  – The **Differential expression analysis report** containing the Test Description, *i.e.* the test that has been used for computing p-values, as well as the selected type of p-value computation (**Asymptotic** or **Permutative**) and multiple testing

correction. This report also provides a result summary that indicates the number of entities satisfying different p-value cut-offs.

– A Venn Diagram that reflects the union and intersection of entities passing the p-value cut-off. You can save any section of the Venn Diagram as a separate entity list by selecting it and clicking the **Save custom list...** button.

Fold-change (Step 6 of 8): Fold change analysis is used to identify genes with expression ratios or differences between a treatment and a control that are outside of a given cutoff or threshold. Fold change is calculated between a condition Condition 1 and one or more other conditions Condition 2 treated as an aggregate. The ratio between Condition 2 and Condition 1 is calculated (Fold change = Condition 2/Condition 1). Fold change gives the absolute ratio of normalized intensities (no log scale) between the average intensities of the samples grouped. The entities satisfying the significance analysis are passed on for the fold change analysis. The wizard shows a table consisting of 3 columns Probe Names, Fold change value and regulation (up or down).The regulation column depicts whether one of the groups has greater or lower intensity values wrt other group. The cut off can be changed using rerun analysis. The default cut off is 2. So it will show all the entities which have fold change values greater than 2. The maximum cut of can be 10. Fold change values cannot be less than 1. A profile plot is then generated. Double click on plot shows the entity inspector giving the annotations corresponding to the selected entity. An entitylist will be created corresponding to entities which satisfied the cut-off in the experiment navigator.

Note: If multiple conditions are selected for condition 2, the fold change for each of the conditions in condition 2 will be calculated.

Fold Change view with the spreadsheet and the profile plot is shown in Figure 7.10.

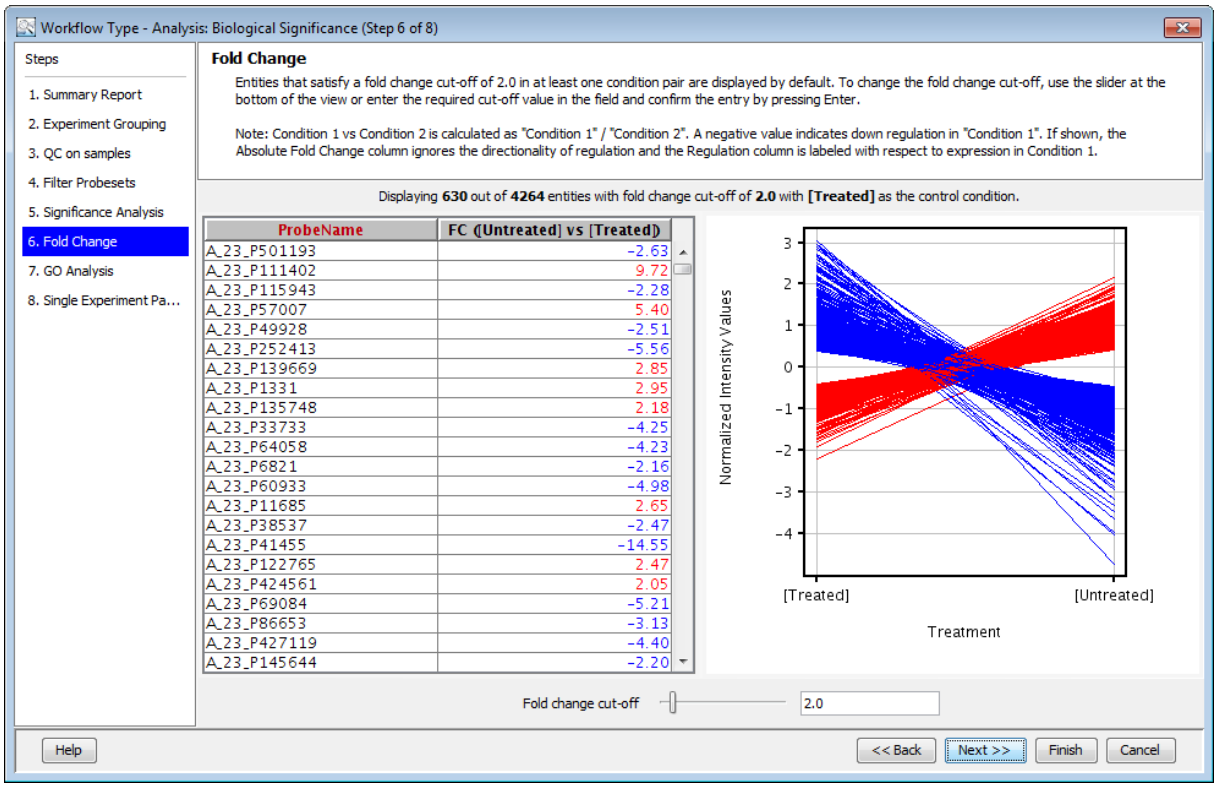


Figure 7.10: Fold Change

Gene Ontology(GO) Analysis (Step 7 of 8): The **Gene Ontology (GO) Consortium** maintains a database of controlled vocabularies for the description of molecular functions, biological processes and cellular components of gene products. The GO terms are displayed in the Gene Ontology column with associated **Gene Ontology Accession** numbers. A gene product can have one or more molecular functions, be used in one or more biological processes, and may be associated with one or more cellular components. Since the Gene Ontology is a Directed Acyclic Graph (DAG), GO terms can be derived from one or more parent terms. The Gene Ontology classification system is used to build ontologies. All the entities with the same GO classification are grouped into the same gene list.

The GO analysis wizard shows two tabs comprising of spreadsheet and a **GO tree**. The **GO Spreadsheet** shows the **GO Accession** and **GO terms** of the selected genes. For each GO term, it shows the number of genes in the selection; and the number of genes in total, along with their percentages. Note that this view is independent of the dataset, is not linked to the master dataset and cannot be lassoed. Thus selection is disabled on this view. However,the data can be exported and views if required from the right-click. The p-value for individual GO terms, also known as the enrichment score, signifies the relative importance or significance of the GO term among the genes in the selection compared the genes in the whole dataset. The default p-value cut-off is set at 0.01 and can be changed to any value between 0 and 1.0. The GO terms that satisfy the cut-off are collected and the all genes contributing to any significant GO term are identified and displayed in the GO analysis results.

The GO tree view is a tree representation of the GO Directed Acyclic Graph (DAG) as a tree view with all GO Terms and their children. Thus there could be GO terms that occur along multiple paths of the GO tree. This GO tree is represented on the left panel of the view. The panel to the right of the GO tree shows the list of genes in the dataset that corresponds to the selected GO term(s). The selection operation is detailed below.

When the GO tree is launched at the beginning of GO analysis, the GO tree is always launched expanded up to three levels. The GO tree shows the GO terms along with their enrichment p-value in brackets. The GO tree shows only those GO terms along with their full path that satisfy the specified p-value cut-off. GO terms that satisfy the specified p-value cut-off are shown in blue, while others are shown in black. Note that the final leaf node along any path will always have GO term with a p-value that is below the specified cut-off and shown in blue. Also note that along an extended path of the tree there could be multiple GO terms that satisfy the p-value cut-off. The search button is also provided on the Go tree panel to search using some keywords

Note : In GeneSpring GX 9.0 GO analysis implementation we consider all the three component Molecular Function, Biological Processes and Cellular location together. Moreover we currently ignore the **part-of** relation in GO graph.

On finishing the GO analysis, the Data Import Wizard view appears and further analysis can be carried out by the user. At any step in the Analysis: Biological Significance workflow, on clicking **Finish**, the analysis stops at that step(creating an entity list if any) and the Data Import Wizard workflow view appears.

Single Experiment Analysis (Step 8 of 8): This step in Analysis: Biological Significance workflow finds relevant WikiPathways associated with the experiment organism from the total number of pathways present in the tool based on similar entities between the pathway and the entity list. The Entity list that is used at this step is the one obtained after the fold change (step 6 of 8). This view displays a table: -

- The table displays the names of all the WikiPathways associated with the experiment organism as well as the number of entities in each pathway. It also shows the number of entities that are similar to the pathway and the entity list (matched entities) and

corresponding p-values. The p-values given in this table show the probability of getting that particular pathway by chance when these set of entities are used.

You have the option to save selected pathways using the ***Custom Save*** option. On clicking, ***Finish*** the main tool window is shown and further analysis can be carried out by the user. The user can view the entity lists and the pathways created as a result of Analysis: Biological Significance workflow on the left hand side of the window under the experiment in the **Project Navigator**. At any step in Analysis Biological Significance workflow, on clicking ***Finish***, the analysis stops at that step (creating an entity list if any). For further details on changing p-value cut-off, see To filter pathway lists by p-value cut-off.

The default parameters used in *Analysis: Biological Significance* workflow is summarized below

	Parameters	**Parameter values**
Expression Data Transformation	Thresholding	Not Applicable
	Normalization	Quantile
	Baseline Transformation	Median of all Samples
	Summarization	RMA
Filter by		
1.Flags	Flags Retained	Not Applicable
2.Expression Values	(i) Upper Percentile cutoff	100
	(ii) Lower Percentile cutoff	20.0
Significance Analysis	p-value computation	Asymptotic
	Correction	Benjamini-Hochberg
	Test	Depends on Grouping
	p-value cutoff	0.05
Fold change	Fold change cutoff	2.0
GO	p-value cutoff	0.1
SIngle Experiment Analysis	p-value cutoff	1.00

Table 7.8: Table of Default parameters for Analysis: Biological Significance workflow

7.3 Data Import Wizard

The *Data Import Wizard* in **GeneSpring** provides tremendous flexibility and power to analyze your microarray data depending upon the technology used, the experimental design and the focus of the study. *Data Import Wizard* provides several choices in terms of of summarization algorithms, normalization routines, baseline transform options and options for flagging spots depending upon the technology. All these choices are available to the user at the time of experiment creation. The choices are specific for each technology (Agilent, Affymetrix, Illumina and Generic Technologies)

and are described under the *Data Import Wizard* section of the respective chapters. Additionally, *Data Import Wizard* also enables the user to create different interpretations to carry out the analysis. Other features exclusive to *Data Import Wizard* are options to choose the p-value computation methods (Asymptotic or permutative), p-value correction types (e.g., Benjamini-Hochberg or Bonferroni), Principal component Analysis (PCA) on the entities, Class Prediction, Gene Set Enrichment Analysis (GSEA), Importing BioPAX pathways and several other utilities. The *Data Import Wizard* can be accessed by choosing the same as the *Workflow Type*, in the *New Experiment* box, at the start of the experiment creation. If the experiment has been created in a *Analysis: Biological Significance* mode, then the user does not have the option to choose the summarization, normalization and baseline transformation, i.e. the experiment creation options. However, one can still access the analysis options available from the *Data Import Wizard*, which opens up after the experiment is created and preliminary analysis done in Analysis: Biological Significance mode.

Described below are the sections of the *Data Import Wizard*:

7.3.1 Experiment Setup

Once an experiment has been created, various workflow options as described below are available from the tab 'Workflow' on the right hand side of the tool desktop.

7.3.1.1 Quick Start Guide

Clicking on this link will take you to the appropriate chapter in the on-line manual giving details about: loading expression files into **GeneSpring**, *Data Import Wizard*, the method of analysis, the details of the algorithms used and the interpretation of results.

7.3.1.2 Experiment Grouping

Experiment Grouping requires the adding of parameters to help define the grouping and replicate structure of the experiment. Parameters can be created by clicking on the *Add parameter* button. Sample values can be assigned by first selecting the desired samples and assigning the value. For removing a particular value, select the sample and click on *Clear*. Press *OK* to proceed. Any number of parameters can be added for analysis in the *Data Import Wizard*.

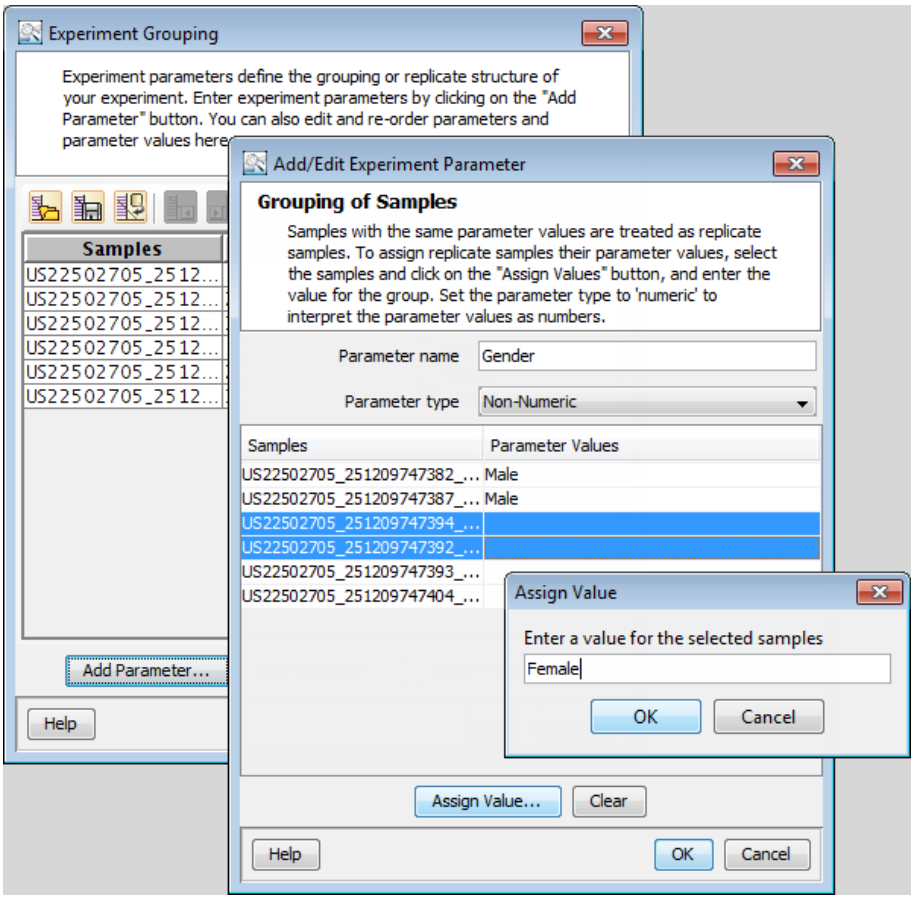


Figure 7.11: Experiment Grouping

Additionally, one can specify the parameter type to be 'Numeric' or 'Non-numeric'. Set parameter type to be 'Numeric' to interpret the parameter values as numbers. (This is particularly useful for a Time Series plot) Numeric parameters are always ordered in ascending order. By default, non-numerical parameters are ordered in alphabetical order.

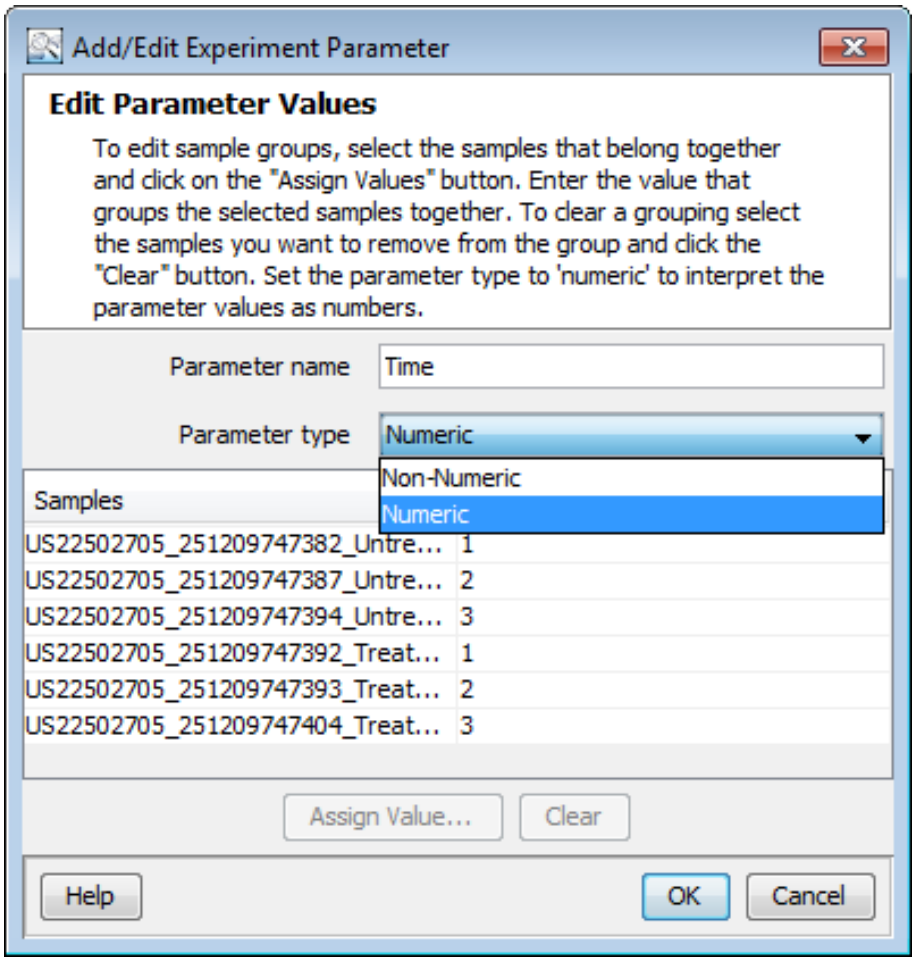


Figure 7.12: Experiment Grouping-selecting parameter type

Experimental parameters can also be loaded, using Load experiment parameters from file icon, from a tab or comma separated text file, containing the *Experiment Grouping* information. The experimental parameters can also be imported from previously used samples, by clicking on Import parameters from samples icon. In case of file import, the file should contain a column containing sample names; in addition, it should have one column per factor containing the grouping information for that factor.

Experimental parameters can be used to create interpretations (See Create Interpretation on page 240) and add Sample parameter plots to a cluster tree dendrogram created from Clustering Analysis (See How to Visualize Sample Metadata in the Dendrogram View on page 680).

Here is an example of a tab separated file.

Sample genotype dosage

A1.txt NT 20

A2.txt T 0
A3.txt NT 20
A4.txt T 20
A5.txt NT 50
A6.txt T 50

Reading this tab file generates new columns corresponding to each factor.

The current set of newly entered experiment parameters can also be saved in a tab separated text file, using Save experiment parameters to file icon. These saved parameters can then be imported and re-used for another experiment as described earlier. In case of multiple parameters, the individual parameters can be re-arranged and moved left or right. This can be done by first selecting a column by clicking on it and using the Move parameter left icon to move it left and Move parameter right icon to move it right. This can also be accomplished using the Right click $\longrightarrow$ *Properties* $\longrightarrow$ columns option. Similarly, parameter values, in a selected parameter column, can be sorted and re-ordered, by clicking on Re-order parameter values icon. Sorting of parameter values can also be done by clicking on the specific column header.

Unwanted parameter columns can be removed by using the Right-click $\longrightarrow$ *Properties* option. The *Delete parameter* button allows the deletion of the selected column. Multiple parameters can be deleted at the same time. Similarly, by clicking on the *Edit parameter* button the parameter name as well as the values assigned to it can be edited.

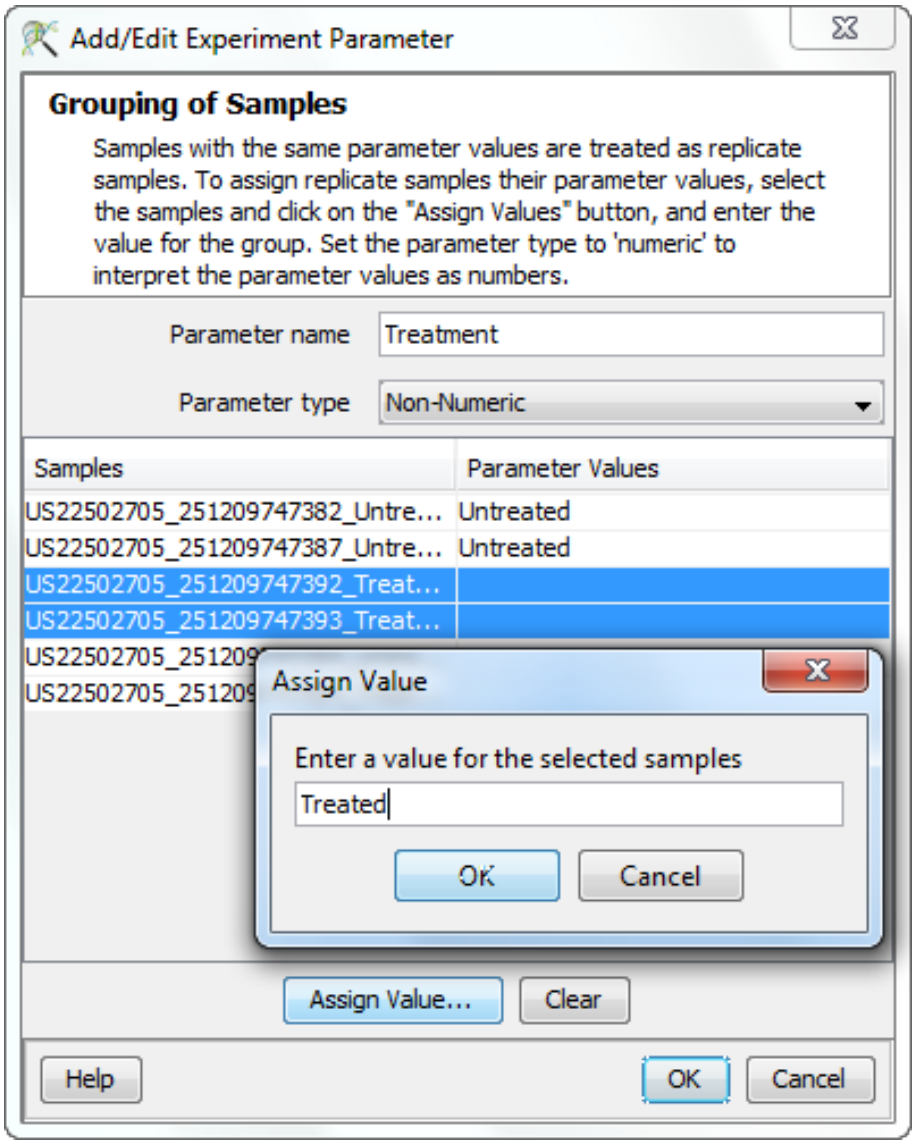


Figure 7.13: Edit or Delete of Parameters

7.3.1.3 Create Interpretation

An interpretation specifies how the samples should be grouped into experimental conditions. The interpretation can be used for both visualization and analysis. Interpretation can be created using the *Create interpretation* wizard which involves the following steps:

Step 1 of 4: Experiment parameters are shown in this step. In case of multiple parameters, all the parameters will be displayed. The user is required to select the parameter(s) using which the interpretation is to be created.

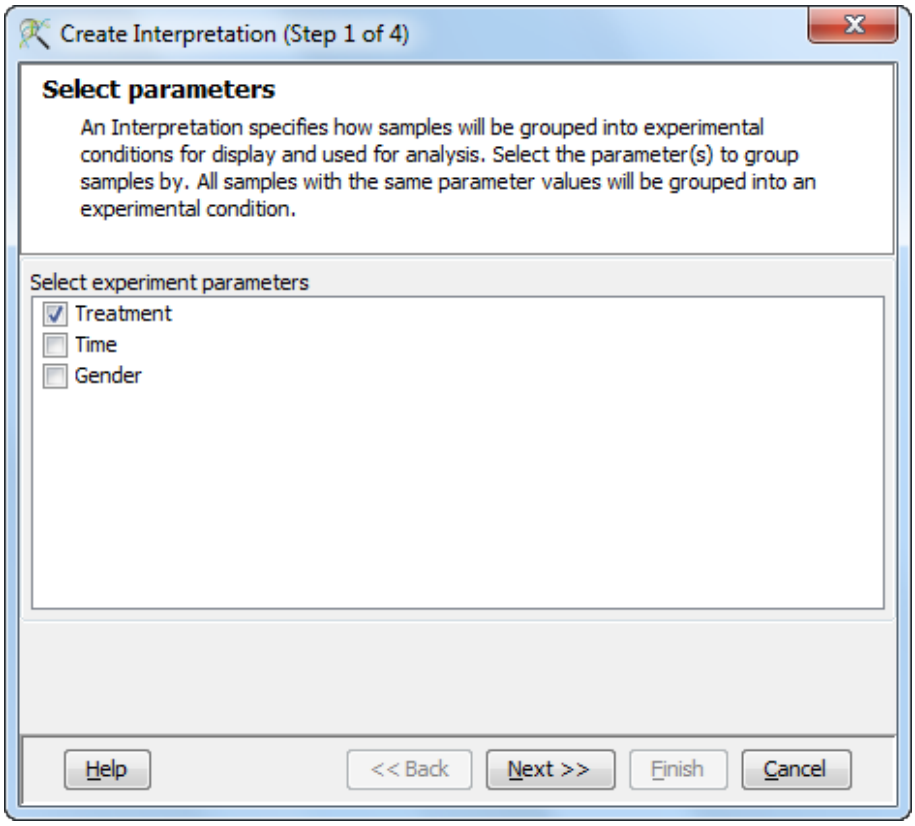


Figure 7.14: Create Interpretation (Step 1 of 4)

Step 2 of 4: The user can select one of two display modes ('Numerical' or 'Categorical') for each parameter selected in the previous step. This display mode applies to only Profile Plot and Box Whisker Plot views.

Numerical Display Mode: In this display mode, expression levels in between conditions are interpolated and each condition is connected to the next. Samples that have the same parameter values are regarded as replicates and the average values for these samples is plotted for each entity.

Categorical Display Mode: In this display mode, distinct parameter values are not connected. It shows parameter values that exist independently of one another, where each value is represented as a discrete point. When a 'Numerical' parameter is used in conjunction with a 'Categorical' parameter, the later separates the data into a series of discrete graphs. Each discrete graph contains all the values of the 'Continuous' parameter.

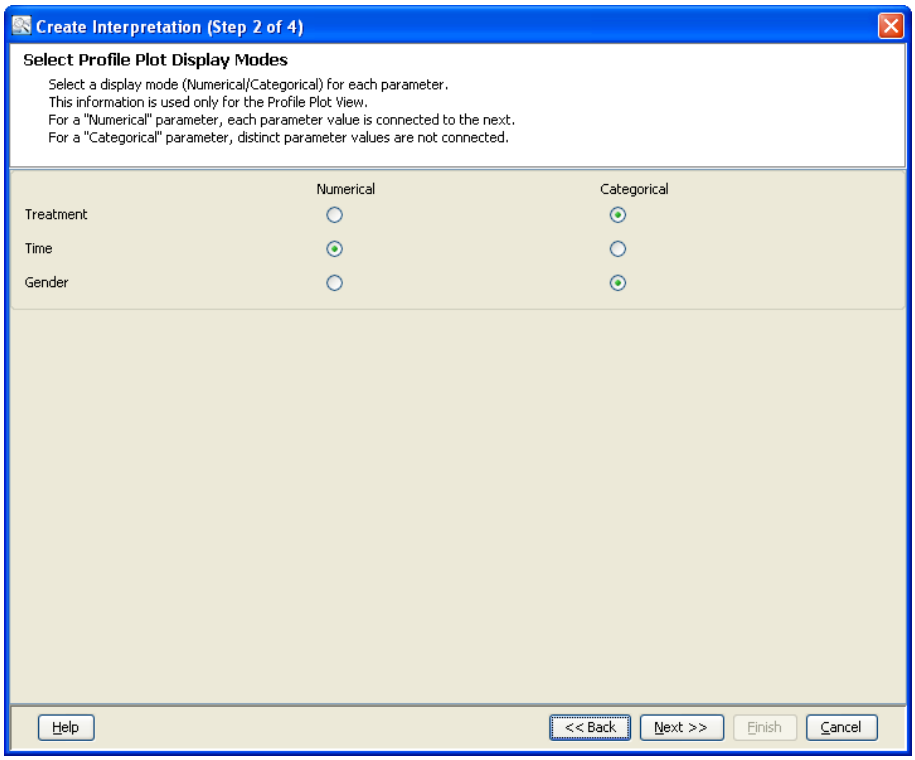


Figure 7.15: Create Interpretation (Step 2 of 4)

Step 3 of 4: Allows the user to select the conditions of the parameters that are to be included in the interpretation. All the conditions (including combinations across the different parameters) are shown. By default all these experimental conditions are selected, click on the box to deselect any unwanted conditions. Any combination of these conditions can be chosen to form an interpretation. If there are multiple samples for a condition, users can choose to create an averaged interpretation instead, or create both a non-averaged and averaged interpretation at the same time by selecting the appropriate option in the dialog. Please note that not all analyses will use an averaged interpretation. For example, while performing statistical analysis, the interpretation that is used is always the non-averaged interpretation. So even if the selected interpretation is the averaged one, **GeneSpring** automatically uses the non-averaged one for the statistical computations.

The user also has an option to exclude entities with flags while creating an interpretation. By default all the flags are included in the interpretation and in case the user wants to exclude any flags, he/she can unselect the same during the interpretation. The user can see the effect of this interpretation on the entity list by viewing the spreadsheet. The effect of excluding flag values on all the samples can be seen by viewing the unaveraged interpretation. This shows the entities and the values which have been excluded, appear blank. The spreadsheet can be viewed by selecting the desired entity list and the interpretation (the selected interpretation appears in bold and the selected entity list is highlighted). However analysis such as clustering, class prediction and PCA take all the flags into account even if specified otherwise in the interpretation.

For more information, on the effect of interpretation on the analysis as well as the way the interpretations are handled in different analysis refer to the section on Conditions and Interpretations.

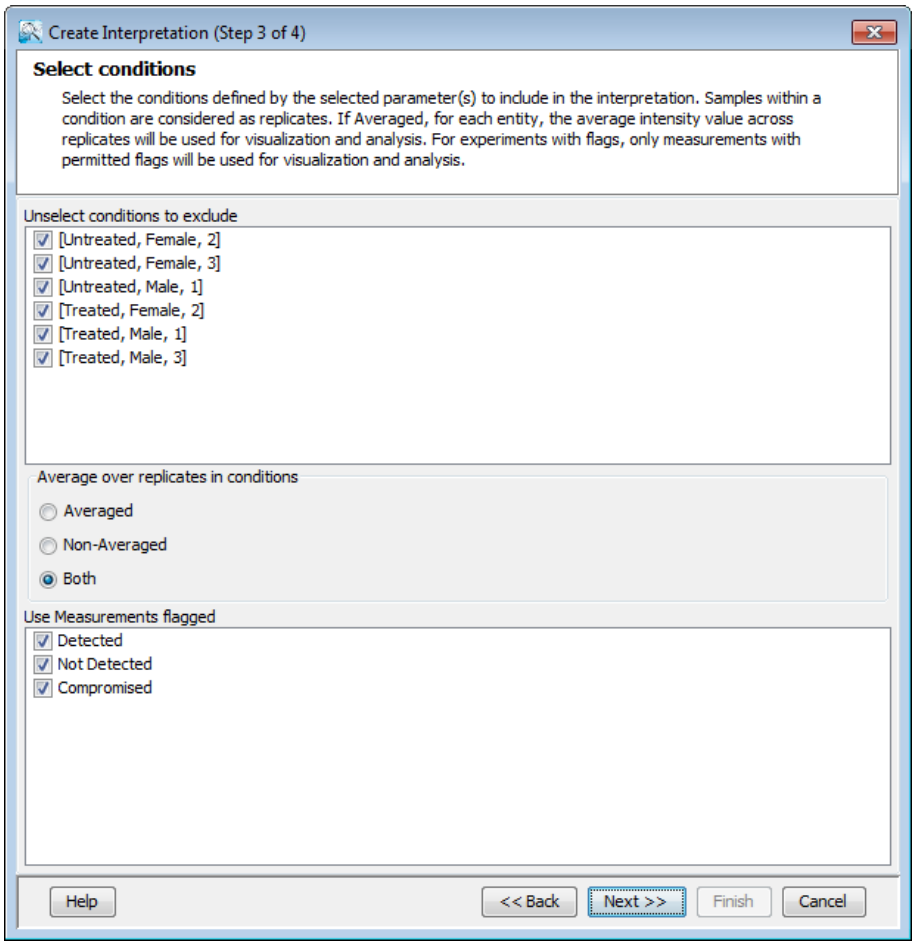


Figure 7.16: Create Interpretation (Step 3 of 4)

Step 4 of 4: This page displays the details of the interpretation created. This includes user editable Name for the interpretation and Notes for description of the interpretation. Descriptions like creation date, last modification date, and owner are also present, but are not editable.

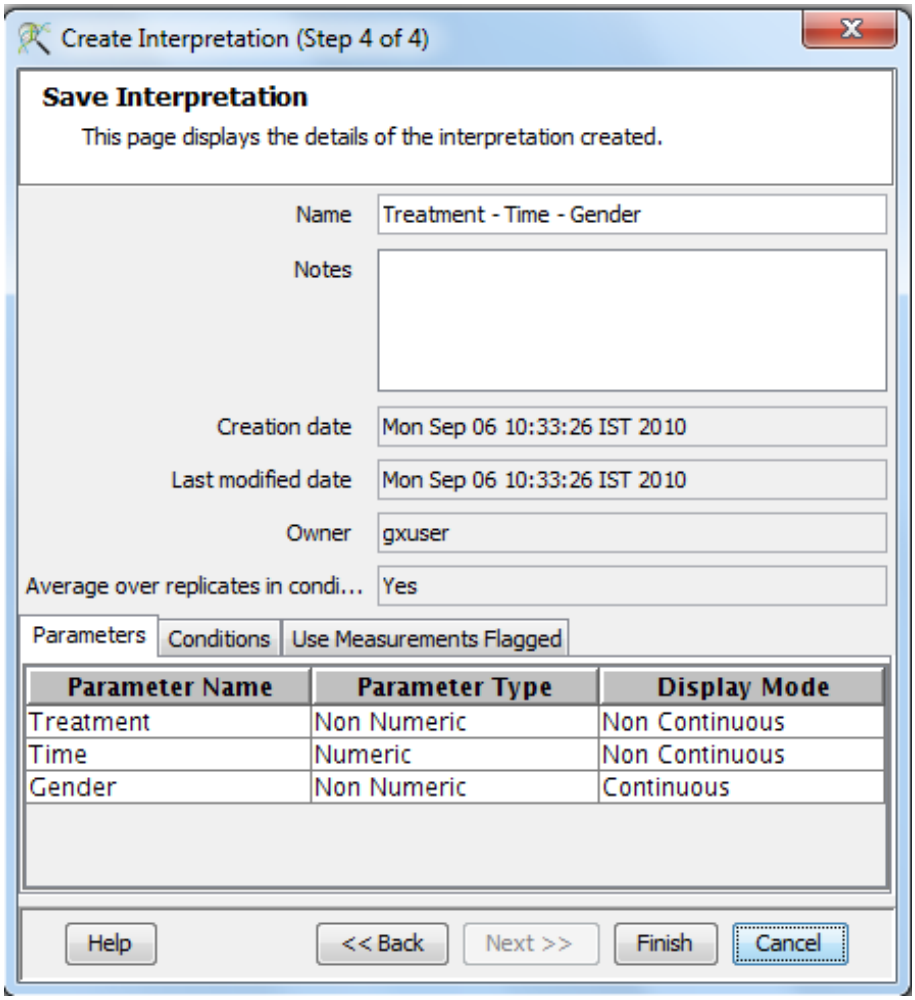


Figure 7.17: Create Interpretation (Step 4 of 4)

7.3.1.4 Create New Gene Level Experiment

Create New Gene-level Experiment is a utility in **GeneSpring** that allows analysis at gene level, even though the signal values are present only at probe level. Suppose an array has 10 different probe sets corresponding to the same gene, this utility allows summarizing across the 10 probes to come up with one signal at the gene level and use this value to perform analysis at the gene level.

Process:

- *Create New Gene-level Experiment* is supported for all those technologies where the **EntrezGeneID** column is available. It takes an entity list as the input and creates a new experiment with all the data for this entity list from the original experiment; even those probes which are not associated with an Entrez Gene ID are retained.
- The entity list taken as input is generated by the 'Filter by flags' process which is the first

step in the gene-level experiment creation. If there are no flags present, then an entity list chooser appears in stead of a 'Filter by Flags' dialog.

- The identifier in the new gene-level experiment will be the Probe IDs concatenated with the Entrez Gene ID; the identifier is only the Probe ID(s) if there was no associated Entrez Gene ID.
- Each new gene-level experiment creation will result in the creation of a new technology on the fly.
- The annotation columns in the original experiment will be carried over except for the following:
  - Chromosome Start Index
  - Chromosome End Index
  - Chromosome Map
  - Cytoband
  - Probe Sequence
- Flag information will also be dropped.
- Raw signal values are used for creating a gene-level experiment; if the original experiment has raw signal values in log scale, the log scale is retained.
- Experiment grouping, if present in the original experiment, will be retained.
- The raw signal values will be averaged by arithmetic mean over the probes (for that Entrez Gene ID) for the new experiment.
- For a Two Color experiment, raw values are summarized for each channel separately and then log ratios are taken.

Create New Gene-level Experiment can be launched from the **Workflow Browser** $\longrightarrow$ **Experiment Set up** $\longrightarrow$ **Create New Gene-level Experiment**.

The first step in creating a new gene-level experiment is to obtain an entity list that will be used to create the experiment. Towards this, the tool first launches the 'Filter probesets by Flags' workflow if there are flags present in the experiment. This four-step wizard-driven workflow varies slightly depending on the technology of the experiment from which it is launched and is described in the respective **Filter probesets by Flags** sections. The entity list that is created at the end of this process is carried forward as the input for creating the gene-level experiment.

If there is no flag information, then an entity list chooser appears instead (see Figure 7.18).

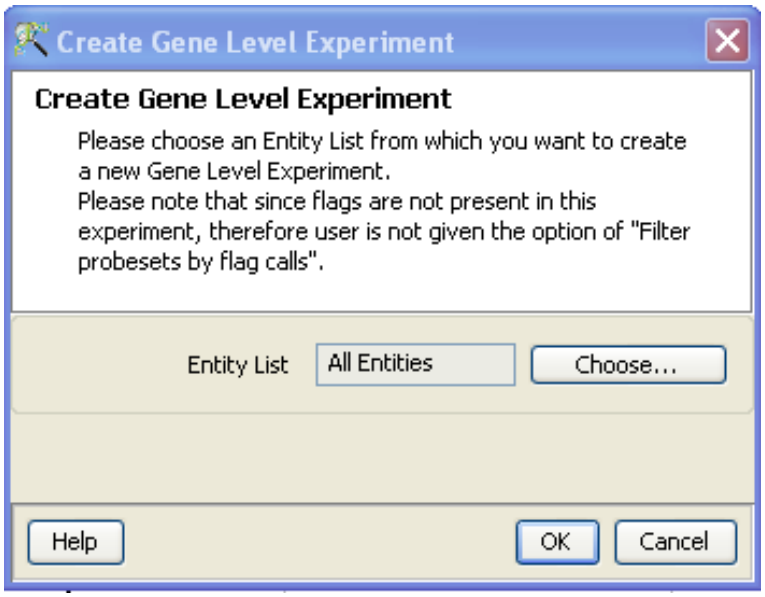


Figure 7.18: Gene-level Experiment - Entity List Chooser

Once an entity list has been obtained, an experiment creation window opens, as illustrated in Figure 7.19; the experiment name and notes can be defined here. **GeneSpring** automatically picks up and describes the entity list taken as input. Note that only the **Data Import Wizard** option is supported for gene-level experiments.

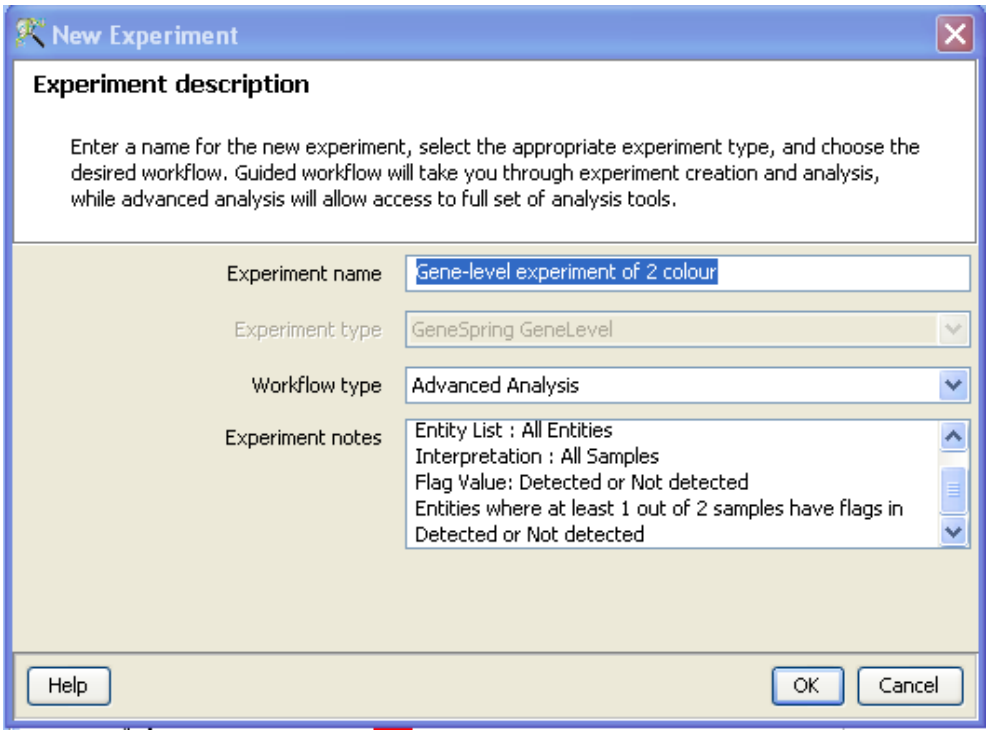


Figure 7.19: Gene-level Experiment Creation

On clicking **OK**, a three-step wizard-driven workflow opens - these steps are the same Normalization and Preprocess Baseline Option steps that are available during the Experiment Creation process, described for each technology in the respective chapters.

- Step 1: Normalization Options
- Step 2: Choose Entities
- Step 3: Preprocess Baseline Options

At the end of this operation, a new experiment titled "Gene-level experiment of [name of original experiment]" is created and all the analysis options available in the original experiment can also be carried out in this experiment.

7.3.2 Quality Control

7.3.2.1 Quality Control on Samples

Quality control (QC) is an important step in micro array data analysis. The data needs to be examined and ambiguous samples should be removed before starting any data analysis. Since microarray technology is varied, quality measures have to be vendor and technology specific. **GeneSpring** packages vendor and technology specific quality measures for quality assessment. It also provides a rich, interactive and dynamic set of visualizations for the user to examine the quality of data. Details of the QC metric used for each technology can be accessed by clicking on the links below.

- Quality Control for Affymetrix Expression
- Quality Control for Exon Expression
- Quality Control for Exon Splicing
- Quality Control for Agilent Single Color
- Quality Control for Agilent Two Color
- Quality Control for Agilent miRNA
- Quality Control for Illumina
- Quality Control for Generic Single Color
- Quality Control for Generic Two Color
- Quality Control for RealTime PCR

7.3.2.2 Filter Probesets by Expression

Entities are filtered based on their signal intensity values. This enables the user to remove very low signal values or those that have reached saturation. Users can decide the proportion of conditions must meet a certain threshold. The *'Filter by Expression* wizard involves the following 4 steps:

Step 1 of 4: Entity list and the interpretation on which filtering is to be done is chosen in this step. Click *Next*.

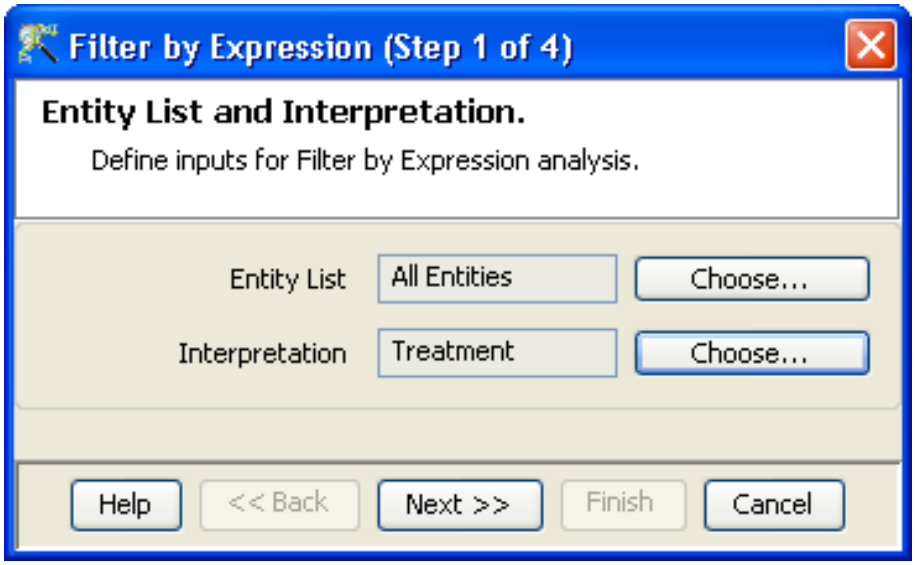


Figure 7.20: Filter probesets by expression (Step 1 of 4)

Step 2 of 4: This step allows the user to select the range of intensity value within which the probe intensities should lie. By lowering the upper percentile cutoff from 100%, saturated probes can be avoided. Similarly increasing the lower percentile cut off, probes biased heavily by background can be excluded. Stringency of the filter can be set in *Retain Entities* box. These fields allow entities that pass the filtering settings in some but not all conditions to be included in the filter results.

With two dye experiments, there are actually 2 values per entity per sample. When *Filter by Expression* is carried out on raw data with two-dye experiments, note that an entity is included in filtered results if *either or both* of the channels pass the defined cut-off.

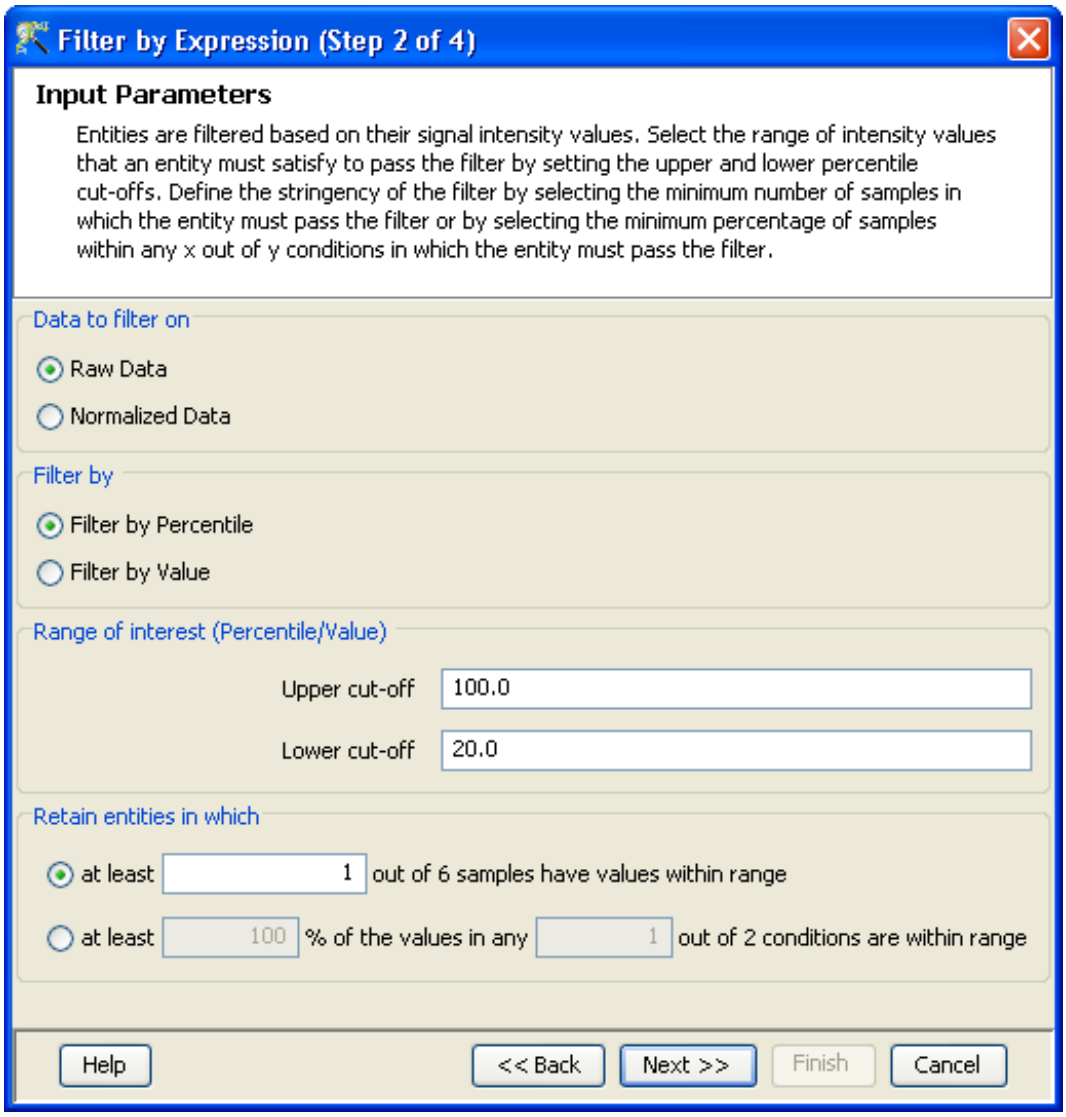


Figure 7.21: Filter probesets by expression (Step 2 of 4)

Step 3 of 4: This window shows the entities which have passed the filter, in the form of a spreadsheet and a profile plot. The number of entities passing the filter is mentioned at the top of the panel. Click *Next*.

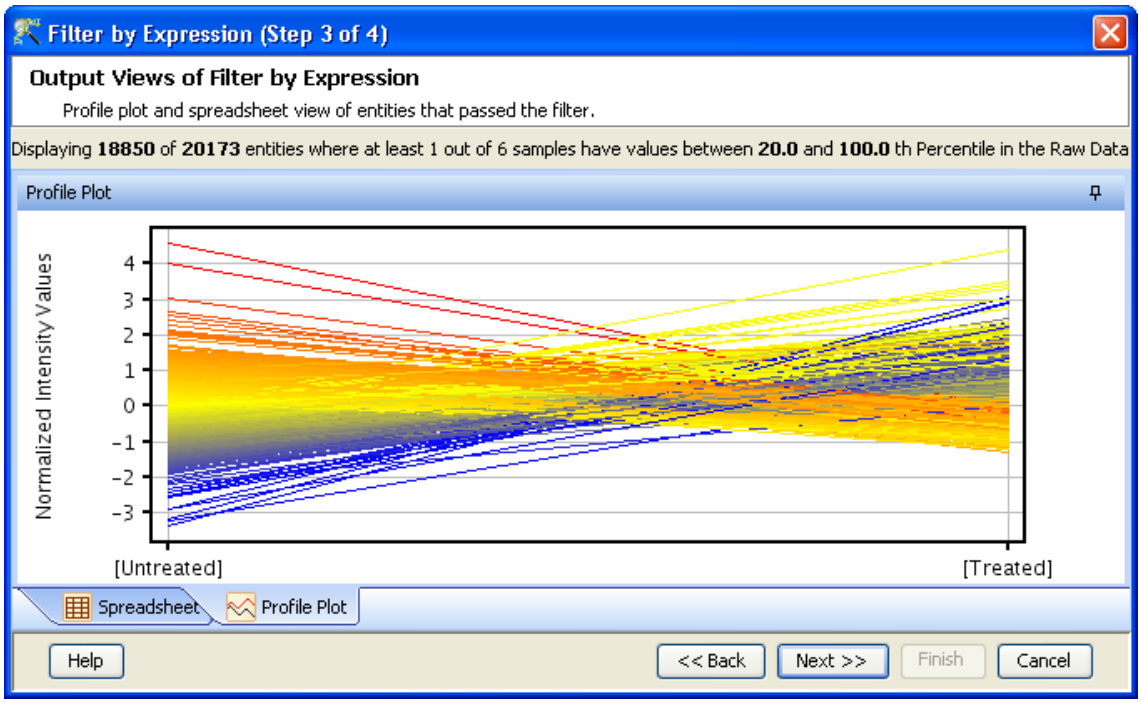


Figure 7.22: Filter probesets by expression (Step 3 of 4)

Step 4 of 4 The last page shows all the entities passing the filter along with their annotations. It also shows the details (regarding Creation date, modification date, owner, number of entities, notes etc.) of the entity list. Click *Finish* and an entity list will be created corresponding to the entities that satisfied the cut-off. Double clicking on an entity in the Profile Plot opens up an *Entity Inspector* giving the annotations corresponding to the selected profile. Additional tabs in the *Entity Inspector* give the raw and the normalized values for that entity. The name of the entity list will be displayed in the experiment navigator. Annotations being displayed here can be configured using *Configure Columns* button.

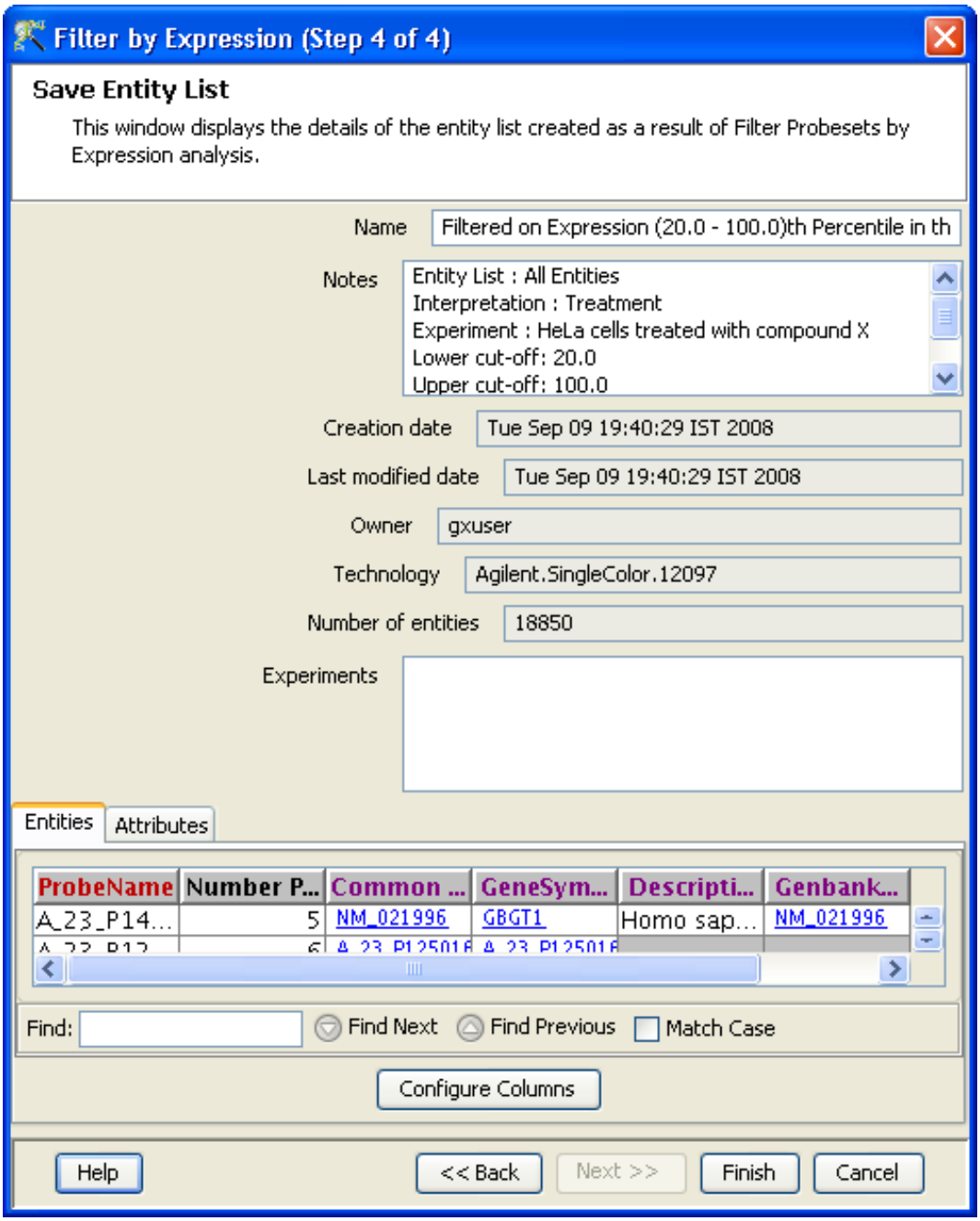


Figure 7.23: Filter probesets by expression (Step 4 of 4)

7.3.2.3 Filter Probesets by Flags

Flags are attributes that denote the quality of the entities. These flags are generally specific to the technology or the array type used. Thus the experiment technology type, i.e., Agilent Single Color, Agilent Two Color, Affymetrix Expression, Affymetrix Exon Expression, and Illumina Bead technology determine the flag notation. For Agilent microarrays, **GeneSpring** now utilizes

Agilent's naming convention of Detected, Compromised, and Not Detected when importing experiments from Agilent's Feature Extraction (FE) software. This replaces the previous naming convention used in **GeneSpring** for Agilent microarray products. Specifically, the Agilent naming convention of Detected replaces the previous Present (P) call, Compromised replaces the Absent call (A), and Not Detected replaces the Marginal call (M). But the [P, A, M] convention is retained for other technologies including Affymetrix. Further, these technology specific flags are described in the respective technology specific section.

For details refer to the corresponding sections:

- Filter probesets for Affymetrix expression
- Filter probesets for Affymetrix Exon expression
- Filter probes for agilent single color
- Filter probes for agilent two color
- Filter probesets for illumina
- Filter probesets for generic single color
- Filter probesets for generic two color

7.3.2.4 Filter Probesets on Data Files

The *Filter Probesets on Data Files* link is present under the *Quality Control* options in the Data Import Wizard. This link allows you to filter entities based on values in a specific column of your original data files. This filter lets you choose any of the columns in your data file and filter on the contents, both numeric and character data.

The Filter Probesets on Data Files wizard-driven workflow has the following four steps:

- The **Input Parameters** window allows the selection of an entity list and an interpretation on which the filtering needs to be performed. This is enabled by selecting the Choose button which shows all the available entity lists and interpretations.
- The **Sample Preview** window shows the first 100 rows of the first sample (as all samples belonging to the same experiment have the same file format) and gives an idea about the columns present in the sample file and their content type. The condition panel allows adding one or more conditions for searching on samples and also to combine search conditions using either AND/OR. The search field shows a dropdown with all the column names and depending on the content of the column (numeric or character) the options for the condition changes. The Search value that needs to be filled up should be the one that is common to the entities of interest. The stringency of the filter can be set in *Retain Entities* box.
- In the **Output Views** window, a spreadsheet and a profile plot appear as two tabs, displaying those probes which have passed the filter conditions. Total number of probes and number of probes passing the filter are displayed on the top of the navigator window. The visualization shows the values after data processing (normalization, baseline transformation etc).

- The **Save Entity List** window shows the details of the entity list that is created as a result of the above analysis. It also shows information regarding Creation date, modification date, owner, number of entities, notes etc. of the entity list. Annotations can be configured using *Configure Columns* button. Selecting Finish results in an entity list being created containing entities which satisfied the cut off. The name of the entity list will be displayed in the experiment navigator.

7.3.2.5 Filter Probesets by Error

This option allows the user to filter on the standard deviation or the coefficient of variation (CV). The option to filter on standard deviation or CV allows the user to filter entities which are above or below the value specified by the user. The user can filter on standard deviation among groups in case the standard deviation is comparable between the groups or he/she can filter on % CV if the standard deviation between the groups is highly varied.

In other words, filtering by CV renders the comparison of standard deviation, mean insensitive. If the condition specified is greater than equal to, then all the entities having values greater than or equal to the specified value are retained and in case the condition specified is lesser than, then the entities having values lesser than the specified value are retained for further analysis. This filtering option can be used for achieving two kinds of objectives:

1. To filter out genes having outlier samples.
2. To filter out genes having low variation in expression values across all samples or in an extreme case-constant values (This can be done by choosing the interpretation **All Samples**).

After selecting the *Filter on Error* option, the user has to go through the following steps:

- The **Entity list** and **Interpretation** window allows the selection of an entity list and an interpretation. This is enabled by selecting the *Choose* button which shows all the available entity lists and interpretations. An unaveraged interpretation is always required for this analysis.
- The **Input Parameters** window allows the selection of either standard deviation or CV as the filtering option. It also allows the stringency of the filter to be set in the *Retain Entities* box.
- In the **Output Views** window, a spreadsheet and a profile plot appear as two tabs, displaying those probes which have passed the filter conditions. Total number of probes and number of probes passing the filter are displayed on the top of the navigator window. The profile plot shows the processed data values and the spreadsheet shows all the entities along with the number of conditions in which they passed the filter criteria and either the CV or the standard deviation values.
- The **Save Entity List** window shows the details of the entity list that is created as a result of the above analysis. It also shows information regarding Creation date, modification date, owner, number of entities, notes etc. of the entity list. Annotations can be configured using *Configure Columns* button. Selecting *Finish* results in an entity list being created containing

entities which satisfied the cut off. The name of the entity list will be displayed in the experiment navigator.

7.3.3 Analysis

7.3.3.1 Statistical Analysis

Statistical analysis is required to determine whether any observed differences in your data are significant. For this purpose, **GeneSpring** offers a variety of parametric and non-parametric statistical tests, including t-test, Welch t-test, Mann-Whitney U Test, Analysis of Variance [ANOVA], Kruskall Wallis Test, Repeated Measures, and Friedman Test. In addition, for ANOVAs performed on more than two groups, **GeneSpring** provides the option to perform the Tukey or SNK post hoc test to identify which pairs of groups are significantly different. **GeneSpring** also provides seven different methods of multiple testing correction when computing p-values, including Bonferroni, Westfall and Young, Benjamini and Hochberg False Discovery Rate. Which of these tests is appropriate for your experiment depends on the design of the experiment (including sample size, sample grouping, whether the data is normally distributed, or variance in the data across groups, etc.).

See Details of Statistical Tests in **GeneSpring** on page 647 for details about each of the available tests and some guidance on when they are typically appropriate. Some examples of sample groupings and corresponding statistical tests have also been provided at the end of this section.

NOTE The information provided in this user guide should not be the only source for your decision to use one test over the other. The wrong choice of statistical test can significantly impact the outcome of your research. If you are not sure about which test to use, please refer to more comprehensive guides on statistical analysis for life sciences or members of your research team before selecting one of the statistical tests during this workflow.

The wizard-driven **Statistical Analysis** workflow available from the **Analysis** section of the **Workflow Browser** consists of nine steps. However, not all steps are required for each of the statistical tests. **GeneSpring** only displays the steps and parameters that are appropriate for the chosen Interpretation and statistical test.

This section describes each step and indicates for which statistical test the step is necessary. Tables 7.9 to 7.16 at the end of the section provide examples for possible sample groupings and the available statistical tests that are appropriate for each scenario.

Step 1 of 9: Choose the entity list and the interpretation on which to perform statistical analysis (see Figure 7.24). **GeneSpring** selects the active entity list and interpretation in the open experiment by default. Click **Choose...** to change either of the default selections. Click **Next** ≫.

NOTE You can only perform Statistical Analysis if all the groups in the chosen Interpretation have replicates. If a group has only one sample, the standard error cannot be calculated.

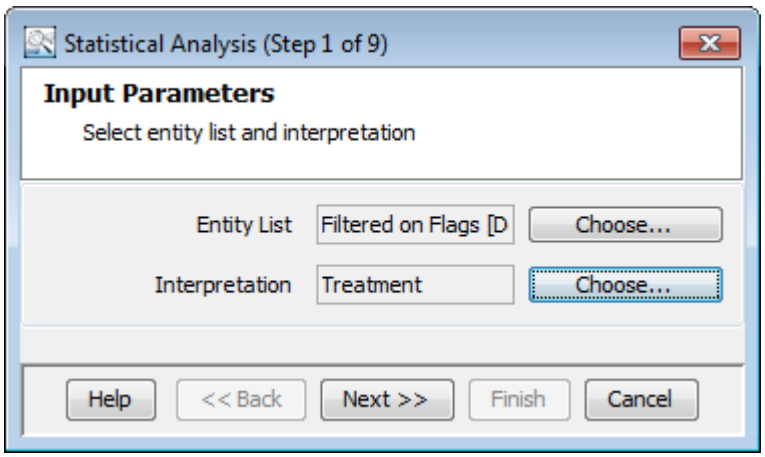


Figure 7.24: Statistical Analysis (Step 1 of 9) - Input Parameters

Step 2 of 9: Choose the desired statistical test from the **Select test** drop-down menu. **GeneSpring** automatically restricts the available options for this drop-down menu based on the experiment grouping defined in the chosen Interpretation in **Step 1**.

If the selected Interpretation in Step 1 contains only one parameter, you also have to determine the pairing of the available experiment groups for comparison in this step, for example "Treated" vs "Untreated" or "Untreated" vs "Treated" (see Figure 7.25). For paired tests, or tests with multiple parameters, pairing options are specified in either **Step 4** or **Step 7**.

Click **Next** $\gg$.

Example: After selecting the Treatment interpretation of the HeLa cells demo experiment that is part of your **GeneSpring** installation in **Step 1** (see Figure 7.24), the list of available tests is restricted to *t*-test and Mann Whitney U Test (see Figure 7.25) because the experiment grouping is based on only one parameter (*i.e.* whether the cell line was treated or untreated). However, it still depends on your experiment design whether you choose the moderated, unpaired, or paired versions of these tests.

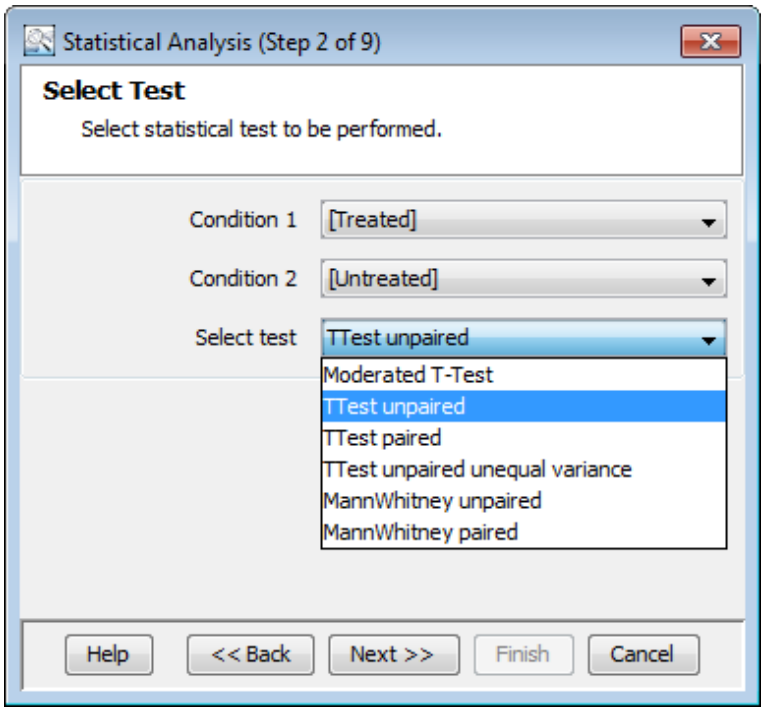


Figure 7.25: Statistical Analysis (Step 2 of 9) - Select Test

Step 3 of 9: This step is only displayed if the **ANOVA** or **ANOVA Unequal Variance** statistical test was selected in **Step 2**.

Select the post-hoc test to be applied from the **Post hoc** drop-down menu. You can choose, **None**, **SNK** (Student-Newman-Keuls), or **Tukey HSD** (Honest Significant Difference). Both tests are used to discover which sample means are significantly different from each other after the ANOVA result indicated the presence of significantly different sample means. Which test you should choose, if any, once again depends on the design of your experiment. See Post Hoc Testing of ANOVA Results on page 654 for more details on each method.

Click **Next** ≫.

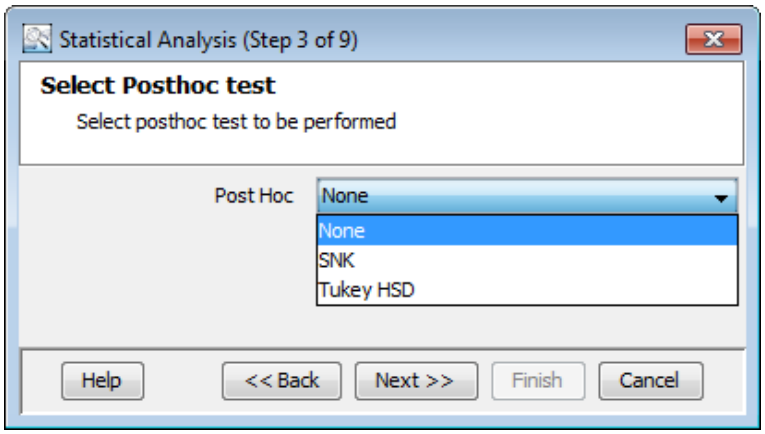


Figure 7.26: Statistical Analysis (Step 3 of 9) - Select Posthoc test

Step 4 of 9: This step is only displayed for the following statistical tests: **t-test paired**, **Mann-Whitney U Test paired**, **Repeated measures**, and **Friedman**.

Specify the pairing of samples across conditions by changing the order in which the samples are listed for each parameter value of your experiment conditions using the **Move Up**, **Move Down**, and **Reset** buttons as required (see Figure 7.27). Samples in the same row are paired for analysis. Click **Next** ≫.

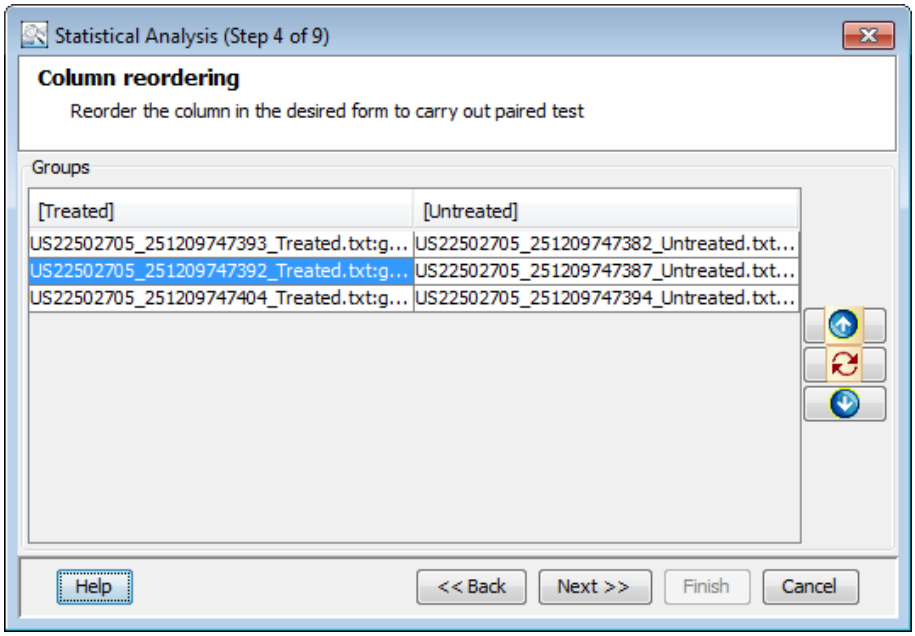


Figure 7.27: Statistical Analysis (Step 4 of 9) - Select Posthoc test

Step 5 of 9: This step for specifying parameters for p-value computation is displayed for the following statistical tests: **n-way ANOVA (n > 1)**, **t-test paired**, **t-test against zero**, **Mann-Whitney U Test paired**, **Mann-Whitney U Test against zero**, **Repeated measures**, and **Friedman**.

In these cases, **GeneSpring** always computes the p-value asymptotically.

Select the **Multiple Testing Correction** method for **GeneSpring** to apply when computing p-values.

For more information on p-value computation and multiple testing correction in **GeneSpring** see Obtaining p-values on page 658 and Adjusting for Multiple Comparisons - Corrected *p*-values on page 660.

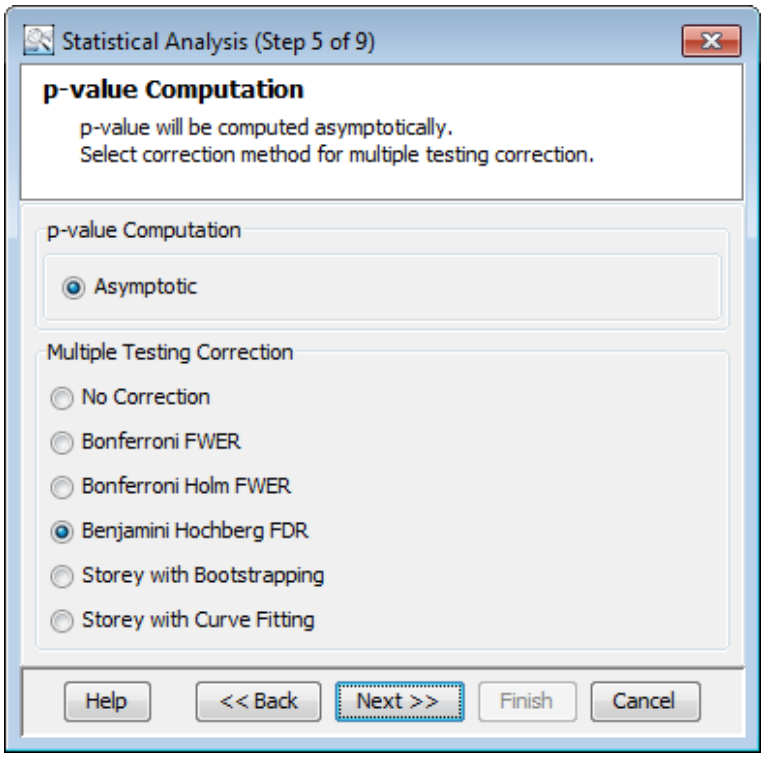


Figure 7.28: Statistical Analysis (Step 5 of 9) - p-value Computation

Step 6 of 9: This step for specifying parameters for p-value computation is displayed for the following statistical tests: **ANOVA**, **ANOVA Unequal Variance**, **Kruskall Wallis**, **Moderated t-test**, **t-test upaired**, and **Mann-Whitney U Test unpaired**.

Choose the desired p-value computation algorithm and the type of multiple testing correction. See Figure 7.29.

NOTE

If permutative p-value computation is chosen, it is recommended that the number of permutations is increased until convergence is reached. Once convergence is reached, the p-values of the entities remain the same for n as well as n+x number of permutations.

For more information on p-value computation and multiple testing correction in **GeneSpring** see Obtaining p-values on page 658 and Adjusting for Multiple Comparisons on page 660.

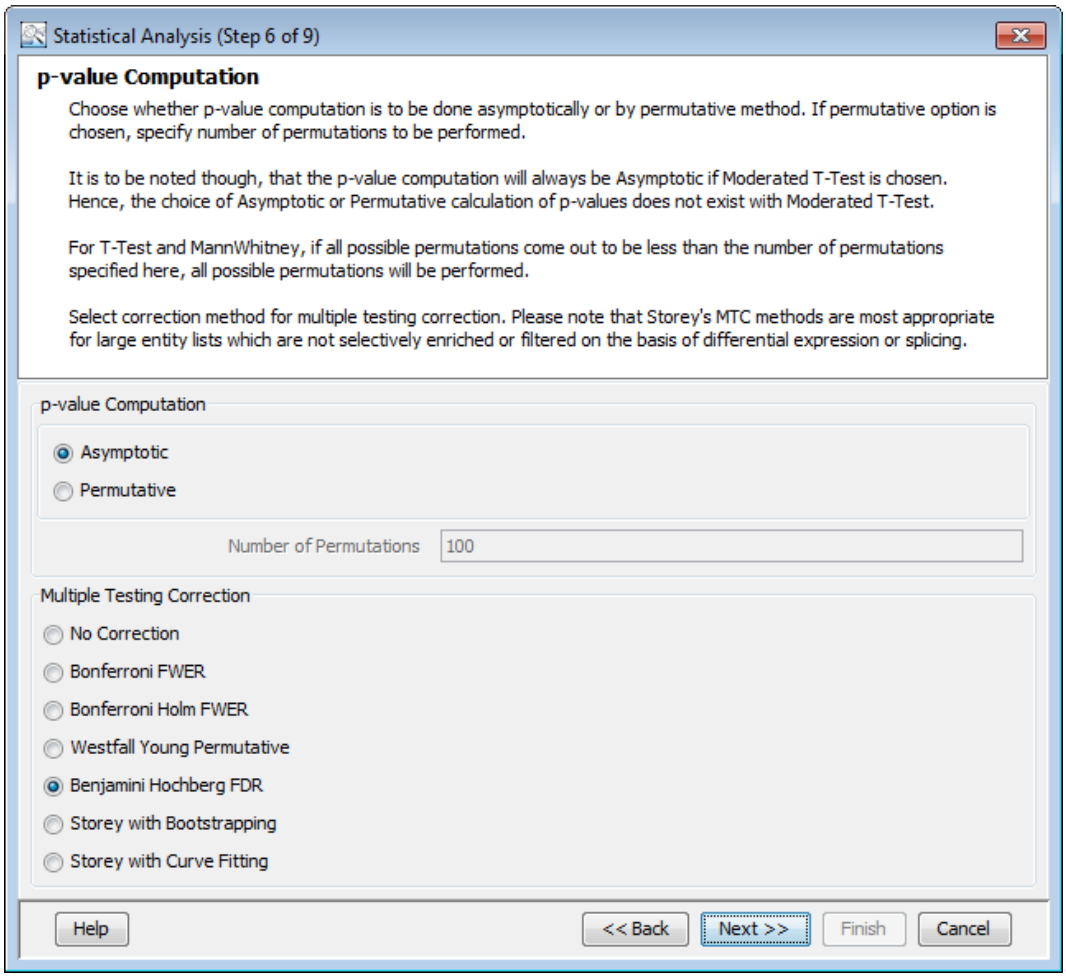


Figure 7.29: Statistical Analysis (Step 6 of 9) - p-value Computation

Step 7 of 9: This step is only displayed for those statistical tests that require you to specify the desired pairing of the samples across conditions in the chosen Interpretation to calculate fold change, namely: **n-way ANOVA**, **ANOVA Unequal Variance**, **Kruskall Wallis**, **Repeated measures**, and **Friedman**.

Using the **Select pairing option** drop-down menu, choose whether you want to select one or more pairs of conditions (select **Pairs of conditions**) or ask **GeneSpring** to pair all conditions against one specified condition (select **All against single condition**). See Figure 7.30.

For the **Pairs of conditions** option, **GeneSpring** displays all the possible combinations of pairing the parameter values of the chosen interpretation. Select the pairs you want to use. You also have the option to reverse the order of one or several pairs so that **Condition 1** becomes **Condition 2** and **Condition 2** becomes **Condition 1** using the **Reverse Order** button. To do this, highlight all the pairs you want to reverse and click the **Reverse Order** button. Note that highlighting is not the same as selecting for use in the analysis. Even an unselected pair can be highlighted and the order will be reversed when the **Reverse Order** button is used.

For the **All against single condition** option, specify the condition against which all other conditions are to be paired using the **Select condition** drop-down menu.

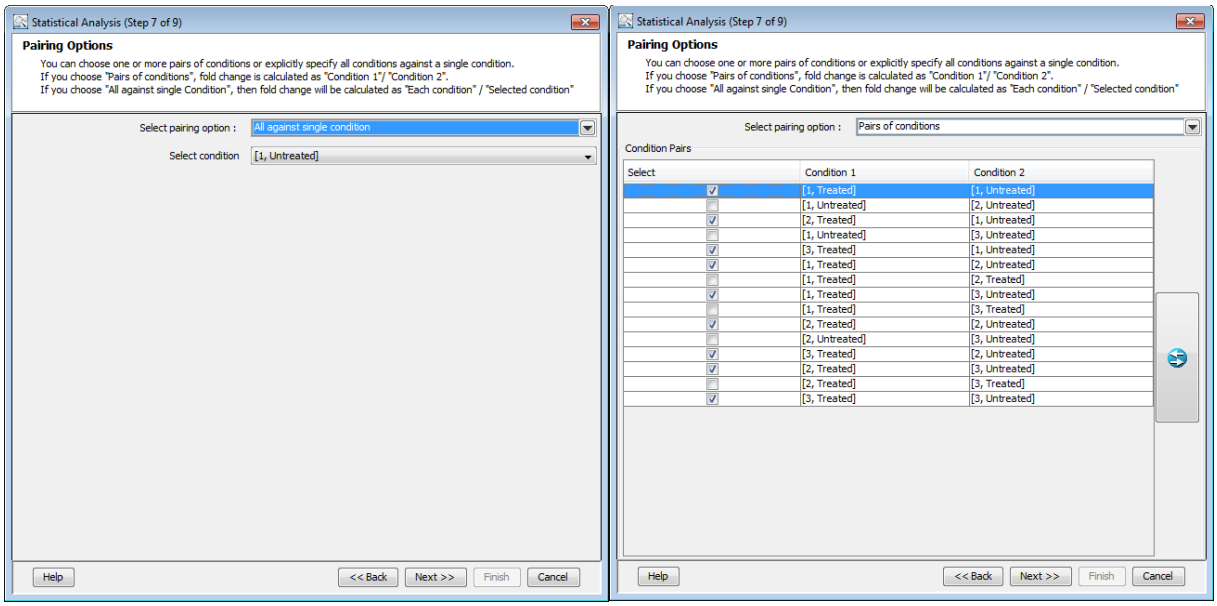


Figure 7.30: Statistical Analysis (Step 7 of 9) - Pairing Options

Step 8 of 9: This step displays the results of the **Statistical Analysis** workflow and is present for all statistical tests. Results are displayed in the form of tiled windows. By default **GeneSpring** displays all the p-values that have passed a p-value cut-off of 0.05. If no significant entities are found at that cut-off, or you would like to set a more restrictive cut-off, then the p-value cut-off can be changed using the **Corrected p-value cut-off** slider at the bottom of the window. The label at the top of the window shows the number of entities that satisfy the given p-value. When you are ready to save the indicated number of entities as a new entity list, click **Next** $\gg$.

The actual views and number of these tiled windows vary depending on the selected statistical test. For example, for t-test, or Mann-Whitney U Test, **GeneSpring** displays the following tiled windows (see Figure 7.31):

- A **p-value table** consisting of Probe Names, p-values, corrected p-values, and Fold change. If an entity is 2-fold up or 2-fold down, it will be called +2.0 or -2.0 respectively. The fold change value is colored in Red or Blue by default for up-regulated and down-regulated entities, respectively.
- The **Differential expression analysis report** containing the Test Description, *i.e.* the test that has been used for computing p-values, as well as the selected type of p-value computation (**Asymptotic** or **Permutative**) and multiple testing correction. This report also provides a result summary that indicates the number of entities satisfying different p-value cut-offs.
- A **Volcano plot** is only present if there are two groups provided in Experiment Grouping. In this plot, the entities are colored depending on their passing the default cut-offs:
  - **Red**: Entities which pass **Corrected p-value** cut-off and are up-regulated.
  - **Blue**: Entities which pass **Corrected p-value** cut-off and are down-regulated.
  - **Grey**: Entities which do not pass the **Corrected p-value** cut-off.

  The Volcano plot shows the negative log_{10} of p-value vs log_2 of fold change. Entities with large fold change and low p-value are easily identifiable on this view.

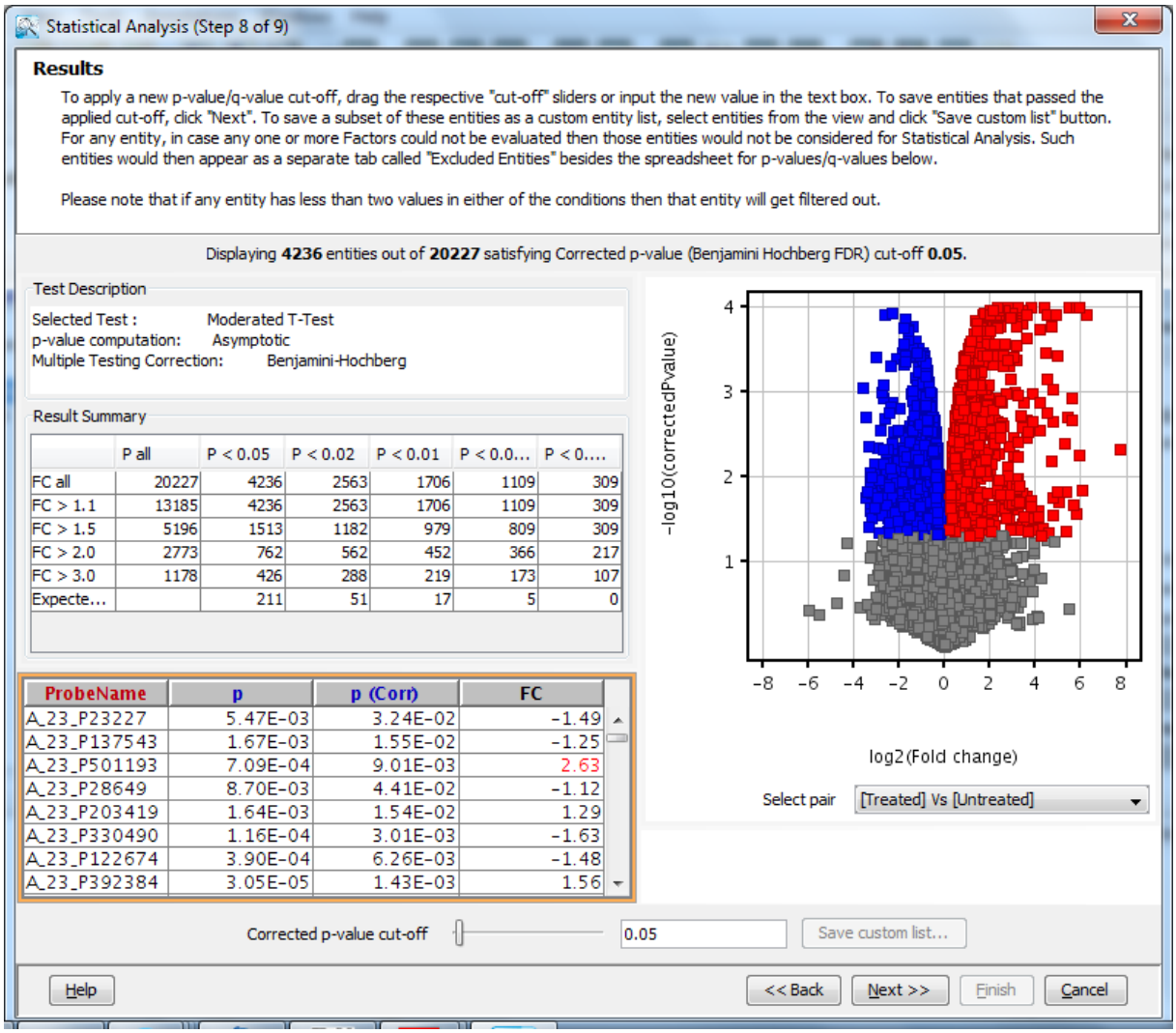


Figure 7.31: Statistical Analysis (Step 8 of 9) - Results

For 2-way ANOVAs, the *p*-value table contains SS ratio columns in addition to the Fold change columns and the Volcano plot is replaced by a Venn Diagram (see Figure 7.32:

- A ***p*-value table** consisting of Probe Names, *p*-values, corrected *p*-values, SS ratio and Fold change. The SS ratio is the mean of the sum of squared deviates (SSD) as an aggregate measure of variability between and within groups.
- The **Differential expression analysis report** containing the Test Description, *i.e.* the test that has been used for computing *p*-values, as well as the selected type of *p*-value computation (**Asymptotic** or **Permutative**) and multiple testing correction. This report also provides a result summary that indicates the number of entities satisfying different *p*-value cut-offs.
- A Venn Diagram that reflects the union and intersection of entities passing the *p*-value cut-off. You can save any section of the Venn Diagram as a separate entity list by selecting it and clicking the **Save custom list...** button.

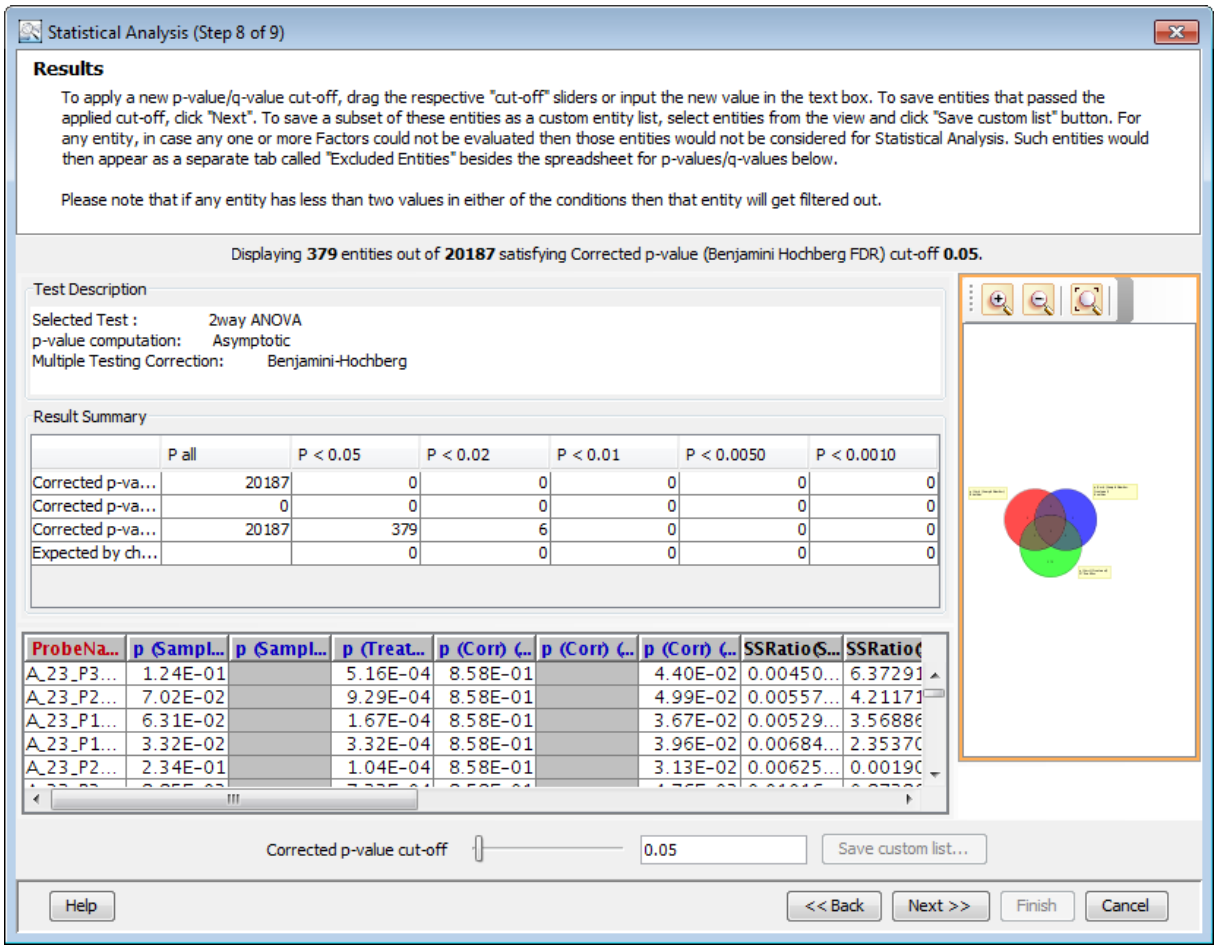


Figure 7.32: Statistical Analysis (Step 8 of 9) - Results for 2-way ANOVA

Step 9 of 9: The last step shows all the entities passing the p-value cut-off along with their annotations and the associated values computed during statistical analysis. It also shows details, like the creation date, modification date, owner, number of entities, notes, etc. of the entity list to be created.

You can add or remove any of the associated value columns using the **Configure Columns** button.

Click **Finish** to save the result as an entity list in the Experiment Navigator.

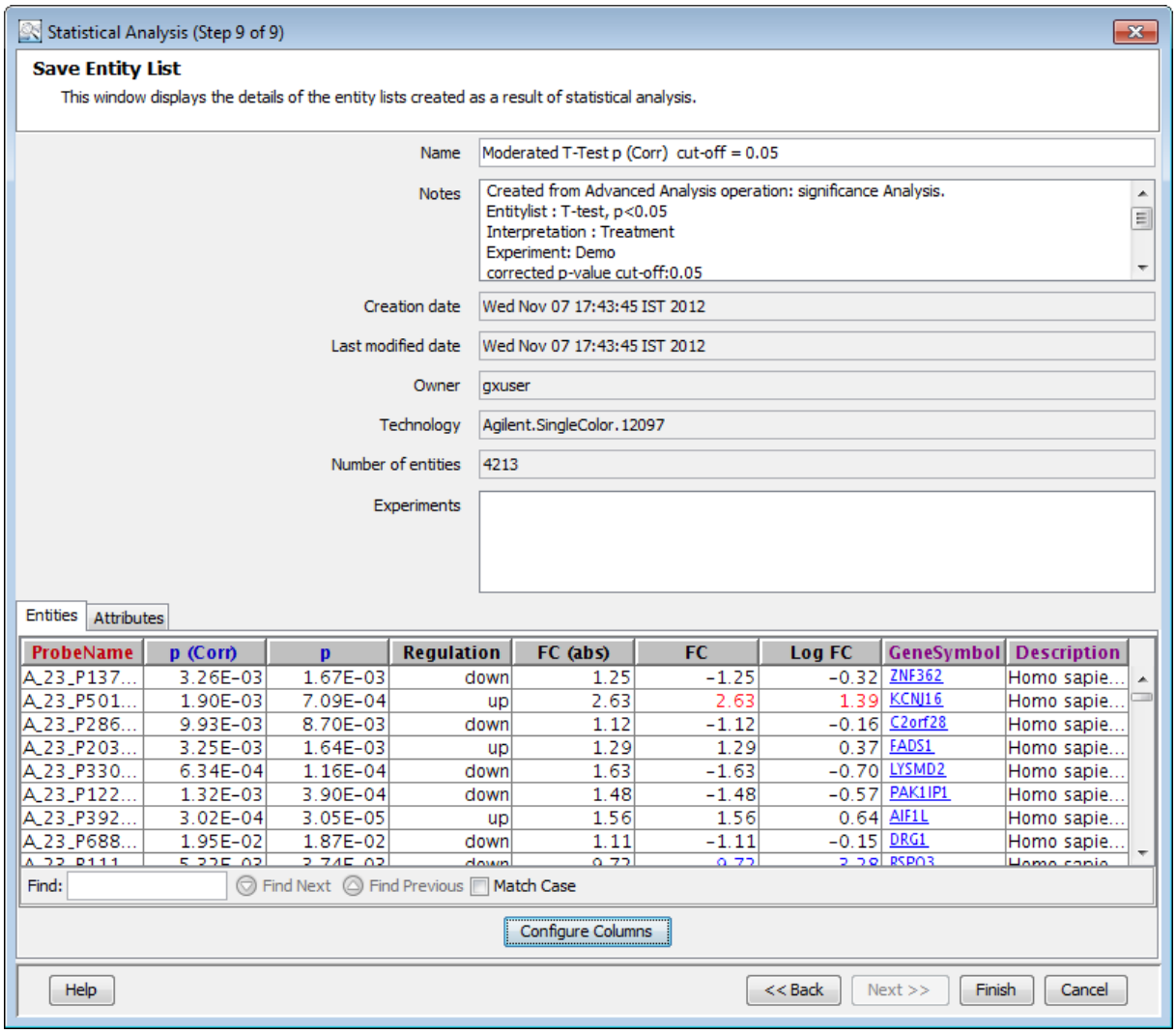


Figure 7.33: Statistical Analysis (Step 9 of 9) - Save Entity List

Example Sample Groupings

One of the criteria for choosing the correct statistical test for your experiment is experiment grouping. This criterion is incorporated in the **GeneSpring Statistical Analysis** workflow and only appropriate statistical tests are available for selection in **Step 2** based on the Interpretation (representing the experiment grouping) you have chosen in **Step 1**.

NOTE

The following examples of sample groupings provide some guidance on which statistical test is most suited to a particular experiment grouping, provided your data is normally distributed. If your data is non-normally distributed, you should choose the non-parametric test instead (for example Mann-Whitney U Test unpaired instead of t-test unpaired, or Mann-Whitney U Test against zero instead of t-test against zero).

- **Example Sample Grouping I:** The example illustrated in Table 7.9 has two groups, Normal and Tumor, each with three replicates. In such a situation, the unpaired t-test is most suitable. In fact, the HeLa cells demo experiment that is part of **GeneSpring** is another example of this type of experiment grouping.

Samples	Grouping
S1	Normal
S2	Normal
S3	Normal
S4	Tumor
S5	Tumor
S6	Tumor

Table 7.9: Sample Grouping and Significance Tests I

- **Example Sample Grouping II:** In the example outlined in Table 7.10 only one group, Tumor, is present. Since there is no condition to compare the data to, the t-test against zero is performed here.

Samples	Grouping
S1	Tumor
S2	Tumor
S3	Tumor
S4	Tumor
S5	Tumor
S6	Tumor

Table 7.10: Sample Grouping and Significance Tests II

- **Example Sample Grouping III:** When three groups are present (for example, Normal, Tumor1 and Tumor2) and one of the groups (Tumour2 in this case) does not have replicates as shown in Table 7.11, statistical analysis cannot be performed.

Samples	Grouping
S1	Normal
S2	Normal
S3	Normal
S4	Tumor1
S5	Tumor1
S6	Tumor2

Table 7.11: Sample Grouping and Significance Tests III

NOTE

In cases where some conditions do not have replicates, statistical analysis can only be performed on conditions with replicates if the samples representing the conditions with out replicates (Sample 6 in example III) are removed from the **Interpretation** used for analysis. Therefore, we recommend that you create a new **Interpretation** (excluding the condition without replicate samples) once experiment creation is complete and re-run statistical analysis using the option available from the **Workflow Browser**.

- **Example Sample Grouping IV:** When there are three groups, each with replicates, within an interpretation as shown in Table 7.12, typically a one-way ANOVA is performed.

Samples	Grouping
S1	Normal
S2	Normal
S3	Tumor1
S4	Tumor1
S5	Tumor2
S6	Tumor2

Table 7.12: Sample Grouping and Significance Tests IV

- **Example Sample Grouping V:** The Table 7.13 shows an example of an experiment where two parameters are present. The presence of two parameters means that **GeneSpring** looks for samples in all the possible parameter value pairings (in this example: Normal/10 min, Normal/50 min, Tumor/10 min, Tumor/50 min). However, there are no samples for the conditions Normal/50 min and Tumor/10 min. Due to the absence of such samples, no statistical significance test is performed.

Samples	Grouping A	Grouping B
S1	Normal	10 min
S2	Normal	10 min
S3	Normal	10 min
S4	Tumor	50 min
S5	Tumor	50 min
S6	Tumor	50 min

Table 7.13: Sample Grouping and Significance Tests V

- **Example Sample Grouping VI:** For experiments with the kind of experiment grouping illustrated in Table 7.14, a two-way ANOVA is performed.

Samples	Grouping A	Grouping B
S1	Normal	10 min
S2	Normal	10 min
S3	Normal	50 min
S4	Normal	50 min
S5	Tumor	50 min
S6	Tumor	50 min
S7	Tumor	10 min
S8	Tumor	10 min

Table 7.14: Sample Grouping and Significance Tests VI

- **Example Sample Grouping VII:** In the example shown in Table 7.15, a two-way ANOVA is performed. However, the result only contains a p-value for each parameter, *i.e.* for Grouping A and Grouping B. The p-value for the combined parameters, Grouping A- Grouping B is not computed. In this particular example, there are six conditions (Normal/10min, Normal/30min, Normal/50min, Tumor/10min, Tumor/30min, Tumor/50min), which is the same as the number of samples. The p-value for the combined parameters can only be computed when the number of samples exceeds the number of possible groupings.

Samples	Grouping A	Grouping B
S1	Normal	10 min
S2	Normal	30 min
S3	Normal	50 min
S4	Tumour	10 min
S5	Tumour	30 min
S6	Tumour	50 min

Table 7.15: Sample Grouping and Significance Tests VII

- **Example Sample Grouping VIII**: In the example shown in Table 7.16, which illustrates an experiment grouping with three parameters, a three-way ANOVA is performed.

Samples	Grouping A	Grouping B	Grouping C
S1	Normal	Female	10
S2	Normal	Male	10
S3	Normal	Male	20
S4	Normal	Female	20
S5	Tumor1	Male	10
S6	Tumor1	Female	10
S7	Tumor1	Female	20
S8	Tumor1	Male	20
S9	Tumor2	Female	10
S10	Tumor2	Female	20
S11	Tumor2	Male	10
S12	Tumor2	Male	20

Table 7.16: Sample Grouping and Significance Tests VIII

7.3.3.2 Filter on Volcano Plot

The **Filter on Volcano Plot** link is present under the Analysis options in the experiment. This option allows you to filter entities on volcano plots, which are constructed using fold change values and p-values. Volcano plots allow you to visualize the relationship between fold-change (magnitude of change) and statistical significance (which takes both magnitude of change and variability into consideration). Volcano plots are used to visually represent differential expression between two different conditions and can be used in publications to provide a visual summary of p-values and fold-change values.

After selecting the Filter on Volcano Plot option, follow these steps:

- The **Input Parameters** window allows the selection of an entity list and an interpretation. This is enabled by selecting the *Choose* button which shows all the available entity lists and interpretations (an option to add a new interpretation is also given). The groups present in the interpretation must have replicates for calculating variance.
- The **Select Test** window allows the selection of the t-test as well as the pair between which the test has to be performed. Choose among the following statistical tests: t-test paired, t-test unpaired, t-test unpaired unequal variance, MannWhitney unpaired, and MannWhitney paired. More information on the above tests is available under section Details of Statistical Tests in **GeneSpring**. The 'Condition 1' and 'Condition 2' drop-down lists allow you to determine the pairing of conditions for calculating the fold change. Fold change is calculated as the ratio between Condition 1 and Condition 2.
- If the statistical test chosen is either t-test paired or MannWhitney paired, then the **Column**

reordering window appears. The reordering can be done by selecting a sample in a column and moving it with the help of the 'up' and 'down' arrow buttons on the side.

- Upon completion of column reordering, the **p-value computation** window appears which allows the selection of a correction method for multiple testing correction. The p-value is computed asymptotically.
- If a statistical test other than the paired tests is chosen, then the window that appears allows the user to select either the Asymptotic or the Permutative option for p-value computation in addition to the options present for multiple testing correction methods. More information on the above options is available in the section Adjusting for Multiple Comparisons - Corrected *p*-values.
- The next step shows the results upon completion of the statistical test with multiple views.

  1. The **Test Description** indicating the chosen statistical test and method for computing p-values and multiple testing correction.
  2. A **p-value table** consisting of Probe Names, p-values and corrected p-values (p (Corr)).
  3. **Volcano plot**: The Volcano plot shows the negative log_{10} of *p*-value vs log_2 of fold change. The entities are colored depending on their passing the default cut-offs:
     - **Red**: Entities which pass both **Corrected *p*-value** and **Fold change** cut-offs and are up-regulated.
     - **Blue**: Entities which pass both **Corrected *p*-value** and **Fold change** cut-offs and are down-regulated.
     - **Green**: Entities which pass the **Corrected *p*-value** cut-off and fail to pass the**Fold change** cut-off.
     - **Grey**: Entities which neither pass the **Corrected *p*-value** cut-off nor the **Fold change** cut-off.

     You can change the default **Corrected *p*-value** and **Fold change** cut-offs by using the sliders.

  You can also select entities of interest from either the *p*-value table, result summary, or the volcano plot and save them as an entity list by selecting the **Save custom list** option.

  The label at the top of the wizard shows the number of entities satisfying the given *p*-value and the fold change cut-offs.

- The **Save Entity List** window shows the details of the entity list that is created as a result of the above analysis. It also shows information regarding Creation date, modification date, owner, number of entities, notes etc. of the entity list. Annotations can be configured using Configure Columns button. Selecting **Finish** results in an entity list being created containing entities which satisfied the cut off. The name of the entity list will be displayed in the experiment navigator.

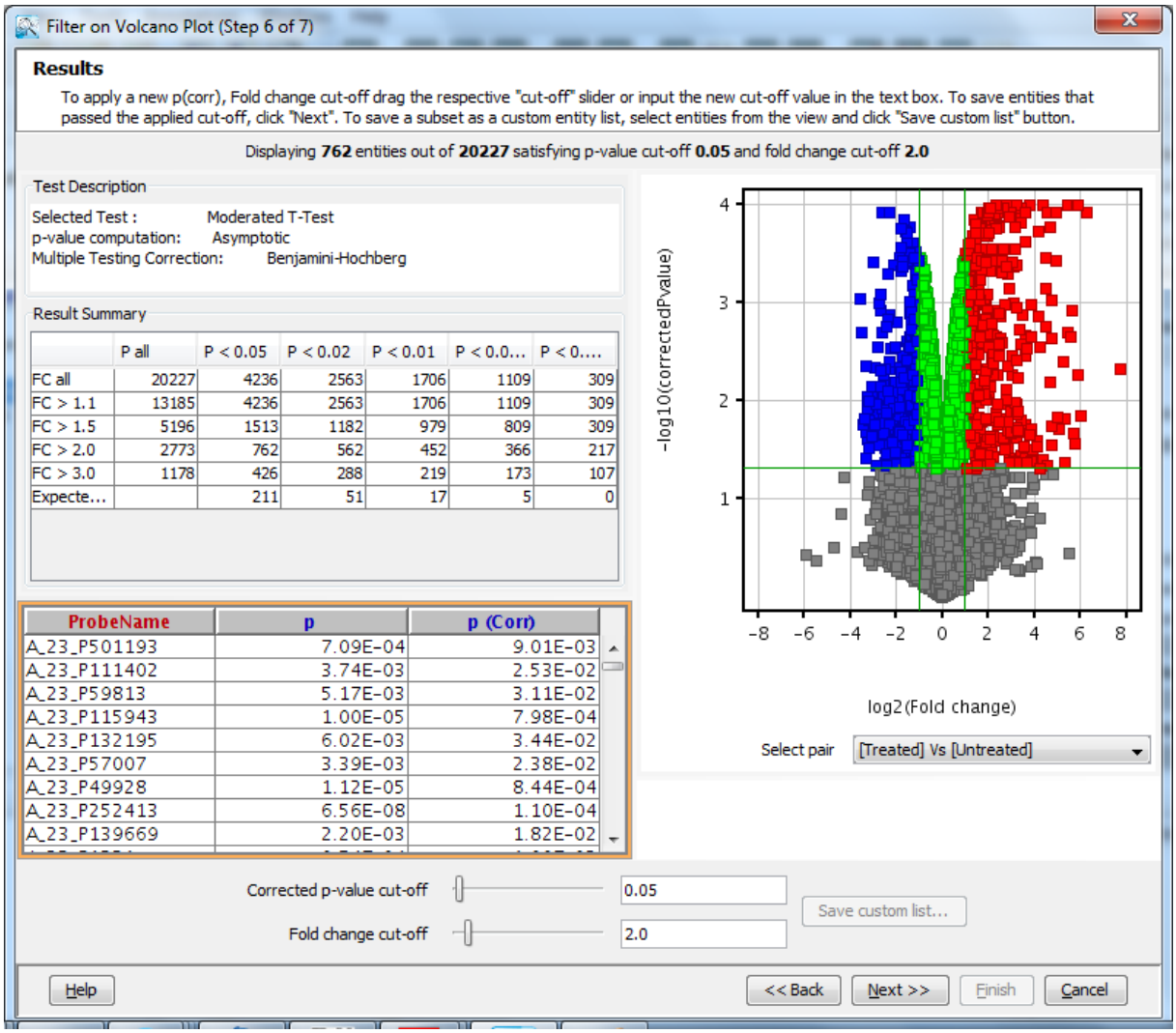


Figure 7.34: Filter on Volcano Plot (Step 6 of 7) - Results

7.3.3.3 Fold change

GeneSpring offers Fold Change Analysis to identify genes with expression ratios or differences between a treatment and a control that are significant to the given cut-off or threshold.

The wizard driven The wizard-driven Fold Change available from the Analysis section of the Workflow Browser consists of four steps.

This section describes each step and Table 7.17 at the end of the section provides an examples for Fold Change computation is **GeneSpring**.

Step 1 of 4: Choose the entity list and the interpretation on which to perform Fold Change

(see Figure 7.35). **GeneSpring** selects the active entity list and interpretation in the open experiment by default. Click **Choose...** to change either of the default selections. Click **Next** ≫.

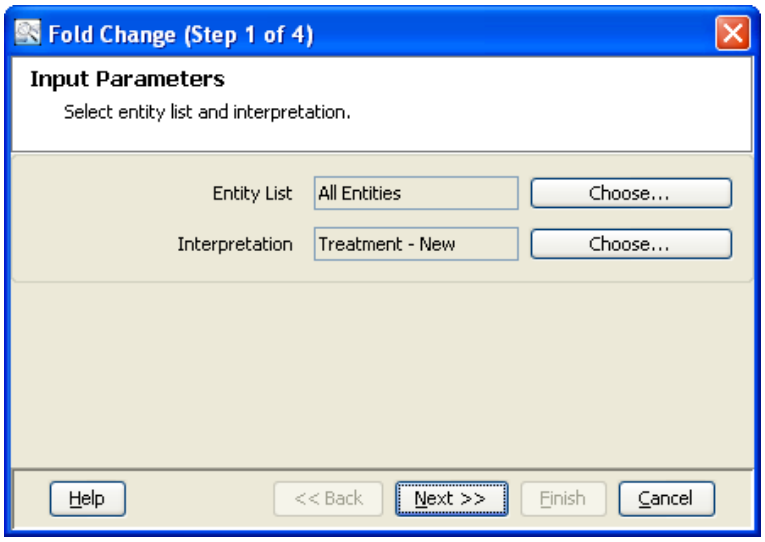


Figure 7.35: Fold Change (Step 1 of 4) - Input Parameters

Step 2 of 4: Select the pairing option to compute Fold Change: **All against single condition** or **Pairs of conditions**.

For **All against single condition**:

1. Select one condition to be compared with other conditions (See Figure 7.36).
2. The selected condition is considered as **Condition 2** for the Fold Change calculation (Refer Fold Change calculation).

For **Pairs of conditions**:

1. Select **Condition Pairs** to compute Fold Change for the selected conditions (See Figure 7.36).
2. You can change the order of conditions using the flip condition button.
3. If you select multiple **Condition Pairs**, Fold Change is computed for all the selected pairs in Step 3.

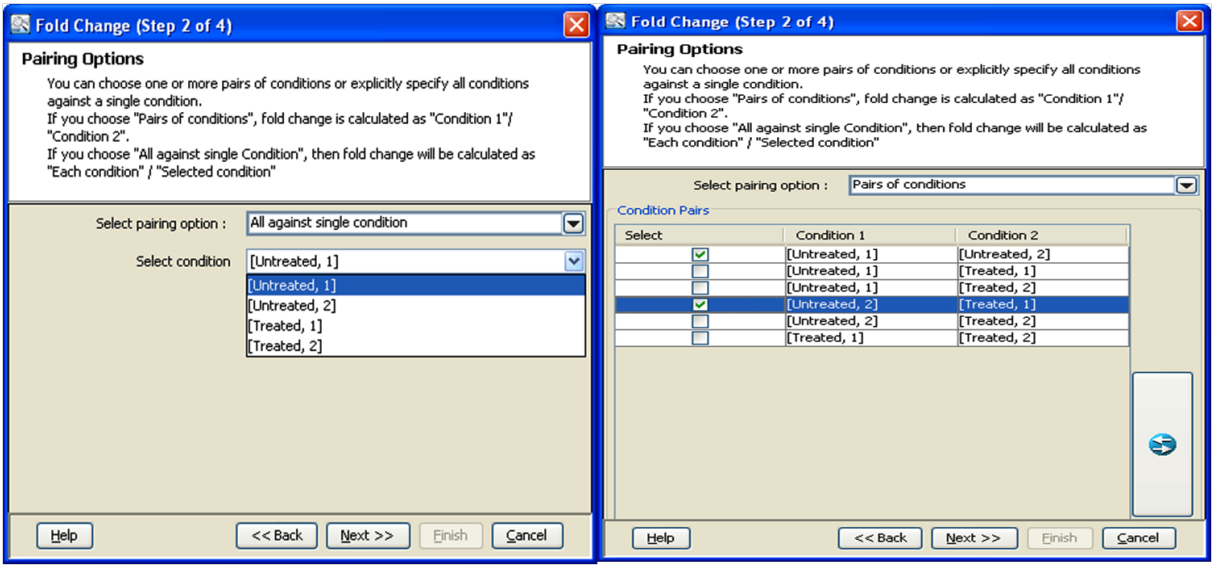


Figure 7.36: Fold Change (Step 2 of 4) - Pairing Options

Step 3 of 4: Results of analysis: Upon completion of Fold Change computation the results are displayed in tabs as a **Profile Plot - By Group** and a **Fold Change** spreadsheet (See Figure 7.37).

- *Profile Plot - By Group consisting of a Profile Plot of Normalized intensity values.* X-axis of the profile plot displays all the conditions from the interpretation in Step 1. The entities are colored in Red or Blue by default for up-regulated and down-regulated fold change values, respectively. Select an entity in the plot and Double-click to launch the Entity Inspector (Refer Entity Inspector).
- *Fold Change* spreadsheet containing Probe Name and Fold Change ratio of the specified conditions in Step 2. If ‘‘All against single condition” pairing option was selected in Step 2, fold change ratio columns will comprise of all conditions (from the selected interpretation in Step 1) against the selected condition (in Step 2). If an entity is 2-fold up or 2-fold down, it will be called +2.0 or -2.0 respectively. The fold change value is colored in Red or Blue by default for up-regulated and down-regulated entities, respectively.

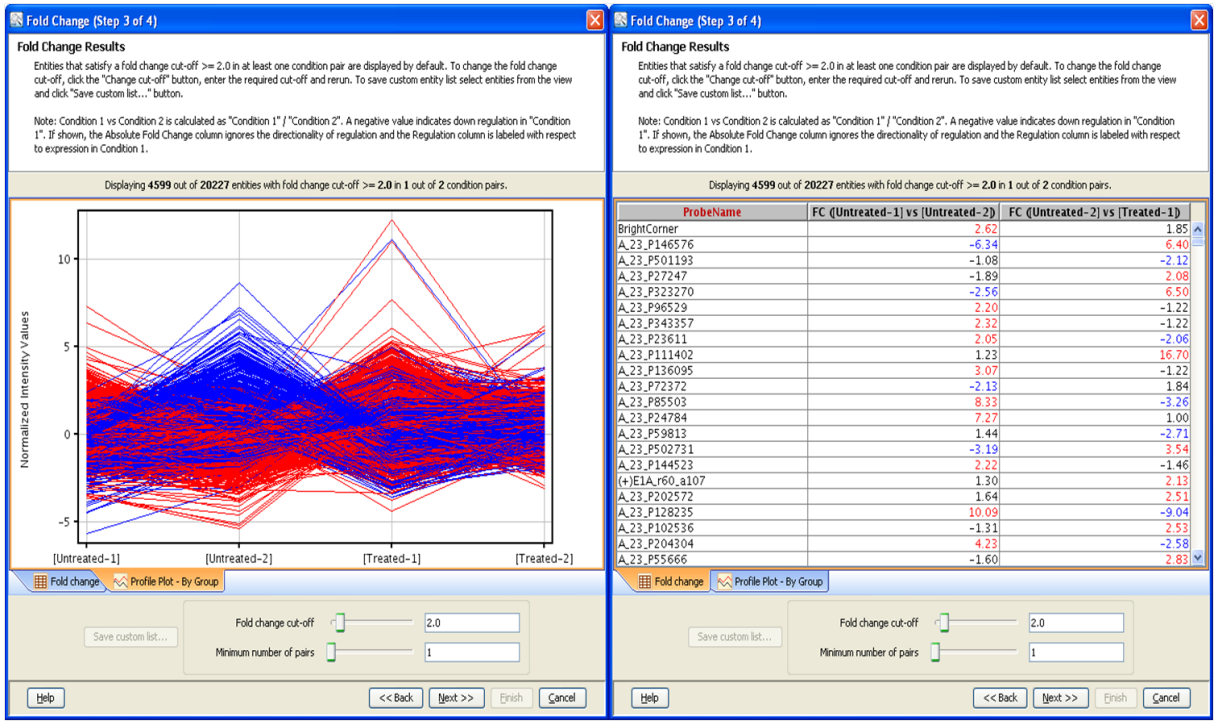


Figure 7.37: Fold Change (Step 3 of 4) - Fold Change Results

How to change the display properties of Fold Change spreadsheet:

Columns of the Fold Change spreadsheet: By default, the columns in the spreadsheet are; Probe Name, Fold Change of the Conditions selected in Step 2 of Fold Change wizard.

1. Right-click on the spreadsheet and select **Columns** tab:
   (a) Select which items you want to include in the spreadsheet by moving the desired items from **Available items** to the **Selected items** field or vice versa using the move right and move left buttons.
   (b) Reorder the selected items, or refresh the order using the up, down, and refresh order buttons.
   (c) Click **OK**.
   GeneSpring applies all the customized settings for the spreadsheet in the Fold Change spreadsheet.

   Or from the **GeneSpring** Tools $\longrightarrow$Options:
   (a) Go to **Tools** $\longrightarrow$**Options** $\longrightarrow$**Miscellaneous** $\longrightarrow$**Fold Change Settings**.
   (b) Select from **Columns to display** options to display the selected columns as default in the Fold Change spreadsheet (See Figure 7.38).
   (c) Click **Apply** to save the changes and **OK** to exit the **Configuration Dialog**.

Configuration of Color in the Fold Change spreadsheet: The fold change value is colored in Red or Blue by default for up-regulated and down-regulated entities, respectively.

1. To change the color settings:

(a) Go to **Tools** ⟶**Options** ⟶**Miscellaneous** ⟶**Fold Change Settings**.

(b) Customize the value for **Fold change color cut-off** to apply the fold change color setting cut-off (See Figure 7.38). All the fold change values above the cut-off (for up-regulated; $+^{ve}$) and below the cut-off (for down-regulated; $-^{ve}$) will be displayed with the respective colors.

(c) To change the display color for **Up-regulation color** and **Down-regulation color**, select an available color from the drop down options (See Figure 7.38).

(d) Click **Apply** to save the changes and **OK** to exit the **Configuration Dialog**.

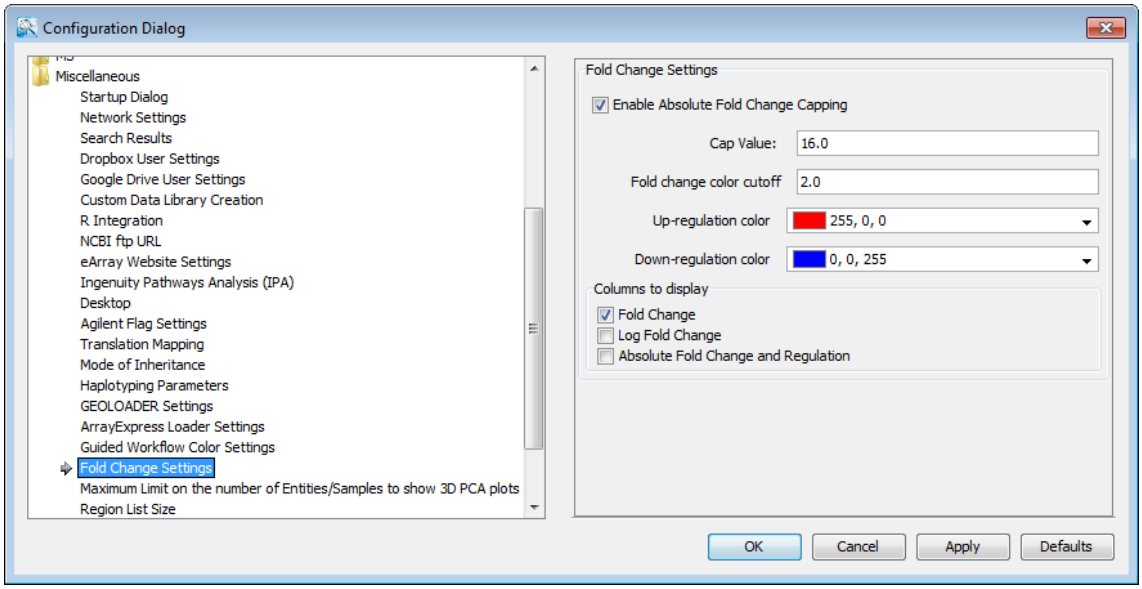


Figure 7.38: Configuration Menu for Fold Change Colour Settings

Step 4 of 4: The last step shows all the entities passing the fold change cut-off along with their annotations and the associated values computed during fold change analysis. It also shows the details like the creation date, modification date, owner, number of entities, notes etc. of the entity list. Click **Finish** to save an entity list in the Experiment navigator. The associated columns can be configured using the **Configure Columns** button.

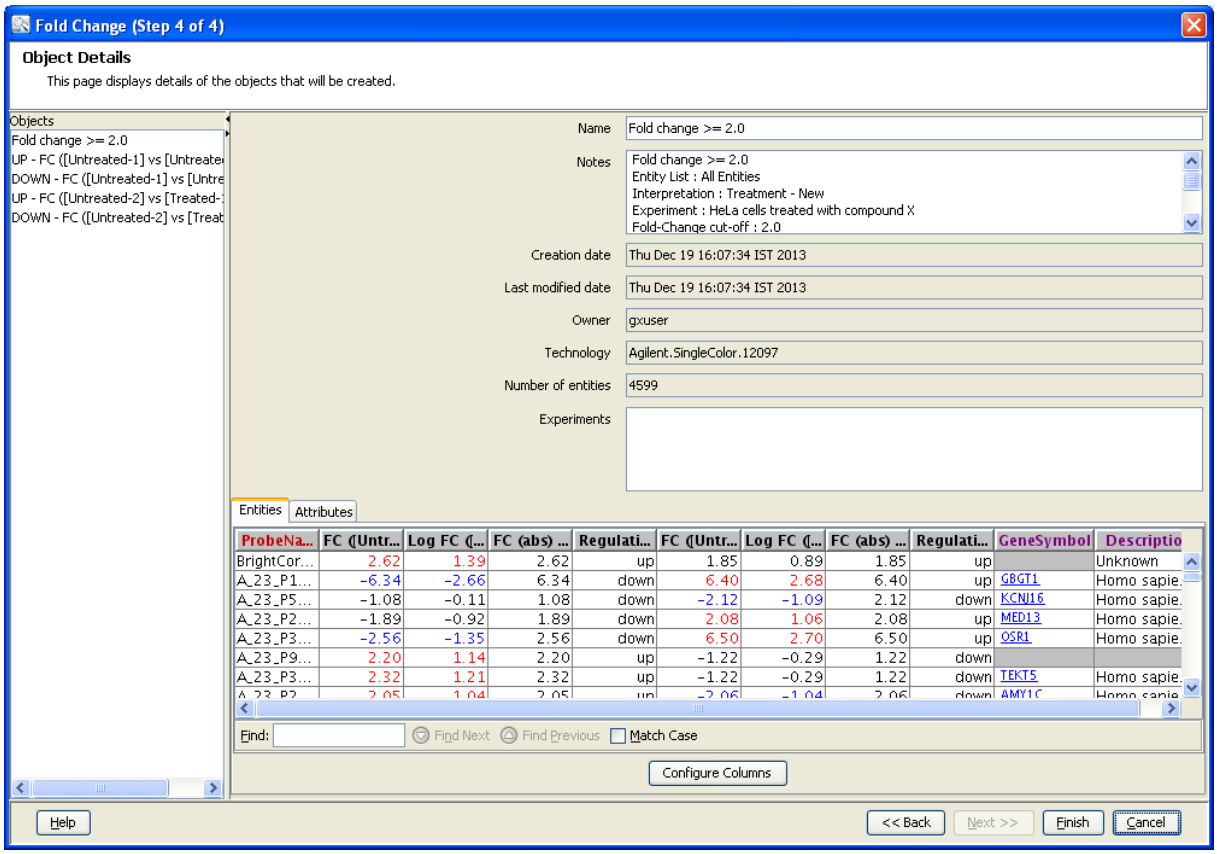


Figure 7.39: Fold Change (Step 4 of 4) - Object Details

NOTE Fold Change values are always ≥ 1.

Note: If multiple conditions are selected for condition one, the fold change for each of the conditions in condition 1 will be calculated.

Fold Change calculation

Fold Change is computed as an aggregate between a condition (Condition 1) and one or more other conditions (Condition 2).

$$Fold\,change = \frac{Condition1}{Condition2}$$

Fold Change value is an absolute ratio of normalized intensities (non log scale) between the average

intensities of the grouped samples.

For example, Table 7.17 has probes 'X' and 'Y' under treated and untreated conditions with their respective intensity values (Normalized, log scale).

	Probe	
	X	Y
Control (log scale)	-0.49592295	1.0647016
Treated (log scale)	0.50534225	-2.2152293
Control (linear scale)	0.709108	2.091737
Treated (linear scale)	1.41946	0.215352
Fold Change (Treated vs Control)	2.001755	9.713094

Table 7.17: Normalized Intensities for Probes X and Y

- For Probe X, fold change for Treated vs Control is computed as 1.41946/0.709108 which equals 2.001755 (up-regulated). The log fold change of this would be 1.001265.
- For Probe Y, fold change for Treated vs Control is computed as 2.091737/0.215352 which equals 9.713094 (down-regulated) and will be displayed as **-9.713094**. The log fold change of this would be -3.279931.

7.3.3.4 Clustering

Clustering Analysis is described in Chapter 25.

7.3.3.5 Correlation

Correlation Analysis is described in Chapter 26.

7.3.3.6 Find similar entities

The above option allows the user to query a specific entity list or the entire data set to find entities whose expression profile matches an entity of interest.

On choosing *Find Similar Entities* under the Analysis section in the workflow, **GeneSpring** takes us through the following steps:

Step 1 of 3: This step allows the user to input parameters that are required for the analysis. Entity list and interpretation are selected here. Next, the entity list displaying the profile of our interest has to be selected in the *Choose Query Entity* box. The similarity metric that can be used in the analysis can be viewed by clicking on the dropdown menu. The options that are provided are:

1. **Euclidean:** Calculates the Euclidean distance where the vector elements are the columns. The square root of the sum of the square of the A and the B vectors for each element is calculated and then the distances are scaled between -1 and +1. Result = (**A-B**).(**A-B**).
2. **Pearson Correlation:** Calculates the mean of all elements in vector **a**. Then it subtracts that value from each element in **a** and calls the resulting vector **A**. It does the same for **b** to make a vector **B**. Result = **A.B**/(—**A**——**B**—)
3. **Spearman Correlation:** It orders all the elements of vector **a** and uses this order to assign a rank to each element of **a**.It makes a new vector **a'** where the i-th element in **a'** is the rank of $\mathbf{a_i}$ in **a** and then makes a vector **A** from **a**' in the same way as **A** was made from **a** in the Pearson Correlation. Similarly, it makes a vector **B** from **b**. Result = **A.B**/(—**A**——**B**—).The advantage of using Spearman Correlation is that it reduces the effect of the outliers on the analysis.

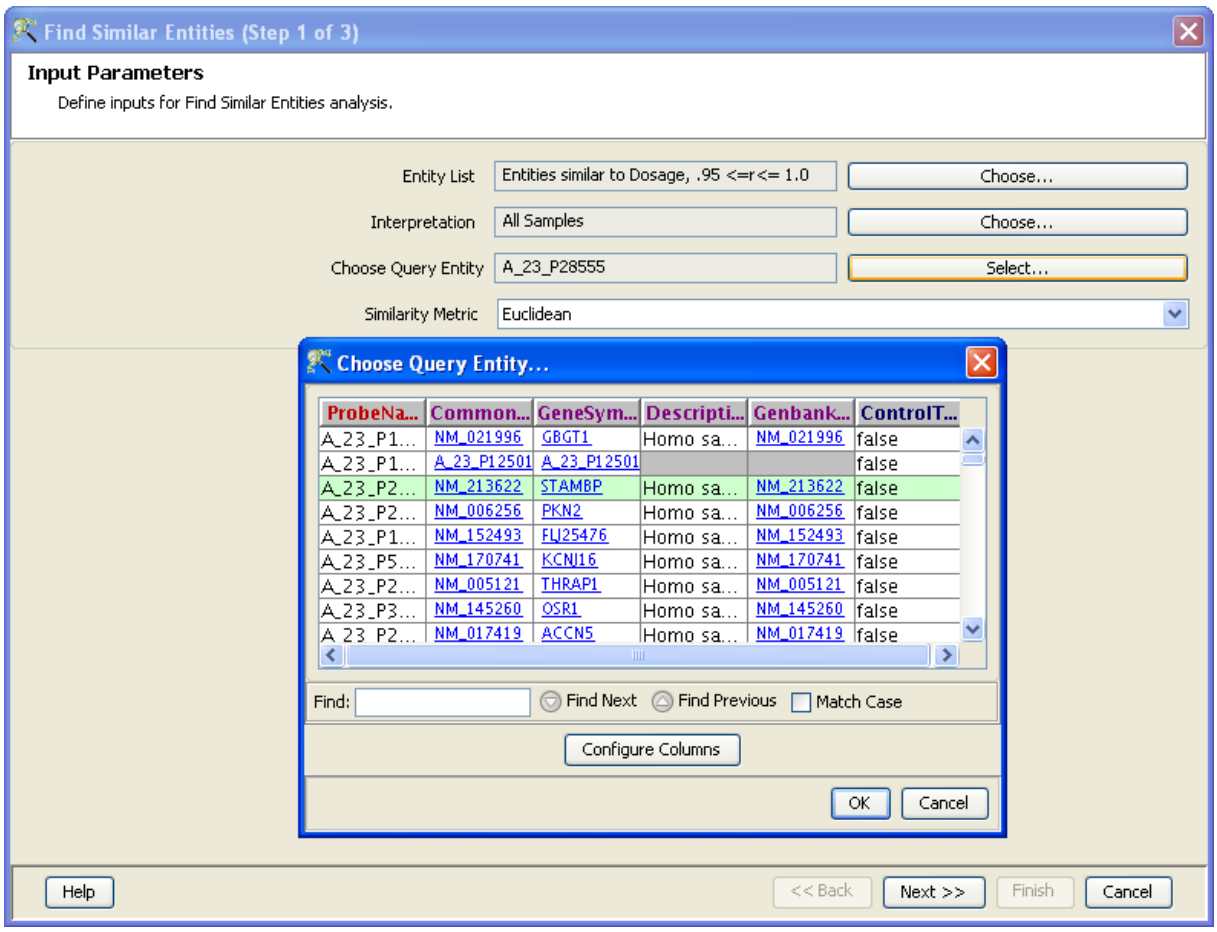


Figure 7.40: Find Similar Entities (Step 1 of 3) - Input Parameters

Step 2 of 3: This step allows the user to visualize the results of the analysis in the form of a profile plot. The plot displays the mean profile of the entities that have passed the similarity cut-off. The default range for the cutoff is Min-0.95 and Max-1.0.The cutoff can be altered by using the Change Cutoff button provided at the bottom of the wizard. After selecting the profiles in the plot,they can be saved as an entity list by using the option *Save Custom List*.

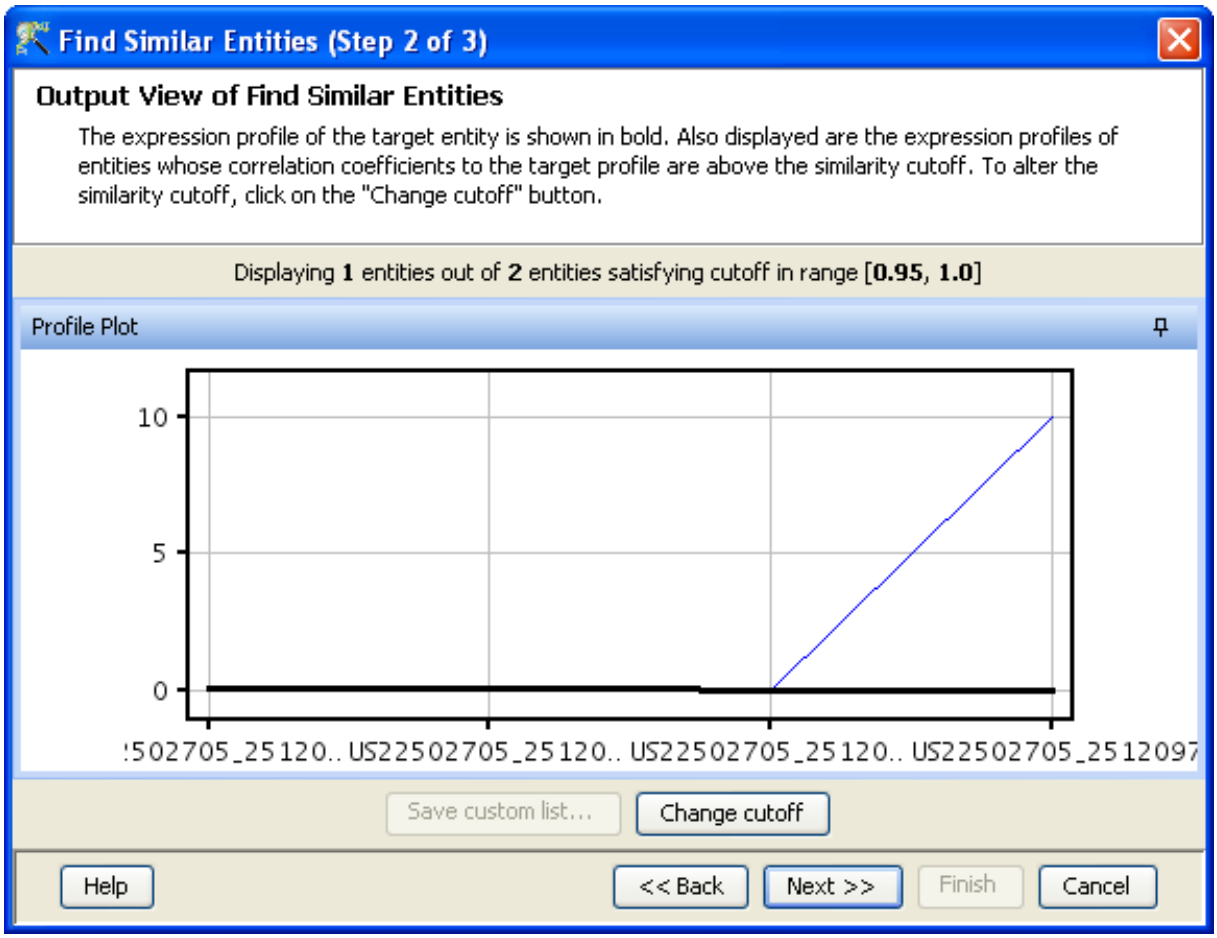


Figure 7.41: Find Similar Entities (Step 2 of 3) - Output View of Find Similar Entities

Step 3 of 3: This step allows the user to save the entity list created as a result of the analysis and also shows the details of the entity list. Option to configure columns that enables the user to add columns of interest from the given list is present. Clicking on *Finish* creates the entity list which can be visualized under the analysis section of the experiment in the project navigator.

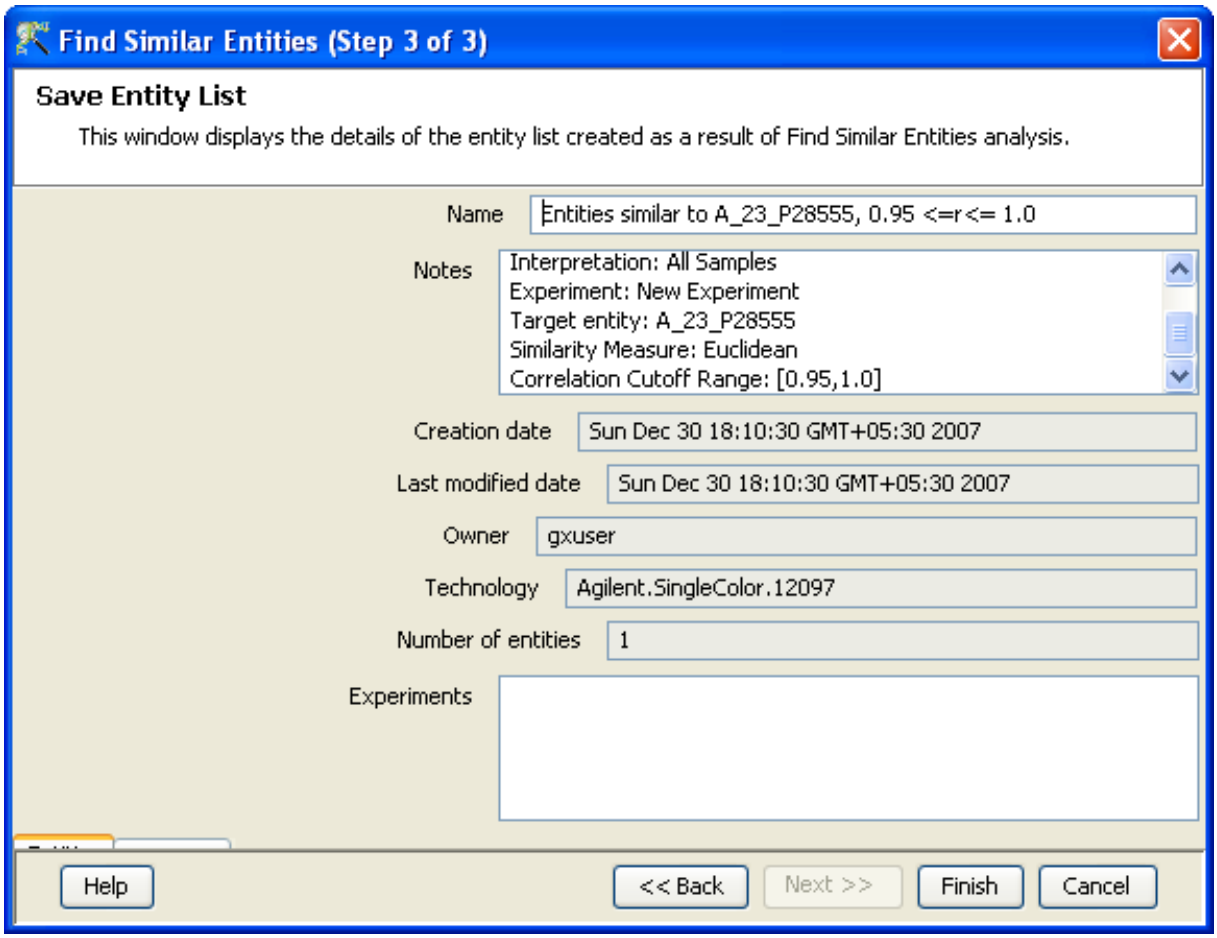


Figure 7.42: Find Similar Entities (Step 3 of 3) - Save Entity List

7.3.3.7 Filter on Parameters

Filter on Parameters calculates the correlation between expression values and parameter values. This filter allows you to find entities that show some correlation with any of the experiment parameters. This filter only works for numerical parameters.

On choosing *Filter on Parameters* under the Analysis section in the workflow,**GeneSpring** takes us through the following steps:

Step 1 of 3: This step allows the user to input parameters that are required for the analysis. The entity list and the interpretation are selected here. Also the experiment parameter of our interest has to be selected in the Parameter box. The similarity metric that can be used in the analysis can be viewed by clicking on the dropdown menu.The options that are provided are:

1. **Euclidean:** Calculates the Euclidean distance where the vector elements are the columns. The square root of the sum of the square of the A and the B vectors for each

element is calculated and then the distances are scaled between -1 and +1.
Result = (**A-B**).(**A-B**)

2. **Pearson Correlation:** Calculates the mean of all elements in vector **a**. Then it subtracts that value from each element in **a** and calls the resulting vector **A**. It does the same for **b** to make a vector **B**.
Result = **A**.**B**/(|**A**||**B**|)

3. Spearman Correlation: It orders all the elements of vector **a** and uses this order to assign a rank to each element of **a**.It makes a new vector **a'** where the i-th element in **a'** is the rank of $\mathbf{a_i}$ in **a** and then makes a vector **A** from **a**' in the same way as **A** was made from **a** in the Pearson Correlation. Similarly, it makes a vector **B** from **b**.
Result = **A**.**B**/(|**A**||**B**|)
The advantage of using Spearman Correlation is that it reduces the effect of outliers on the analysis.

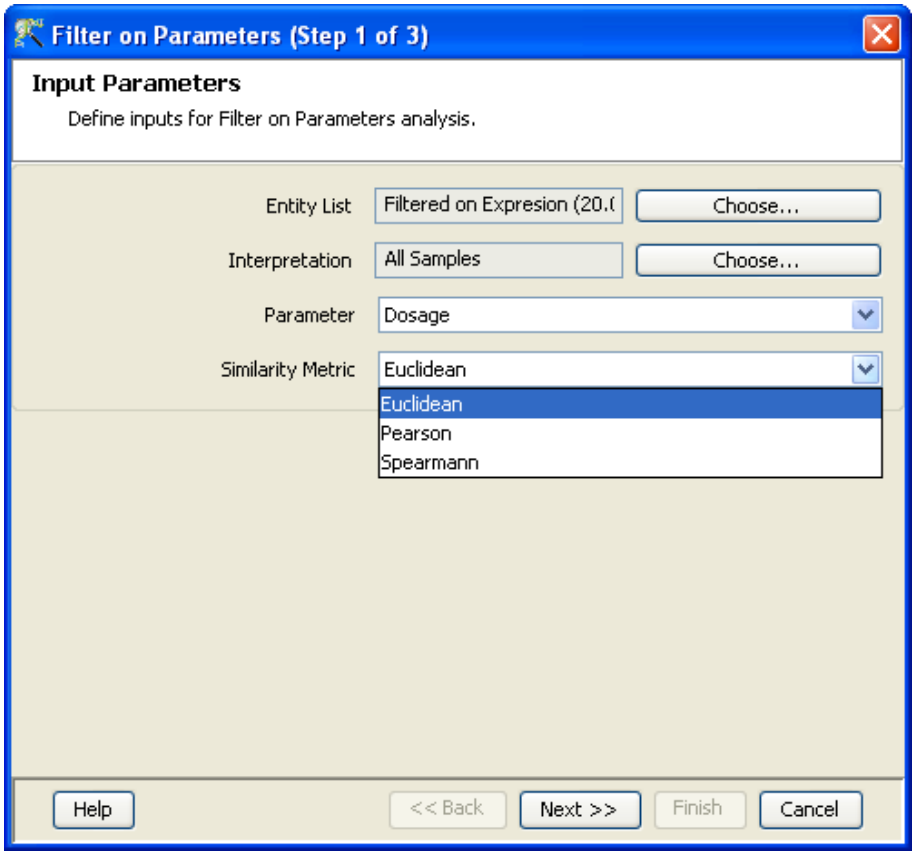


Figure 7.43: Filter on Parameters (Step 1 of 3) - Input Parameters

Step 2 of 3: This step allows the user to visualize the results of the analysis in the form of a profile plot. The plot displays the mean profile of the entities that have passed the similarity cut-off. The default range for the cutoff is Min - 0.95 and Max - 1.0.The cutoff can be altered by using the *Change Cutoff* button provided at the bottom of the wizard.Also after selecting the profiles in the plot, they can be saved as an entity list by using the option *Save Custom List*.

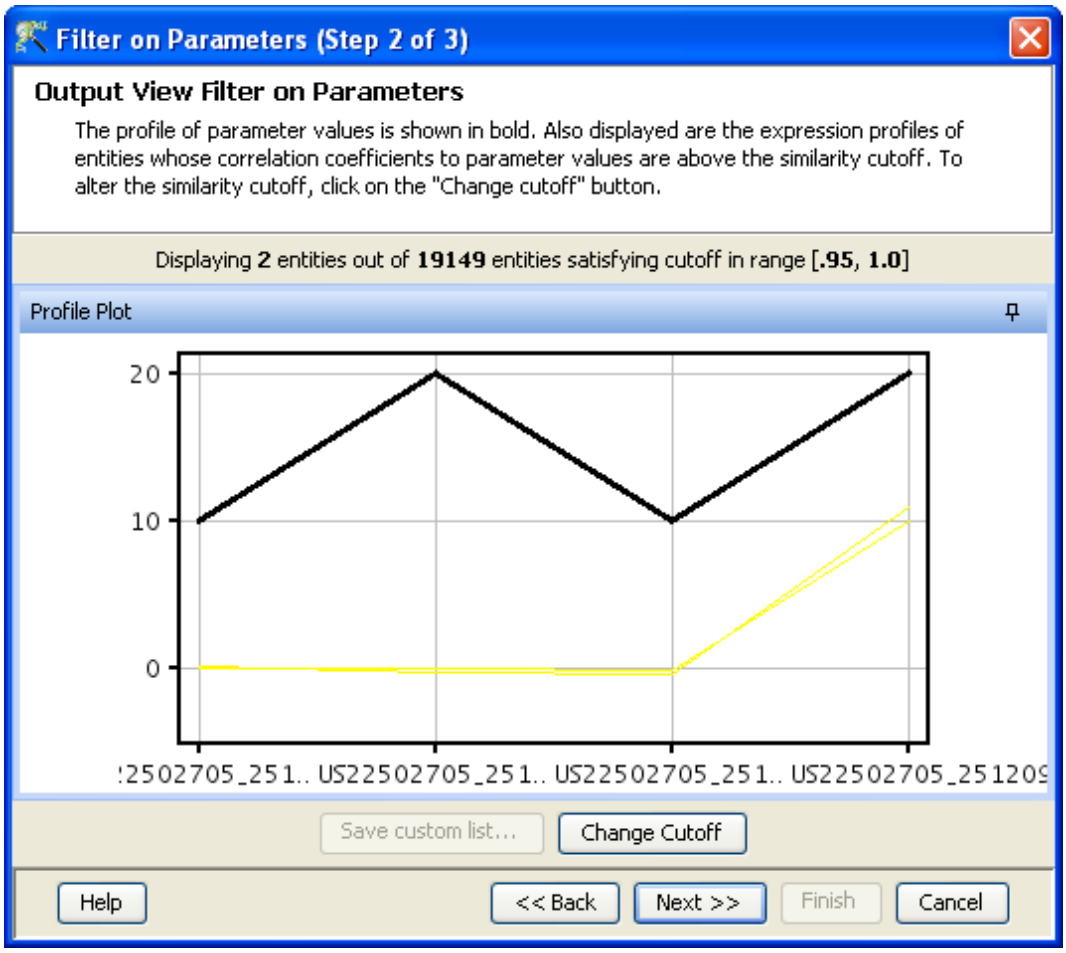


Figure 7.44: Filter on Parameters (Step 2 of 3) - Output View of Filter on Parameters

Step 3 of 3: Here, the created entity list and its details as a result of the analysis is displayed. There is also an option to configure columns that enables the user to add columns of interest from the given list. Clicking on *Finish* creates the entity list which can be visualized in the project navigator.

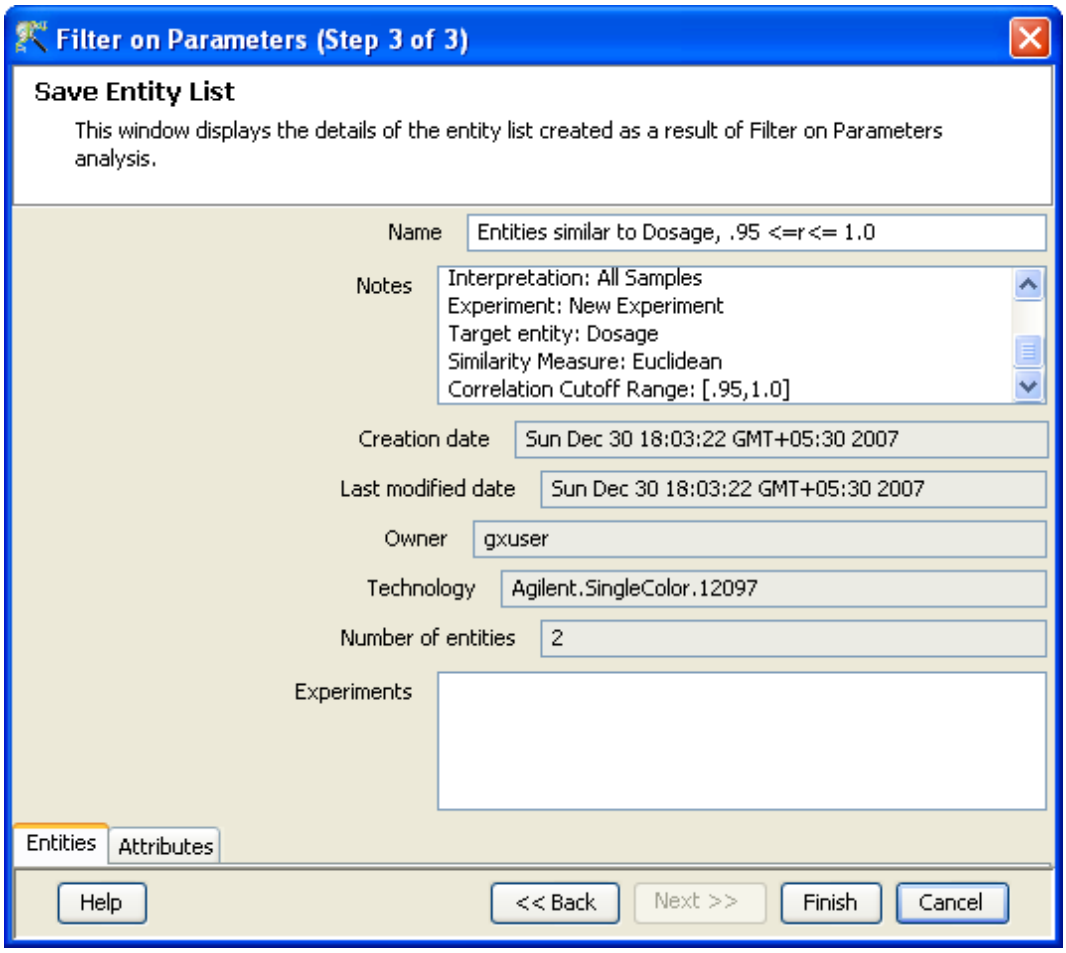


Figure 7.45: Filter on Parameters (Step 3 of 3) - Save Entity List

7.3.3.8 Principal Component Analysis

Viewing Data Separation using Principal Component Analysis:

In visualizing gene expression data of several thousand genes for each sample between various tumour types, there is enough redundancy among the large collection. This redundancy can be used to some advantage in order to reduce the dimensionality of the input data. Visualizing data in 2 or 3 dimensions is much easier than doing so in higher dimensions and the aim of dimensionality reduction is to effectively reduce the number of dimensions to 2 or 3.

This can be achieved in two ways, by;

1. Dropping lesser important dimensions, or,
2. Combining several dimensions to yield a smaller number of dimensions.

Principal Components Analysis (PCA) essentially does the latter by taking linear combinations of dimensions. Each linear combination is in fact an Eigen Vector of the similarity matrix associated with the dataset. These linear combinations (called Principal Axes) are ordered in decreasing order of associated Eigen Value. Ideally, two or three of the top few linear combinations in this ordering serve as very good set of dimensions to project and view the data in. These dimensions capture most of the information in the data.

GeneSpring supports a quick PCA implementation along with an interactive 3D viewer for the projected points in the smaller dimensional space. It clearly brings out the separation between different groups of entities (rows) or conditions (columns) whenever such separations are present in the experimental data.

Perform **Principal Component Analysis** using the following steps:

Step 1 of 4: Select an Entity List and an interpretation for performing PCA.

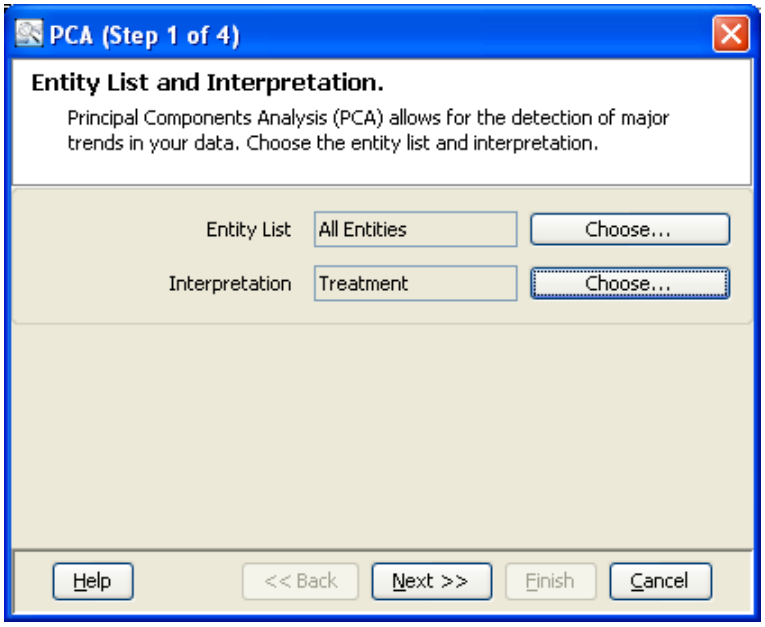


Figure 7.46: PCA (Step 1 of 4) - Entity List and Interpretation

Step 2 of 4: Input parameters for PCA in this step. You can run PCA on entities (rows) or conditions (columns) of the dataset.

You can also provide pruning options for running the PCA in this step. Ideally, only the first few eigenvectors (principal components) capture most of the variation in the data. The execution speed of PCA algorithm can be greatly enhanced when only a few eigenvectors are computed as compared to all.

You can provide the number of eigenvectors to be computed by specifying the **Number of Principal Components**. Or else, select **Total percentage variation** and specify a value for the algorithm to compute as many eigenvectors as required to capture the specified Total percentage variation in the data.

The normalization option to 'mean center' (zero mean) is applied by default. You can deselect the option to **Scale** (to unit standard deviation) the experimental data.

NOTE If the range of values in the data columns vary widely, do not deselect the **Scale** option.

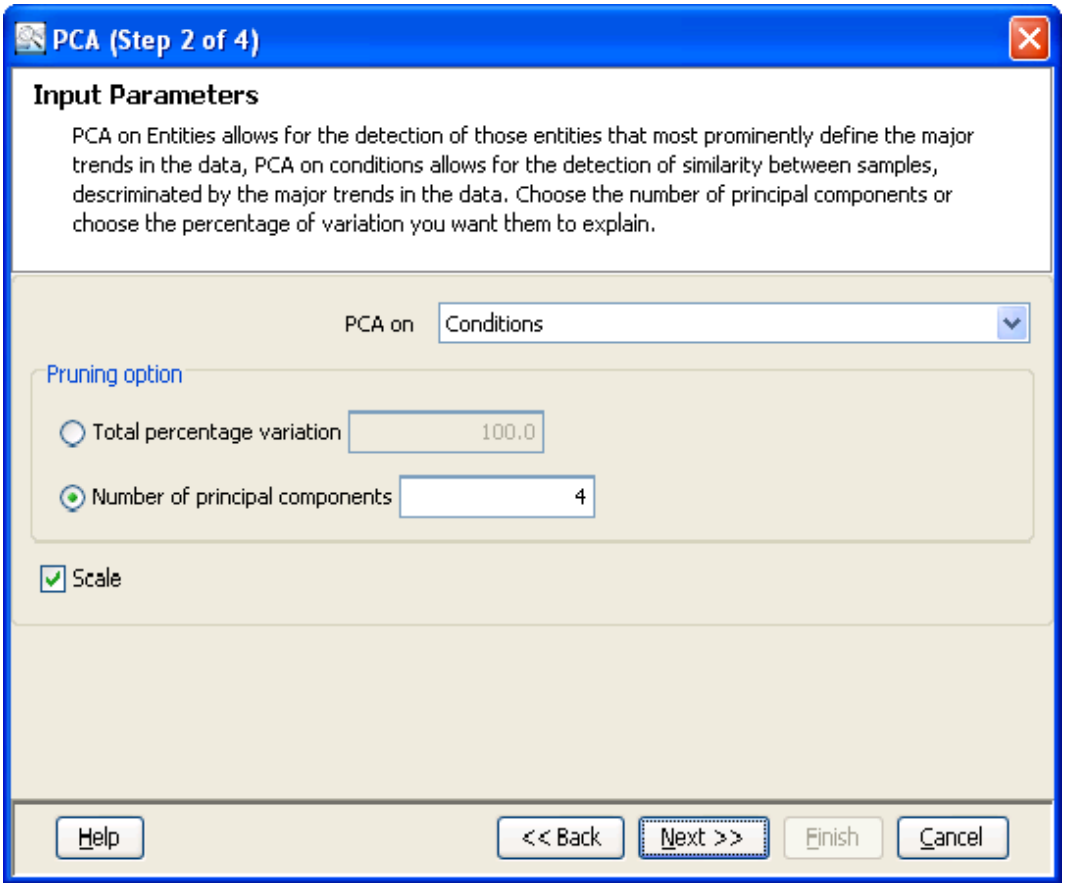


Figure 7.47: PCA (Step 2 of 4) - Input Parameters

Step 3 of 4: This step shows the output of **Principal Component Analysis**.

The six output views of PCA shown in multiple tabs are:

- **Eigen Values:** This is a plot of the Eigen values (Component 1, Component 2, etc.) on X-axis against their respective percentage contribution (Y-axis). The minimum number of principal axes required to capture most of the information in the data can be gauged from this plot. The red line indicates the actual variation captured by each eigen-value, and the blue line indicates the cumulative variation captured by all eigen values up to that point. The minimum value for PCA Eigen values is $(1 * 10^{-3})$ / (total number of Principal components) and the maximum value is the square root of the maximum float value handled by the machine.
- **PCA Scores:** This is a scatter plot of data projected along the principal axes (eigenvectors). By default, the first and second PCA components are plotted to begin with, which capture the maximum variation of the data. If the dataset has a class label column, the points are colored with respect to that column, and it is possible to visualize the separation (if any) of classes in the data. You can select different PCA components using the drop-down menu for the X-Axis and Y-Axis. Mouse over the plot to know more details of the components.

In this PCA Scores scatter plot, a T2 Hotelling Ellipse with 95% confidence is displayed on these data points (entities or samples). All data points that lie out side the ellipse can be treated as outliers. You can change the confidence level of the T2 Hotelling Ellipse from **Right-click** $\longrightarrow$**Properties** $\longrightarrow$**T2 Hotteling Ellipse** $\longrightarrow$**Confidence Percentage**.

From **Properties** $\longrightarrow$**T2 Hotteling Ellipse**, you can also choose to:

- Change the default behavior by choosing not to show a T2 Hotelling Ellipse on the PCA Scores
- Change the color and width of the T2 Hotelling Ellipse

- **3D PCA Scores:** 3D view of PCA Scores plot in which the first, second, and the third PCA components are plotted by default. This view captures the maximum variation of the data.
- **PCA Loadings:**

  For PCA on Entities, each entity is measured across, say n, samples. So imagine each entity to be a point in n dimensional space, where the axes correspond to the various samples. To visualize these entities in 2 or 3 dimensional space, the PCA algorithm first rotates the above n dimensional space so that the first 2 or 3 coordinates capture most of the variation in the data, and then restricts attention to just these 2 or 3 coordinates. The PCA Scores plot shows the entities plotted in these 2 or 3 dimensions. Each of these rotated coordinates is a linear combination of the original coordinates (recall each original coordinate corresponds to a sample), and this linear combination can be viewed in the loadings plot. For instance, if rotated coordinate 1 is $a_1x_1 + a_2x_2 + \cdots + a_nx_n$ then in the loadings plot, the x coordinate for sample 1 will be a_1, for sample 2 a_2, and so on. And if a_1, a_2 are much bigger than the rest of the a_i's in absolute value, then it means that the measurements show substantially greater variation in samples 1 and 2 than they do in the remaining samples. Thus the loadings plot can help identify which samples show greater variation in these entity measurements.

  As mentioned earlier, each principal component (or eigenvector) is a linear combination of the selected columns. The relative contribution of each column to an eigenvector is called its loading and is depicted in the PCA Loadings plot. The X-Axis consists of columns, and the Y-Axis denotes the weight contributed to an eigenvector by that column. Each eigenvector is plotted as a profile, and it is possible to visualize whether there is a certain subset of columns which overwhelmingly contribute (large absolute value of weight) to an important eigenvector; this would indicate that those columns are important distinguishing features in the whole data.
- **3D PCA Loadings:** 3D view of the PCA Loadings plot where the three axis are the three principal components (first three by default). each of the columns is represented in this space by their weight contributions to the eigen vectors corresponding to the three axis.
- **C-C Plot:** This is a Correlation - Covariance Plot. It is only enabled for PCA on conditions.

  In these Output views, entities can be selected and saved using an active **Save Custom List** button or using **Right-click** $\longrightarrow$ **Save entities** where applicable.

NOTE If you have more than 1000 entities in the selected entity list, this plot is not displayed by default. Go to **Tools** $\longrightarrow$**Options** $\longrightarrow$**Miscellaneous** $\longrightarrow$**Maximum Limit on the number of Entities/Samples to be displayed for 3D PCS plots** and change the default value.

- **Legend:** This shows the legend for the respective active tab.

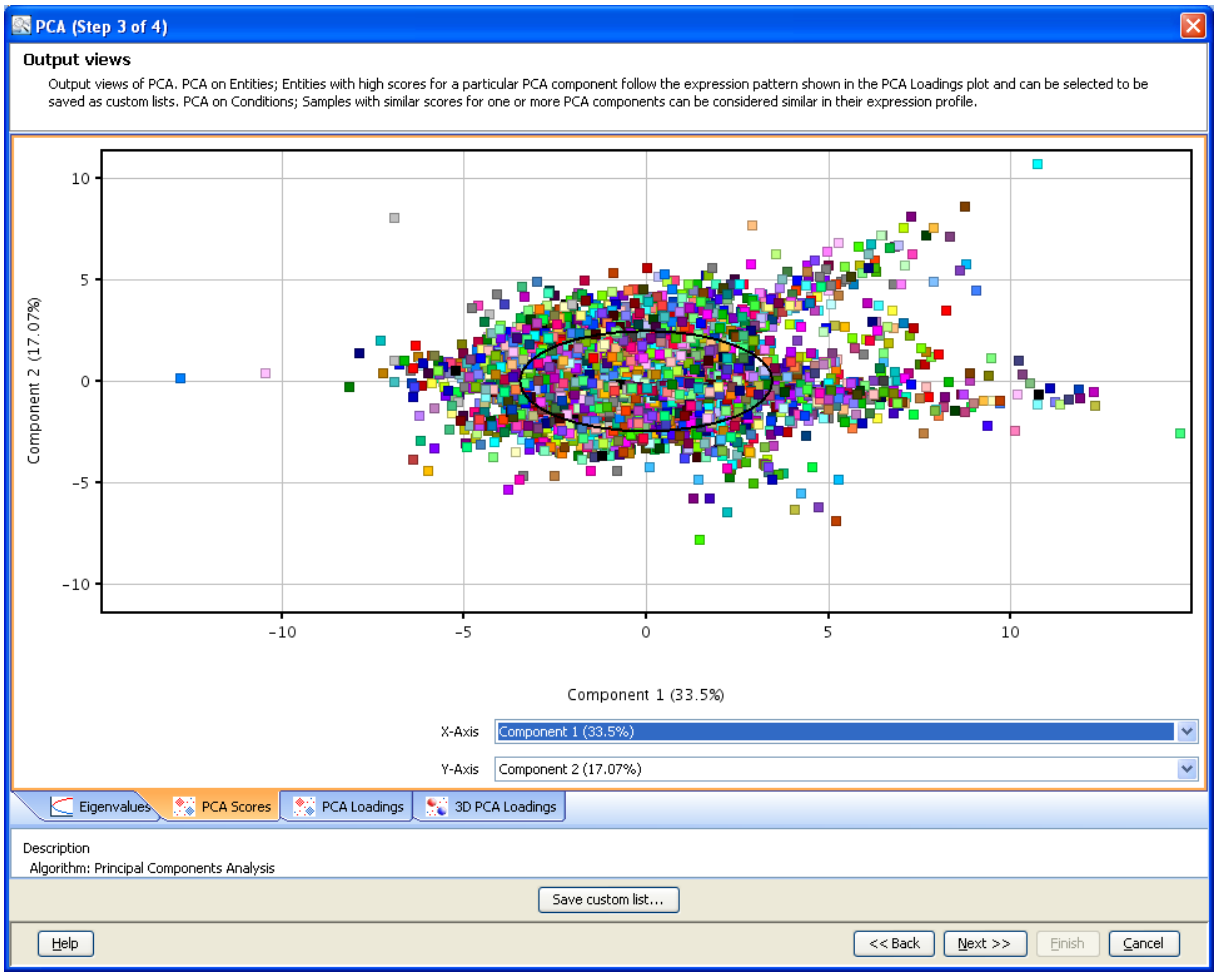


Figure 7.48: PCA (Step 3 of 4) - Output Views

Step 4 of 4 This window allows saving the output of *Principal Components Analysis*.

Finish creates a node at the same level of the input entity list used for PCA. Double-clicking on the node will re-launch the Eigenvalues plot, Score plot, loading plot, corresponding 3D plots, and the C-C plot created during PCA analysis. Right-clicking on the node and selecting 'Expand as an entity list' creates an entity list with component wise loadings value for each entity.

7.3.4 Class Prediction

GeneSpring has a variety of prediction models that include Decision Tree (DT), Neural Network (NN), Support Vector Machine (SVM), and Naive Bayesian (NB) algorithms. You can build prediction any of these prediction models on the current active experiment that will use the expression values in an entity list to predict the conditions of the interpretation in the current experiment. Once the model has been built satisfactorily, these models can be used to predict the condition given the expression values. Such prediction are being explored for diagnostic purposes from gene expression data.

7.3.4.1 Build Prediction model

For further details refer to section Build Prediction Model.

7.3.4.2 Run prediction

For further details refer to section Run Prediction.

7.3.5 Results Interpretation

This section contains algorithms that help in the interpretation of the results of statistical analysis. You may have arrived at a set of genes, or an entity list that are significantly expressed in your experiment. **GeneSpring** provides algorithms for analysis of your entity list with gene ontology terms. It also provides algorithms for Gene Set Enrichment Analysis or GSEA, which helps you compare your entity list with standard gene sets of known functionality or with your own custom gene sets. In this section, there are also algorithms that help you find entity lists that are similar to a chosen entity list.

7.3.5.1 GO Analysis

Gene Ontology Analysis provides algorithms to explore the Gene Ontology terms associated with the entities in your entity list and calculates enrichment scores for the GO terms associated with your entity list. For a detailed description of GO analysis in the refer to the chapter on GO Analysis.

7.3.5.2 GSEA

Gene set enrichment analysis is discussed in a separate chapter called Gene Set Enrichment Analysis.

7.3.5.3 Find Similar Entity Lists

Similar entity lists are entity lists that contain a significant number of overlapping entities with the one selected, the significance quantified by *p-value*. Select an entity list and start the search by defining the target entity lists and the type of target. The search can be performed even across experiments and projects and on entities belonging to different organisms and technology, provided translation is possible.

A custom search can also be performed by defining conditions based on which target entity lists will be pulled out across projects and used for search. Different conditions can be combined using 'OR' and 'AND' feature. This wide search area allows you to harness novel information on entities across populations.

The wizard to perform this operation has three steps:

1. Step 1 of 3 of *Find Similar Entity Lists*: Choose the entity list for which you want to find similar entity lists. Then choose the target entity list and the type of target on which the search will be performed. Under 'Target Entity List', if you choose 'Custom', you can define your own set of target entity lists in Step 2 of the wizard. In all other cases the wizard proceeds directly to Step 3.
2. Step 2 of 3 of *Find Similar Entity Lists*: This step is shown only if the 'Custom' option has been chosen in Step 1 under Target Entity Lists.

   On clicking *'Choose Entity Lists'*, a two-step search wizard opens up.

   (a) Step 1 of 2 of *'EntityList Search Wizard'*: In the table, choose the required search field, search condition and the value. Multiple searches can be combined by defining 'OR' or 'AND' feature from the drop down 'Combine Search Conditions by'. The conditions can also be defined based on user attributes after checking the item 'Show User Attributes'.

   (b) Step 2 of 2 of *'EntityList Search wizard'*: The results are shown in the form of a table here. Choose thoseentity lists that need to be searched on and click *Finish* to exit this wizard. The chosen entity lists are now shown as targets in the *'Find Similar Entity Lists'* wizard in step 2.
3. Step 3 of 3: Here the results are presented in the form of two tables. One table showing significant entity lists and the other showing non significant entity lists. The columns in the table list Experiment, query Entity list, Numberof entities matching with technology and the query, and p-value. The p-value is calculated using the hypergeometric distribution. This equation calculates the probability of overlap corresponding to k or more entities between an entity list of n entities whose significance we are trying to assess compared against an entity list of m entities, when randomly sampled from a universe of u genes.

   This equation calculates the probability of overlap corresponding to k or more entities between an entity list of n entities compared against an entity list of m entities when randomly sampled from a universe of u genes:

   $$P(X \geq k) = \frac{1}{\binom{u}{n}} \sum_{i=k}^{n} \binom{m}{i} \binom{u-m}{n-i}. \tag{7.1}$$

   The p-value cut-off can be changed using *Change Cutoff* button.

   To import significant entity list into the experiment, select the entity list and click *Custom Save* button. Click *Finish* and all the similar entity lists will be imported into the active experiment.

7.3.6 Pathway Analysis

In **GeneSpring**, you can visualize your experimental data on imported curated pathways and interaction networks, or create your own interaction networks using **GeneSpring**'s proprietary

Natural Language Processing (NLP) algorithm and searches based on Medical Subject Headings (MeSH) terms. You can import curated pathways from the WikiPathways portal (http://www.wikipathways.org), the BioCyc portal (http://www.biocyc.org/), from the KEGG Pathway Database (http://www.genome.jp/kegg/pathway.html), or in the BioPAX (Level 2 and Level 3) format, the GPML format, or in a Text format.

Once pathways have been imported into **GeneSpring**, or created in **GeneSpring**, they can be added to different experiments to highlight matching experimental data in the pathway. Matching entities between the pathway and the active entity list in the Experiment Navigator are highlighted by a blue border. You can also perform Multi-Omic Analysis (MOA) or Single Experiment Analysis (SEA) to find significant pathways for your research. Both SEA and MOA allow you to compare your data with pathways from different organisms than your experiment. MOA, as the name suggests, even allows you to compare data from different omics experiments (for example Expression Analysis and Metabolomics) to find significant pathways.

In addition, **GeneSpring** allows you to connect to other popular pathway analysis tools like MetaCore, Ingenuity Pathways Analysis (IPA), or Cytoscape to further analyze your data.

You can find more information on how to get a license for the Pathway Analysis Module and all the available features under the Pathway Analysis and NLP Networks chapters in this manual.

7.3.7 Utilities

This section explains utilities that are useful for the analysis. These items can be found in the workflow browser in the 'Utilities' Section.

7.3.7.1 Import Entity List from File

This option allows the user to bring any entity list of interest into the tool. Typically the entity list is a list of probeset IDs or of gene symbols. This functionality is useful when the user wants to view the expression profiles of a select set of genes in any experiment. It can also be used to see the superimposition with pathways or to explore associated GO terms. The entity list should be either in .txt, .xls, or .tsv formats. Once imported, this list will be added as a child to 'All Entities' list in the Experiment Navigator. The Entity List could be in the form of gene symbols or Probe set IDs or any other annotation which matches with the technology of the active experiment. Import Entity List dialog can be started either from the Utilies section of the workflow or by clicking on the Import Entity List from File icon on the toolbar. The dialog consists of four fields:

Choose File - This asks the user to specify the path of the file to be imported.

Choose column to match - Here the user has to choose a column that is present in the imported file. This is needed to merge the file with the dataset.

Identifier mark - The column to be imported can be either the probeset ID, Unigene Id or any other annotation. Choose the appropriate mark from the drop-down menu.

Columns to be imported - any other annotation columns to be imported from the Entity List file can be specified here. These additional columns can be brought in only if the Entity List has a Technology Identifier column, otherwise the imported column will be seen as blank.

7.3.7.2 Analysis: Biological Significance

Analysis: Biological Significance: Clicking on this option launches the Analysis: Biological Significance: Wizard. This allows the user to switch to Analysis: Biological Significance workflow from the *Data Import Wizard* when desired.

7.3.7.3 Remove Entities with Missing Signal Values

This option allows the user to remove entities which have missing values in the data file. This usually occurs in the case of custom files. This is important as Clustering and Class Prediction analysis require entity lists with 'no' missing values.

7.3.7.4 Filter on Entity List

This utility allows user to filter an Entity list using its annotations and list associated values. The filter can be set by defining a search field, a search condition like 'equals' or 'starts with', and a value for the search field, as applicable. Multiple searches can be combined using **OR** or **AND** condition.

The *Filter on Entity List* dialog can be opened from the Utilities section of the workflow.

Filter on Entity List opens a four step wizard.

Step 1 of 4 Allows selection of entity list

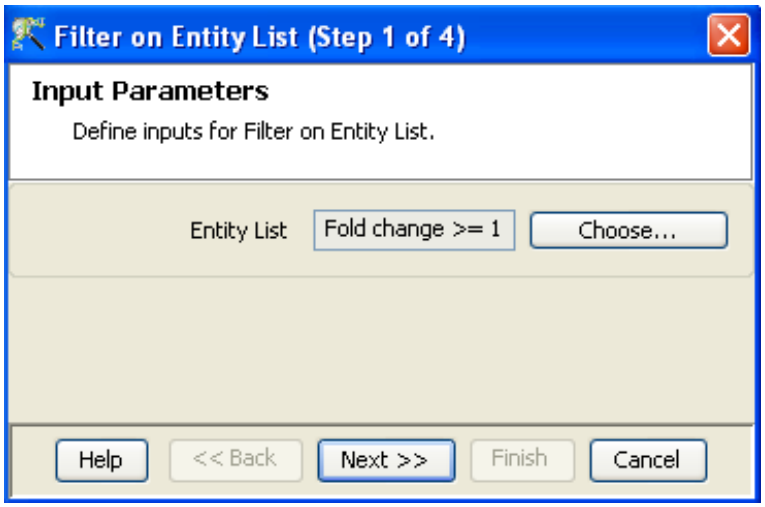


Figure 7.49: Filter on Entity List (Step 1 of 4) - Input Parameters

Step 2 of 4 Allows defining the filter conditions using three fields: Search field, Condition and Search value.

1. *Search field* Shows all the annotations and list associated values as drop down options
2. *Condition* If the selected search field is a string, the self-explanatory conditions "equals, does not equal, starts with, ends with, or includes" appear as drop down options. If the selected search field is a numerical field, (for example - Fold change), the options under Condition are their numerical equivalents, $=$, $\neq$, $\leq$, $\geq$ and 'in the range'.
3. *Search value* Allows the desired value (either string or a number, depending on the search field) to be input.

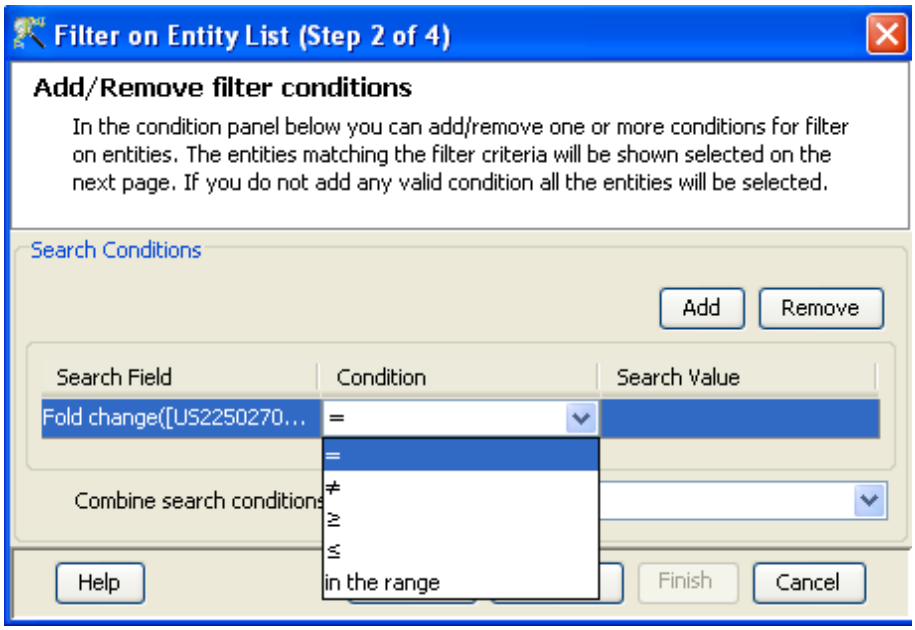


Figure 7.50: Filter on Entity List (Step 2 of 4) - Add/Remove Filter Conditions

More search conditions can be added or removed using the Add/Remove button. There is also a functionality to combine different search conditions using **OR** or **AND** conditions.

Step 3 of 4 The filter results are displayed as a table in this step. Those entities that satisfy the filter conditions are selected by default. All the entities will be selected if the filter conditions are not valid. The selections in the result page can be modified by Ctrl-click.

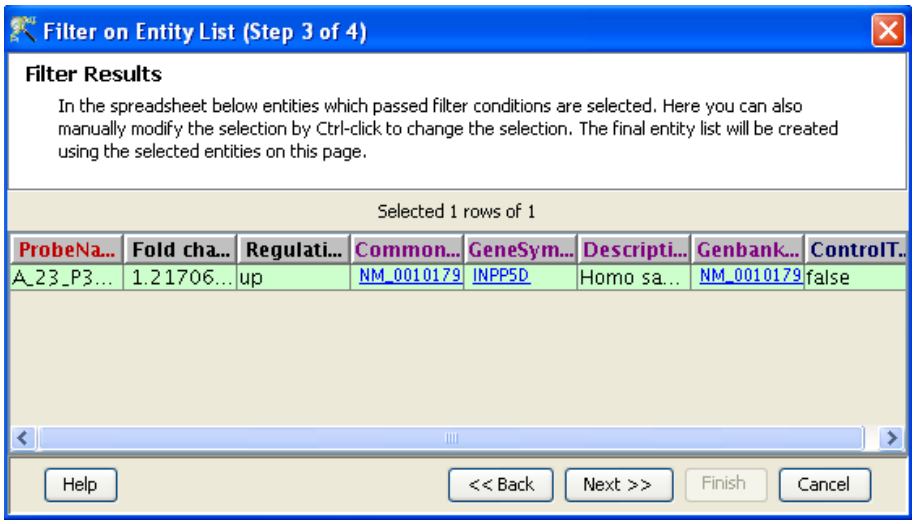


Figure 7.51: Filter on Entity List (Step 3 of 4) - Filter Results

Step 4 of 4 Allows saving the filtered entity list. Here, the columns in the entity list can be configured before saving. **Finish** will save the filtered entity list as a child node under the original entity list in the experiment.

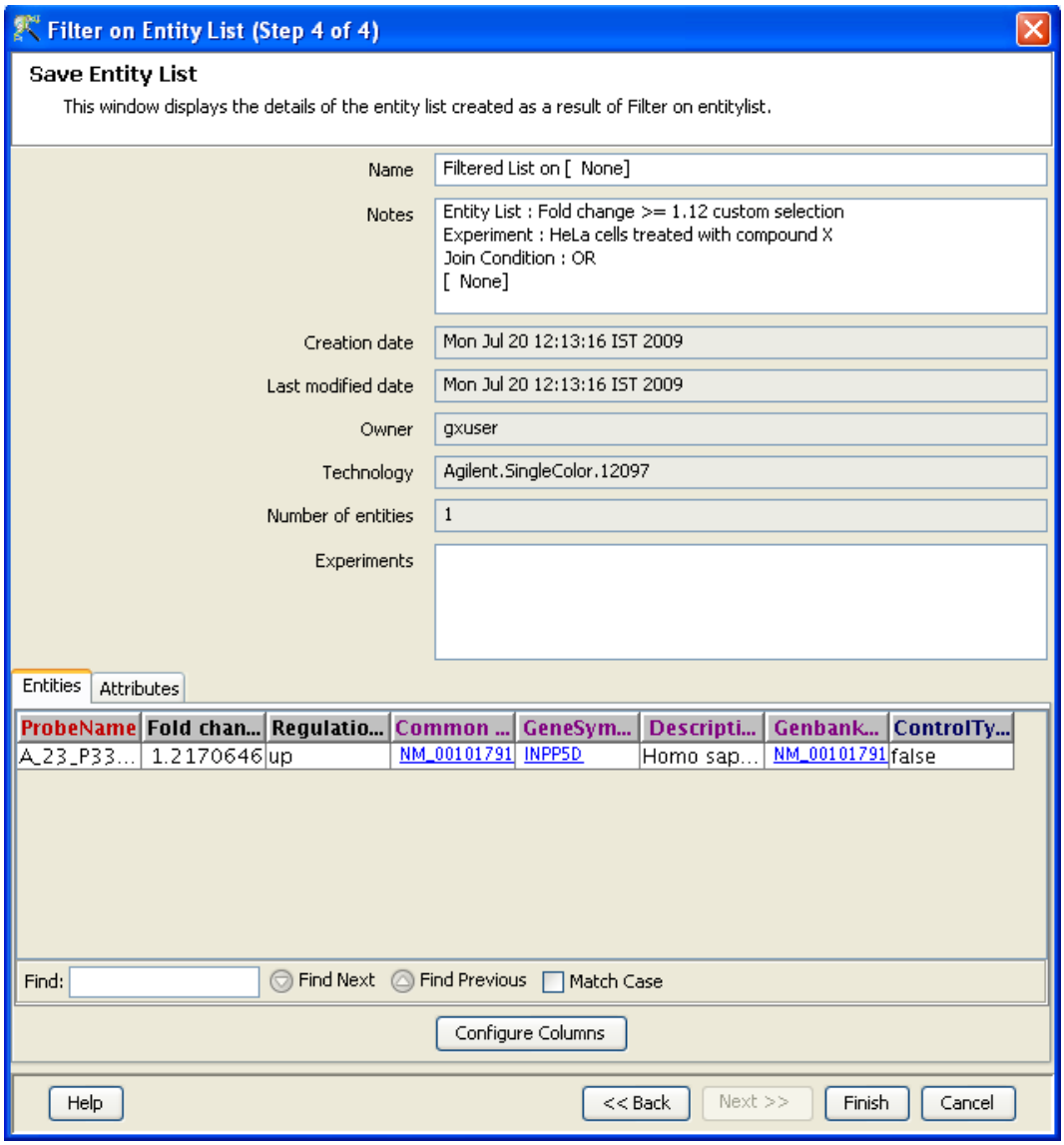


Figure 7.52: Filter on Entity List (Step 4 of 4) - Save Entity List

Chapter 8

Analyzing Agilent Single Color Expression Data

GeneSpring supports Agilent Single Color technology. The data files are in .txt format and are obtained from Agilent Feature Extraction(FE) 8.5 and 9.5.3 onwards. When the data file is imported into **GeneSpring** the following columns get imported for the purpose of experiment creation: ControlType, ProbeName, Signal and Feature Columns.

An Agilent Single Color Workflow can be used if either a single color experiment is performed or if a two color experiment is performed but subsequent analysis requires the splitting of the channel into 2 individual channels. These 2 channels can then be treated as 2 single color samples. For the latter situation, see the section on Analyzing Agilent Two Color data in Agilent Single Color Experiment Type.

The Agilent Single Color Workflow supports most of the Standard Agilent technologies. The Agilent custom arrays and the files from FE other than 8.5 and 9.5.3 can be analyzed by creating a Generic Single Color technology using the corresponding workflow. In order to do so, certain column markings should be indicated (which are automatically done with standard technologies). These details can be found in the section on Custom Agilent Arrays, while the Generic Single Color technology creation is available in Chapter 13 in the section Creating Technology.

8.1 Running the Agilent Single Color Workflow

Follow the instructions in section Running the Expression Experiment Workflow for creating an Agilent Single Colour Experiment. Figures 8.1, 8.2, 8.3, 8.4 show the process of choosing analysis type, experiment type, loading data, choosing samples and re-ordering the data files.

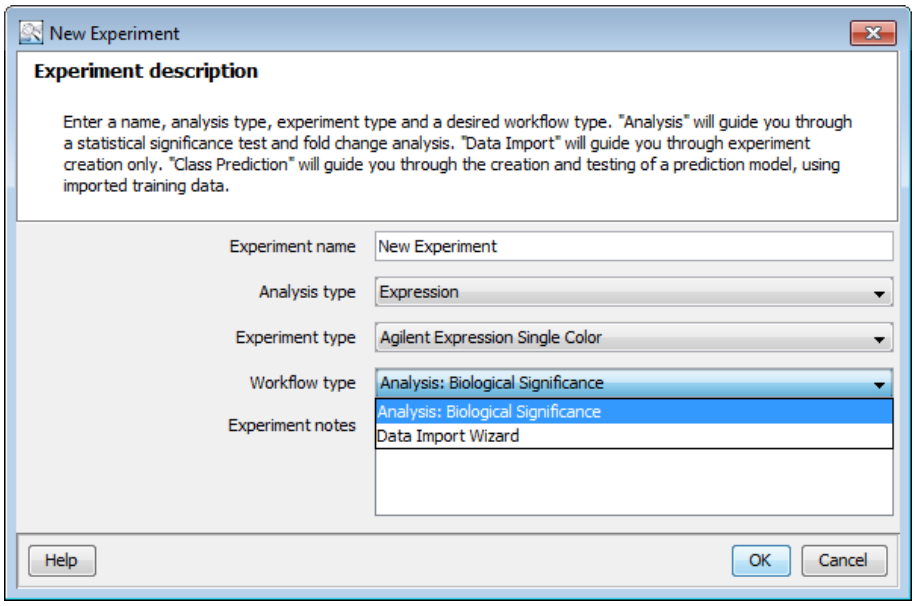


Figure 8.1: Experiment Description

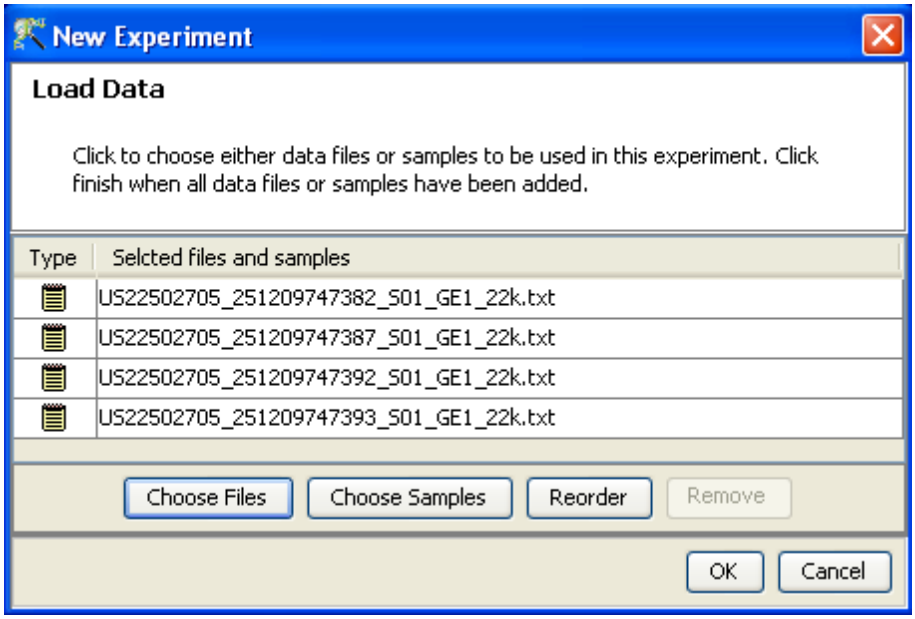


Figure 8.2: Load Data

Figure 8.3: Choose Samples

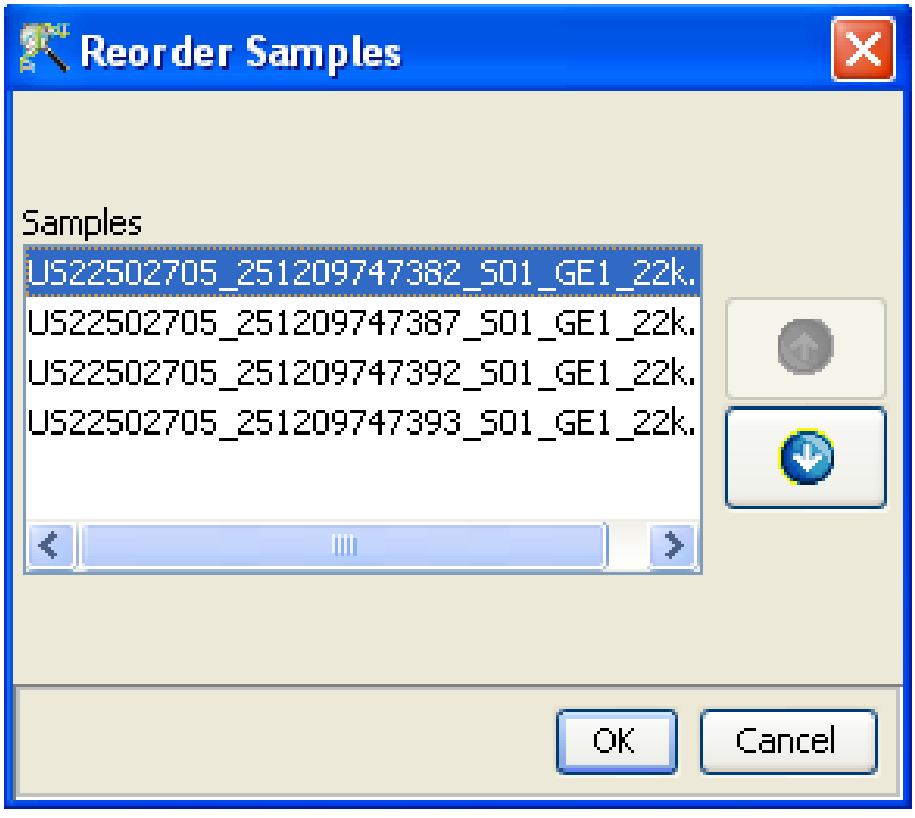


Figure 8.4: Reordering Samples

8.1.1 Analyzing Agilent Two Color data in Agilent Single Color Experiment Type

Essentially a Two Color technology can be used to analyze two samples within one slide or multiple samples in different arrays of a slide. This can be done in the following experimental designs: Imagine you have the following samples of cy3/cy5: a/b, b/c, c/d, and d/a (loop design). Although you don't have sample a/c, you can still make that comparison through b. This allows you to make more comparison pairs using less chips. However, a loop design needs special handling from an analysis perspective. So, an overly simplistic approach is to split the channels and treat them as one-color data. Another experimental design where the channel-splitting can be done is cy3/cy5: a/b, c/d, e/f and g/h. Here 'a' can be compared with either b, d, f or h.

The Two Color data from Agilent FE is Lowess normalized. After the channel-splitting, it is recommended to perform either Quantile or Median Shift normalization as well.

When an Agilent Single Color experiment is created using an Agilent Two Color file as input, a message appears asking the user if a Single Color experiment needs to be created. Clicking on *OK* splits the channels and an experiment is created. This is seen in Figure 8.5.

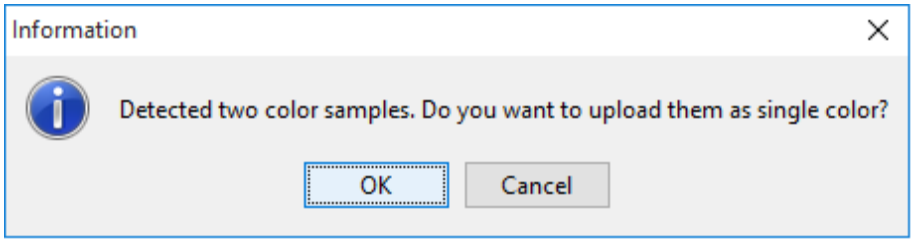


Figure 8.5: Confirmation Dialog Box

Upon clicking *OK*, the Agilent Single Color workflow appears.

Analysis: Biological Significance wizard appears with the sequence of steps on the left hand side with the current step being highlighted. The workflow allows the user to proceed in schematic fashion and does not allow the user to skip steps.

8.2 Data Processing for Agilent Single Color arrays

- **File formats**: The data files should be in text (.txt) format and obtained from Agilent Feature Extraction (FE).
- **Raw Signal Values**: The term "raw" signal values refer to the linear data after thresholding and summarization Summarization is performed by computing the geometric mean.
- **Normalized Signal Values**: "Normalized" value is the value generated after log2 transformation and normalization (Percentile Shift, Scale, Normalize to control genes or Quantile) and baseline transformation.

- **Flag values**: Flags are attributes that denote the quality of the entities. These flags are generally specific to the technology or the array type used. For Agilent microarrays, **GeneSpring** now utilizes Agilent's naming convention of Detected, Compromised, and Not Detected when importing experiments from Agilent's Feature Extraction (FE) software. This replaces the previous naming convention used in GeneSpring for Agilent microarray products. Specifically, the Agilent naming convention of Detected replaces the previous Present (P) call, Compromised replaces the Absent call (A), and Not Detected replaces the Marginal call (M).

  The resultant flag value of any probe is decided by the following logic shown. For each probe with multiple flags, the order of importance is Compromised>Not Detected>Detected. If there is even one 'Compromised', then the resultant flag is 'Compromised'. If there is no Compromised, but 'Not Detected' and 'Detected', then 'Not Detected' is assigned. If there are only 'Detected' then only the resultant flag is assigned as 'Detected'. If there are equal number of 'Not Detected' and 'Detected', then 'Not Detected'is assigned. At the end of this exercise, each probe is assigned one flag.

- **Treatment of on-chip replicates**: For each probe replicate with multiple flags, the above flag logic is applied. To get the overall flag across all replicates, **GeneSpring** excludes 'Compromised' flag and assigns the majority considering the remaining ones. If there are only 'Compromised' flags, only then the overall flag becomes 'Compromised'.

  Agilent Flag Logic:
  Probes are first considered horizontally across the 5 columns to assign a resultant flag value to each probe. For assigning overall flag for a given probe P1, the resultant flags are then considered vertically across all replicates. In Example 1, we exclude or ignore the 'Compromised' signal(1) and assign the majority flag which is: Not Detected. Hence the overall signal = $(2 + 3 + 4 + 5)/4 = 3.5$

  The following two examples illustrate this.

Replicate Probes	Signal	F1	F2	F3	F4	F5	Resultant Flag
P1	1	Detected	Not Detected	Compromised	Not Detected	Detected	Compromised
P1	2	Detected	Not Detected	Not Detected	Not Detected	Detected	Not Detected
P1	3	Detected	Detected	Detected	Detected	Detected	Detected
P1	4	Not Detected	Not Detected	Detected	Detected	Not Detected	Not Detected
P1	5	Not Detected	Detected	Not Detected	Not Detected	Detected	Not Detected

Table 8.1: Agilent Flag Logic: Example 1

In the below example, even if there is one 'Compromised' flag, the resultant flag becomes 'Compromised'. To get the overall flag across replicates, if there is a compromised flag, we normally exclude or ignore it. But in this example, since there are only 'Compromised' flags, the overall flag is assigned 'Compromised' for this probe. Overall signal = (1+2+3+4+5)/5 = 3

- **Treatment of Control probes**: The control probes are included while performing normalization. However there should be an exact match between the control probes in the

Replicate Probes	Signal	F1	F2	F3	F4	F5	Resultant Flag
P1	1	Detected	Not Detected	Compromised	Not Detected	Detected	Compromised
P1	2	Detected	Not Detected	Not Detected	Compromised	Detected	Compromised
P1	3	Detected	Detected	Compromised	Detected	Detected	Compromised
P1	4	Not Detected	Not Detected	Detected	Compromised	Not Detected	Compromised
P1	5	Not Detected	Detected	Compromised	Not Detected	Detected	Compromised

Table 8.2: Agilent Flag Logic: Example 2

technology and the sample for the probes to be utilized, as the comparison between the identifier columns is case-sensitive.

- **Empty Cells**: Not Applicable.
- **Sequence of events**: The sequence of events involved in the processing of the data files is: Thresholding $\longrightarrow$ Summarization (summarization is performed by computing the geometric mean) $\longrightarrow$ log transformation$\longrightarrow$ Normalization$\longrightarrow$ Baseline Transformation.

8.3 Analysis: Biological Significance

The Analysis: Biological Significance workflow has 8 steps:

1. Summary Report
2. Experiment Grouping
3. QC on samples
4. Filter Probes
5. Significance analysis
6. Fold Change
7. GO analysis
8. Pathway Analysis

These steps are explained in section Analysis: Biological Significance. Details specific to Affymetrix Expression Data are explained here.

Quality Control (Step 3 of 8):

The 3rd step in the Analysis: Biological Significance workflow is the QC on samples which is displayed in the form of four tiled windows. They are as follows:

- Quality controls Metrics- Report and Experiment grouping tabs
- Quality Controls Metrics- Plot
- 3D PCA scores.
- Legend

QC on Samples generates four tiled windows as seen in Figure 8.6.

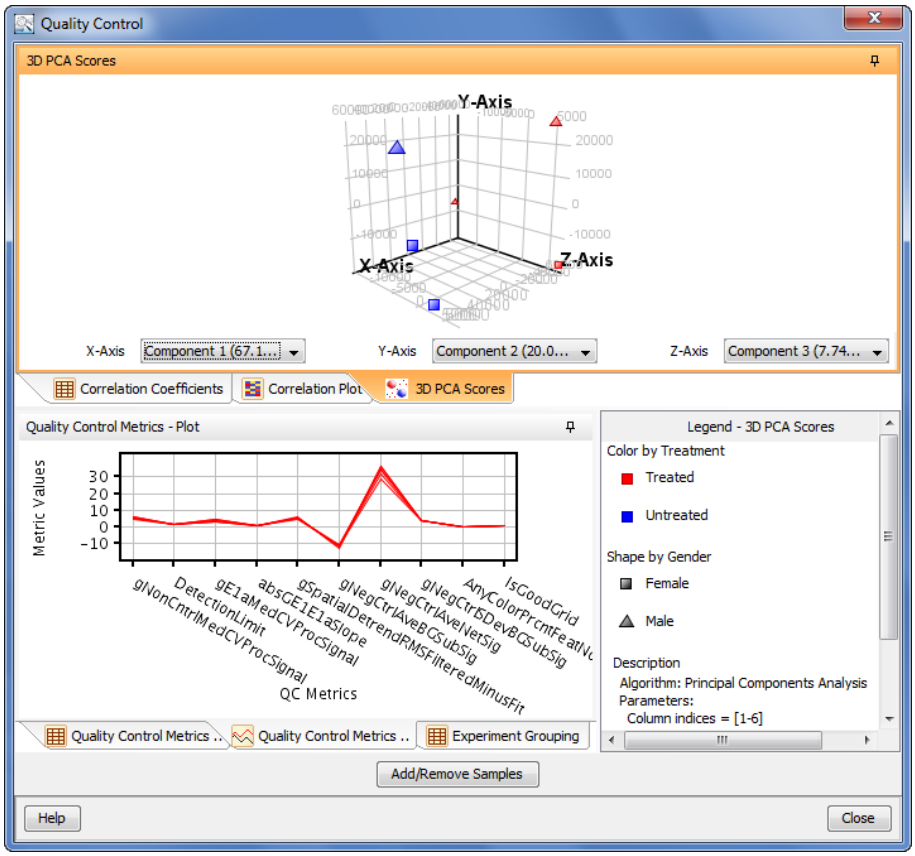


Figure 8.6: Quality Control on Samples

The *Metrics Report* has statistical results to help you evaluate the reproducibility and reliability of your single color microarray data.

Quality controls *Metrics Plot* shows the QC metrics present in the QC report in the form of a plot. Please refer to Agilent Feature Extraction Software Reference Guide (v10.7.3 or higher) for detailed descriptions of the QC Metrics.

Principal Component Analysis (PCA) calculates the PCA scores and visually represents them in a 3D scatter plot. The scores are used to check data quality. It shows one point per array and is colored by the *Experiment Factors* provided earlier in the *Experiment Groupings* view. This allows viewing of separations between groups of replicates. Ideally, replicates within a group should cluster together and separately from arrays in other groups. The PCA components, represented in the X, Y and Z axes are numbered 1, 2, 3... according to their decreasing significance. The 3D PCA scores plot can be customized via ***Right-Click*** $\longrightarrow$ ***Properties***. To zoom into a 3D Scatter plot, press the Shift key and simultaneously hold down the left mouse button and move the mouse upwards. To zoom out, move the mouse downwards instead. To rotate, press the Ctrl key, simultaneously hold down the left mouse button and move the mouse around the plot.

The *Add/Remove* samples allows the user to remove the unsatisfactory samples and to add the samples back if required. Whenever samples are removed or added back, normalization as well as baseline transformation is performed again on the samples. Click on *OK* to proceed.

The fourth window shows the legend of the active QC tab.

8.4 Data Import Wizard

The *Data Import Wizard* offers a variety of choices to the user for the analysis. Flag options can be changed and raw signal thresholding can be altered. Additionally there are options for baseline transformation of the data and for creating different interpretations. To create and analyze an experiment using the *Data Import Wizard*, load the data as described earlier. In the **New Experiment Dialog**, choose the **Workflow Type** as Data Import Wizard. Clicking on ***OK*** will open a new experiment wizard which then proceeds as follows:

1. **Step 1 of 5: Load Data** As in case of *Analysis: Biological Significance*, either data files can be imported or else pre-created samples can be used.

   - For loading new txt files, use *Choose Files*.
   - If the txt files have been previously used in **GeneSpring** experiments *Choose Samples* can be used.

   Step 1 of Experiment Creation, the 'Load Data' window, is shown in Figure 8.7.

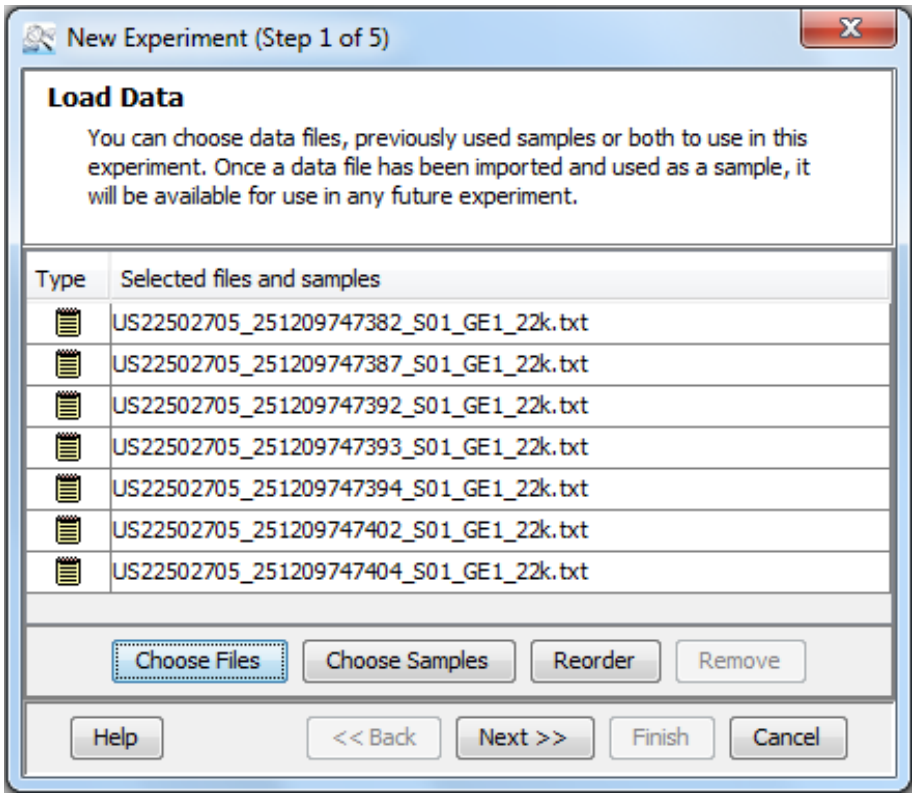


Figure 8.7: Load Data

2. **Step 2 of 5: Advanced Flag Import** This gives the options for importing flag information. The information is derived from the Feature columns in data file. User has the option of changing the default flag settings that appear in this step. The 'Save as Default' handle allows saving the current flag settings under the tool configuration. When a file is imported, **GeneSpring** will show these saved default setting in this step, by default. The settings can be changed either in this wizard or from *Tools* → *Options* → *Miscellaneous* → *Agilent Flag Settings*.

   Step 2 of Experiment Creation, the 'Advanced flag Import' window, is depicted in the Figure 8.8.

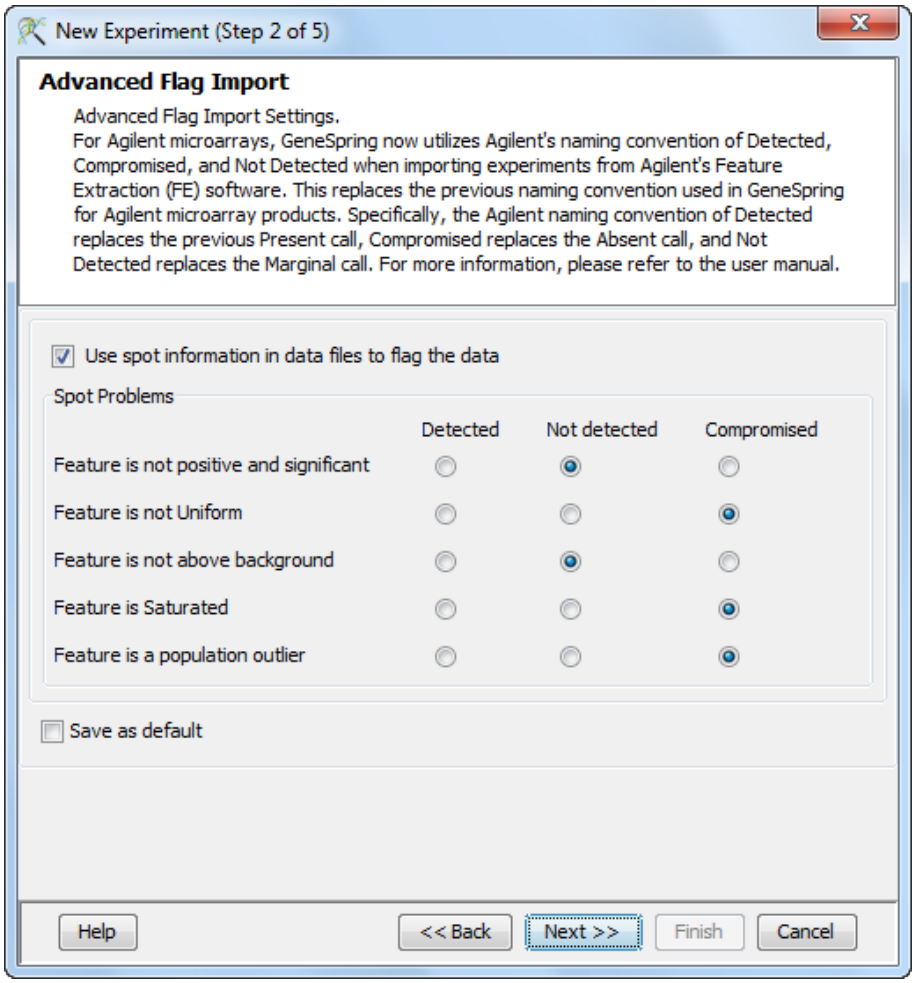


Figure 8.8: Advanced flag Import

3. **Step 3 of 5: Normalization Options**

   Criteria for preprocessing of input data is set here. It allows the user to threshold raw signals to chosen values and select normalization algorithms (None, Percentile Shift, Scale, Quantile, Normalize to control genes or Normalize to External Value).

   - **None:** No normalization is done. The threshold value for raw signals is set to 1.0 and can be changed by the user and all other options are disabled.
   - **Percentile Shift:** On selecting this normalization method, the **Percentile Target** box gets enabled allowing the user to enter a specific percentile value. The default value is set at 75^th percentile.
   - **Scale:** On selecting this normalization method, the user is presented with an option to scale it to the median of percentile target/mean of means of 'all' or 'control' samples. On choosing the latter, the user has to select the control samples from the available samples in the **Choose Samples** box. The **Percentile Target** box is set at a default value of 75.
   - **Normalize to control genes:** After selecting this option, the user has to specify the control genes in the next wizard. The **Percentile Target** box is disabled.
   - **Normalize to External Value:** This option will bring up a table listing all samples and a default scaling factor of '1.0' against each of them. The user can use the *'Assign*

Value' button at the bottom to assign a different scaling factor to each of the sample; multiple samples can be chosen simultaneously and assigned a value. The **Percentile Target** box is disabled.

For details on the above normalization methods, refer to section Normalization Algorithms.

Figure 8.9 illustrates the Step 3 of Experiment Creation.

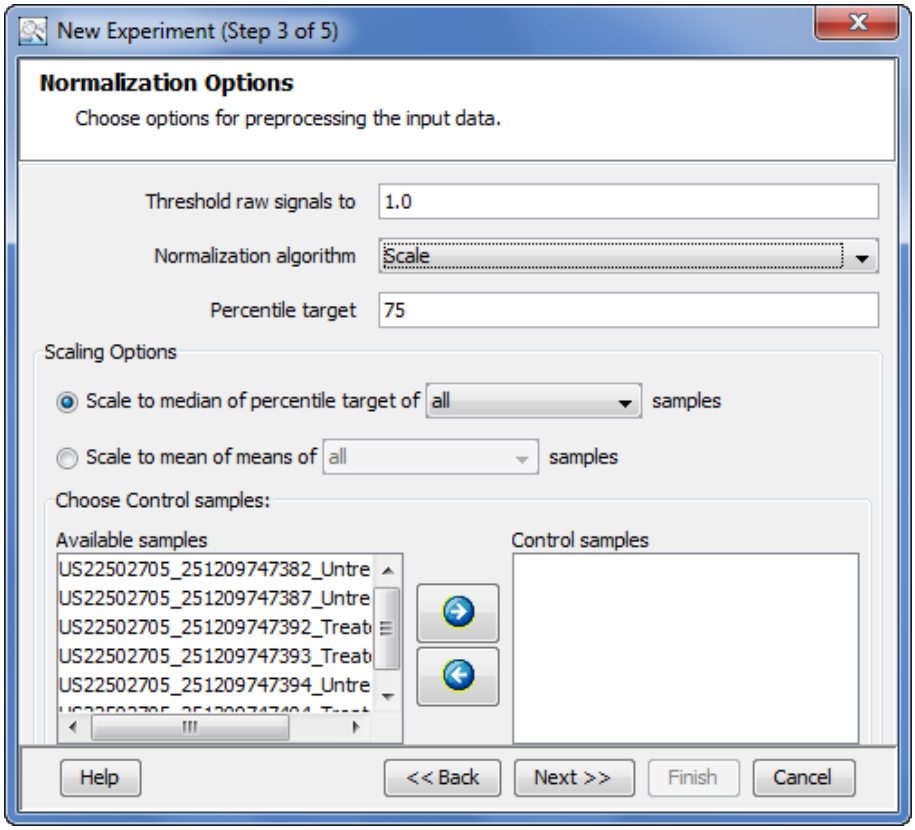


Figure 8.9: Preprocess Options

4. **Step 4 of 5: Choose entities**

   If the **Normalize to control genes** option was chosen in step 3, then the list of control entities can be specified in the following ways in this wizard:

   - By choosing a file(s) (txt, csv or tsv) which contains the control entities of choice denoted by their probe id. Any other annotation will not be suitable.
   - By searching for a particular entity by using the ***Choose Entities*** option. This leads to a search wizard in which the entities can be selected. All the annotation columns present in the technology are provided and the user can search using terms from any of the columns. The user has to select the entities that he/she wants to use as controls when they appear in the **Output Views** page and then click ***Finish***. This will result in the entities getting selected as control entities and will appear in the wizard.

   The user can choose either one or both the options to select his/her control genes. The chosen genes can also be removed after selecting the same. See Figure 8.10.

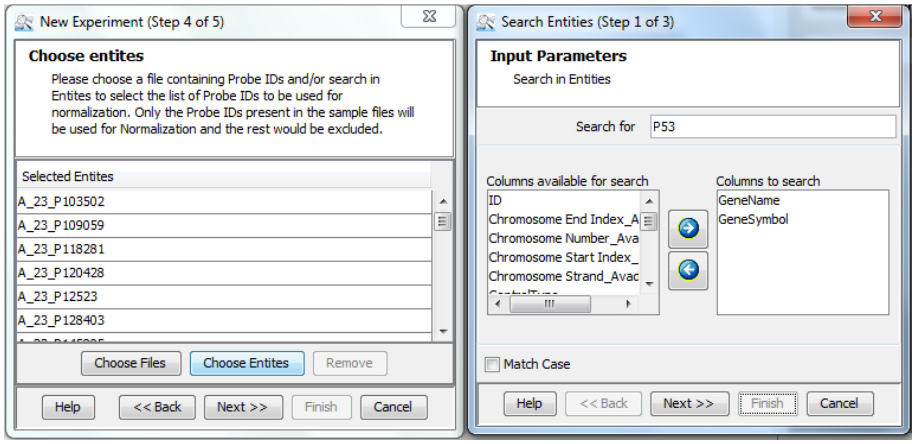


Figure 8.10: Normalize to control genes

In case the entities chosen are not present in the technology or sample, they will not be taken into account during experiment creation. The entities which are present in the process of experiment creation will appear under matched probe ids whereas the entities not present will appear under unmatched probe ids in the experiment notes in the experiment inspector.

5. **Step 5 of 5: Preprocess Baseline Options**

   This step allows the user to perform baseline transformation. See Figure 8.11.

   The baseline options include:

   - ***Do not perform baseline*** No transformation is done.
   - ***Baseline to median of all samples:*** For each probe the median of the log summarized values from all the samples is calculated and subtracted from each of the samples.
   - ***Baseline to median of control samples:*** For each sample, an individual control or a set of controls can be assigned. Alternatively, a set of samples designated as controls can be used for all samples. For specifying the control for a sample, select the sample and click on ***Assign value***. This opens up the ***Choose Control Samples*** window. The samples designated as Controls should be moved from the *Available Items* box to the *Selected Items* box. Click on ***Ok***. This will show the control samples for each of the samples.

     In *Baseline to median of control samples*, for each probe the median of the log summarized values from the control samples is first computed and then this is subtracted from the sample. If a single sample is chosen as the control sample, then the probe values of the control sample are subtracted from its corresponding sample.

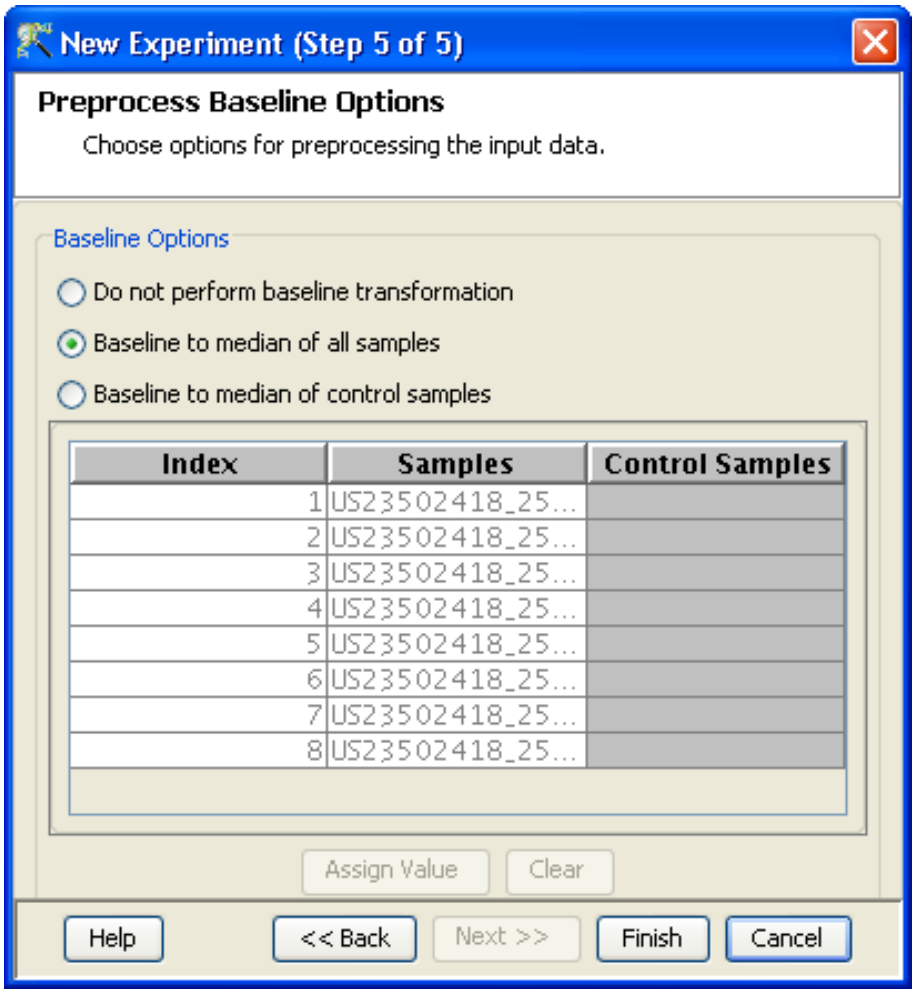


Figure 8.11: Baseline Transformation Options

Clicking *Finish* creates an experiment, which is displayed as a Box Whisker plot in the active view. Alternative views can be chosen for display by navigating to *View* in Toolbar.

8.4.1 Experiment Setup

Once an experiment is created, the *Data Import Wizard* steps appear on the right hand side. This gives options to view quick start guide, create experiment grouping and interpretation and the utility to create gene level experiment from probe level experiment. See section Experiment Setup for complete details.

8.4.2 Quality Control

- **Quality Control on Samples:**

Quality Control (QC) allows you decide which samples are ambiguous and which are passing the quality criteria. Based on the QC results, unreliable you can remove samples from further downstream analysis. The QC dialog displays three tiled windows:

- 3D PCA scores, Correlation Analysis - Samples, and Correlation Coefficients tab.
- Quality Control Metrics - Report, Quality Control Metrics - Plot, and Experiment Grouping.
- Legend.

Figure 8.12 has tiled windows which shows the QC on Samples.

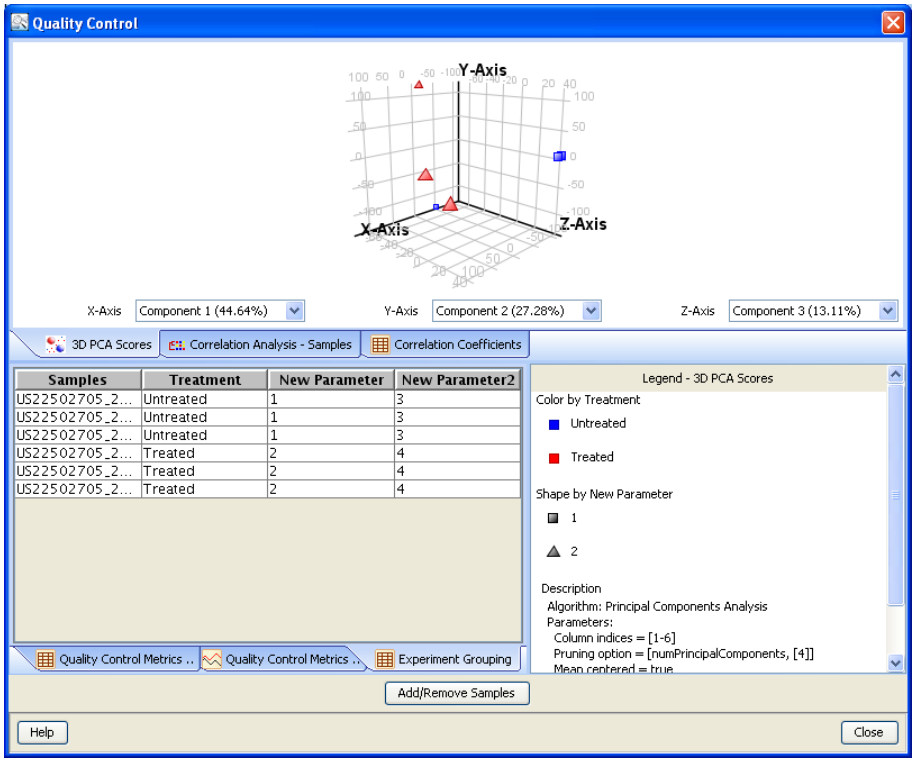


Figure 8.12: Quality Control Dialog

In the **Principal Component Analysis (PCA)** tab, PCA scores are visually represented in a 3D scatter plot. These scores are used to check data quality. It shows one point per array and is colored by the *Experiment Factors* provided earlier in the *Experiment Groupings* view. This allows viewing of separations between groups of replicates. Ideally, replicates within a group should cluster together and separately from arrays in other groups. The PCA components, represented in the X, Y, and Z axes are numbered 1, 2, 3... according to their decreasing significance. The 3D PCA scores plot can be customized via ***Right-Click*⟶*Properties***. To zoom into a 3D Scatter plot, press the Shift key and simultaneously hold down the left mouse button and move the mouse upwards. To zoom out, move the mouse downwards instead. To rotate, press the Ctrl key, simultaneously hold down the left mouse button and move the mouse around the plot.

The **Correlation Analysis - Samples** tab shows the correlation analysis heatmap across arrays (samples) (See How to Perform Correlation Analysis on Samples on page 750). The correlation coefficient for each pair of arrays are displayed in the **Correlation Coefficients**

tab. Correlation coefficients are calculated using Pearson Similarity Metric (see Pearson on page 799).

Pearson Similarity Metric: Calculates the mean of all elements in vector **a**. Then it subtracts that value from each element in **a** and calls the resulting vector **A**. It does the same for **b** to make a vector **B**. Correlation Coefficient is then computed as **A.B**/(—**A**——**B**—).

In a Correlation Analysis heatmap:

- Sample Order- In a correlation heatmap, the initial order of samples is the same as the input sample order at the time of experiment creation.
- Clustering - Perform clustering analysis on the correlation heatmap using the **Run clustering** icon from the correlation analysis heatmap toolbar (See To run clustering on a correlation heatmap on page 760). Samples are clustered based on correlation coefficients using hierarchical clustering algorithm and plotted in the order they cluster together. The linkage rule used is average and the distance metrics used is Euclidean.

For more information on the available toolbar icons in a correlation analysis heatmap, see Correlation Heatmap View Toolbar on page 776. See Properties on 792 to change the properties of the correlation analysis heatmap.

NOTE Correlation coefficients are computed on raw,non-normalized data and in linear scale.

The **Quality Control Metrics - Report** include the statistical results to help you evaluate the reproducibility and reliability of your single microarray data.

Quality controls Metrics - Plot shows the QC metrics present in the QC report in the form of a plot. Please refer to Agilent Feature Extraction Software Reference Guide (v10.7.3 or higher) for detailed descriptions of the QC Metrics.

Experiment Grouping shows the parameters and parameter values for each sample.

The third window shows the legend of the active QC tab.

Unsatisfactory samples or those that have not passed the QC criteria can be removed from further analysis, at this stage, using *Add/Remove Samples* button. Once a few samples are removed, re-normalization and baseline transformation of the remaining samples is carried out again. The samples removed earlier can also be added back. Click on **OK** to proceed.

- **Filter Probe Set by Expression:** Entities are filtered based on their signal intensity values. For details refer to the section on Filter Probesets by Expression
- **Filter Probe Set by Flags:** In this step, the entities are filtered based on their flag values, Detected, Not Detected and Compromised. Users can set what proportion of conditions must meet a certain threshold. The flag values that are defined at the creation of the new experiment (Step 2 of 3) are taken into consideration while filtering the entities. The filtration is done in 4 steps:

  1. Step 1 of 4 : *Entity list and interpretation* window opens up. Select an entity list by clicking on *Choose Entity List* button. Likewise by clicking on *Choose Interpretation* button, select the required interpretation from the navigator window. This is seen in Figure 8.13

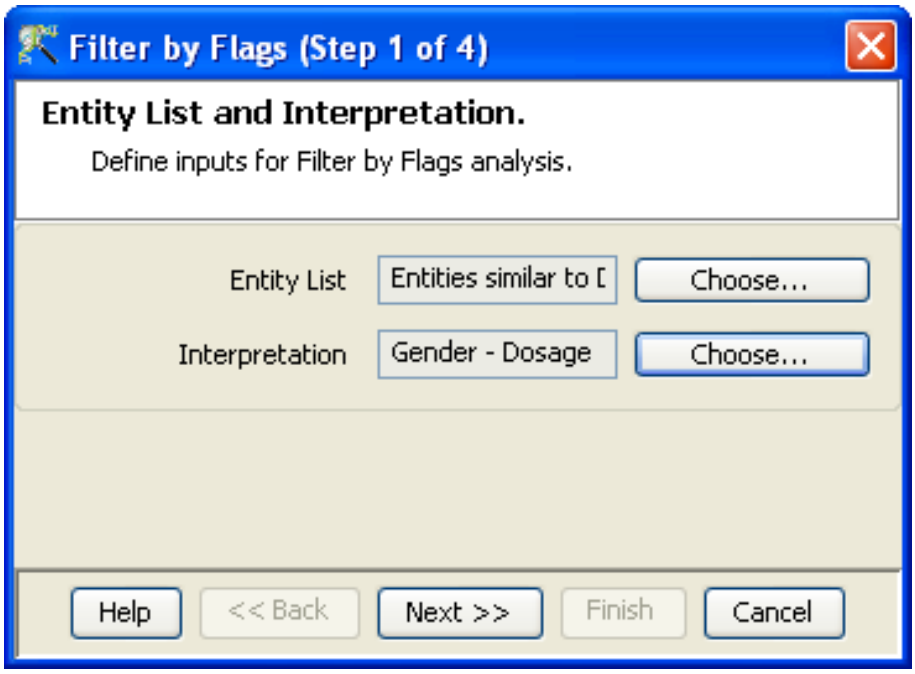


Figure 8.13: Entity List and Interpretation

2. Step 2 of 4: This step is used to set the Filtering criteria and the stringency of the filter. Select the flag values that an entity must satisfy to pass the filter. By default, the Detected and Not Detected flags are selected. Stringency of the filter can be set in *Retain Entities* box (See Figure 8.14).

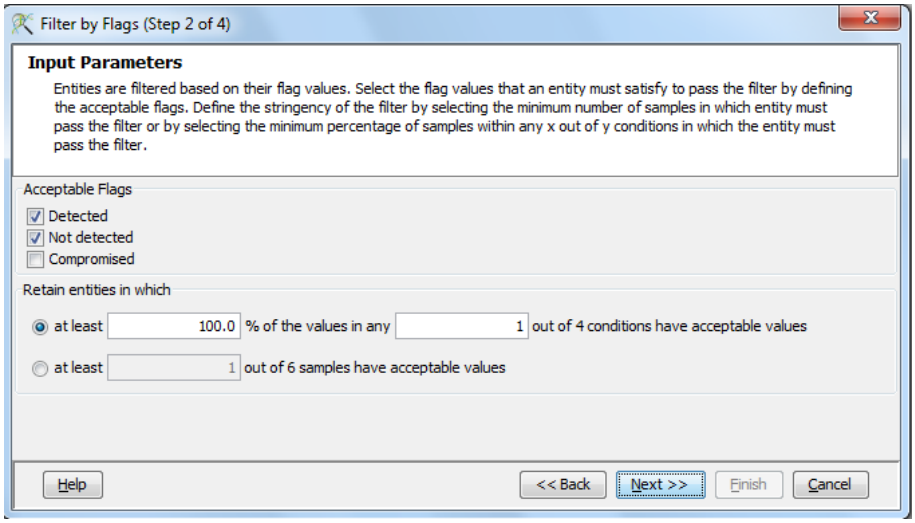


Figure 8.14: Input Parameters

3. Step 3 of 4: A spreadsheet and a profile plot appear as 2 tabs, displaying those probes which have passed the filter conditions. Baseline transformed data is shown here. Total number of probes and number of probes passing the filter are displayed on the top of the navigator window (See Figure 8.15).

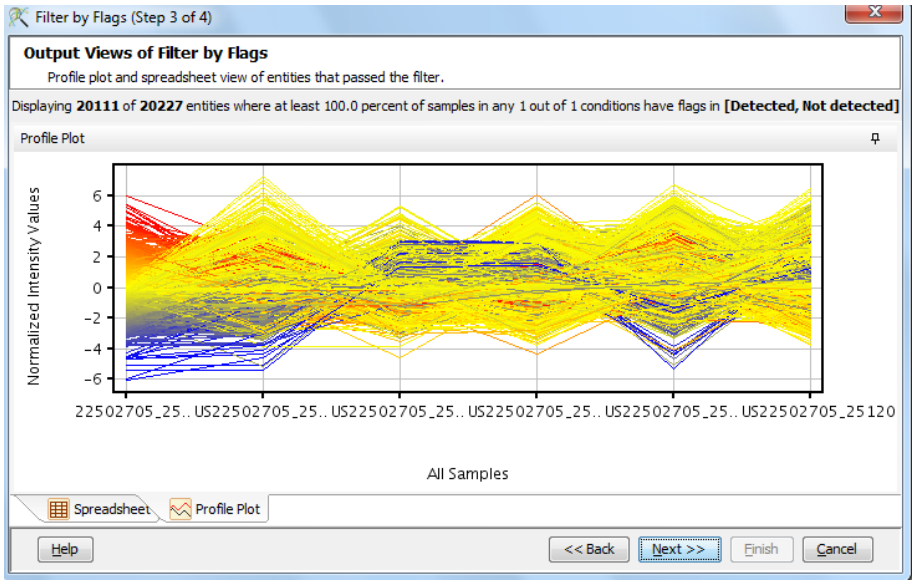


Figure 8.15: Output Views of Filter by Flags

4. Step 4 of 4: Click *Next* to annotate and save the entity list. See Figure 8.16.

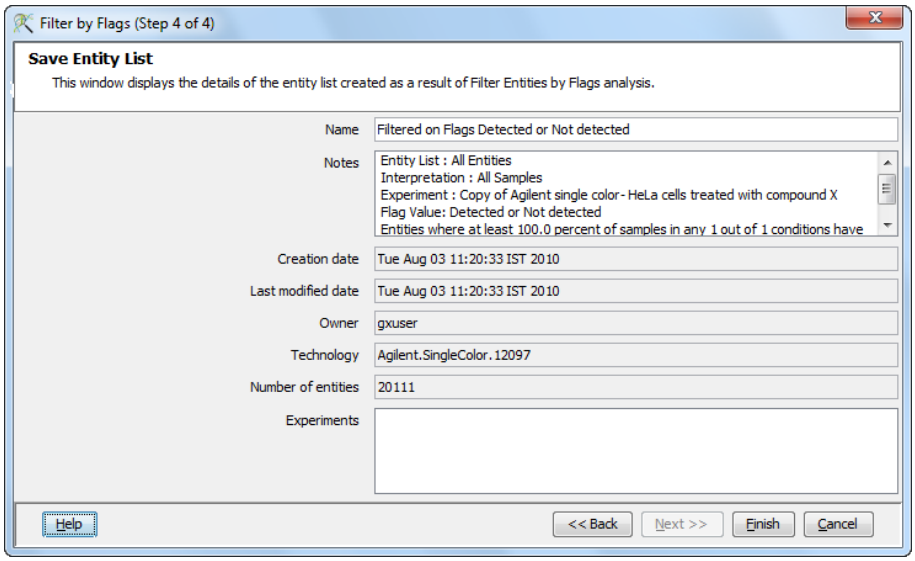


Figure 8.16: Save Entity List

- **Filter Probesets on Data Files:** Entities can be filtered based on values in a specific column of the original data files. For details refer to the section on Filter Probesets on Data Files.
- **Filter Probesets by Error:** Entities can be filtered based on the standard deviation or coefficient of variation using this option. For details refer to the section on Filter Probesets by Error.

8.4.3 Analysis

- **Statistical Analysis**
  For details refer to section Statistical Analysis in the Data Import Wizard section.
- **Filter on Volcano Plot**
  For details refer to section Filter on Volcano Plot.
- **Fold Change**
  For details refer to section Fold change.
- **Clustering**
  For details refer to chapter Clustering Analysis.
- **Correlation**
  For details refer to chapter Correlation Analysis.
- **Find Similar Entities**
  For details refer to section Find similar entities.
- **Filter on Parameters**
  For details refer to section Filter on Parameters.
- **Principal Component Analysis**
  For details refer to section PCA.

8.4.4 Class Prediction

- **Build Prediction Model** For details refer to section Build Prediction Model.
- **Run Prediction** For details refer to section Run Prediction.

8.4.5 Results

- **Gene Ontology (GO) analysis**
  For more details see GO Analysis.
- **Gene Set Enrichment Analysis (GSEA)**
  For more details see Performing GSEA in **GeneSpring**.
- **Gene Set Analysis (GSA)**
  For more details see Gene Set Analysis.
- **Find Similar Entity Lists**
  This feature is discussed in a separate section called Find Similar Entity Lists.
- **Single Experiment Analysis or Multi-Omic Analysis**
  This feature is discussed in detail in the section Single Experiment and Multi-Omic Analysis.

- **Export to IPA**

  This feature is discussed in detail in the section Ingenuity Pathways Analysis (IPA) Connector.

- **Import IPA Entity List**

  This feature is discussed in detail in the section Import IPA Entity List.

- **NLP Network Discovery**

  This feature is discussed in detail in the section NLP Network Discovery.

- **MeSH Network Builder**

  This feature is discussed in detail in the section MeSH Network Builder.

- **Extract Relations via NLP**

  This feature is discussed in detail in the section Extract Relations via NLP.

8.4.6 Utilities

- **Import Entity list from File**

  For details refer to section Import Entity list from File.

- **Analysis: Biological Significance**

  For details refer to section Analysis: Biological Significance.

- **Filter On Entity List**

  For further details refer to section Filter On Entity List.

- **Remove Entities with missing signal values**

  For details refer to section Remove Entities with missing signal values.

Chapter 9

Analyzing Agilent Two Color Expression Data

GeneSpring supports Agilent Two Color technology, with data files in .txt or .gpr formats. The data files in .txt format are obtained from Agilent Feature Extraction(FE) 8.5 and 9.5.3. When the data file is imported into **GeneSpring** the following columns get imported: ControlType, ProbeName, Signal (2 columns) and feature columns (2 sets). With files in .gpr formats, DesignID information is required; if present in the file, it is automatically recognized for import, or the user is prompted to input the DesignID. Note that if the design ID is not correct, there may be errors while processing the data.

Agilent Two Color Workflow supports most of the Standard Agilent technologies. The Agilent custom arrays other than .gpr formats, and the files from FE other than 8.5 and 9.5.3 can be analyzed by creating a Generic Two Color technology and using the corresponding workflow. In order to do so, certain column markings should be indicated (which are automatically done with standard technologies). These details can be found in the section on Custom Agilent Arrays, while the Generic Two Color technology creation is available in Chapter 14 in the section Creating Technology. Agilent Two Color files can be also split into single channels and analyzed as single color files. For the above situation, see the section on Analyzing Agilent Two Color data in Agilent Single Color Experiment Type

9.1 Running the Agilent Two Color Workflow

Follow the instructions in section Running the Expression Experiment Workflow for creating an Agilent Two Color Experiment.

Figures 9.1, 9.7, 9.3, 9.4 show the process of choosing analysis type, experiment type, loading data, choosing samples and re-ordering the data files.

There is an additional step for Agilent Two Color Experiments. This step gives the option of

performing Dye-Swap on selected samples. Data/Sample files chosen in previous step are shown here and the user can select those arrays that were dye-swapped while performing the experiment. Accordingly, **GeneSpring** will swap the data between cy5 and cy3 for these arrays. (See Figure 9.5)

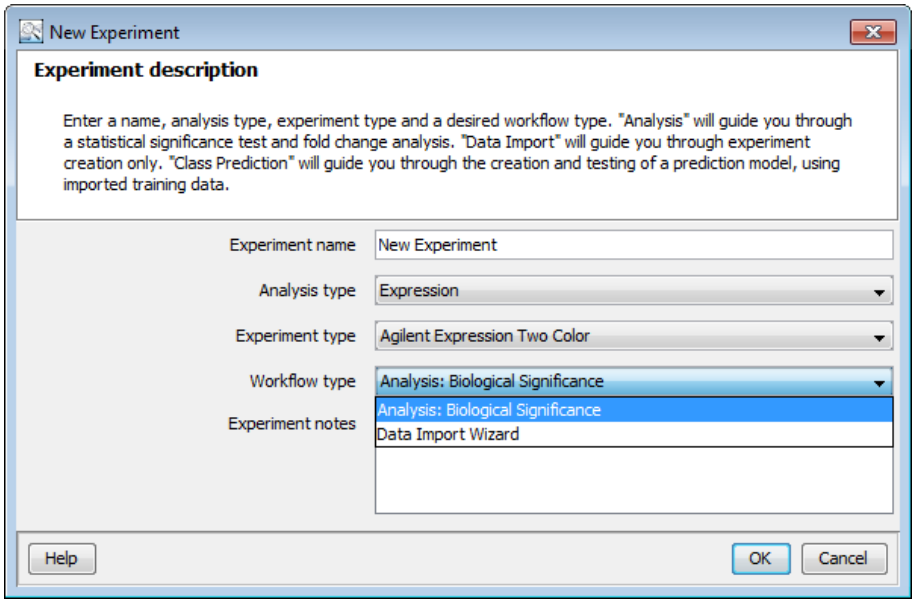


Figure 9.1: Experiment Description

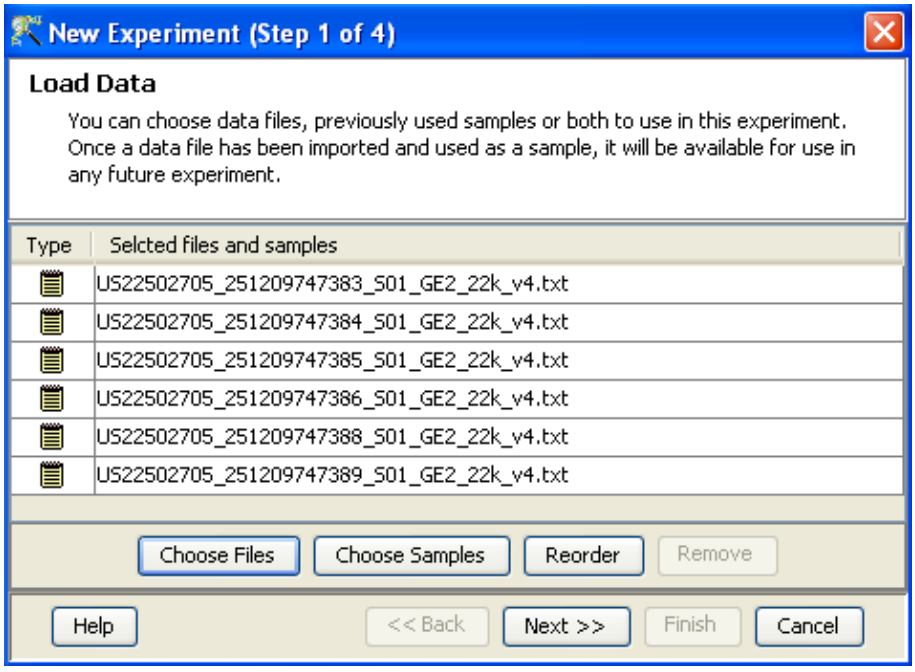


Figure 9.2: Load Data

Figure 9.3: Choose Samples

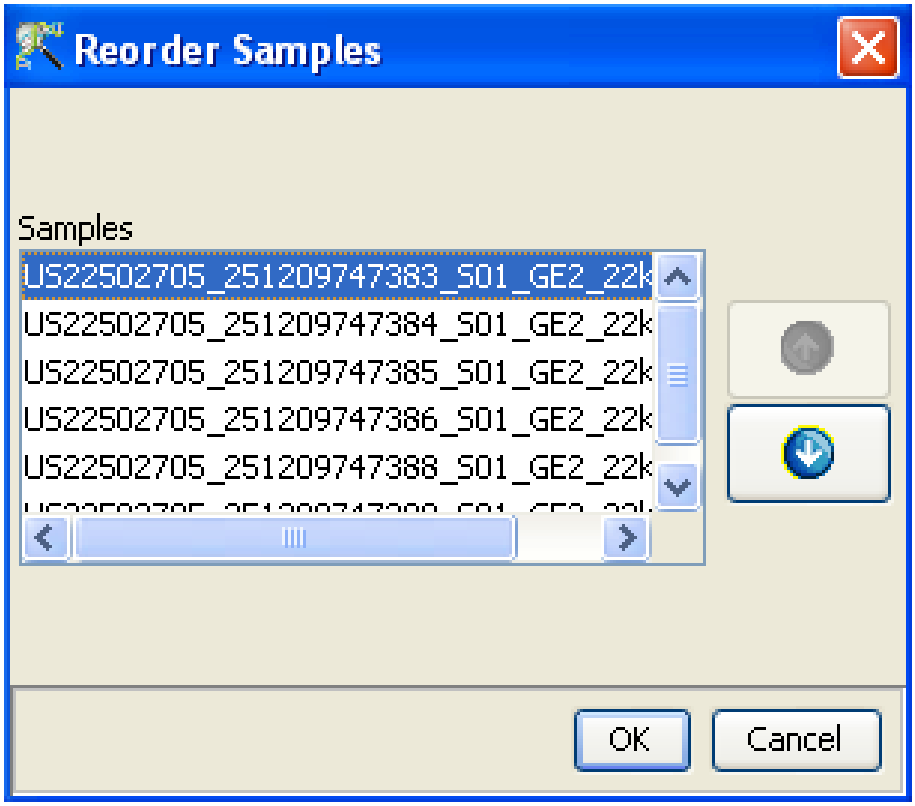


Figure 9.4: Reordering Samples

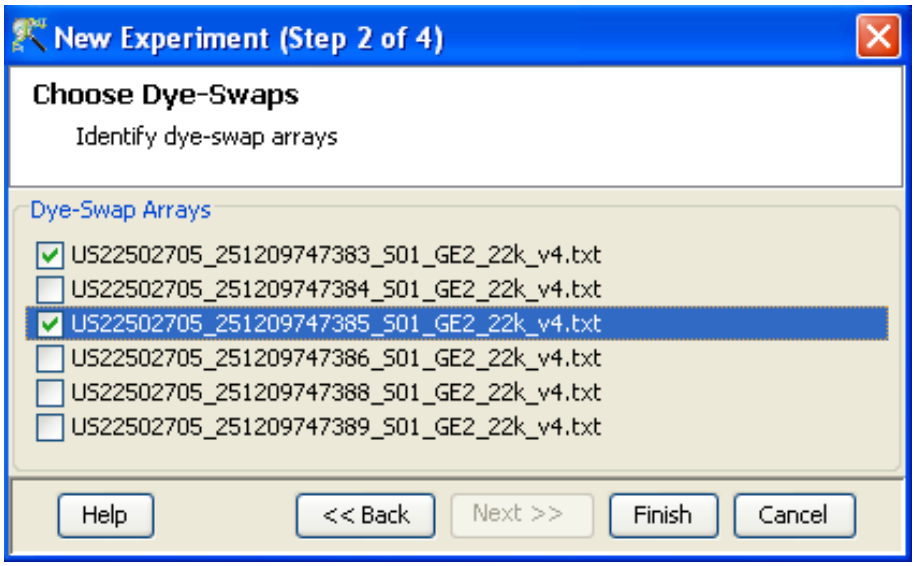


Figure 9.5: Dye Swap

Analysis: Biological Significance wizard appears with the sequence of steps on the left hand side with the current step being highlighted. The workflow allows the user to proceed in schematic fashion and does not allow the user to skip steps.

9.2 Data Processing for Agilent Two Color arrays

- **File formats**: The data files should be in text (.txt) format (obtained from Agilent Feature Extraction (FE) 8.5 and 9.5.3) or in .gpr format. Please note that the cy3 and cy5 columns taken from the .txt files are already lowess normalized whereas the signal columns considered from the gpr files are not lowess normalized.
- **Raw Signal Values**: The term ‘‘raw’’ signal values refer to the linear data after thresholding and summarization for the individual channels (cy3 and cy5). Summarization is performed by computing the geometric mean.
- **Normalized Signal Values**: The term ‘‘Normalized” signal value refers to the data after ratio computation, log2 transformation and Baseline Transformation.
- **Flag values**: The resultant flag value of any probe is decided by the following logic. For each probe with multiple flags, the order of importance is Compromised>Not Detected>Detected. If there is even one ’Compromised’, then the resultant flag is ‘Compromised’. If there is no Compromised, but ‘Not Detected’ and ‘Detected’, then ‘Not Detected’ is assigned. If there are only ‘Detected’ then only the resultant flag is assigned as ‘Detected’. If there are equal number of ‘Not Detected’ and ’Detected’, then ‘Not Detected’ gets preference. At the end of this exercise, each probe is assigned one flag.
- **Treatment of on-chip replicates**: For each probe replicate with multiple flags, the above flag logic is applied. To get the overall flag across all replicates of the probe, **GeneSpring** excludes ‘Compromised’ flag and assigns the majority considering the remaining ones. If there are only ‘Compromised’ flags, only then the overall flag becomes ‘Compromised’.

For more information on Flag logic, please refer to examples in section on: Agilent Flag Logic.

- **Treatment of Control probes**: The control probes are included while performing normalization. However there should be an exact match between the control probes in the technology and the sample for the probes to be utilized, as the comparison between the identifier columns is case-sensitive.
- **Empty Cells**: Not Applicable.
- **Sequence of events**: The sequence of events involved in the processing of the data files is: Thresholding⟶ Summarization⟶ dye swap⟶ ratio computation⟶ log transformation⟶ Baseline Transformation.

9.3 Analysis: Biological Significance

Analysis: Biological Significance workflow has 8 steps:

1. Summary Report
2. Experiment Grouping
3. QC on samples
4. Filter Probes
5. Significance analysis
6. Fold Change
7. GO analysis
8. Pathway Analysis

These steps are explained in section Analysis: Biological Significance. Details specific to Agilent Two Color Data are explained here.

Quality Control (Step 3 of 8):

The 3rd step in the Analysis: Biological Significance workflow is the QC on samples which is displayed in the form of four tiled windows.

Note that for experiments created using .gpr file formats, the *Quality Control* step is skipped.

The four tiled windows are as follows:

- Quality controls Metrics- Report and Experiment grouping tabs

- Quality controls Metrics- Plot
- PCA scores
- Legend

QC on Samples generates four tiled windows as illustrated in Figure 9.6.

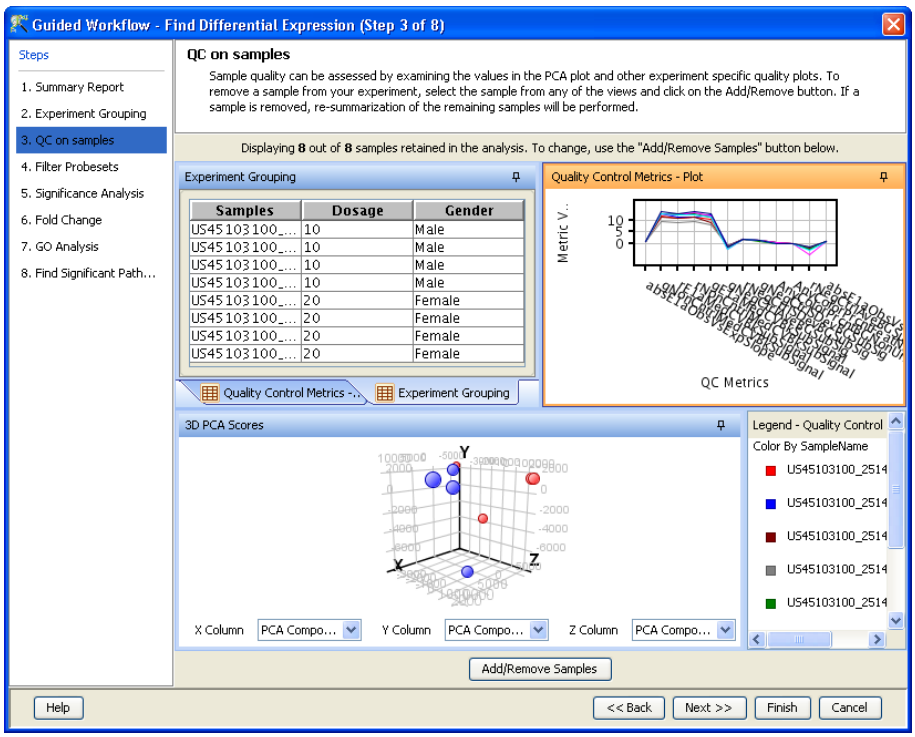


Figure 9.6: Quality Control on Samples

The metrics report include statistical results to help you evaluate the reproducibility and reliability of your microarray data. Please refer to Agilent Feature Extraction Software Reference Guide (v10.7.3 or higher) for detailed descriptions of the QC Metrics.

Quality controls Metrics Plot shows the QC metrics present in the QC report in the form of a plot. *Principal Component Analysis (PCA)* calculates the PCA scores and visually represents them in a 3D scatter plot. The scores are used to check data quality. It shows one point per array and is colored by the *Experiment Factors* provided earlier in the *Experiment Groupings* view. This allows viewing of separations between groups of replicates. Ideally, replicates within a group should cluster together and separately from arrays in other groups. The PCA components, represented in the X, Y and Z axes are numbered 1, 2, 3... according to their decreasing significance. The 3D PCA scores plot can be customized via ***Right-Click*** $\longrightarrow$ ***Properties***. To zoom into a 3D Scatter plot, press the Shift key and simultaneously hold down the left mouse button and move the mouse upwards. To zoom out, move the mouse downwards instead. To rotate, press the Ctrl key, simultaneously hold down the left mouse button and move the mouse around the plot.

The *Add/Remove* samples allows the user to remove the unsatisfactory samples and to add the samples back if required. Whenever samples are removed or added back, summarization as well as baseline transformation is performed on the samples. Click on *OK* to proceed.

The fourth window shows the legend of the active QC tab.

9.4 Data Import Wizard

The *Data Import Wizard* offers a variety of choices to the user for the analysis. Flag options can be changed and raw signal thresholding can be altered. Additionally there are options for baseline transformation of the data and for creating different interpretations. To create and analyze an experiment using the *Data Import Wizard*, load the data as described earlier. In the *New Experiment Dialog*, choose the Workflow Type as *Data Import Wizard*. Click *OK* will open a new experiment wizard which then proceeds as follows:

1. **Step 1 of 5: Load Data** As in case of *Analysis: Biological Significance* workflow, either data files can be imported or else pre-created samples can be used.

   - For loading new txt files or gpr files, use Choose Files.
   - If these data files have been previously used in **GeneSpring** experiments *Choose Samples* can be used.

   The *Load Data* window is shown in Figure 9.7.

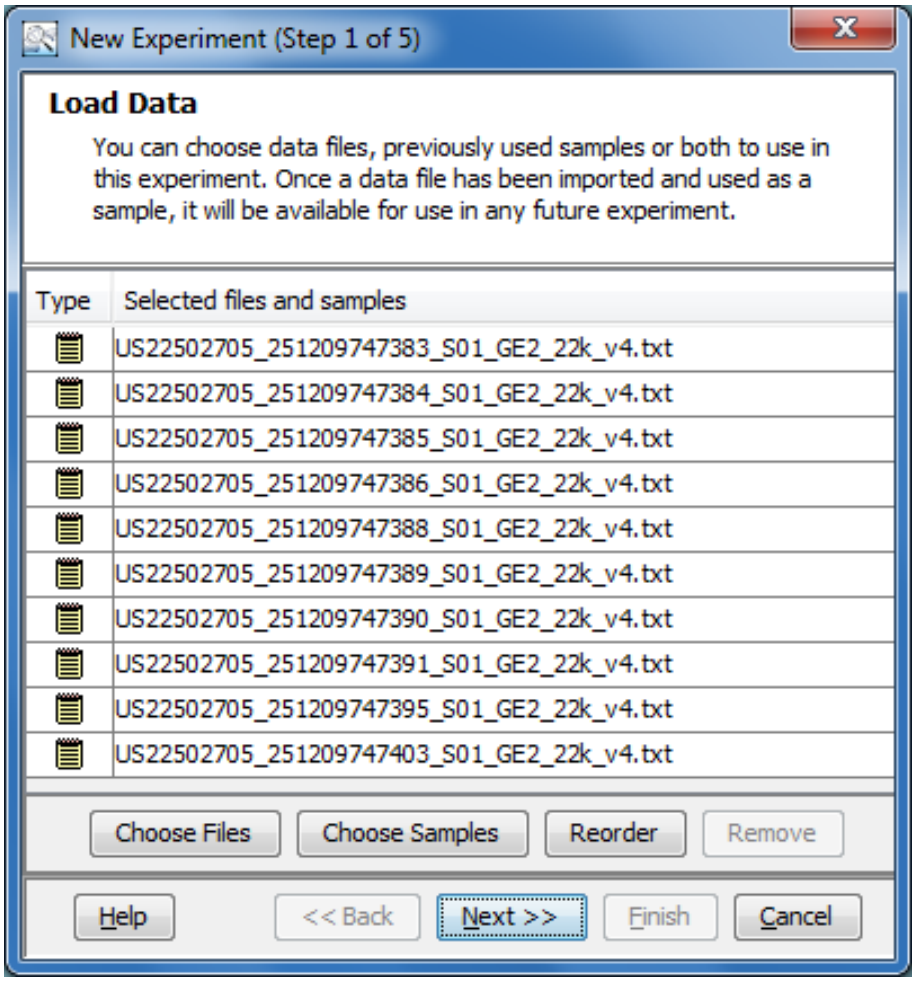


Figure 9.7: Load Data

2. **Step 2 of 5 : Samples Validation**

   This step is shown only if there is mismatch in technology between the gpr files input in step 1. **GeneSpring** requires that the files input for any particular experiment be of the same technology. The work around is go back to step 1 and remove those sample files that are of different technology.

   The *Samples Validation* window is shown in Figure 9.8.

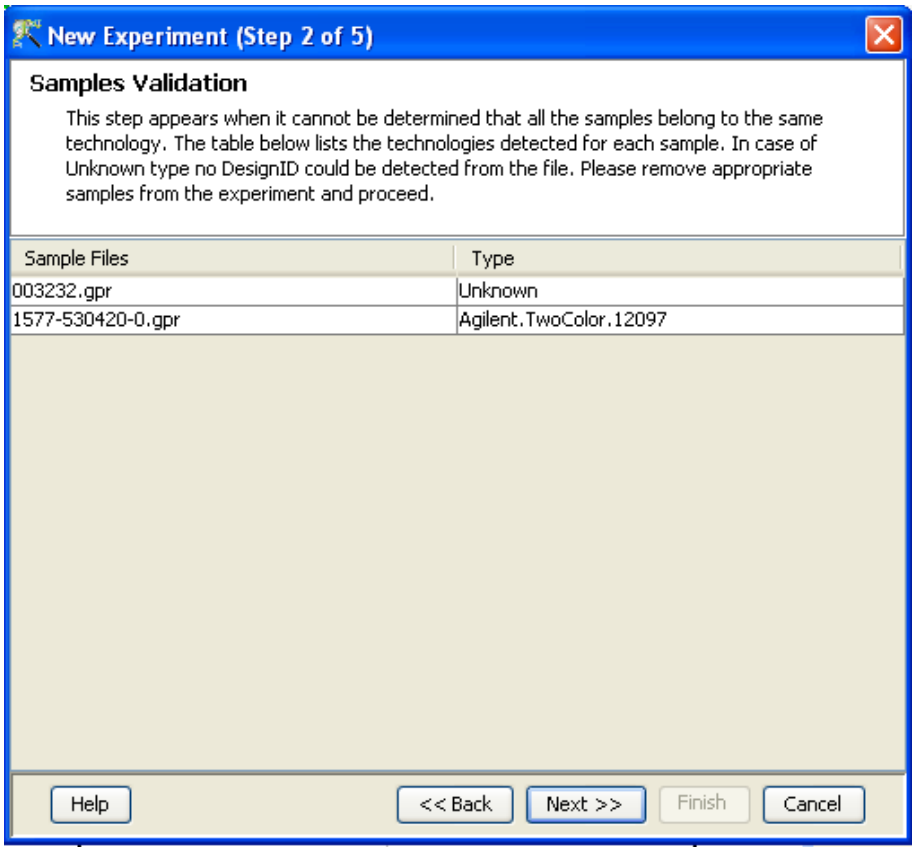


Figure 9.8: Samples Validation

3. **Step 3 of 5: Choose Dye-swaps** Dye-swapping accounts for dye related bias (as both dyes bind DNA with different affinities) that cannot be removed by standard normalization methods.

   In **GeneSpring**, during the creation of two color experiments, users can specify sample(s) that have been dye swapped. The experiment creation wizard has an additional step (step 3 of 5) that lists all the sample(s) chosen by the user for experiment creation. At this step users can select sample(s) that have been dye swapped. For sample(s) that have been marked as dye-swapped, **GeneSpring** treats Cy3 as‘‘signal’’ and Cy5 as ‘‘control” and computes the signal as Cy3/Cy5. For non-dye swapped sample(s), **GeneSpring** treats Cy5 as ‘‘signal’’ and Cy3 as ‘‘control” and computes the signal as Cy5/Cy3. The *Choose Dye Swaps* window is depicted in the Figure 9.9.

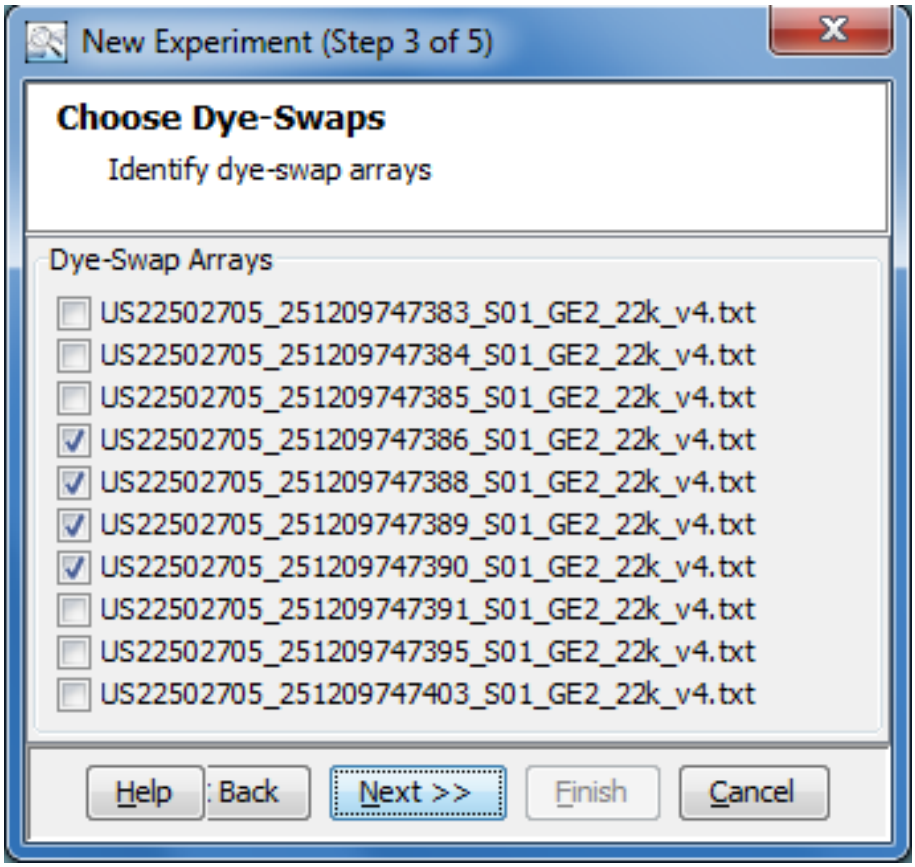


Figure 9.9: Choose Dye-Swaps

4. **Step 4 of 5: Advanced Flag Import**

   This gives the options for importing flag information. The information is derived from the Feature columns in data file. User has the option of changing the default flag settings that appear in this step. The 'Save as Default' handle allows saving the current flag settings under the tool configuration. When a file is imported, **GeneSpring** will show these saved default setting in this step, by default. The settings can be changed either in this wizard or from *Tools* → *Options* → *Miscellaneous* → *Agilent Flag Settings*.

   This step is skipped for files in .gpr formats.

   Figure 9.10 shows the Step to import flags in Experiment Creation.

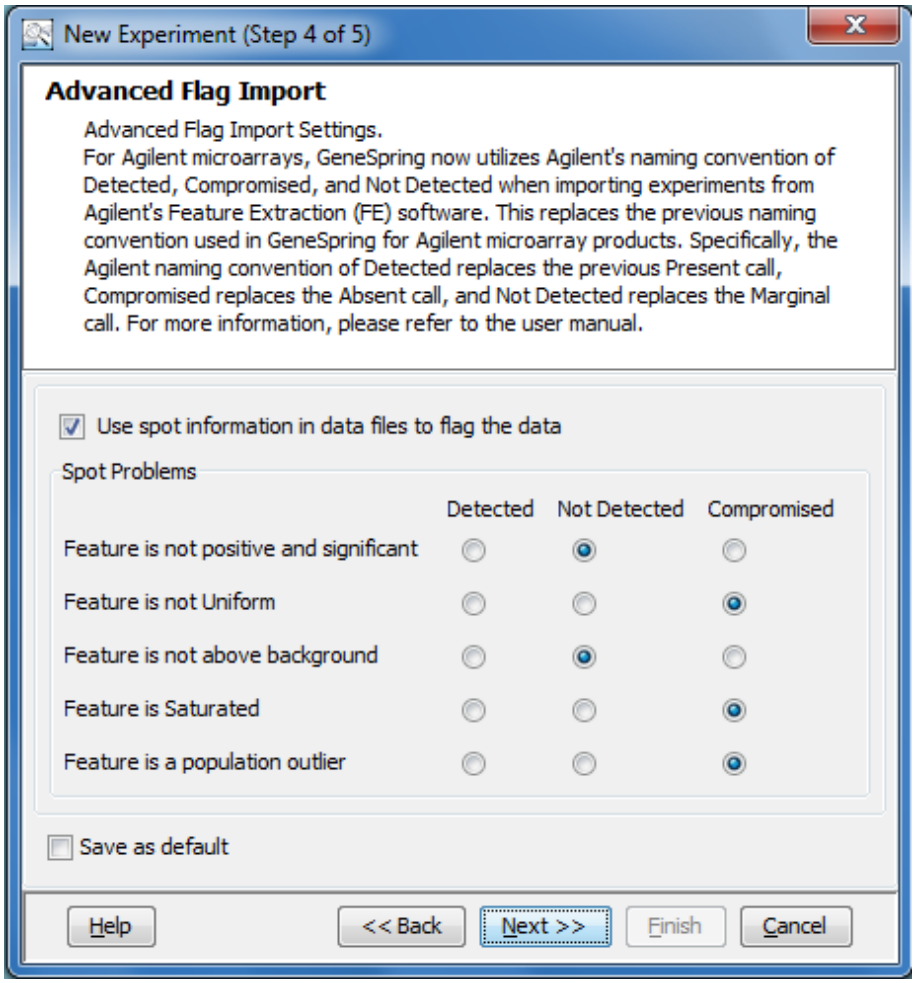


Figure 9.10: Advanced flag Import

5. **Step 5 of 5 : Preprocess Baseline Options**

   The final step of Experiment Creation is shown in Figure 9.11.

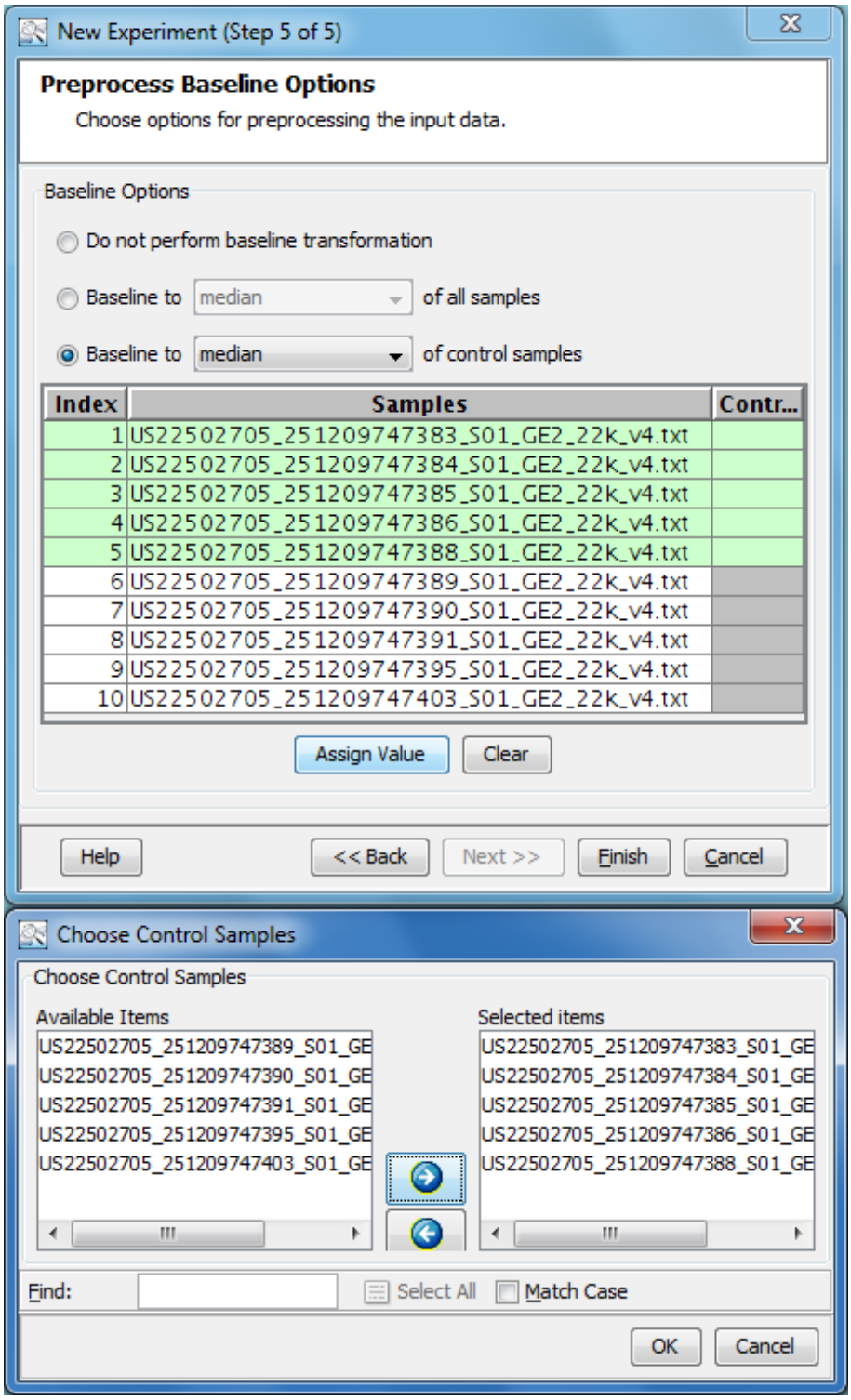


Figure 9.11: Preprocess Options

Criteria for preprocessing of input data is set here. It allows the user to choose the appropriate baseline transformation option.

The baseline options include:

- ***Do not perform baseline***
- ***Baseline to median of all samples:*** For each probe the median of the log summarized values from all the samples is calculated and subtracted from each of the samples.
- ***Baseline to median of control samples:*** For each sample, an individual control or a set of controls can be assigned. Alternatively, a set of samples designated as controls can be used for all samples. For specifying the control for a sample, select the sample and click on ***Assign value***. This opens up the ***Choose Control Samples*** window. The samples designated as Controls should be moved from the *Available Items* box to the *Selected Items* box. Click on ***Ok***. This will show the control samples for each of the samples.

  In *Baseline to median of control samples*, for each probe the median of the log summarized values from the control samples is first computed and then this is subtracted from the sample. If a single sample is chosen as the control sample, then the probe values of the control sample are subtracted from its corresponding sample.

9.4.1 Experiment Setup

Once an experiment is created, the *Data Import Wizard* steps appear on the right hand side. This gives options to view quick start guide, create experiment grouping and interpretation and the utility to create gene level experiment from probe level experiment. See section Experiment Setup for complete details.

9.4.2 Quality Control

- **Quality Control on Samples:**

  Quality Control (QC) allows you to decide which samples are ambiguous and which are passing the quality criteria. Based on the QC results, you can remove unreliable samples from any downstream analysis. The QC dialog displays three tiled windows.

  The view shows three tiled windows:

  - 3D PCA scores
  - Quality Control Metrics - Report, Quality Control Metrics - Plot, and Experiment Grouping tabs.
  - Legend

  Figure 9.12 has the three tiled windows which reflect the QC performed on samples.

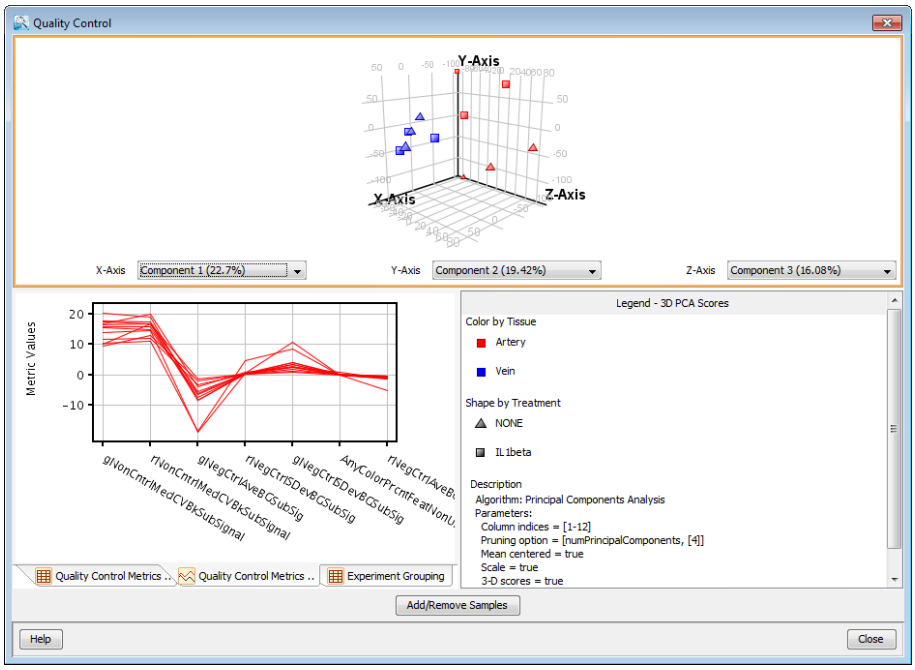


Figure 9.12: Quality Control

NOTE This view does not work if the experiment was created from Agilent two color files in $*.gpr$ format.

Principal Component Analysis (PCA) calculates the PCA scores and visually represents them in a 3D scatter plot. The scores are used to check data quality. It shows one point per array and is colored by the *Experiment Factors* provided earlier in the *Experiment Groupings* view. This allows viewing of separations between groups of replicates. Ideally, replicates within a group should cluster together and separately from arrays in other groups. The PCA components, represented in the X, Y and Z axes are numbered 1, 2, 3... according to their decreasing significance. The 3D PCA scores plot can be customized via ***Right-Click*⟶*Properties***. To zoom into a 3D Scatter plot, press the Shift key and simultaneously hold down the left mouse button and move the mouse upwards. To zoom out, move the mouse downwards instead. To rotate, press the Ctrl key, simultaneously hold down the left mouse button and move the mouse around the plot.

The **Quality Control Metrics - Report** include the statistical results to help you evaluate the reproducibility and reliability of your single microarray data.

textbfQuality controls Metrics Plot shows the QC metrics present in the QC Report in the form of a plot. Please refer to Agilent Feature Extraction Software Reference Guide (v10.7.3 or higher) for detailed descriptions of the QC Metrics.

Experiment Grouping shows the parameters and parameter values for each sample.

The third window shows the legend of the active QC tab.

The *Add/Remove* samples allows the user to remove the unsatisfactory samples and to add the samples back if required. Whenever samples are removed or added back, summarization

as well as baseline transformation is performed on the samples. Click on *OK* to proceed.

- **Filter Probes by Expression:** Entities are filtered based on their signal intensity values. For details refer to the section on Filter Probes by Expression
- **Filter Probes by Flags:** In this step, the entities are filtered based on their flag values, Detected, Not Detected and Compromised. Users can set what proportion of conditions must meet a certain threshold. The flag values that are defined at the creation of the new experiment (Step 3 of 4) are taken into consideration while filtering the entities. The filtration is done in 4 steps:

  1. Step 1 of 4 : *Entity list and interpretation* window opens up. Select an entity list by clicking on *Choose Entity List* button. Likewise by clicking on *Choose Interpretation* button, select the required interpretation from the navigator window.

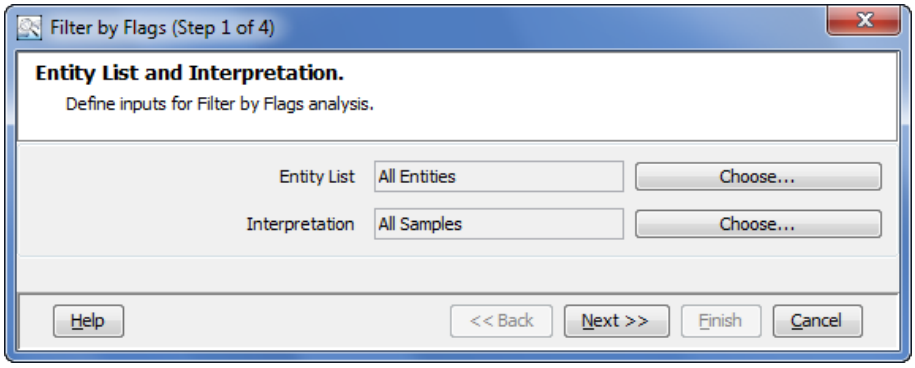


Figure 9.13: Entity list and Interpretation

  2. Select the flag values that an entity must satisfy to pass the filter. By default, the Detected and Not Detected flags are selected.
  3. Step 2 of 4: This step is used to set the filtering criteria and the stringency of the filter. Select the flag values that an entity must satisfy to pass the filter. By default, the Detected and Not Detected flags are selected. Stringency of the filter can be set in *Retain Entities* box.

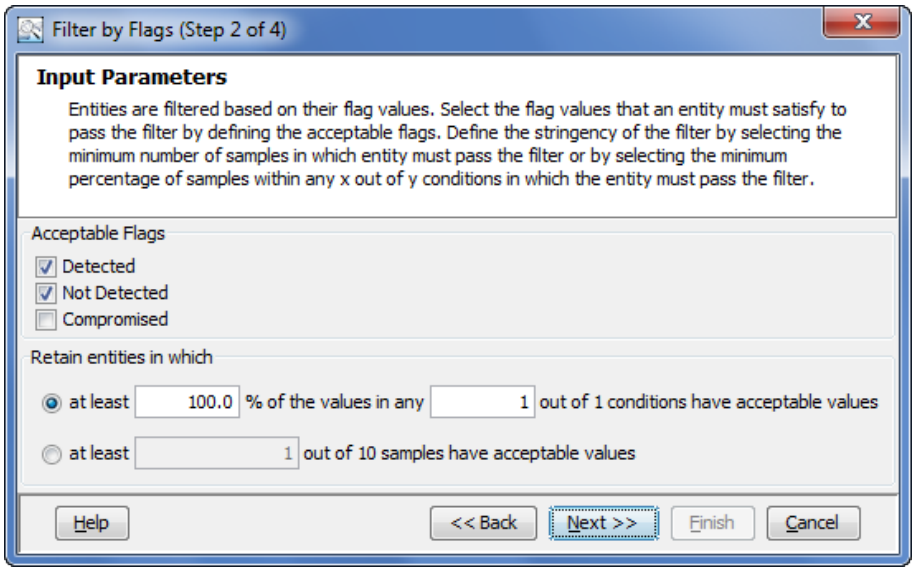


Figure 9.14: Input Parameters

4. Step 3 of 4: A spreadsheet and a profile plot appear as 2 tabs, displaying those probes which have passed the filter conditions. Baseline transformed data is shown here. Total number of probes and number of probes passing the filter are displayed on the top of the navigator window (See Figure 9.15).

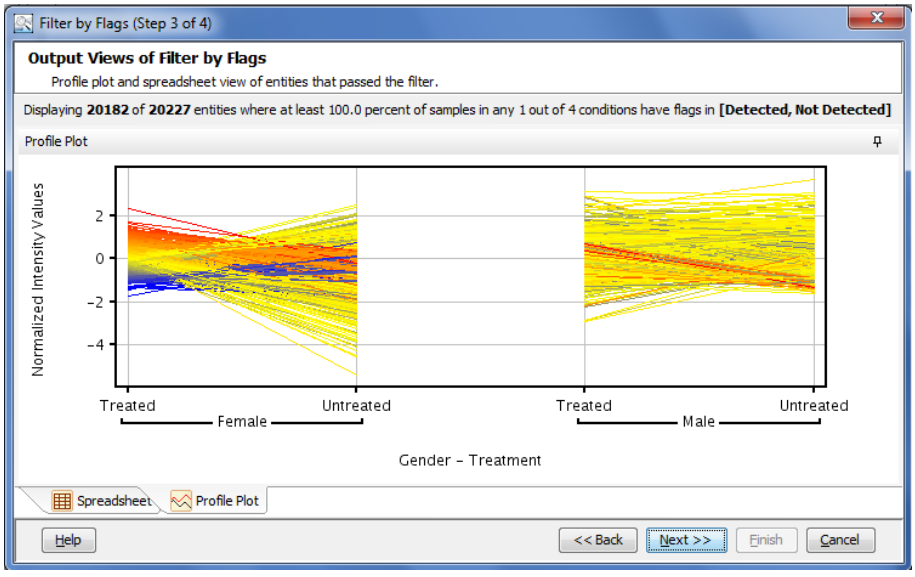


Figure 9.15: Output Views of Filter by Flags

5. Step 4 of 4: Click *Next* to annotate and save the entity list. (See Figure 9.16)

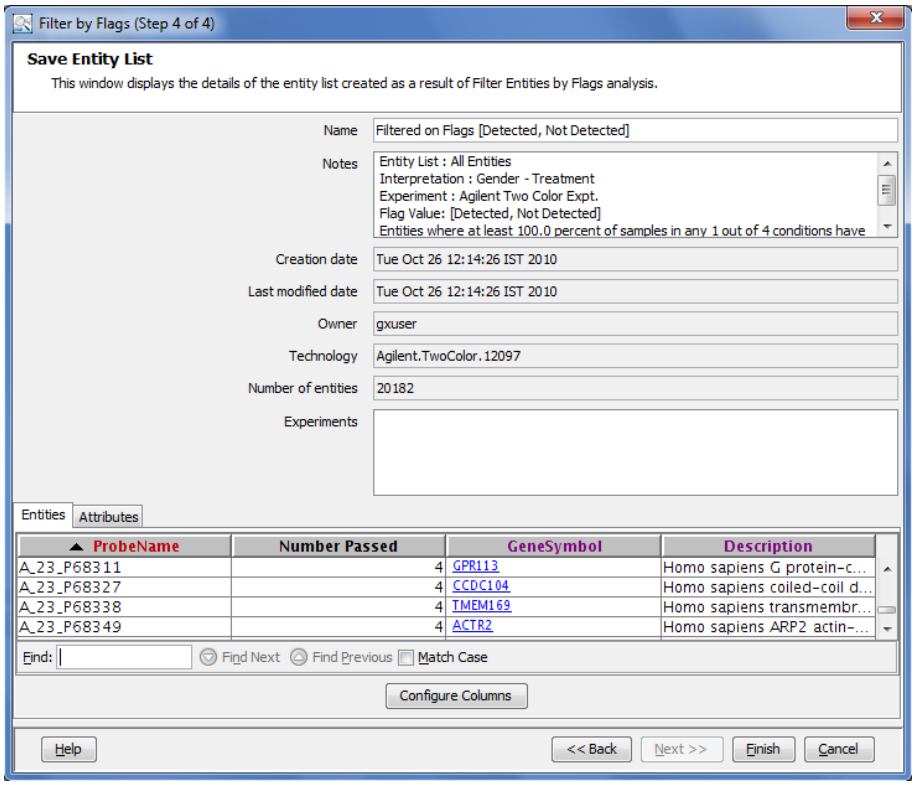


Figure 9.16: Save Entity List

- **Filter Probes on Data Files:** Entities can be filtered based on values in a specific column of the original data files. For details refer to the section on Filter Probes on Data Files
- **Filter Probes by Error:** Entities can be filtered based on the standard deviation or coefficient of variation using this option. For details refer to the section on Filter Probes by Error

9.4.3 Analysis

- **Statistical Analysis**

  For details refer to section Statistical Analysis in the Data Import Wizard section.

- **Filter on Volcano Plot**

  For details refer to section Filter on Volcano Plot.

- **Fold Change**

  For details refer to section Fold change.

- **Clustering**

  For details refer to chapter Clustering Analysis.

- **Correlation**

  For details refer to chapter Correlation Analysis.

- **Find Similar Entities**
  For details refer to section Find similar entities.
- **Filter on Parameters**
  For details refer to section Filter on Parameters.
- **Principal Component Analysis**
  For details refer to section PCA.

9.4.4 Class Prediction

- **Build Prediction Model** For details refer to section Build Prediction Model.
- **Run Prediction** For details refer to section Run Prediction.

9.4.5 Results

- **Gene Ontology (GO) analysis**
  For more details see GO Analysis.
- **Gene Set Enrichment Analysis (GSEA)**
  For more details see Performing GSEA in **GeneSpring**.
- **Gene Set Analysis (GSA)**
  For more details see Gene Set Analysis.
- **Find Similar Entity Lists**
  This feature is discussed in a separate section called Find Similar Entity Lists.
- **Single Experiment Analysis or Multi-Omic Analysis**
  This feature is discussed in detail in the section Single Experiment and Multi-Omic Analysis.
- **Export to IPA**
  This feature is discussed in detail in the section Ingenuity Pathways Analysis (IPA) Connector.
- **Import IPA Entity List**
  This feature is discussed in detail in the section Import IPA Entity List.
- **NLP Network Discovery**
  This feature is discussed in detail in the section NLP Network Discovery.
- **MeSH Network Builder**
  This feature is discussed in detail in the section MeSH Network Builder.
- **Extract Relations via NLP**
  This feature is discussed in detail in the section Extract Relations via NLP.

9.4.6 Utilities

- **Import Entity list from File**

  For details refer to section Import Entity list from File.

- **Analysis: Biological Significance**

  For details refer to section Analysis: Biological Significance.

- **Filter On Entity List**

  For further details refer to section Filter On Entity List.

- **Remove Entities with missing signal values**

  For details refer to section Remove Entities with missing signal values.

9.5 Custom Agilent Arrays

The standard Agilent arrays can be analyzed using the Agilent Single and the Two Color Experiment types. In addition, **GeneSpring** also allows the user to analyze Custom Agilent arrays using Generic Single/Two Color Experiment types. To perform a Generic Single/Two Color analysis using Agilent arrays, the files can be an output from any FE, so long as they are in a tabular format. An annotation file is also required. Analysis through the Generic Two Color and Single Color workflows involves creation of a custom technology (Refer to chapters 13 and 14 on Creating Technology in Generic Single Color and Creating Technology in Generic Two Color) and specific markings of columns to perform GO, GSEA, or to view in Genome Browser. miRNA files can also be analyzed similarly.

Chapter 10

Analyzing Affymetrix Expression Data

GeneSpring supports the Affymetrix GeneChip technology and most of the Affymetrix GeneChips can be analyzed using **GeneSpring** . To obtain a list of the supported chips, go to *Annotations* $\longrightarrow$ *Create Technology* $\longrightarrow$ *From Agilent Server*. This will display a list of supported chip types. Affymetrix technology can also be created if a custom CDF is being used. For more details refer to the section on Affymetrix Technology creation using Custom CDF.

10.1 Running the Affymetrix Workflow

Follow the instructions in section Running the Expression Experiment Workflow for creating an Affymetrix Expression Experiment. Figures 10.1, 10.2, 10.3, 10.4 show the process of choosing analysis type, experiment type, loading data, choosing samples and re-ordering the data files.

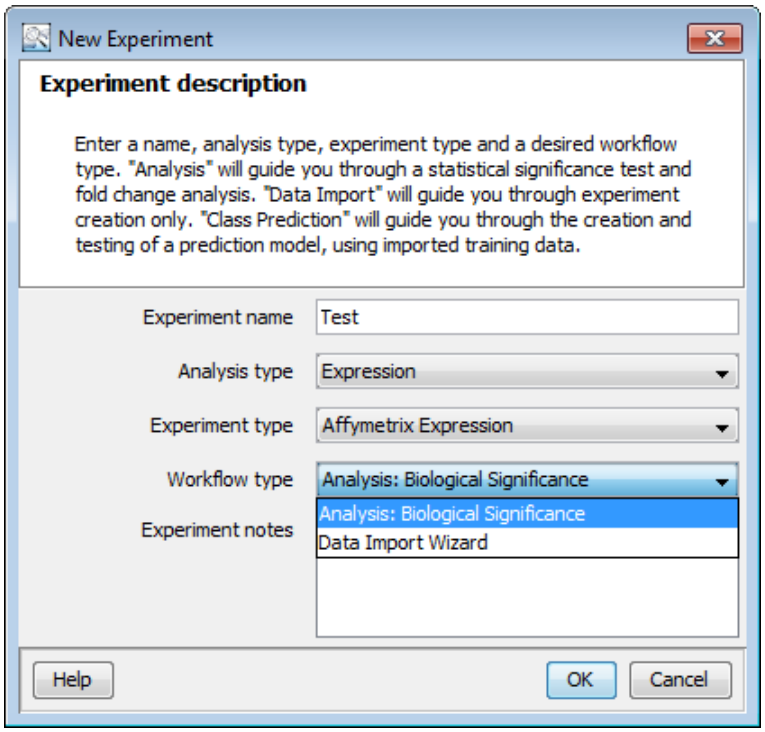


Figure 10.1: Experiment Description

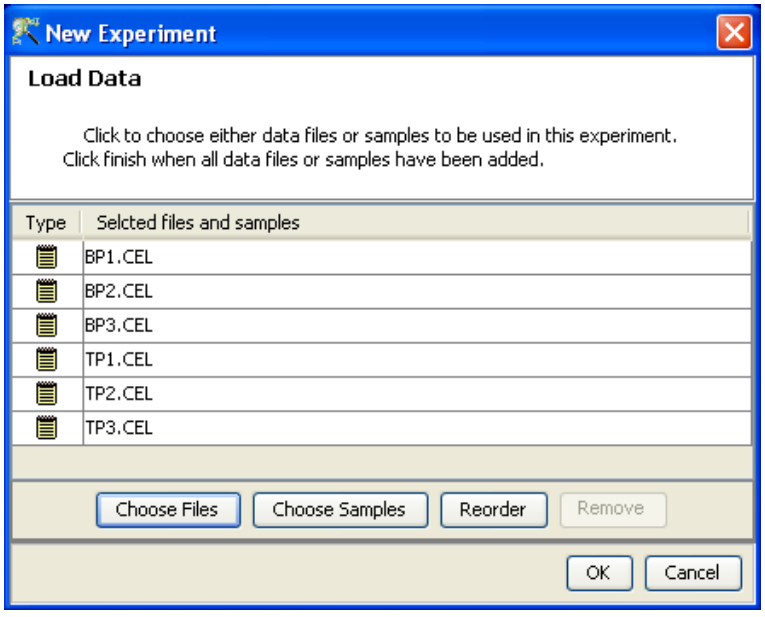


Figure 10.2: Load Data

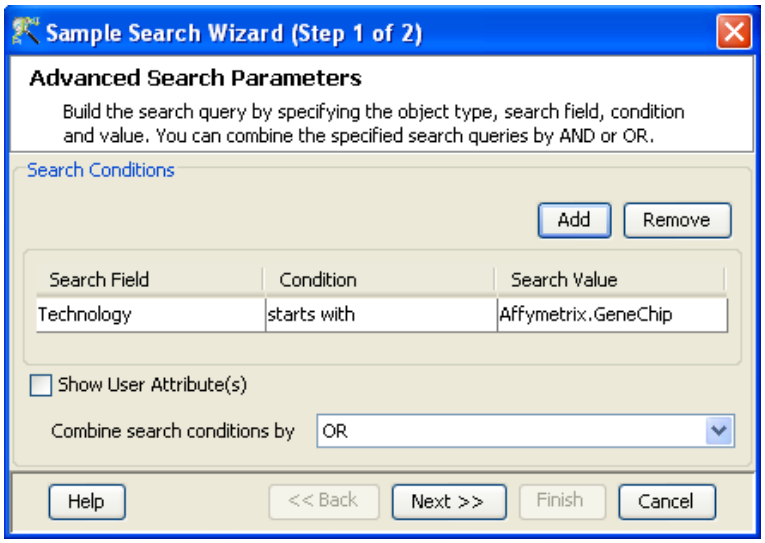


Figure 10.3: Choose Samples

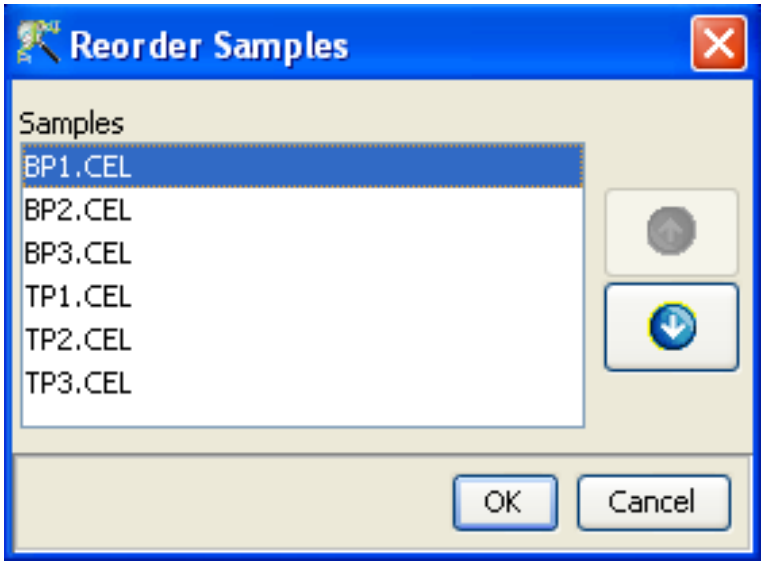


Figure 10.4: Reordering Samples

10.2 Data Processing

1. **File formats**: The data file should be present either as a CEL file or a CHP file or a TEXT file. However while creating an experiment; only one type of file (CEL/CHP.TXT) can be used.

2. **Raw signal values (CEL files)**: In an Affymetrix Expression experiment, the term ''raw'' signal values refer to the linear data after summarization using a summarization algorithm (RMA, PLIER, GCRMA, LiWong and MAS5).

3. **Raw signal values (CHP files)**: In an Affymetrix Expression experiment, the term "raw" files refers to the linear data obtained from the CHP files. In an Affymetrix Expression experiment, **GeneSpring** does not handle input data from CHP files if they are present in the log scale as the tool assumes that the data is in the linear scale and performs another log transformation.

4. **Normalized signal values (CEL files)**: ''Normalized'' values are generated after log transformation and baseline transformation.

5. **Normalized signal values (CHP files)**: The term "Normalized" refers to values generated after log transformation, normalization (Percentile Shift, Scale and Normalize to control genes) and baseline transformation.

6. **Treatment of on-chip replicates**: Not Applicable.

7. **Flag values**: The flag values are calculated only when MAS5 algorithm is chosen for summarization and is inclusive of the algorithm.

8. **Treatment of Control probes**: Not Applicable.

9. **Empty Cells**: Not Applicable.

10. **Sequence of events (CEL files)**: The sequence of events involved in the processing of a CEL file is Summarization$\longrightarrow$ log transformation$\longrightarrow$ baseline transformation.

11. **Sequence of events (CHP files)**: The sequences of events involved in the processing of a CHP file are log transformation$\longrightarrow$ normalization$\longrightarrow$ baseline transformation. If the data in the CHP file is already log transformed, then **GeneSpring** detects it and proceeds with the normalization step.

12. **Sequence of events (TXT files)**: The sequences of events involved in the processing of a TXT file are log transformation$\longrightarrow$ normalization$\longrightarrow$ baseline transformation. The **GeneSpring** prompts the user to specify if the if the data in the Text file is already log transformed or not; User can then specity options for thresholding and normalization.

10.3 Analysis: Biological Significance

Analysis: Biological Significance has 8 steps:

1. Summary Report
2. Experiment Grouping
3. QC on samples
4. Filter Probe Sets
5. Statistical Analysis
6. Fold Change
7. GO analysis
8. Pathway Analysis

These steps are explained in section Analysis: Biological Significance. Details specific to Affymetrix Expression Data are explained here.

Quality Control on Samples (Step 3 of 8):

The 3rd step in *Analysis: Biological Significance* is the QC on samples which is displayed in the form of four tiled windows.

This window is disabled for TXT files.

- Internal controls and experiment grouping tabs
- Hybridization controls
- PCA scores.
- Legend

QC generates four tiled windows as seen in Figure 10.5.

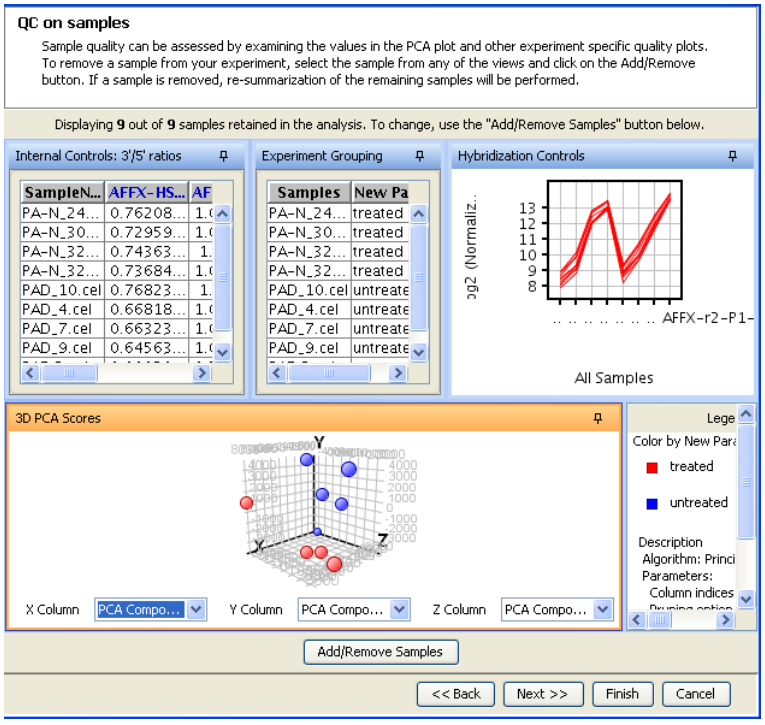


Figure 10.5: Quality Control on Samples

The views in these windows are lassoed i.e., selecting the sample in any of the view highlights the sample in all the views.

Internal Controls view shows RNA sample quality by showing 3'/5' ratios for a set of specific probesets which include the actin and GAPDH probesets. The 3'/5' ratio is output for each such probeset and for each array in the experiment. The ratios for actin and GAPDH should be no more than 3. A ratio of more than 3 indicates sample degradation and is shown in the table in red color. The *Experiment Grouping* tab, present in the same view shows the samples and the parameters assigned.

Hybridization Controls view depicts the hybridization quality. Hybridization controls are composed of a mixture of biotin-labelled cRNA transcripts of bioB, bioC, bioD, and cre prepared in staggered concentrations (1.5, 5, 25, and 100pm respectively). This mixture is spiked-in into the hybridization cocktail. bioB is at the level of assay sensitivity and should be called Present at least 50% of the time. bioC, bioD and cre must be present all of the time and must appear in increasing concentrations. The X-axis in this graph represents the controls and the Y-axis,the log of the Normalized Signal Values.

Principal Component Analysis (PCA) calculates the PCA scores and visually represents them in a 3D scatter plot. The scores are used to check data quality. It shows one point per array and is colored by the *Experiment Factors* provided earlier in the *Experiment Groupings* view. This allows viewing of separations between groups of replicates. Ideally, replicates within a group should cluster together and separately from arrays in other groups. The PCA components, represented in the X, Y and Z axes are numbered 1, 2, 3... according to their decreasing significance. The 3D PCA scores plot can be customized via ***Right-Click*** $\longrightarrow$ ***Properties***. To zoom into a 3D Scatter plot, press the Shift key and simultaneously hold down the left mouse button and move the mouse upwards. To zoom out, move the mouse downwards instead. To rotate, press the Ctrl key, simultaneously hold down the left mouse button and move the mouse around the plot.

The *Add/Remove Samples* button allows the user to remove the unsatisfactory samples and to add the samples back if required. Whenever samples are removed or added back, summarization as well as baseline transformation is performed again on the newer sample set. Click on *OK* to proceed.

The fourth window shows the legend of the active QC tab.

10.4 Data Import Wizard

The **Data Import Wizard** offers a variety of choices to the user for the analysis.

- Several different summarization algorithms are available for probeset summarization.
- There are options for baseline transformation of the data and for creating different interpretations.
- Supports import of TXT files through templates.

To create and analyze an experiment using the **Data Import Wizard**, load the data as described earlier. In the **New Experiment Dialog**, choose the **Workflow Type** as Data Import Wizard. Clicking ***OK*** will open a New Experiment Wizard, which then proceeds as follows:

10.4.1 Creating an Affymetrix Expression Experiment

An analysis using the **Data Import Wizard** can be done using either CEL or CHP or TXT files. However, a combination of the file types are not allowed.

The following steps describe how to import a CEL/CHP/TXT file into **GeneSpring** .

Note that while importing text files, **GeneSpring** will automatically check with available templates and try to import based on a template. Standard files created in GCOS and Expression Console are available as templates in **GeneSpring** . There are two such templates available for each of GCOS and Expression Console - Metrics file where each sample is a file and Pivot file where multiple samples are in a file.

In order to create experiment from pivot files, GeneSpring assumes that the input files are in certain format. Column names ending with 'Signal' are treated as signal columns and column names ending with 'Detection' are treated as Flag columns. If the file contains column names like 'T1_Signal' and 'T1_Detection', then a sample 'T1' is created with 'T1_Signal' as signal column and 'T1_Detection' as flag column.

When a new TXT file is input, it is checked against these standard templates. If it matches any of these standard, it is imported based on that template. But if it does not match with any of these templates, then the user is taken through a custom template creation procedure. Templates created and saved by the user are added to the list of available templates, which then can be chosen as standard template while importing TXT files.

Step 1 of 10 : Load data As in case of **Analysis: Biological Significance**, either data files can be imported or else pre-created samples can be used.

- For loading new CEL/CHP/TXT files, use ***Choose Files***.
- If the CEL/CHP/TXT files have been previously used in experiments ***Choose Samples*** can be used.

The **Load Data** window is shown in Figure 10.6.

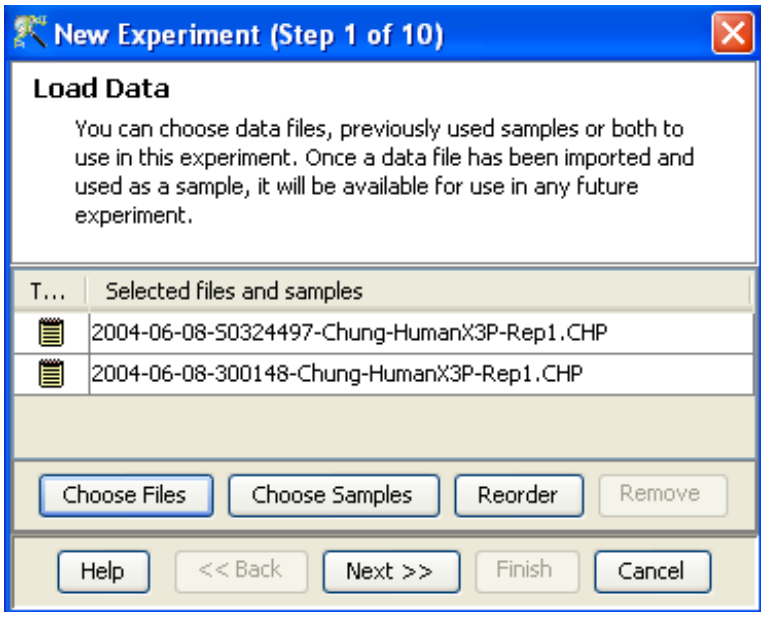


Figure 10.6: Load Data

Step 2 of 10: Choose Technology and Template This step comes up only for sample files in TXT format. The *Select Technology* drop down lists all the Affymetrix technologies available while *Choose a Template* shows available templats (those prepackaged in the tool and those saved by the user) as well as option to choose 'Custom Template'.

If 'Custom Template' is chosen, the user has to specify a sample data file to be considered as template and the number of samples in that data file. A 'Template name' also has to be input so that the custom template can be saved for future use. If a custom template is chosen, the wizard goes through steps 3 to 5, specific to creation of custom template. These steps are skipped if the TXT file is of the standard template form.

The **Choose Technology and Template** window is shown in Figure 10.7.

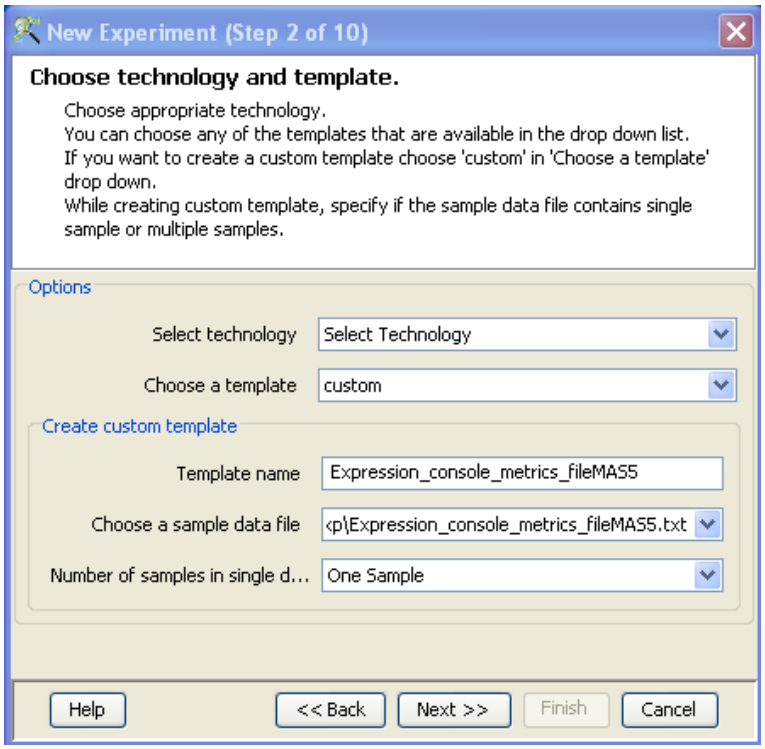


Figure 10.7: Choose Technology and Template

Skip the custom template creation steps and Go to Step 6 of 10 : Select ARR Files

Note that steps 3 to 5 of this wizard are applicable only when custom template has to be created to import non-standard TXT files.

Step 3 of 10 : Select Row Scope for Import This window shows the first few rows of the chosen template file (by default, it is 100 rows; can be changed from *Tools* → *Options menu*). The user can define the scope of the import here.

The **Select Row Scope for Import** window is shown in Figure 10.8.

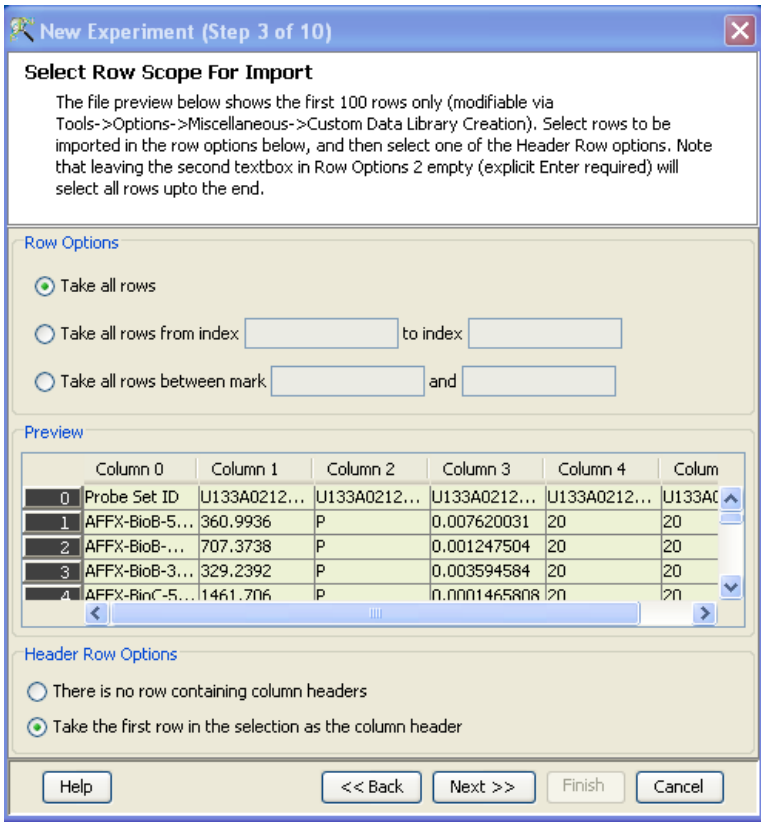


Figure 10.8: Select Row Scope for Import

Step 4 of 10 : Choose Identifier and Signal Column This window allows the user to define the Identifier column, the background signal column and the Flag column from the chosen template file. The flags can be configured.

This step is shown only if the chosen template file has only one sample in the file; for multiple samples in single file, step 5 is shown.

The **Choose Identifier and Signal Column** window is shown in Figure 10.9.

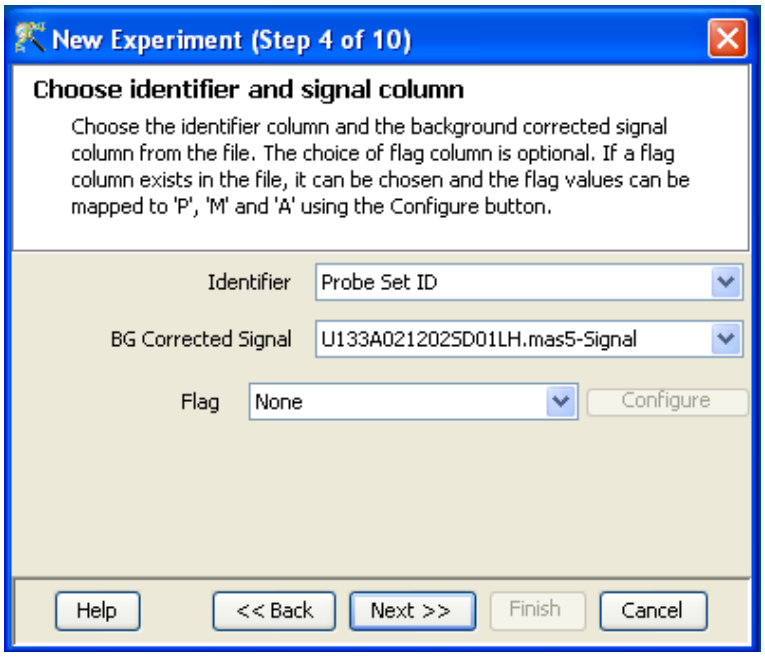


Figure 10.9: Choose Identifier and Signal Column

Step 5 of 10 : Single Colour Many Samples in one File Selection This is the equivalent of Step 4 for files with multiple samples. This window provides drop downs to choose Identifier column, the Signal and Flag columns.

There is an option to identify the signal and flag colums using keywords or the user can choose any column and mark it explicitly as signal or flag column. If 'keyword' option is chosen, the user has to *'Refresh'* so that the columns with the specified key words get picked up and shown in the table at the bottom as signal and flag columns.

User can also choose a representative flag column and configure it.

This window is shown in Figure 10.10.

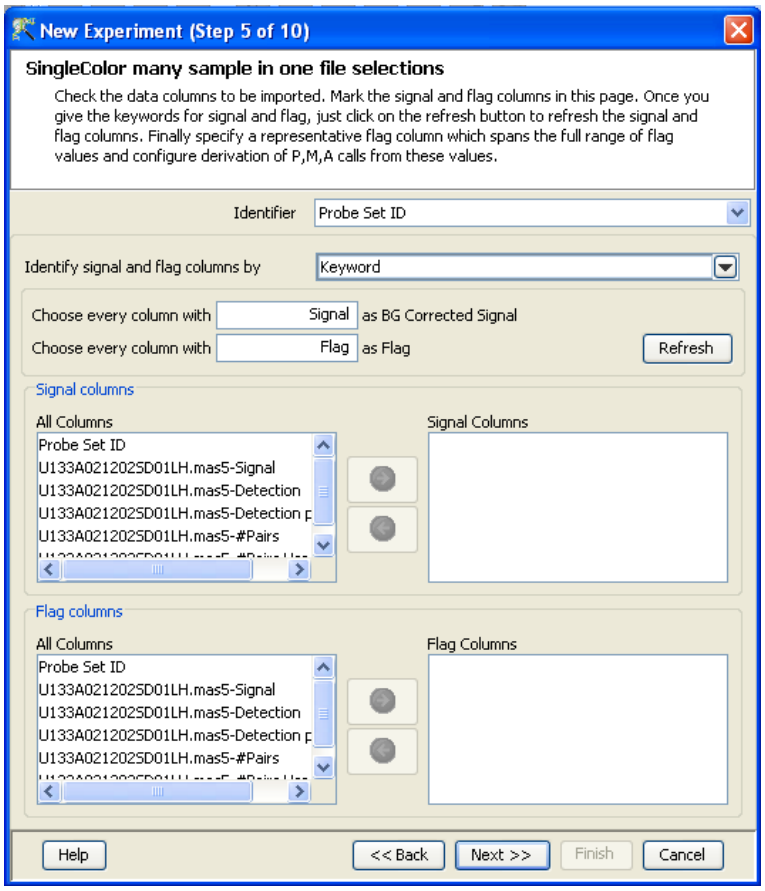


Figure 10.10: Single Color Multiple Samples in one File Selection

Step 6 of 10 : Select ARR files ARR files are Affymetrix files that hold annotation information for each sample CEL and CHP file and are associated with the sample based on the sample name. These are imported as annotations to the sample. Click on ***Next*** to proceed to the next step.

Note that this step is skipped for TXT files.

The **Select ARR files** window is depicted in the Figure 10.11.

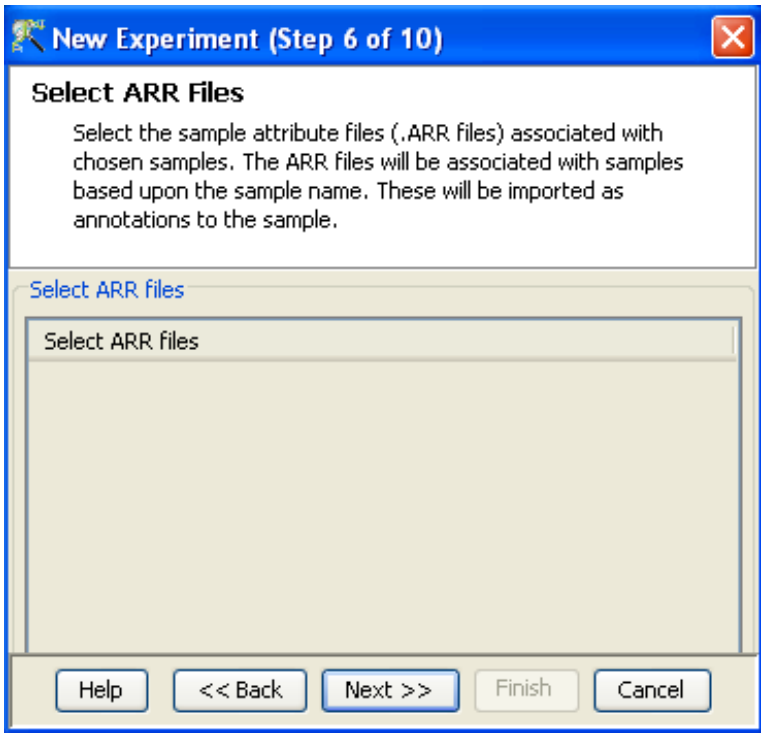


Figure 10.11: Select ARR files

Step 7 of 10 : Preprocess Baseline options This step is specific for CEL files. Any one of the Summarization algorithms provided from the drop down menu can be chosen to summarize the data. The available summarization algorithms are:

- The RMA algorithm due to Irazarry et al. [Ir1, Ir2, Bo].
- The MAS5 algorithm, provided by Affymetrix [Hu1].
- The PLIER algorithm due to Hubbell [Hu2].
- The LiWong (dChip) algorithm due to Li and Wong [LiW].
- The GCRMA algorithm due to Wu et al. [Wu].

See Chapter Probe Summarization Algorithms for details on the above algorithms.
Subsequent to probeset summarization, baseline transformation of the data can be performed. The baseline options include

- ***Do not perform baseline***
- ***Baseline to median of all samples:*** For each probe the median of the log summarized values from all the samples is calculated and subtracted from each of the samples.
- ***Baseline to median of control samples:*** For each sample, an individual control or a set of controls can be assigned. Alternatively, a set of samples designated as controls can be used for all samples. For specifying the control for a sample, select the sample and click on ***Assign value***. This opens up the ***Choose Control Samples*** window. The samples designated as Controls should be moved from the *Available Items* box to the *Selected Items* box. Click on ***OK***. This will show the control samples for each of the samples.

  In *Baseline to median of control samples*, for each probe the median of the log summarized values from the control samples is first computed and then this is subtracted from

the sample. If a single sample is chosen as the control sample, then the probe values of the control sample are subtracted from its corresponding sample.

Figure 10.12 shows the Step to perform base line operations for CEL file in Experiment Creation.

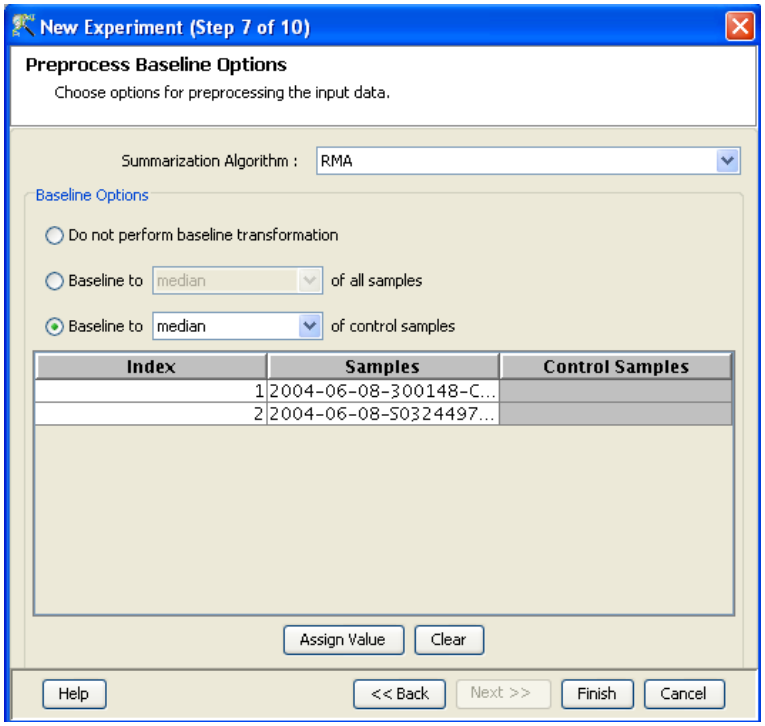


Figure 10.12: Summarization Algorithm

Step 8 of 10 : Normalization This step is specific for CHP files only. See Figure 10.13.

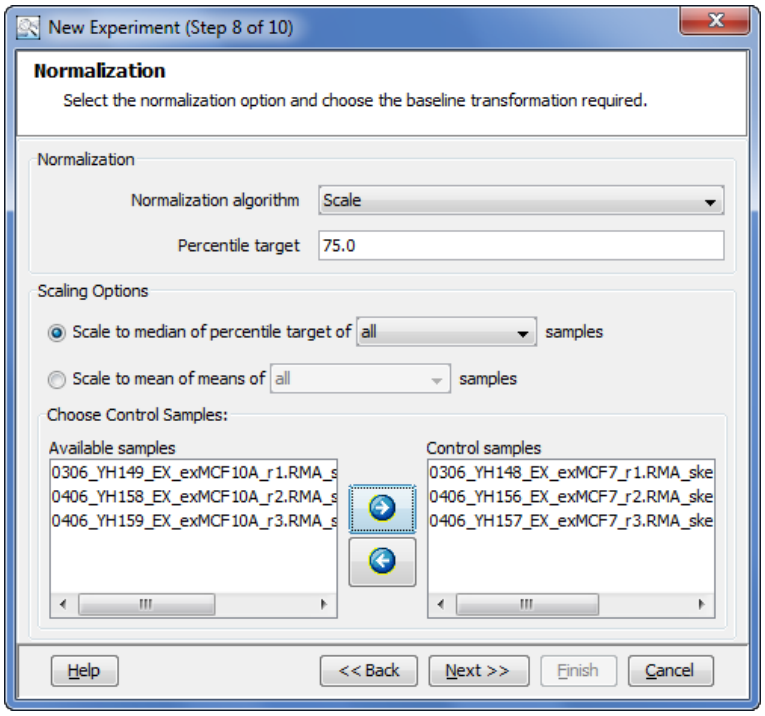


Figure 10.13: Normalization and Baseline Transformation

It gives the user the following normalization options.

- **Percentile Shift:** On selecting this normalization method, the **Percentile Target** box gets enabled allowing the user to enter a specific percentile value. The default value is set at 75^th percentile.
- **Scale:** On selecting this normalization method, the user is presented with an option to scale it to the median of percentile target/mean of means of 'all' or 'control' samples. On choosing the latter, the user has to select the control samples from the available samples in the **Choose Samples** box. The **Percentile Target** box is set at a default value of 75.
- **Normalize to control genes:** After selecting this option, the user has to specify the control genes in the next wizard. The **Percentile Target** box is disabled.
- **Normalize to External Value:** This option will bring up a table listing all samples and a default scaling factor of '1.0' against each of them. The user can use the *'Assign Value'* button at the bottom to assign a different scaling factor to each of the sample; multiple samples can be chosen simultaneously and assigned a value. The **Percentile Target** box is disabled.

  See Chapter Normalization Algorithms for details on normalization algorithms.

Step 9 of 10 : Choose entities If the **Normalize to control genes** option is chosen in the previous step, then the list of control entities can be specified in the following ways in this wizard:

- By choosing a file(s) (txt, csv or tsv) which contains the control entities of choice denoted by their probe id. Any other annotation will not be suitable.
- By searching for a particular entity by using the ***Choose Entities*** option. This leads to a search wizard in which the entities can be selected. All the annotation columns

present in the technology are provided and the user can search using terms from any of the columns. The user has to select the entities that he/she wants to use as controls, when they appear in the **Output Views** page and then click ***Finish***. This will result in the entities getting selected as control entities and will appear in the wizard.

The user can choose either one or both the options to select his/her control genes. The chosen genes can also be removed after selecting the same. See Figure 10.14.

In case the entities chosen are not present in the technology or sample, they will not be taken into account during experiment creation. The entities which are present in the process of experiment creation will appear under matched probe IDs whereas the entities not present will appear under unmatched probe ids in the experiment notes in the experiment inspector.

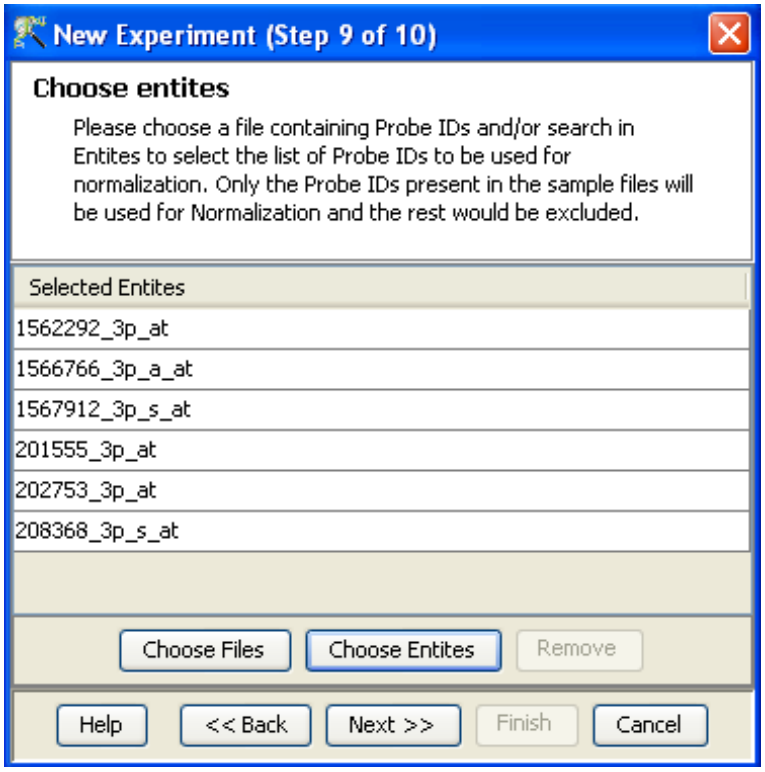


Figure 10.14: Normalize to control genes

Step 10 of 10 : Preprocess Baseline Options (for CHP files) This step allows the user to perform baseline transformation, with CHP files, after normalization. See Figure 10.15. The methods available are the same as those used for CEL files in Step 7 of 10.

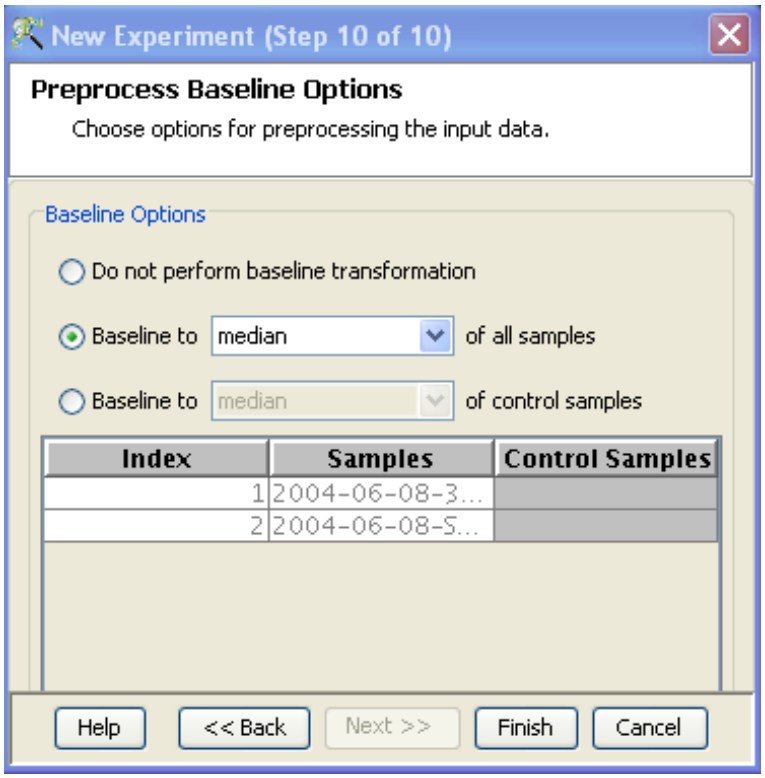


Figure 10.15: Baseline Transformation

Clicking ***Finish*** creates an experiment, which is displayed as a Box Whisker plot in the active view. Alternative views can be chosen for display by navigating to **View** in Toolbar.

10.4.2 Experiment Setup

Once an experiment is created, the *Data Import Wizard* steps appear on the right hand side. This gives options to view quick start guide, create experiment grouping and interpretation and the utility to create gene level experiment from probe level experiment. See section Experiment Setup for complete details.

10.4.3 Quality Control

- **Quality Control on Samples:**

  Quality Control (QC) or the Sample QC allows you to decide which samples are ambiguous and which are passing the quality criteria. Based upon the QC results, the unreliable samples can be removed from the analysis.

NOTE Quality Control on Samples is not supported for sample files in *.txt* format.

The QC dialog shows three tiled windows:

- 3D PCA scores, Correlation Analysis - Samples, and a Correlation Coefficients tab.
- Internal Controls 3′/5′ ratios, Experiment Grouping, Hybridization Controls.
- Legend.

Figure 10.16 has the tiled windows showing the components of QC on samples.

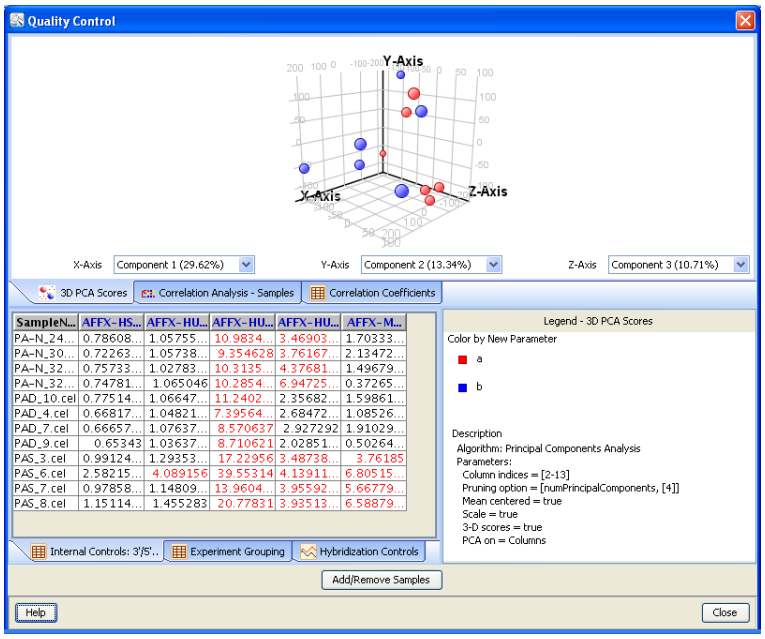


Figure 10.16: Quality Control Dialog

In the **Principal Component Analysis (PCA)** tab, PCA scores are visually represented in a 3D scatter plot. These scores are used to check data quality. It shows one point per array and is colored by the *Experiment Factors* provided earlier in the *Experiment Groupings* view. This allows viewing of separations between groups of replicates. Ideally, replicates within a group should cluster together and separately from arrays in other groups. The PCA components, represented in the X, Y, and Z axes are numbered 1, 2, 3... according to their decreasing significance. The 3D PCA scores plot can be customized via ***Right-Click*⟶*Properties***. To zoom into a 3D Scatter plot, press the Shift key and simultaneously hold down the left mouse button and move the mouse upwards. To zoom out, move the mouse downwards instead. To rotate, press the Ctrl key, simultaneously hold down the left mouse button and move the mouse around the plot.

The **Correlation Analysis - Samples** tab shows the correlation analysis heatmap across arrays (samples) (See How to Perform Correlation Analysis on Samples on page 750). The correlation coefficient for each pair of arrays are displayed in the **Correlation Coefficients** tab. Correlation coefficients are calculated using Pearson Similarity Metric (see Pearson on page 799).

Pearson Similarity Metric: Calculates the mean of all elements in vector **a**. Then it subtracts that value from each element in **a** and calls the resulting vector **A**. It does the same for **b** to make a vector **B**. Correlation Coefficient is then computed as **A.B**/(—**A**——**B**—).

In a Correlation Analysis heatmap:

- Sample Order- In a correlation heatmap, the initial order of samples is the same as the input sample order at the time of experiment creation.
- Clustering - Perform clustering analysis on the correlation heatmap using the **Run clustering** icon from the correlation analysis heatmap toolbar (See To run clustering on a correlation heatmap on page 760). Samples are clustered based on correlation coefficients using hierarchical clustering algorithm and plotted in the order they cluster together. The linkage rule used is average and the distance metrics used is Euclidean.

For more information on the available toolbar icons in a correlation analysis heatmap, see Correlation Heatmap View Toolbar on page 776. See Properties on 792 to change the properties of the correlation analysis heatmap.

NOTE Correlation coefficients are computed on raw, non-normalized data and in linear scale.

The **Internal Controls** view depicts RNA sample quality by showing 3'/5' ratios for a set of specific probesets which include the actin and GAPDH probesets. The 3'/5' ratio is output for each such probeset and for each array. The ratios for actin and GAPDH should be no more than 3 (though for Drosophila, it should be less than 5). A ratio of more than 3 indicates sample degradation and is indicated in the table in red color.

Experiment Grouping tab shows the parameters and parameter values for each sample.

The **Hybridization Controls** view depicts the hybridization quality. Hybridization controls are composed of a mixture of biotin-labelled cRNA transcripts of bioB, bioC, bioD, and cre prepared in staggered concentrations (1.5, 5, 25, and 100 pm respectively). This mixture is spiked-in into the hybridization cocktail. bioB is at the level of assay sensitivity and should be present at least 50% of the time. bioC, bioD and cre must be Present all of the time and must appear in increasing concentrations. The *Hybridization Controls* shows the signal value profiles of these transcripts (only 3' probesets are taken) where the X axis represents the Biotin labelled cRNA transcripts and the Y axis represents the log of the Normalized Signal Values.

The third window shows the legend of the active QC tab.

Unsatisfactory samples or those that have not passed the QC criteria can be removed from further analysis, at this stage, using *Add/Remove Samples* button. Once a few samples are removed, re-summarization of the remaining samples is carried out again. The samples removed earlier can also be added back. Click on **OK** to proceed.

- **Filter Probe Set by Expression:** Entities are filtered based on their signal intensity values. For details refer to the section on Filter Probesets by Expression
- **Filter Probe Set by Flags:**

  This step is specific for analysis where MAS5.0 summarization has been done on samples. MAS5.0 generates flag values, the P(present), M(marginal) and A(absent), for each row in

each sample. In the *Filter Probe Set by Flags* step, entities can be filtered based on their flag values. This is done in 4 steps:

1. Step 1 of 4 : *Entity list and interpretation* window opens up. Select an entity list by clicking on *Choose Entity List* button. Likewise by clicking on *Choose Interpretation* button, select the required interpretation from the navigator window.

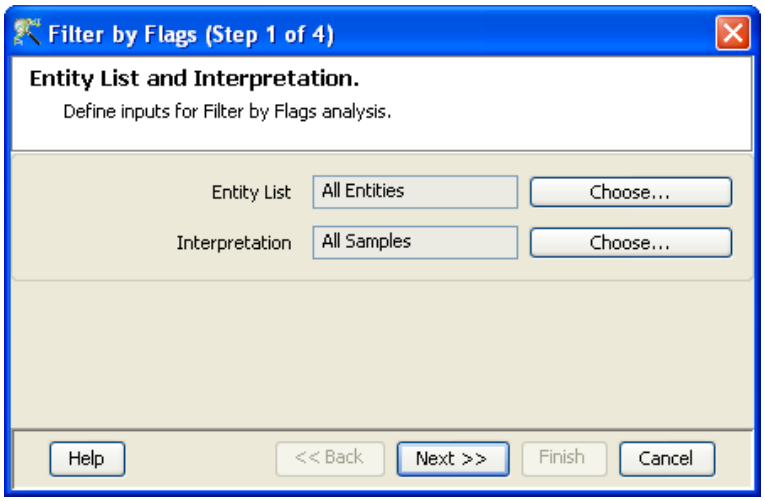


Figure 10.17: Entity List and Interpretation

2. Step 2 of 4: This step is used to set the Filtering criteria and the stringency of the filter. Select the flag values that an entity must satisfy to pass the filter. By default, the Present and Marginal flags are selected. Stringency of the filter can be set in *Retain Entities* box.

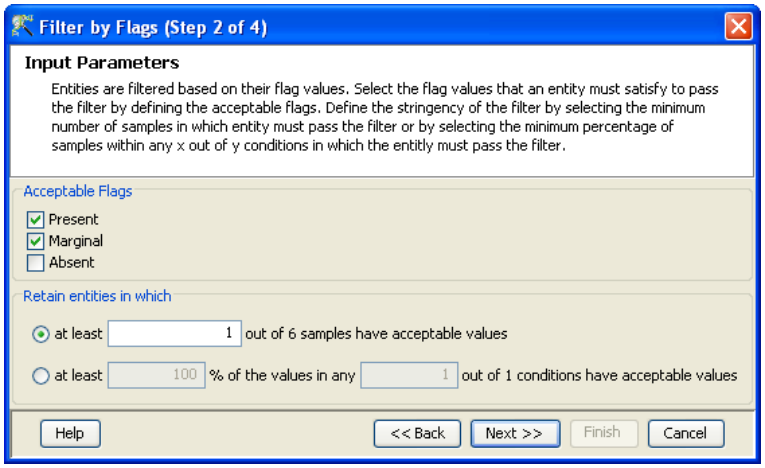


Figure 10.18: Input Parameters

3. Step 3 of 4: A spreadsheet and a profile plot appear as two tabs, displaying those probes which have passed the filter conditions. Baseline transformed data is shown here. Total number of probes and number of probes passing the filter are displayed on the top of the navigator window. (See Figure 10.19).

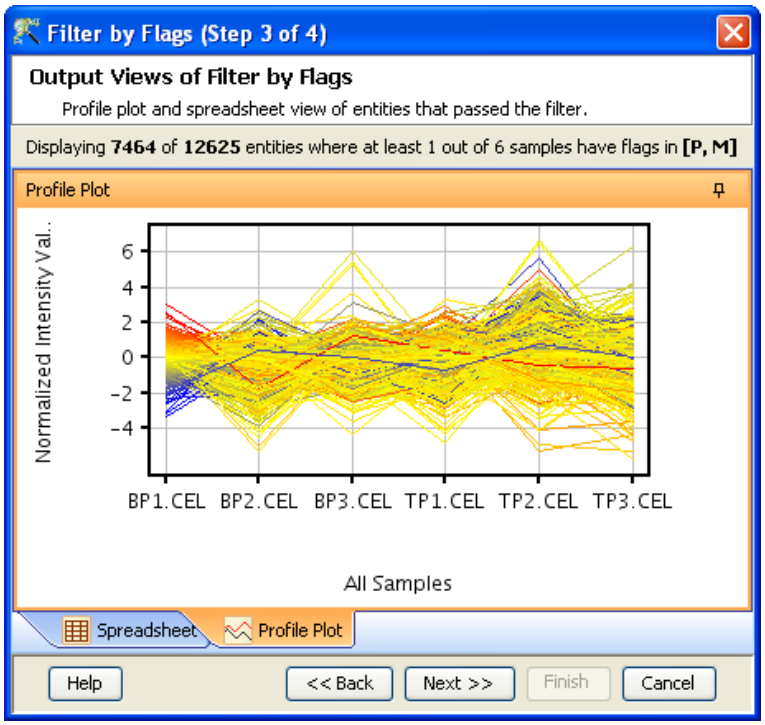


Figure 10.19: Output Views of Filter by Flags

4. Step 4 of 4: Click *Next* to annotate and save the entity list. (See Figure 10.20).

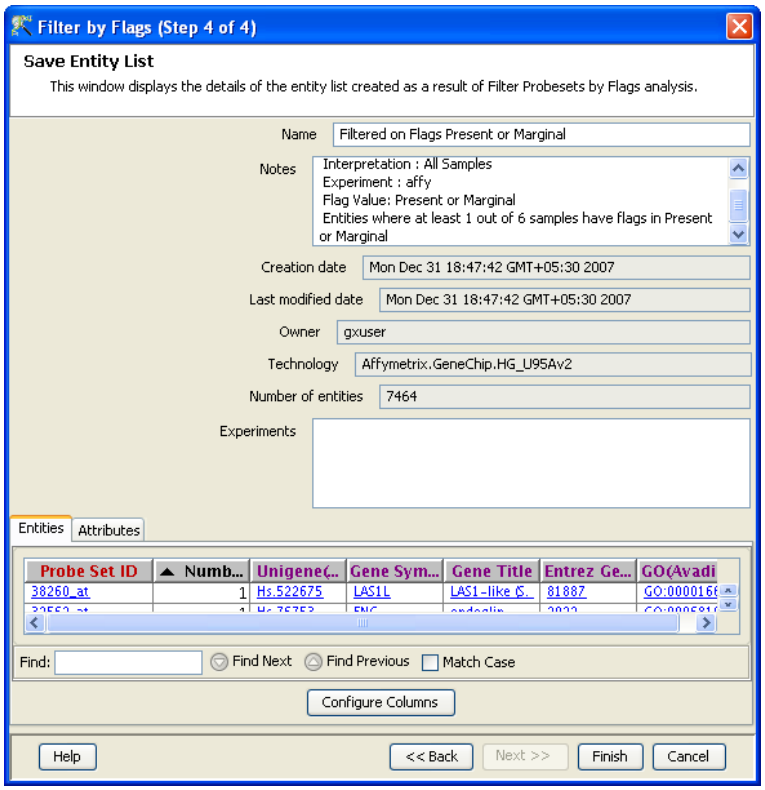


Figure 10.20: Save Entity List

- **Filter Probesets by Error:** Entities can be filtered based on the standard deviation or coefficient of variation using this option. For details refer to the section on Filter Probesets by Error

10.4.4 Analysis

- **Statistical Analysis**

  For details refer to section Statistical Analysis in the Data Import Wizard section.

- **Filter on Volcano Plot**

  For details refer to section Filter on Volcano Plot.

- **Fold Change**

  For details refer to section Fold change.

- **Clustering**

  For details refer to chapter Clustering Analysis.

- **Correlation**

  For details refer to chapter Correlation Analysis.

- **Find Similar Entities**

  For details refer to section Find similar entities.

- **Filter on Parameters**
  For details refer to section Filter on Parameters.
- **Principal Component Analysis**
  For details refer to section PCA.

10.4.5 Class Prediction

- **Build Prediction Model** For details refer to section Build Prediction Model.
- **Run Prediction** For details refer to section Run Prediction.

10.4.6 Results

- **Gene Ontology (GO) analysis**
  For more details see GO Analysis.
- **Gene Set Enrichment Analysis (GSEA)**
  For more details see Performing GSEA in **GeneSpring**.
- **Gene Set Analysis (GSA)**
  For more details see Gene Set Analysis.
- **Find Similar Entity Lists**
  This feature is discussed in a separate section called Find Similar Entity Lists.
- **Single Experiment Analysis or Multi-Omic Analysis**
  This feature is discussed in detail in the section Single Experiment and Multi-Omic Analysis.
- **Export to IPA**
  This feature is discussed in detail in the section Ingenuity Pathways Analysis (IPA) Connector.
- **Import IPA Entity List**
  This feature is discussed in detail in the section Import IPA Entity List.
- **NLP Network Discovery**
  This feature is discussed in detail in the section NLP Network Discovery.
- **MeSH Network Builder**
  This feature is discussed in detail in the section MeSH Network Builder.
- **Extract Relations via NLP**
  This feature is discussed in detail in the section Extract Relations via NLP.

10.4.7 Utilities

- **Import Entity list from File**

  For details refer to section Import Entity list from File.

- **Analysis: Biological Significance**

  For details refer to section Analysis: Biological Significance.

- **Filter On Entity List**

  For further details refer to section Filter On Entity List.

- **Remove Entities with missing signal values**

  For details refer to section Remove Entities with missing signal values.

10.4.8 Affymetrix Technology creation using Custom CDF

Creating a Technology using Affymetrix Custom CDF:

GeneSpring offers the user a facility to create Custom Affymetrix expression (GeneChip) technology if you have a Custom CDF file. This happens in situations where you have a Custom Affymetrix array or might want to use a Custom CDF for a Standard technology, for e.g., the ones obtained from `http://brainarray.mbni.med.umich.edu/brainarray/Database/CustomCDF/genomic_curated_CDF.asp`

Following are the steps for creating an Affymetrix Custom technology:

1. Go to *Annotations* $\longrightarrow$ *Create Technology* $\longrightarrow$ *Affymetrix Expression.*
2. For creating a Custom Affymetrix technology, the CDF file is mandatory. The PSI, CIF, Probe Tab and annotation files are optional. If however, Probe Tab is also being used, then make sure that the system has 'R' package installed and that its path has been set, by going to *Tools* $\longrightarrow$ *Options* $\longrightarrow$ *Miscellaneous*$\longrightarrow$ *R path.* The Bioconductor packages makecdfenv, matchprobes and gcrma also need to be installed before the Probe Tab file can be used. See Figures 10.21 and 10.22.

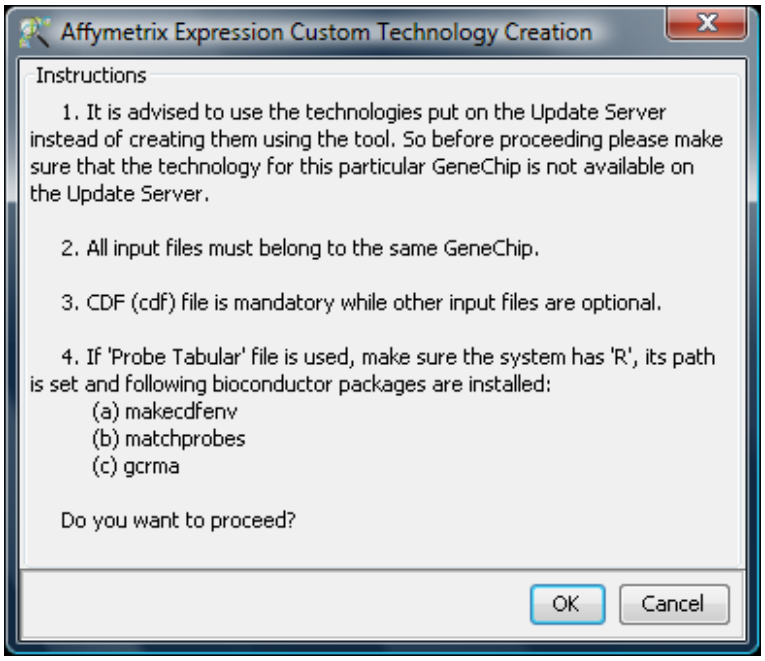


Figure 10.21: Confirmation Dialog Box

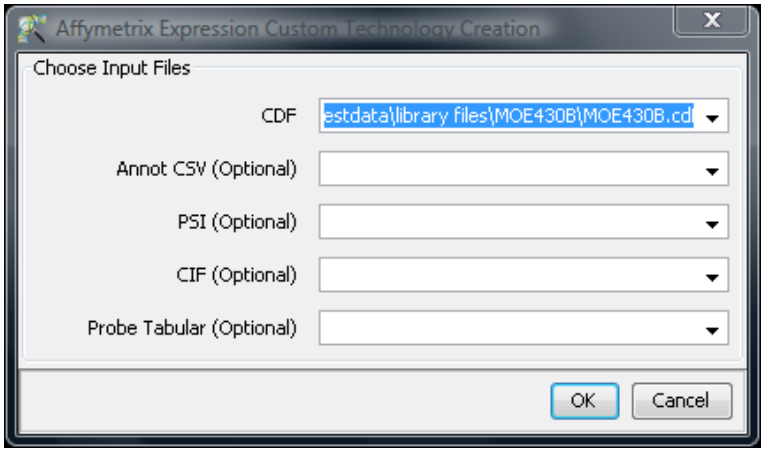


Figure 10.22: Choose Input Files

3. The CDF file name should reflect the GeneChip name for which it is being used. If the Custom CDF is derived from a Standard technology, then it should be renamed to that of the Standard technology for e.g., the Custom CDF file HS95Av2_HS_UG_1.cdf derived from GeneChip HG_U95Av2 should be renamed to HG_U95Av2.cdf. This is necessary because to ensure that no errors occur, **GeneSpring** tries to match the CDF/technology name with the GeneChip name from the data file, during the process of experiment creation. Taking again the example of HS95Av2_HS_UG_1.cdf (for the GeneChip HG_U95Av2), if the CDF is not renamed, an experiment created using the HG_U95Av2 CEL files will use the the Standard Affymetrix.GeneChip.HG_U95Av2 technology, instead of the newly created Affymetrix.GeneChip.HS95Av2_HS_UG_1 technology.

4. In case of Custom CDF derived from a Standard technology, refer to either *Search* $\longrightarrow$

Technologies or go to *Annotations* $\longrightarrow$ *Create Technology* $\longrightarrow$ *From Agilent Server* to get the exact name of the GeneChip (It is case sensitive).

5. The technology created will automatically take the name of the Custom CDF and will be named as Affymetrix.GeneChip.<CDF_file_name>. If a technology with the same name, standard or custom already exists, then **GeneSpring** overwrites it with the new one after the user's confirmation.

Notes:

1. In **GeneSpring**, for a given Affymetrix GeneChip, at any point of time, you cannot have more than one technology.
2. If you replace an older Affymetrix technology with the new one, then the behavior of the experiments created with the older technology is unpredictable. So it is advised to remove/delete those.
3. If an experiment needs to be analyzed now with the standard technology for which a Custom CDF had been used, it will be necessary to delete the technology created using the Custom CDF. Otherwise, the standard technology cannot be downloaded.

Chapter 11

Affymetrix Summarization Algorithms

This section describes technical details of the various probe summarization algorithms, normalization using spike-in and housekeeping probesets, and computing absolute calls.

11.0.1 Probe Summarization Algorithms

Probe summarization algorithms perform the following 3 key tasks: Background Correction, Normalization, and Probe Summarization (i.e. conversion of probe level values to probeset expression values in a robust, i.e., outlier resistant manner. The order of the last two steps could differ for different probe summarization algorithms. For example, the RMA algorithm does normalization first, while MAS5 does normalization last. In RMA and GCRMA the summarization is inherently on log scale, whereas in PLIER and MAS5 summarization works on linear scale. Further, the methods mentioned below fall into one of two classes -- the PM based methods and the $PM - MM$ based methods. The $PM - MM$ based methods take $PM - MM$ as their measure of background corrected expression while the PM based measures use other techniques for background correction. MAS5, MAS4, and Li-Wong are $PM - MM$ based measures while RMA and GCRMA are PM based measures. For a comparative analysis of these methods, see [33, 34].

A brief description of each of the probe summarization options available in **GeneSpring** is given below. Some of these algorithms are native implementations within **GeneSpring** and some are directly based on the Affymetrix codebase. The exact details are described in the table below.

RMA with only pm probes	Implemented in **GeneSpring**	Validated against R with bgversion=2
GCRMA	Implemented in **GeneSpring**	Validated against default GCRMA in R
MAS5	Licensed from Affymetrix	Validated against Affymetrix Data
PLIER	Summarization licensed from Affymetrix, Normalization implemented in **GeneSpring**	Validated against Affymetrix Data
LiWong	Implemented in **GeneSpring**	Validated against R
Absolute Calls	Licensed from Affymetrix	Validated against Affymetrix Data

Masked Probes and Outliers. Finally, note that CEL files have masking and outlier information about certain probes. These masked probes and outliers are removed.

11.0.1.1 The RMA (Robust Multichip Averaging) Algorithm

The RMA method was introduced by Irazarry et al. [33, 34] and is used as part of the RMA package in the Bioconductor suite. In contrast to MAS5, this is a PM based method. It has the following components.

Background Correction. The RMA background correction method is based on the distribution of PM values amongst probes on an Affymetrix array. The key observation is that the smoothened histogram of the $\log(PM)$ values exhibits a sharp normal-like distribution to the left of the mode (i.e., the peak value) but stretches out much more to the right, suggesting that the PM values are a mixture of non-specific binding and background noise on one hand and specific binding on the other hand. The above peak value is a natural estimate of the average background noise and this can be subtracted from all PM values to get background corrected PM values. However, this causes the problem of negative values. Irizarry et al. [33, 34] solve the problem of negative values by imposing a positive distribution on the background corrected values. They assume that each observed PM value O is a sum of two components, a signal S which is assumed to be exponentially distributed (and is therefore always positive) and a noise component N which is normally distributed. The background corrected value is obtained by determining the expectation of S conditioned on O which can be computed using a closed form formula. However, this requires estimating the decay parameter of the exponential distribution and the mean and variance of the normal distribution from the data at hand. These are currently estimated in a somewhat ad-hoc manner.

Normalization. The RMA method uses Quantile normalization. Each array contains a certain distribution of expression values and this method aims at making the distributions across various arrays not just similar but identical! This is done as follows. Imagine that the expression values from various arrays have been loaded into a dataset with probesets along rows and arrays along columns. First, each column is sorted in increasing order. Next, the value in each row is replaced with the average of the values in this row. Finally, the columns are unsorted (i.e., the effect of the sorting step is reversed so that the items in a column go back to wherever they came from). Statistically, this method seems to obtain very sharp normalizations [11]. Further, implementations of this method run very fast.

GeneSpring uses all arrays to perform normalization on the raw intensities, irrespective of their variance.

Probe Summarization. RMA models the observed probe behavior (i.e., $\log(PM)$ after background correction) on the log scale as the sum of a probe specific term, the actual expression value on the log scale, and an independent identically distributed noise term. It then estimates the actual expression value from this model using a robust procedure called *Median Polish*, a classic method due to Tukey.

11.0.1.2 The GCRMA Algorithm

This algorithm was introduced by Wu et al [59] and differs from RMA only in the background correction step. The goal behind its design was to reduce the bias caused by not subtracting MM in the RMA algorithm. The GCRMA algorithm uses a rather technical procedure to reduce this bias and is based on the fact that the non-specific affinity of a probe is related to its base sequence. The algorithm computes a background value to be subtracted from each probe using its base sequence

11.0.1.3 The Li-Wong Algorithm

There are two versions of the Li-Wong algorithm [40], one which is $PM - MM$ based and the other which is PM based. Both are available in the dChip software. **GeneSpring**has only the $PM - MM$ version.

Background Correction. No special background correction is used by the **GeneSpring** implementation of this method. Some background correction is implicit in the $PM - MM$ measure.

Normalization. While no specific normalization method is part of the Li-Wong algorithm as such, dChip uses *Invariant Set* normalization. An *invariant set* is a a collection of probes with the most conserved ranks of expression values across all arrays. These are identified and then used very much as spike-in probesets would be used for normalization across arrays. In **GeneSpring**, the current implementation uses Quantile Normalization [11] instead, as in RMA.

Probe Summarization. The Li and Wong [40] model is similar to the RMA model but on a linear scale. Observed probe behavior (i.e., $PM - MM$ values) is modelled on the linear scale as a *product* of a probe affinity term and an actual expression term along with an additive normally distributed independent error term. The maximum likelihood estimate of the actual expression level is then determined using an estimation procedure which has rules for outlier removal. The outlier removal happens at multiple levels. At the first level, outlier arrays are determined and removed. At the second level, a probe is removed from all the arrays. At the third level, the expression value for a particular probe on a particular array is rejected. These three levels are performed in various iterative cycles until convergence is achieved. Finally, note that since $PM - MM$ values could be negative and since **GeneSpring** outputs values always on the logarithmic scale, negative values are thresholded to 1 before output.

11.0.1.4 The Average Difference and Tukey-BiWeight Algorithms

These algorithms are similar to the MAS4 and MAS5 methods [30] used in the Affymetrix software, respectively.

Background Correction. These algorithm divide the entire array into 16 rectangular zones and the second percentile of the probe values in each zone (both PM's and MM's combined) is chosen as the background value for that region. For each probe, the intention now is to reduce the expression level measured for this probe by an amount equal to the background level computed for the zone containing this probe. However, this could result in discontinuities at zone boundaries. To make these transitions smooth, what is actually subtracted from each probe is a weighted combination of the background levels computed above for all the zones. Negative values are avoided by thresholding.

Probe Summarization. The one-step Tukey Biweight algorithm combines together the background corrected $\log(PM - MM)$ values for probes within a probe set (actually, a slight variant of MM is used to ensure that $PM - MM$ does not become negative). This method involves finding the median and weighting the items based on their distance from the median so that items further away from the median are down-weighted prior to averaging.

The Average Difference algorithm works on the background corrected $PM - MM$ values for a probe. It ignores probes with $PM - MM$ intensities in the extreme 10 percentiles. It then computes the mean and standard deviation of the $PM - MM$ for the remaining probes. Average of $PM - MM$ intensities within 2 standard deviations from the computed mean is thresholded to 1 and converted to the log scale. This value is then output for the probeset.

Normalization. This step is done after probe summarization and is just a simple scaling to equalize means or trimmed means (means calculated after removing very low and very high intensities for robustness).

11.0.1.5 The PLIER Algorithm

This algorithm was introduced by Hubbell [31] and introduces a integrated and mathematically elegant paradigm for background correction and probe summarization. The normalization performed is the same as in RMA, i.e., Quantile Normalization. After normalization, the PLIER procedure runs an optimization procedure which determines the best set of weights on the PM and MM for each probe pair. The goal is to weight the PMs and MMs differentially so that the weighted difference between PM and MM is non-negative. Optimization is required to make sure that the weights are as close to 1 as possible. In the process of determining these weights, the method also computes the final summarized value.

11.0.1.6 Comparative Performance

For comparative performances of the above mentioned algorithm, see [33, 34] where it is reported that the RMA algorithm outperforms the others on the GeneLogic spike-in study [24].

11.0.2 Computing Absolute Calls

GeneSpring uses code licenced from Affymetrix to compute calls. The Present, Absent and Marginal Absolute calls are computed using a Wilcoxon Signed Rank test on the (PM-MM)/(PM+MM) values for probes within a probeset. This algorithm uses the following parameters for making these calls:

- The *Threshold Discrimination Score* is used in the Wilcoxon Signed Rank test performed on (PM-MM)/(PM+MM) values to determine signs. A higher threshold would decrease the number of false positives but would increase the number of false negatives.
- The second and third parameters are the *Lower Critical p-value* and the *Higher Critical p-value* for making the calls. Genes with p-value in between these two values will be called Marginal, genes with p-value above the Higher Critical p-value will be called Absent and all other genes will be called Present.

11.0.2.1 Parameters for Summarization Algorithms and Calls

The algorithms MAS5 and PLIER and the Absolute Call generation procedure use parameters which can be seen at *File* $\longrightarrow$ *Configuration*. However, modifications of these parameters are not currently available in **GeneSpring**. These should be available in the future versions.

Chapter 12

Analyzing Illumina Data

GeneSpring supports the Illumina Single Color (Direct Hyb) experiments. **GeneSpring** supports only those projects from GenomeStudio which were created using the bgx manifest files. To generate the data file, the Sample Probe Profile should be exported out from GenomeStudio in **GeneSpring** format. These text files can then be imported into **GeneSpring** . From these text file, the

- Probe ID,
- Average Signal values and the
- detection p-value columns

are automatically extracted and used for project creation. Typically, a single Illumina data file contains multiple samples.

GenomeStudio provides the option of performing normalization on the data, therefore if the data is already normalized, the workflow to be chosen is Data Import Wizard. This is because, *Data Import Wizard* allows the user to skip normalization steps whereas in *Analysis: Biological Significance workflow*, normalization is performed by default.

Projects from GenomeStudio created using .xml files can still be analyzed in **GeneSpring**, via the Custom technology creation or as Generic Single Color experiments. For more details on the same, see the section on Illumina Custom Technology creation

12.1 Running the Illumina Workflow:

Follow the instructions in section Running the Expression Experiment Workflow to start the experiment. Figures 12.1, 12.5, 12.3 illustrate the process of choosing analysis type, experiment type, loading data, and choosing samples for Illumina Experiment.

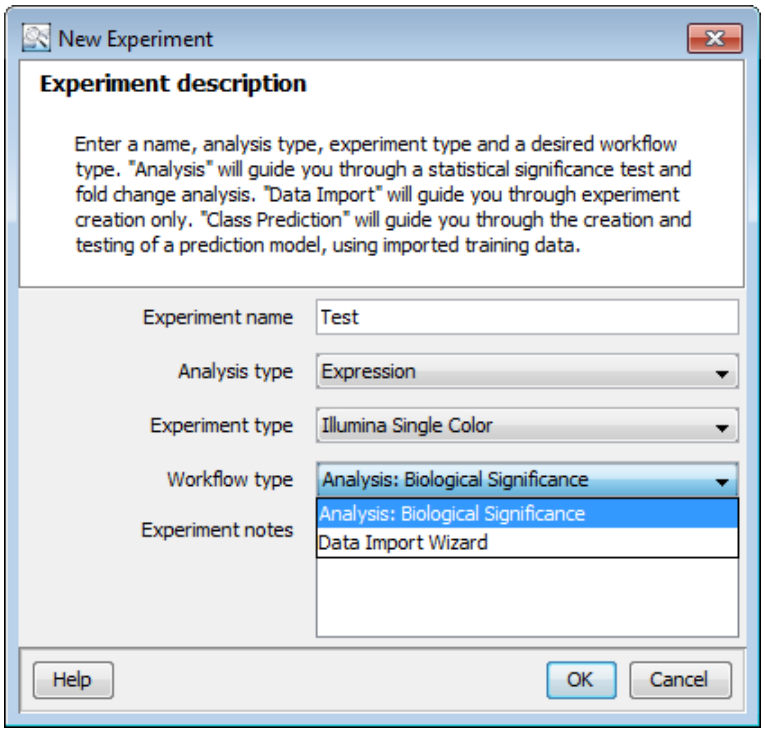


Figure 12.1: Experiment Description

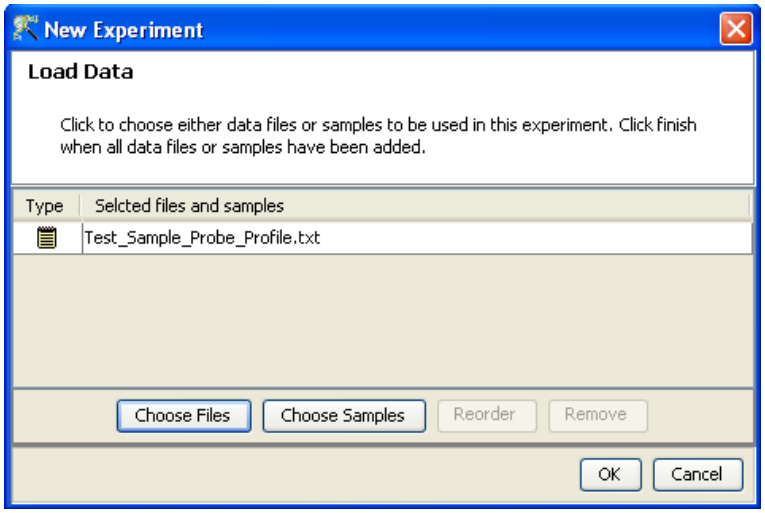


Figure 12.2: Load Data

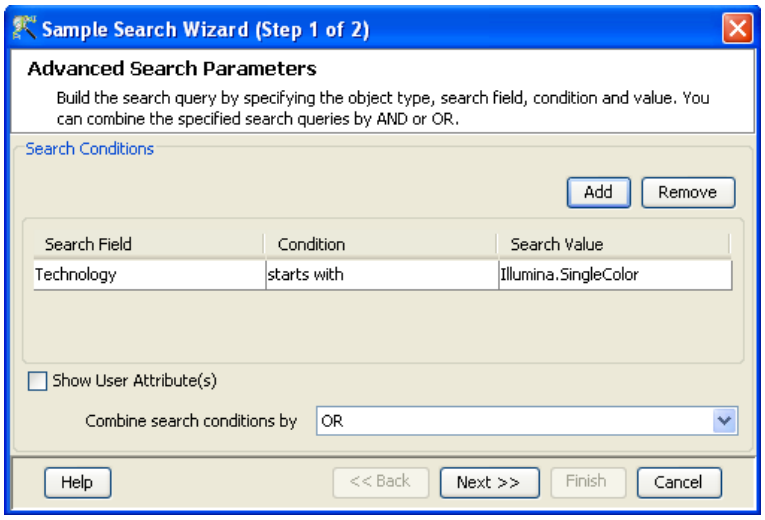


Figure 12.3: Choose Samples

Analysis: Biological Significance wizard appears with the sequence of steps on the left hand side with the current step being highlighted. The Workflow allows the user to proceed in schematic fashion and does not allow the user to skip steps.

12.2 Data Processing for Illumina arrays

- **File formats**: The data file (.txt format) should be the Sample Probe Profile that is exported out from GenomeStudio in **GeneSpring** format. Also note that the samples in the data file should not have special characters such as `! * # ; ? \ / : " < > |` and space in their names.
- **Raw Signal Values**: The term ''raw'' signal values refer to the linear data that is present in the data file that is imported into **GeneSpring** from GenomeStudio.
- **Normalized Signal Values**: ''Normalized'' value is the value generated after thresholding, log transformation and normalization (Percentile Shift, Scale, Normalize to control genes or Quantile) and Baseline Transformation.
- **Treatment of on-chip replicates**: It is not applicable as the data obtained from GenomeStudio is already summarized.
- **Flag values**: The flag values are calculated based on the detection p-value column (from GenomeStudio) and the flag settings defined by the user in the second step of experiment creation in the Data Import Wizard. (In Analysis: Biological Significance, default settings are used)
- **Treatment of Control probes**: The control probes are included while performing normalization.
- **Empty Cells**: Not Applicable.
- **Sequence of events**: The sequence of events involved in the processing of the text data files is: Thresholding$\longrightarrow$ log transformation$\longrightarrow$ Normalization$\longrightarrow$ Baseline Transformation

12.3 Analysis: Biological Significance

The Analysis: Biological Significance workflow has 8 steps:

1. Summary Report
2. Experiment Grouping
3. QC on samples
4. Filter Probe Sets
5. Significance analysis
6. Fold Change
7. GO analysis
8. Pathway Analysis

These steps are explained in section Analysis: Biological Significance. Details specific to Illumina Expression Data are explained here.

Quality Control (Step 3 of 8):

The 3rd step in the Analysis: Biological Significance is the QC on samples which is displayed in the form of four tiled windows. They are as follows:

- Correlation coefficients table and Experiment grouping tabs
- Correlation coefficients plot
- PCA scores.
- Legend

QC on Samples generates four tiled windows as seen in Figure 12.4.

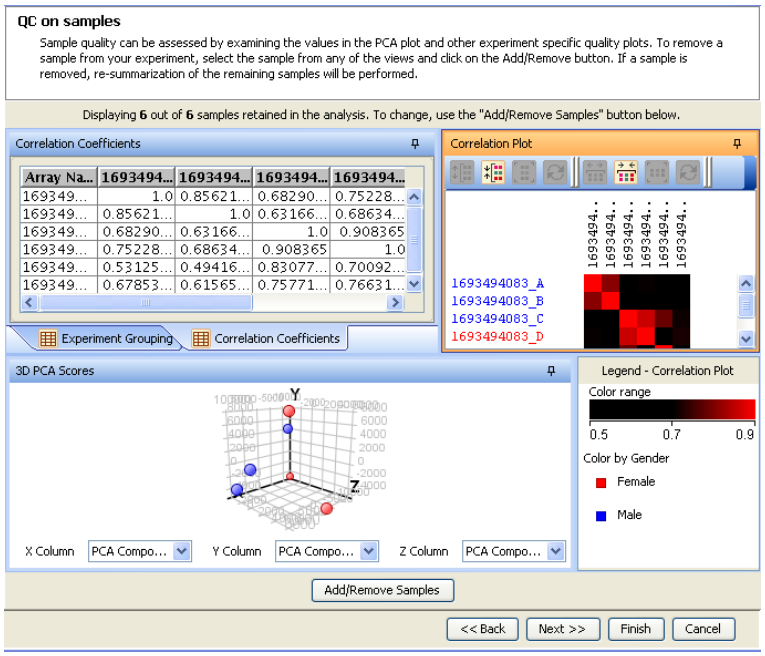


Figure 12.4: Quality Control on Samples

The views in these windows are lassoed i.e., selecting the sample in any of the view highlights the sample in all the views.

The *Correlation Plots* shows the correlation analysis across arrays. It finds the correlation coefficient for each pair of arrays and then displays these in textual form as a correlation table as well as in visual form as a heatmap. The correlation coefficient is calculated using Pearson Correlation Coefficient.

Pearson Correlation: Calculates the mean of all elements in vector **a**. Then it subtracts that value from each element in **a** and calls the resulting vector **A**. It does the same for **b** to make a vector **B**. Result = **A.B**/(—**A**——**B**—)

Samples are clustered based on correlation coefficients using hierarchical clustering and plotted in the order they cluster together by default. The order of the samples in the heatmap is based on either of the following:

- Clustering - This ordering is done using the hierarchical clustering model and is the default option. The linkage rule used is average and the distance metrics used is Euclidean.
- Sample Order- This ordering reflects the input sample order.
- Experiment Grouping- This option clusters the samples together based on the experiment grouping.

This reordering option can be accessed via Right click⟶ Properties⟶ Sample Order.

The heatmap is colorable by Experiment Factor information via Right-Click⟶ Properties. The intensity levels in the heatmap can also be customized here.

Principal Component Analysis (PCA) calculates the PCA scores and visually represents them in a 3D scatter plot. The scores are used to check data quality. It shows one point per array and is colored by the *Experiment Factors* provided earlier in the *Experiment Groupings* view. This allows viewing of separations between groups of replicates. Ideally, replicates within a group should cluster together and separately from arrays in other groups. The PCA components, represented in the X, Y and Z axes are numbered 1, 2, 3... according to their decreasing significance. The 3D PCA scores plot can be customized via ***Right-Click⟶ Properties***. To zoom into a 3D Scatter plot, press the Shift key and simultaneously hold down the left mouse button and move the mouse upwards. To zoom out, move the mouse downwards instead. To rotate, press the Ctrl key, simultaneously hold down the left mouse button and move the mouse around the plot.

The *Add/Remove* samples allows the user to remove the unsatisfactory samples and to add the samples back if required. Whenever samples are removed or added back, normalization as well as baseline transformation is performed again on the samples. Click on *OK* to proceed.

The fourth window shows the legend of the active QC tab.

12.4 Data Import Wizard

The *Data Import Wizard* offers a variety of choices to the user for the analysis. The detection p-value range can be selected to decide on Present and Absent calls, raw signal thresholding can be altered and either Median Shift or Quantile Normalization can be chosen. Additionally there are options for baseline transformation of the data and for creating different interpretations. To create and analyze an experiment using the *Data Import Wizard*, load the data as described earlier. In the *New Experiment Dialog*, choose the *Workflow Type* as Advanced. Click *OK* will open a new experiment wizard which then proceeds as follows:

1. **New Experiment (Step 1 of 5):** As in case of *Analysis: Biological Significance*, either data files can be imported or else pre-created samples can be used.

   - For loading new text files, use *Choose Files*.
   - If the txt files have been previously used in **GeneSpring** experiments *Choose Samples* can be used.

   Step 1 of 3 of Experiment Creation, the 'Load Data' window, is shown in Figure 12.5.

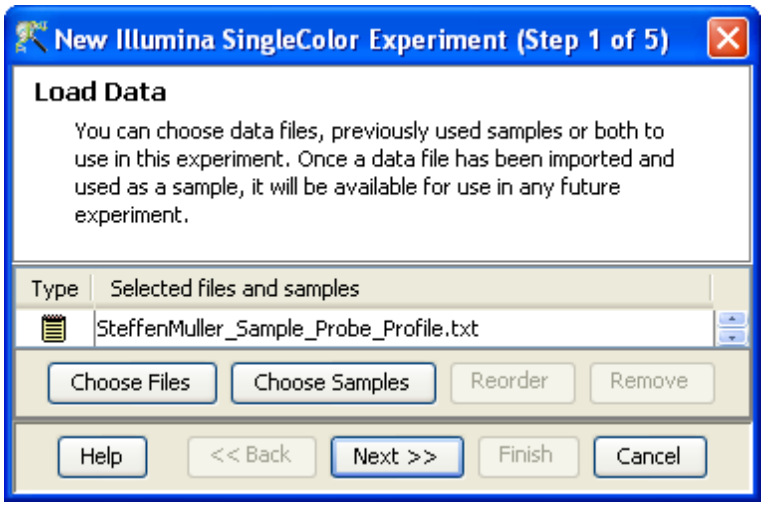


Figure 12.5: Load Data

2. **New Experiment (Step 2 of 5):** This step allows the user to determine the detection p-value range for Present and Absent flags. The Intermediate range will be taken as Marginal. The default values that are given for Present and Absent flags are 0.8 (lower cut-off) and 0.6 (upper cut-off) respectively. Step 2 of 3 of Experiment Creation, the Identify Calls Range window, is depicted in the Figure 12.6.

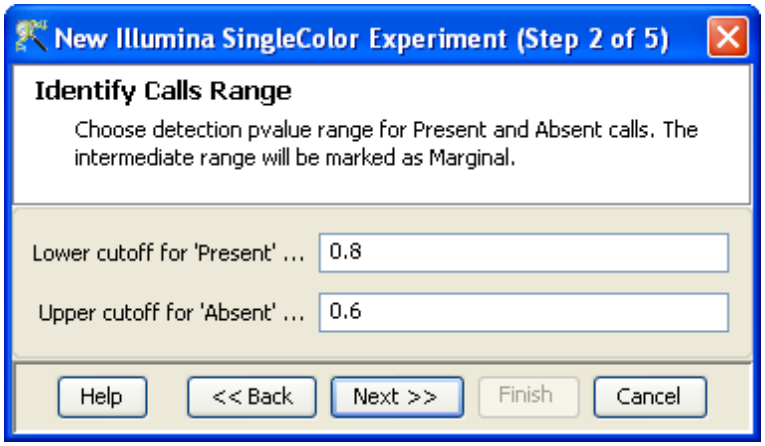


Figure 12.6: Identify Calls Range

3. **New Experiment (Step 3 of 5):** Criteria for preprocessing of input data is set here. It allows the user to threshold raw signals to chosen values and to select normalization algorithms(Quantile, Percentile Shift, Scale and Normalize to control genes).

   - **Percentile Shift:** On selecting this normalization method, the **Percentile Target** box gets enabled allowing the user to enter a specific percentile value. The default value is set at 75^th percentile.
   - **Scale:** On selecting this normalization method, the user is presented with an option to scale it to the median of percentile target/mean of means of 'all' or 'control' samples. On choosing the latter, the user has to select the control samples from the available samples in the **Choose Samples** box. The **Percentile Target** box is set at a default value of 75.

- **Normalize to control genes:** After selecting this option, the user has to specify the control genes in the next wizard. The **Percentile Target** box is disabled.
- **Normalize to External Value:** This option will bring up a table listing all samples and a default scaling factor of '1.0' against each of them. The user can use the *'Assign Value'* button at the bottom to assign a different scaling factor to each of the sample; multiple samples can be chosen simultaneously and assigned a value. The **Percentile Target** box is disabled.

For details on the above normalization methods, refer to section Normalization Algorithms.

Figure 12.7 shows the Step 3 of 5 of Experiment Creation.

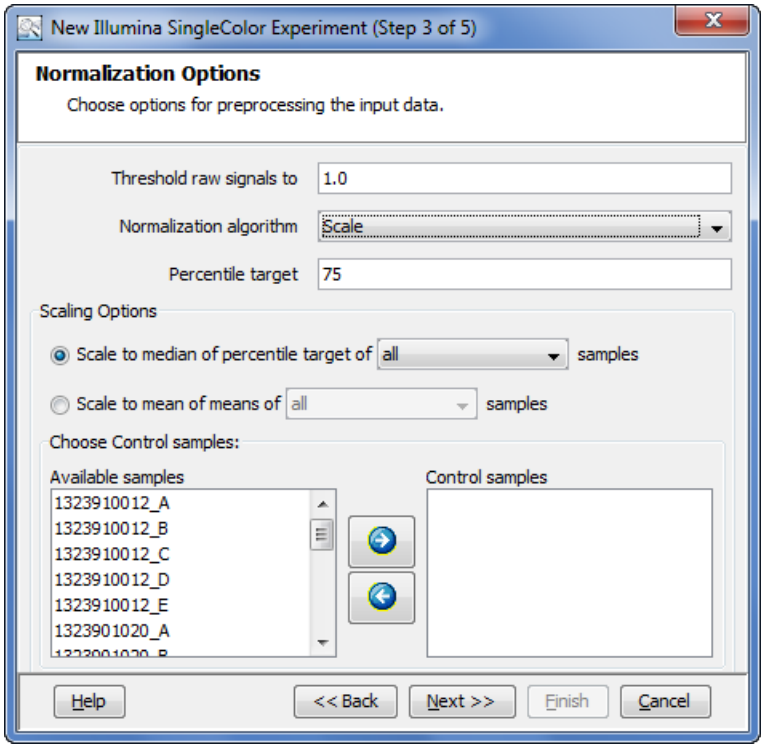


Figure 12.7: Preprocess Options

4. **New Experiment (Step 4 of 5):**

   If the **Normalize to control genes** option is chosen, then the list of control entities can be specified in the following ways in this wizard:

   - By choosing a file(s) (txt, csv or tsv) which contains the control entities of choice denoted by their probe id. Any other annotation will not be suitable.
   - By searching for a particular entity by using the ***Choose Entities*** option. This leads to a search wizard in which the entities can be selected. All the annotation columns present in the technology are provided and the user can search using terms from any of the columns. The user has to select the entities that he/she wants to use as controls when they appear in the **Output Views** page and then click ***Finish***. This will result in the entities getting selected as control entities and will appear in the wizard.

   The user can choose either one or both the options to select his/her control genes. The chosen genes can also be removed after selection is over. See Figure 12.8.

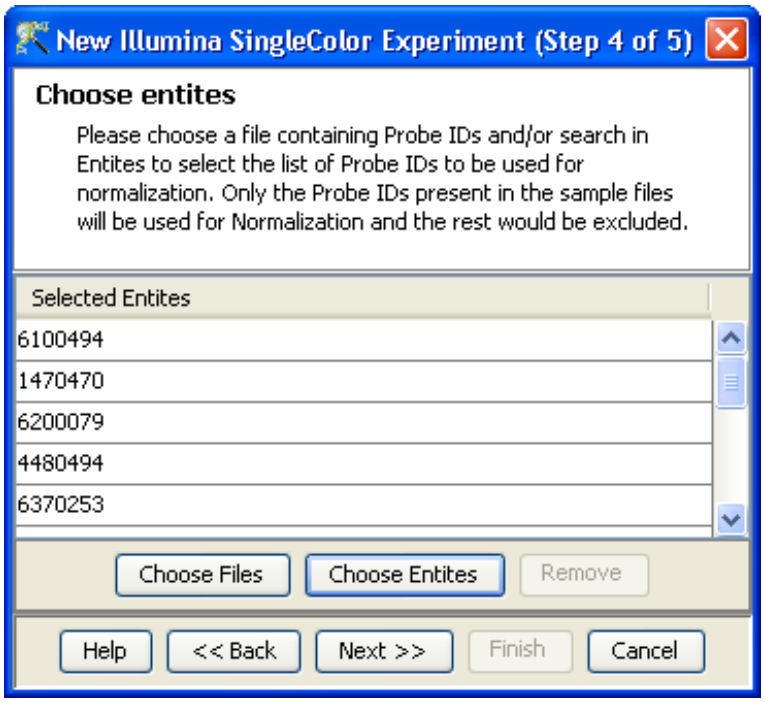


Figure 12.8: Choose Entities

In case the entities chosen are not present in the technology or sample, they will not be taken into account during experiment creation. The entities which are present in the process of experiment creation will appear under matched probe ids whereas the entities not present will appear under unmatched probe ids in the experiment notes in the experiment inspector.

5. **New Experiment (Step 5 of 5):**

   This step allows the user to perform baseline transformation. See Figure 12.9. The baseline options include:

   - ***Do not perform baseline***
   - ***Baseline to median of all samples:*** For each probe the median of the log summarized values from all the samples is calculated and subtracted from each of the samples.
   - ***Baseline to median of control samples:*** For each sample, an individual control or a set of controls can be assigned. Alternatively, a set of samples designated as controls can be used for all samples. For specifying the control for a sample, select the sample and click on ***Assign value***. This opens up the ***Choose Control Samples*** window. The samples designated as Controls should be moved from the *Available Items* box to the *Selected Items* box. Click on ***Ok***. This will show the control samples for each of the samples.

     In *Baseline to median of control samples*, for each probe the median of the log summarized values from the control samples is first computed and then this is subtracted from the sample. If a single sample is chosen as the control sample, then the probe values of the control sample are subtracted from its corresponding sample.

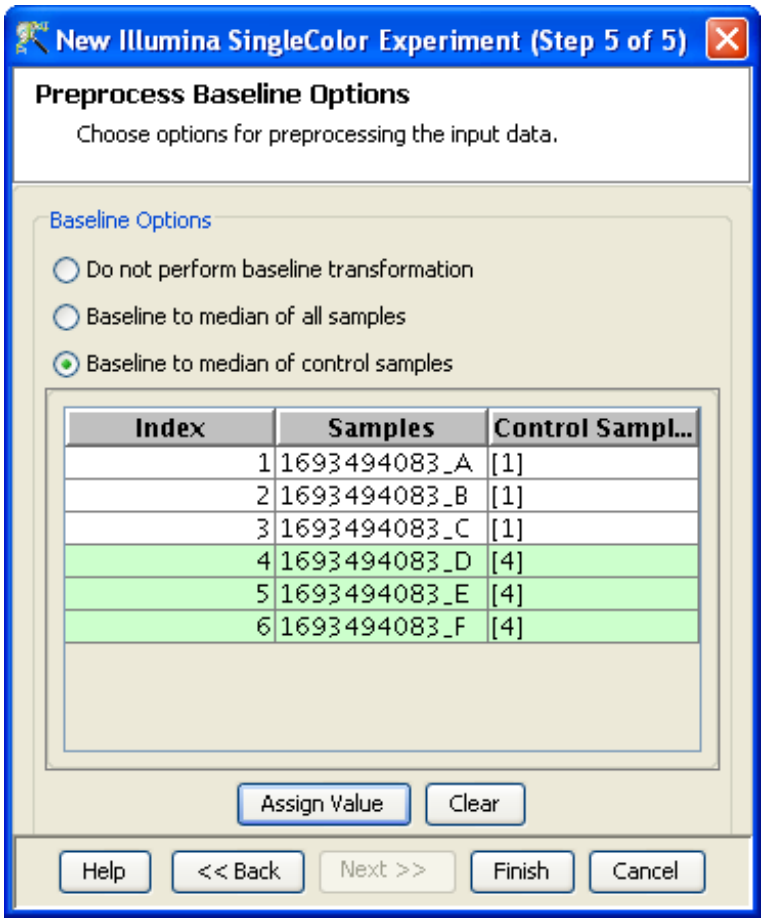


Figure 12.9: Preprocess Baseline Options

12.4.1 Experiment Setup

Once an experiment is created, the *Data Import Wizard* steps appear on the right hand side. This gives options to view quick start guide, create experiment grouping and interpretation and the utility to create gene level experiment from probe level experiment. See section Experiment Setup for complete details.

12.4.2 Quality control

- **Quality Control on samples:**

  Quality Control (QC) allows you decide which samples are ambiguous and which are passing the quality criteria. Based on the QC results, unreliable you can remove samples from further downstream analysis. The QC dialog displays three tiled windows:

  – 3D PCA scores, Correlation Analysis - Samples, and a Correlation Coefficients tab.

  – Experiment Grouping.

– Legend.

Figure 12.10 illustrates the tiled windows for QC on samples.

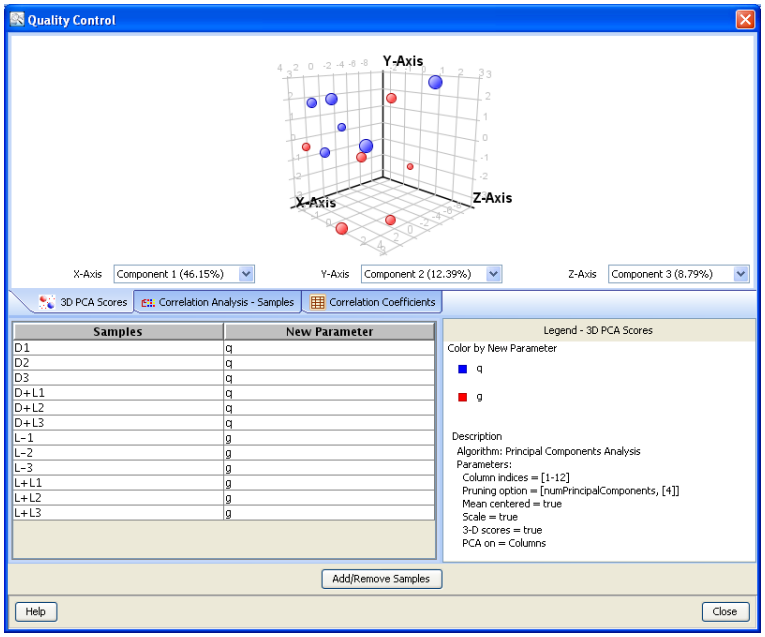


Figure 12.10: Quality Control Dialog

In the **Principal Component Analysis (PCA)** tab, PCA scores are visually represented in a 3D scatter plot. These scores are used to check data quality. It shows one point per array and is colored by the *Experiment Factors* provided earlier in the *Experiment Groupings* view. This allows viewing of separations between groups of replicates. Ideally, replicates within a group should cluster together and separately from arrays in other groups. The PCA components, represented in the X, Y, and Z axes are numbered 1, 2, 3... according to their decreasing significance. The 3D PCA scores plot can be customized via ***Right-Click*⟶*Properties***. To zoom into a 3D Scatter plot, press the Shift key and simultaneously hold down the left mouse button and move the mouse upwards. To zoom out, move the mouse downwards instead. To rotate, press the Ctrl key, simultaneously hold down the left mouse button and move the mouse around the plot.

The **Correlation Analysis - Samples** tab shows the correlation analysis heatmap across arrays (samples) (See How to Perform Correlation Analysis on Samples on page 750). The correlation coefficient for each pair of arrays are displayed in the **Correlation Coefficients** tab. Correlation coefficients are calculated using Pearson Similarity Metric (see Pearson on page 799).

Pearson Similarity Metric: Calculates the mean of all elements in vector **a**. Then it subtracts that value from each element in **a** and calls the resulting vector **A**. It does the same for **b** to make a vector **B**. Correlation Coefficient is then computed as **A.B**/(—**A**——**B**—).

In a Correlation Analysis heatmap:

– Sample Order- In a correlation heatmap, the initial order of samples is the same as the input sample order at the time of experiment creation.

– Clustering - Perform clustering analysis on the correlation heatmap using the **Run clustering** icon from the correlation analysis heatmap toolbar (See To run clustering

on a correlation heatmap on page 760). Samples are clustered based on correlation coefficients using hierarchical clustering algorithm and plotted in the order they cluster together. The linkage rule used is average and the distance metrics used is Euclidean.

For more information on the available toolbar icons in a correlation analysis heatmap, see Correlation Heatmap View Toolbar on page 776. See Properties on 792 to change the properties of the correlation analysis heatmap.

NOTE Correlation coefficients are computed on raw, non-normalized data and in linear scale.

Experiment Grouping shows the parameters and parameter values for each sample.

The third window shows the legend of the active QC tab.

Unsatisfactory samples or those that have not passed the QC criteria can be removed from further analysis, at this stage, using *Add/Remove Samples* button. Once a few samples are removed, re-normalization and baseline transformation of the remaining samples is carried out again. The samples removed earlier can also be added back. Click on **OK** to proceed.

- **Filter Probe Set by Expression:** Entities are filtered based on their signal intensity values. For details refer to the section on Filter Probesets by Expression
- **Filter Probe Set by Flags:** In this step, the entities are filtered based on their flag values, the P(present), M(marginal) and A(absent). Users can set what proportion of conditions must meet a certain threshold. The flag values that are defined at the creation of the new experiment (Step 2 of 3) are taken into consideration while filtering the entities. The filtration is done in 4 steps:

  1. Step 1 of 4 : *Entity list and interpretation* window opens up. Select an entity list by clicking on *Choose Entity List* button. Likewise by clicking on *Choose Interpretation* button, select the required interpretation from the navigator window.

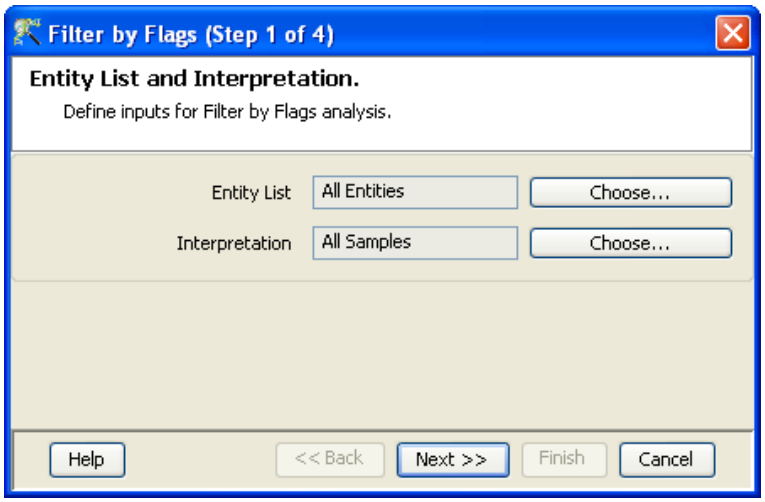


Figure 12.11: Entity List and Interpretation

2. Step 2 of 4: This step is used to set the Filtering criteria and the stringency of the filter. Select the flag values that an entity must satisfy to pass the filter. By default, the Present and Marginal flags are selected. Stringency of the filter can be set in *Retain Entities* box.

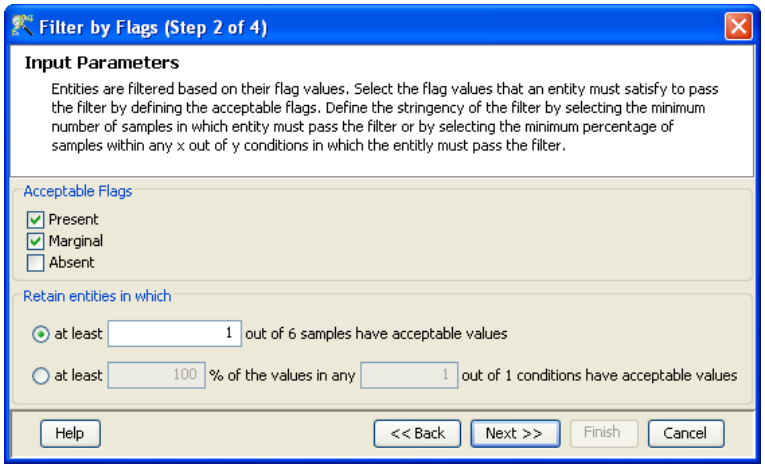


Figure 12.12: Input Parameters

3. Step 3 of 4: A spreadsheet and a profile plot appear as 2 tabs, displaying those probes which have passed the filter conditions. Baseline transformed data is shown here. Total number of probes and number of probes passing the filter are displayed on the top of the navigator window (See Figure 12.13).

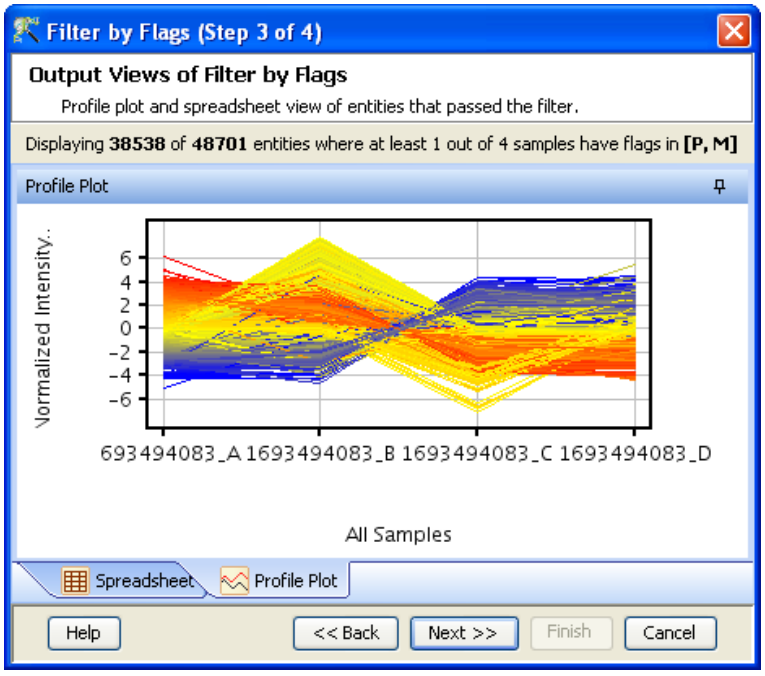


Figure 12.13: Output Views of Filter by Flags

4. Step 4 of 4: Click *Next* to annotate and save the entity list (See Figure 12.14).

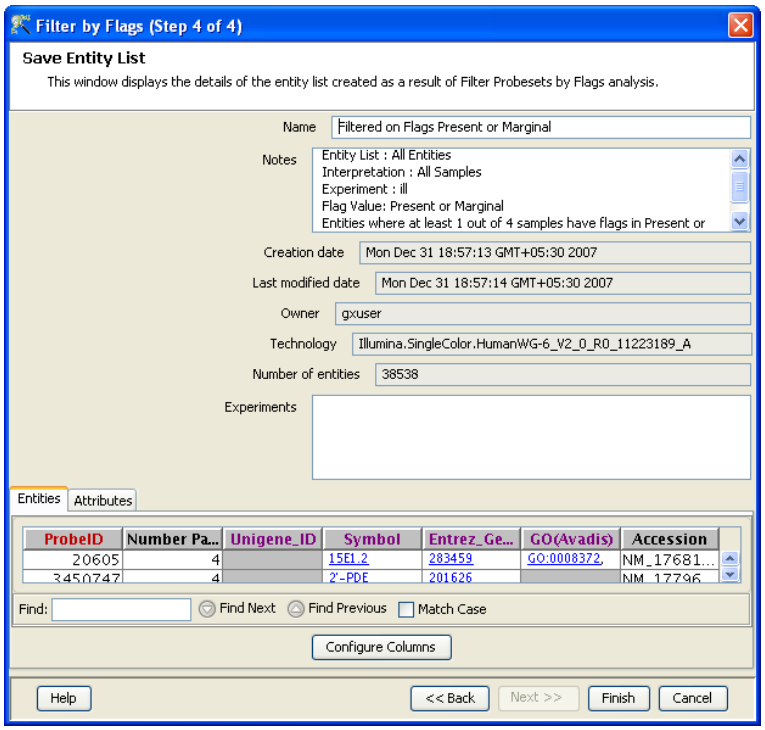


Figure 12.14: Save Entity List

- **Filter Probesets on Data Files:** Entities can be filtered based on values in a specific column of the original data files. For details refer to the section on Filter Probesets on Data Files
- **Filter Probesets by Error:** Entities can be filtered based on the standard deviation or coefficient of variation using this option. For details refer to the section on Filter Probesets by Error

12.4.3 Analysis

- **Statistical Analysis**

  For details refer to section Statistical Analysis in the Data Import Wizard section.

- **Filter on Volcano Plot**

  For details refer to section Filter on Volcano Plot.

- **Fold Change**

  For details refer to section Fold change.

- **Clustering**

  For details refer to chapter Clustering Analysis.

- **Correlation**

  For details refer to chapter Correlation Analysis.

- **Find Similar Entities**
  For details refer to section Find similar entities.
- **Filter on Parameters**
  For details refer to section Filter on Parameters.
- **Principal Component Analysis**
  For details refer to section PCA.

12.4.4 Class Prediction

- **Build Prediction Model** For details refer to section Build Prediction Model.
- **Run Prediction** For details refer to section Run Prediction.

12.4.5 Results

- **Gene Ontology (GO) analysis**
  For more details see GO Analysis.
- **Gene Set Enrichment Analysis (GSEA)**
  For more details see Performing GSEA in **GeneSpring**.
- **Gene Set Analysis (GSA)**
  For more details see Gene Set Analysis.
- **Find Similar Entity Lists**
  This feature is discussed in a separate section called Find Similar Entity Lists.
- **Single Experiment Analysis or Multi-Omic Analysis**
  This feature is discussed in detail in the section Single Experiment and Multi-Omic Analysis.
- **Export to IPA**
  This feature is discussed in detail in the section Ingenuity Pathways Analysis (IPA) Connector.
- **Import IPA Entity List**
  This feature is discussed in detail in the section Import IPA Entity List.
- **NLP Network Discovery**
  This feature is discussed in detail in the section NLP Network Discovery.
- **MeSH Network Builder**
  This feature is discussed in detail in the section MeSH Network Builder.
- **Extract Relations via NLP**
  This feature is discussed in detail in the section Extract Relations via NLP.

12.4.6 Utilities

- **Import Entity list from File**

  For details refer to section Import Entity list from File.

- **Analysis: Biological Significance**

  For details refer to section Analysis: Biological Significance.

- **Filter On Entity List**

  For further details refer to section Filter On Entity List.

- **Remove Entities with missing signal values**

  For details refer to section Remove Entities with missing signal values.

12.4.7 Illumina Custom Technology creation

The number of standard technologies available for Illumina in **GeneSpring** can be obtained from *Annotations* $\longrightarrow$ *Create Technology* $\longrightarrow$ *From Agilent Server*. Illumina projects can also be analyzed by creating a custom technology (*Annotations*$\longrightarrow$ *Create Technology*$\longrightarrow$ *Custom from file*) and then using the Generic Single Color workflow. This is done in either of the 2 cases:

- If you have projects created using Illumina technologies, which are not supported by the Illumina Single Color Importer.
- If you need additional annotation columns (over and above which comes when you do a GeneSpring Format export) from GenomeStudio like the probe sequence, probe coordinates etc..

To create a Custom Technology using an Illumina GenomeStudio project, follow the steps outlined below:

- Create a project in GenomeStudio using either the .xml or the .bgx Content Descriptor file.
- Once a project is created, four spreadsheets-the Sample Probe profile, Sample Gene profile, Group Probe profile, and Group Gene profile are generated. These contain the Intensity values and some annotations. More annotations can be brought into these by going to Column Chooser in GenomeStudio. This allows you to either show or hide additional columns. This file, either with or without additional annotations can be exported out as a text file.
- The file can be exported either in GeneSpring format by going to *File*$\longrightarrow$ *Export in GeneSpring Format* or in a tab delimited text format by clicking on the Export Displayed Data to File icon in GenomeStudio.
- These text files can then be imported into **GeneSpring** to create a Generic Single Color experiment. For details on creating a Generic Single Color experiment, refer to the Chapter 13 on Creating Technology. In the process of technology creation, use the data file containing the annotations as both the data and annotation file.
- Proceed with the rest of the steps as usual.

Chapter 13

Analyzing Generic Single Color Expression Data

GeneSpring supports Generic Single Color technology. Any custom array with single color technology can be analyzed here. However, a technology first needs to be created, based upon the file format being imported.

13.1 Creating Technology

Technology creation is a step common to both Generic Single Color and Two color experiments. Technology creation enables the user to specify the columns (Signals, Flags, Annotations etc.) in the data file and their configurations which are to be imported. Different technologies need to be created for different file formats. Custom technology can be created by navigating to *Annotations* in the menu bar and selecting *Create Technology* $\longrightarrow$ *Custom from file*. The process uses one data file as a sample file to mark the columns. Therefore, it is important that all the data files being used to create an experiment should have identical formats.

The *Create Custom Technology* wizard has multiple steps. While steps 1, 2, 3 and 9 are common to both the Single color and Two Color, the remaining steps are specific to either of the two technologies.

- (Step 1 of 9)

  User input details, i.e., Technology type, Technology name, Organism, Sample data file location, Number of samples in a single data file and the Annotation file location are specified here. Files with a single sample or with multiple samples can be used to create the technology. Click *Next*. See Figure 13.1.

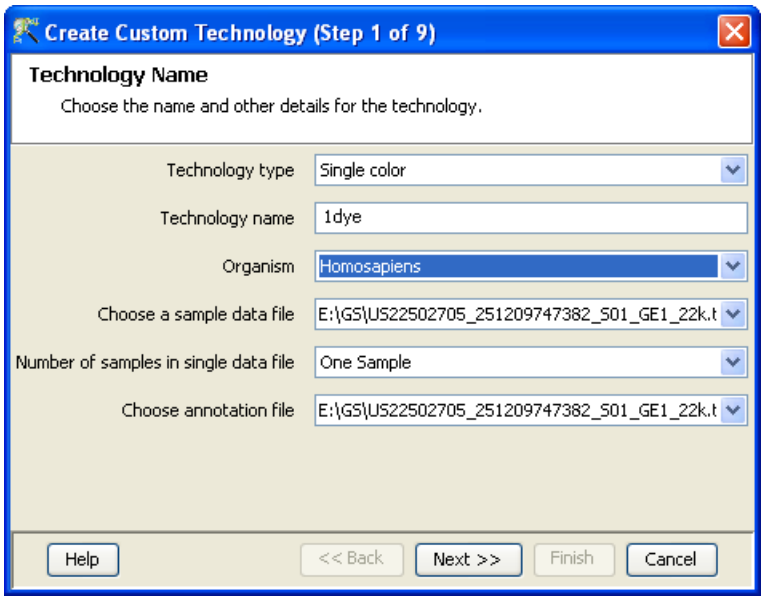


Figure 13.1: Technology Name

- (Step 2 of 9) This allows the user to specify the data file format. For this operation, four options are provided, namely, the *Separator*, the *Text qualifier*, the *Missing Value Indicator* and the *Comment Indicator*. The *Separator* option specifies if the fields in the file to be imported are separated by tab, comma, space etc. New separators can be defined by scrolling down to *'Enter New'* and providing the appropriate symbol in the textbox. *Text qualifier* is used for indicating characters used to delineate full text strings. This is typically a single or double quote character. The *Missing Value Indicator* is for declaring a string that is used whenever a value is missing. This applies only to cases where the value is represented explicitly by a symbol such as N/A or NA. The *Comment Indicator* specifies a symbol or string that indicates a comment section in the input file. Comment Indicators are markers at the beginning of the line which indicate that the line should be skipped (typical examples is the # symbol). See Figure 13.2.

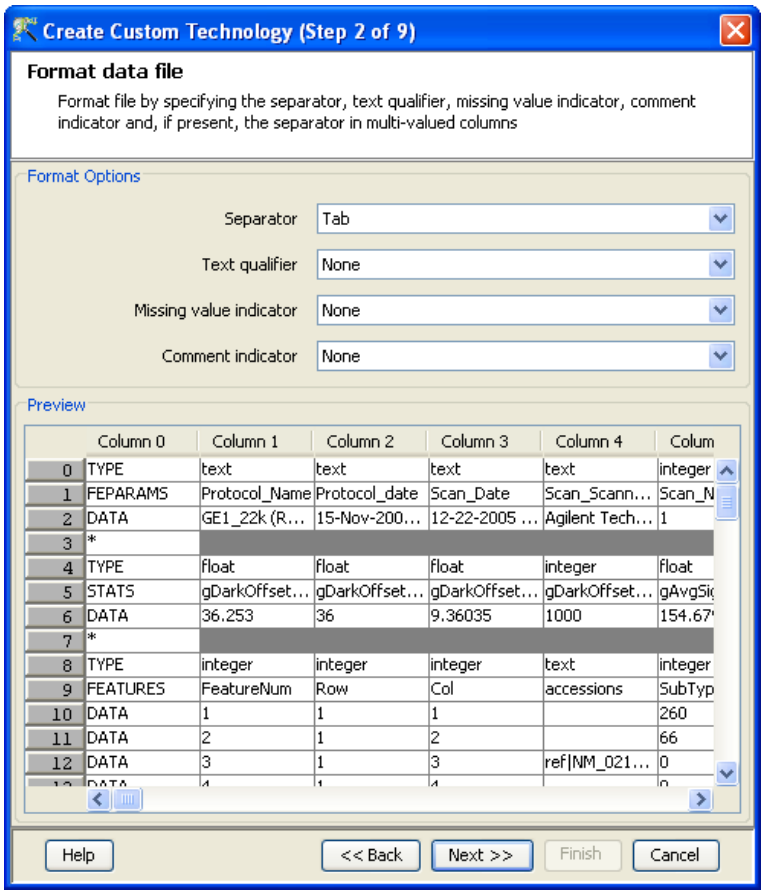


Figure 13.2: Format data file

- (Step 3 of 9) The data files typically contain headers which are descriptive of the chip type and are not needed for the analysis. Only those rows containing the data values are required. The purpose of this step is to identify which rows need to be imported. The rows to be imported must be contiguous in the file. The rules defined for importing rows from this file will then apply to all other files to be imported using this technology. Three options are provided for selecting rows: The default option is to select all rows in the file. Alternatively, one can choose to take a block of rows between specific row numbers (use the preview window to identify row numbers) by entering the row numbers in the appropriate textboxes. Remember to press the Enter key before proceeding. In addition, for situations where the data of interest lies between specific text markers, those text markers can be indicated. Note also that instead of choosing one of the options from the radio buttons, one can choose to select specific contiguous rows from the preview window itself by using Left-Click and Shift-Left-Click on the row header. The panel at the bottom should be used to indicate whether or not there is a header row; in the latter case, dummy column names will be assigned. See Figure 13.3.

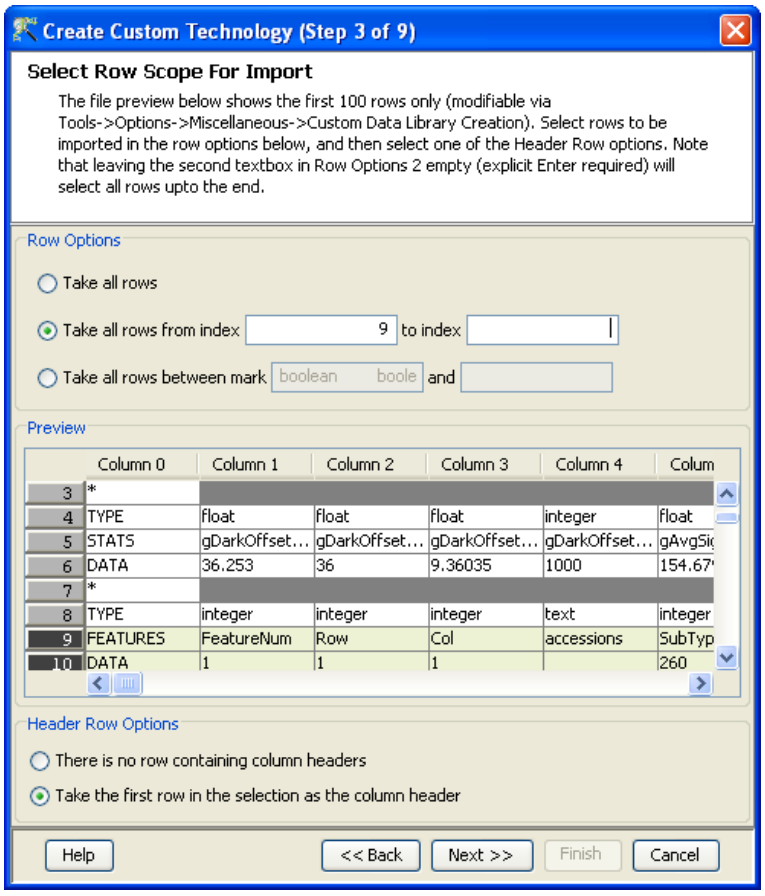


Figure 13.3: Select Row Scope for Import

- (Step 4 of 9) This step is specific for file formats which contain a single sample per file. Identifier, background(BG) corrected signal and the flag columns are indicated here. Please note that the identifier column should not be numeric. Flag column can be configured using the *Configure* button to designate Present(P), Absent(A) or Marginal(M) values. See Figure 13.4.

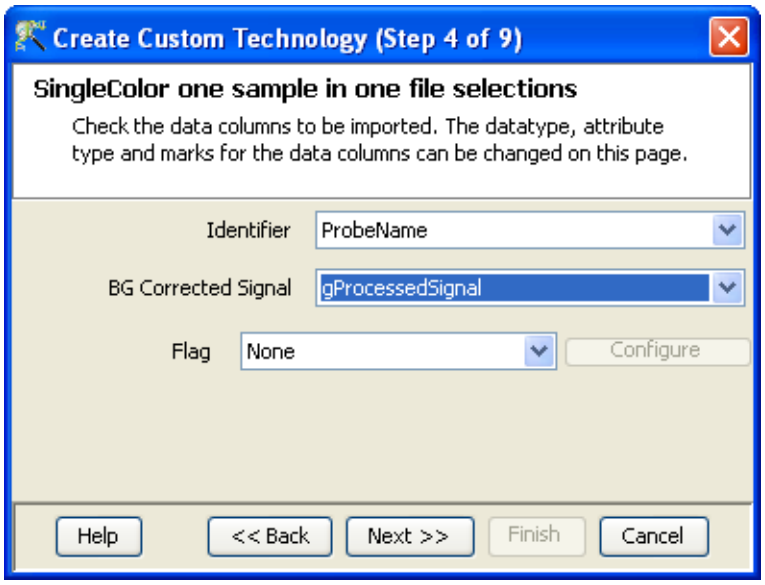


Figure 13.4: Single Color one sample in one file selections

- (Step 5 of 9)

  This step is specific for file formats which contain multiple samples per file. Such file formats typically contain a single column having the identifier and multiple columns representing the samples (one data column per sample). Please note that the samples in the data file should not have special characters such as `! * # ; ? \ / : " < > |` and space in their names. In this step, the Identifier column has to be indicated (Identifier column should not be numeric). The signal and flag columns for each sample also should be identified here and moved from *All columns* to *Signal columns* and *Flag columns* box respectively. This can be done either by putting in the *Keyword* for the Signal and the Flag columns and clicking *Refresh* or by choosing *Custom* and selecting the columns as either Signal or Flag in the *Identify signal and flag columns by* option. After marking 2 columns, the user can utilize the option *Guess the Rest* for marking the other columns. The tool tries to match the names of the selected columns with the rest and marks those columns which have similar names to the selected ones. The *Choose representative flag* allows the user to choose one of the flag columns for configuring the flag settings. See Figures 13.5 and 13.6.

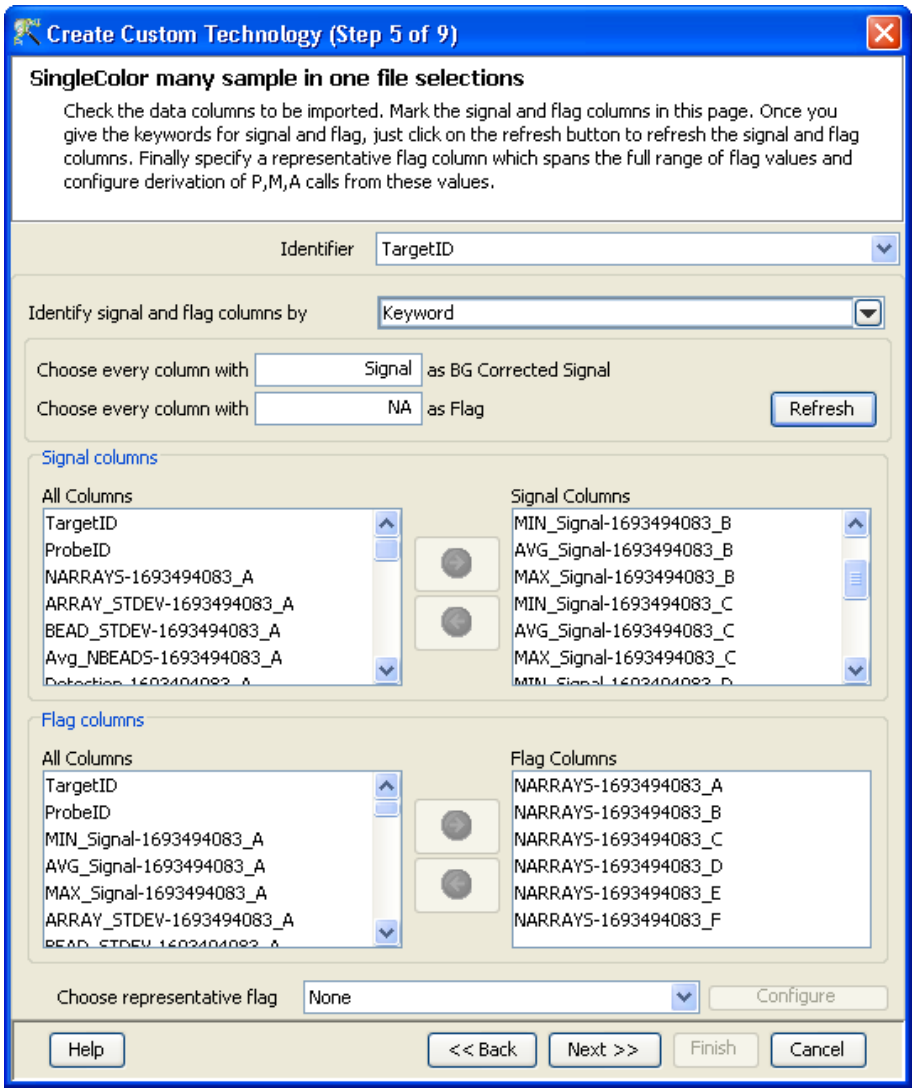


Figure 13.5: Single Color-Multiple Samples Per File-Keyword Selection

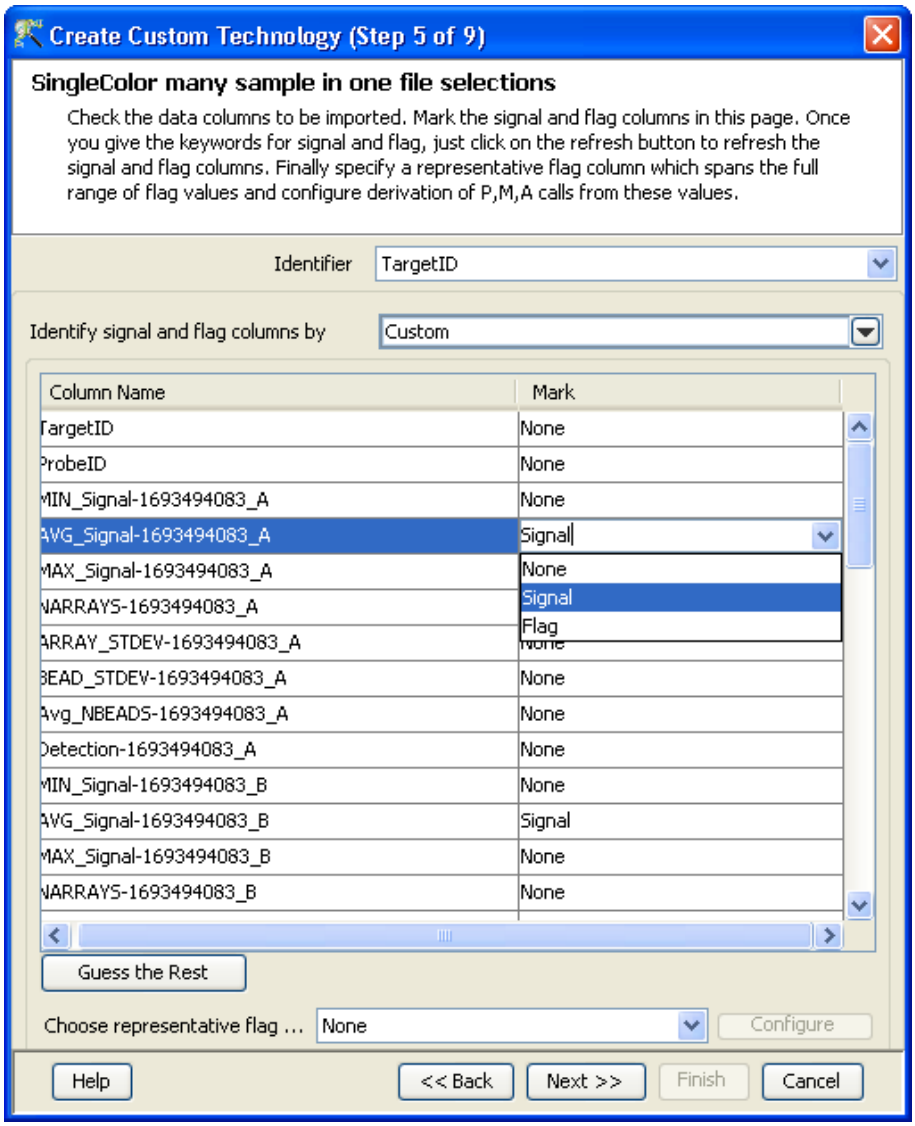


Figure 13.6: Single Color-Multiple Samples Per File-Custom Selection

- (Steps 6 of 9)

  This step of the wizard is used in case of technology creation for 2-dye or 2-color samples.

- (Steps 7 of 9)

  This step is similar to the step 2 of 9 and is used to format the annotation file. If a separate annotation file does not exist, then the same data file can be used as an annotation file, provided it has the annotation columns.

- (Step 8 of 9) Identical to step 3 of 9, this allows the user to select row scope for import in the annotation file.

- (Step 9 of 9) The Step 9 of technology creation is an extremely important step which allows the user to mark the columns appropriately. Proper marking of the various columns in the annotation file will enable the various functionalities like GO, GSEA, Genome Browser, Pathway Analysis to proceed smoothly. The markings to be given for all these functions are elaborated below:

- GSEA: The annotation file should contain a column containing the Gene Symbol. This column should be marked as *Gene Symbol* from the drop-down menu.
- GSA: The annotation file should contain a column containing the Gene Symbol. This column should be marked as *Gene Symbol* from the drop-down menu.
- GO: For carrying out GO analysis, the annotation file can either contain a single column with all the GO IDs in it, separated by a separator or it can contain separate columns for the different GO processes. The single column with multiple GO IDs should be marked as *Gene Ontology accession* from the drop-down menu. Instead if columns containing individual GO processes(Biological Process, Cellular Component and Molecular Function) are present, they should be marked accordingly in the dropdown menu.
- Genome Browser: In order to view the data in Genome Browser, the annotation file should contain a Chromosome Start Index, Chromosome End Index, Strand and Chromosome Number columns. Provide the column mark for Chromosome Start index, Chromosome End index, Strand, Chromosome number respectively, from the drop-down menu.

  Note: The Chromosome Start index < Chromosome End index. For viewing Profile track only, in the Genome Browser, chromosome start index and chromosome number are needed. The labelling of the chromosome numbers should follow this convention-chr1, chr2 .....i.e. the word starts with chr followed by the chromosome number (without any space). For viewing data track, all four Chromosome Start index, Chromosome End index, Strand, Chromosome number are needed.

  - If a custom technology is being created using an **Illumina** data and annotation file, then for the Genome Browser functionality, the column markings have to be handled as follows:

    For viewing using the Genome Browser, the annotation files has three columns which have values for all four (Chromosome Start Index, Chromosome End Index and Chromosome Number and Strand). Therefore before creating the custom experiment the user needs to parse these columns and create three new columns as follows :

    Probe_Chr_Orientation-- This column can be taken as it is. It should be marked as Strand.

    Chromosome -- A new column must be created wherein a 'chr' should be appended to each entry in the Chromosome column and this new column should be marked as Chromosome Number.

    Probe_Coordinates-- This column has each entry in the format a-b where a ¡ b. Two new columns need to be created. one which has only the a values, (it should be marked as Chromosome Start Index) one which has only the b values (it should be marked as Chromosome End Index).
  - If a custom technology is being created using an **Agilent** data and annotation file, then for the Genome Browser functionality, the column markings have to be handled as follows:

    The annotation files have a single column 'Map' which has values for all four Chromosome Start Index, Chromosome End Index and Chromosome Number and Strand. Therefore before creating the custom experiment the user needs to parse the file and separate the four columns as Chromosome Start Index, Chromosome End Index Chromosome Number and Strand.

    Each entry in the Map column is typically in the format chrQ:a..b

  if a < b, the corresponding Chromosome Number is chrQ; the corresponding Chromosome Start Index is a; the corresponding Chromosome End Index is b; the corresponding Strand is + .

if a > b the corresponding Chromosome Number is chrQ; the corresponding Chromosome Start Index is b; the corresponding Chromosome End Index is a; the corresponding Strand is - .

For example, a Map value of chr14:34101457..34101398 corresponds to a Chromosome Start Index of 34101398, a Chromosome End Index of 34101457, a Chromosome Number of chr14 and a Strand of - (because in chrX:a..b a>b)

For example, a Map value of chr6:46222041..46222100 corresponds to a Chromosome Start Index of 46222041, a Chromosome End Index of 46222100, a Chromosome Number of chr6 and a Strand of +(because in chrX:a..b a<b)

- Import BioPAX pathways: Pathways being imported should be in .owl format. During custom technology creation, provide the column mark for Entrez Gene ID/SwissProt from the drop-down menu. Only after this mark is provided can the proteins involved in a particular pathway be highlighted.
- Pathway Analysis: The annotation file should contain any of the annotations listed under transcriptomics in Table 31.1.
- Translation: This operation can be performed between organisms listed in the **Homologene table** in section Translation. Entrez Gene ID column has to be marked for performing translation.

See Figure 13.7.

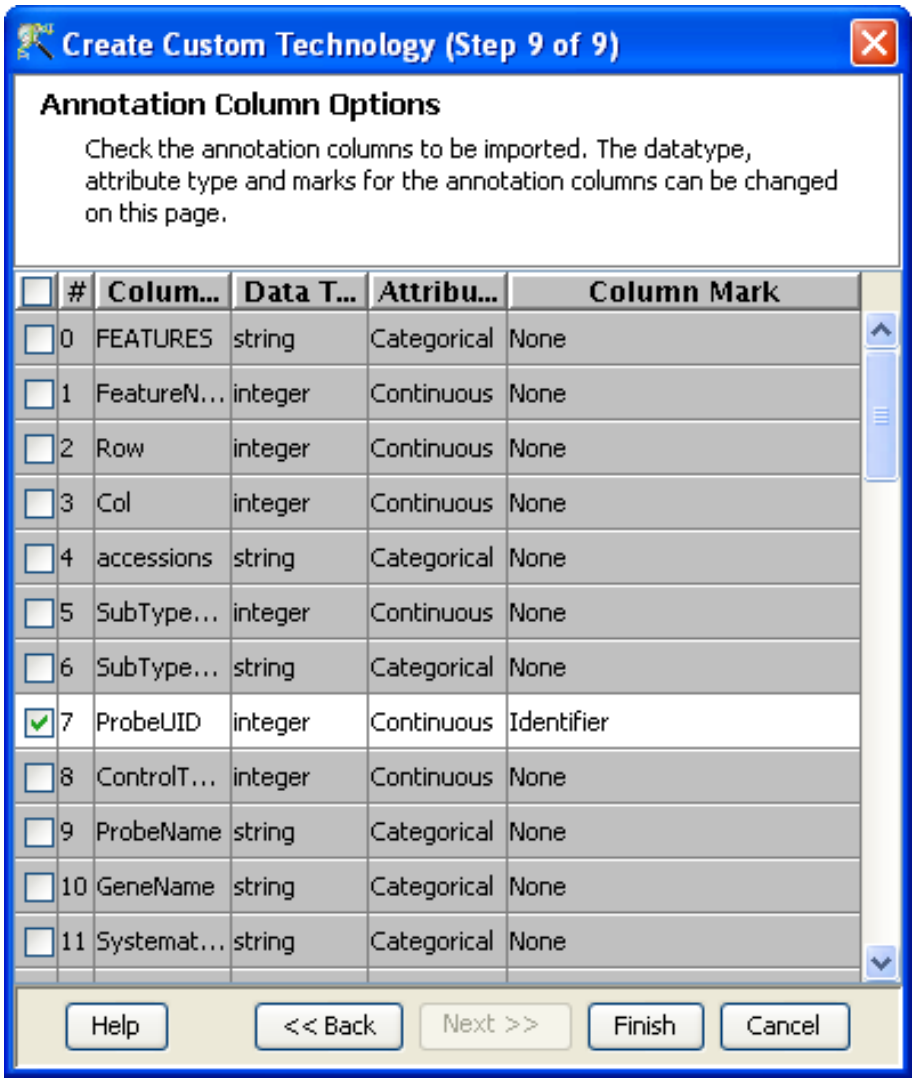


Figure 13.7: Annotation Column Options

The types of Data and Attribute marks available for the annotation columns are

- **Categorical:** A column marked as a ‘‘categorical” column means that the values in this column fall into certain finite distinct categories.
- **Continuous:** A column marked as a ‘‘continuous” column means that the values in this column can vary, potentially, over any large range.
- **String:** A continuous sequence of symbols or digits, not including a space.
- **Float:** A real number, i.e a number which can be given by a decimal representation.

Please note that the Entrez ID annotation column needs to be first specified as categorical and then as string in the dropdown.

The annotation marks are colored on the basis of their functionality in the tool. The meaning of the various colors are illustrated in the Figure 14.6.

Note:This figure is provided only for illustration purpose and is not available in the tool.

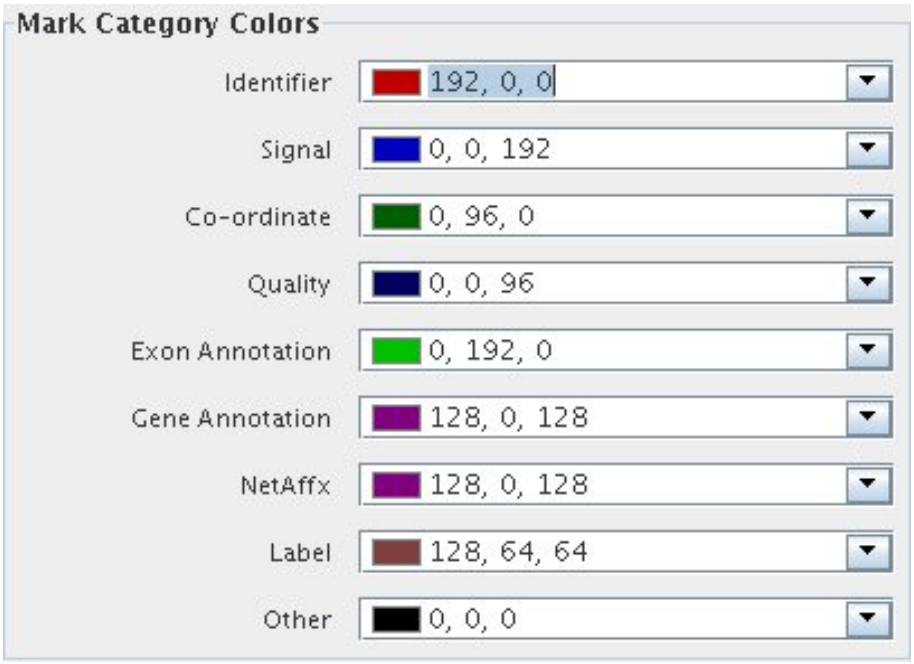


Figure 13.8: Annotation Mark Colors

Click ***Finish*** to exit the wizard.

13.1.1 Project and Experiment Creation

After technology creation, data files satisfying the file format can be used to create an experiment. The following steps will guide you through the process of experiment creation.

Upon launching **GeneSpring**, the startup is displayed with 3 options.

1. **Create new project**
2. **Open existing project**
3. **Open recent project.**

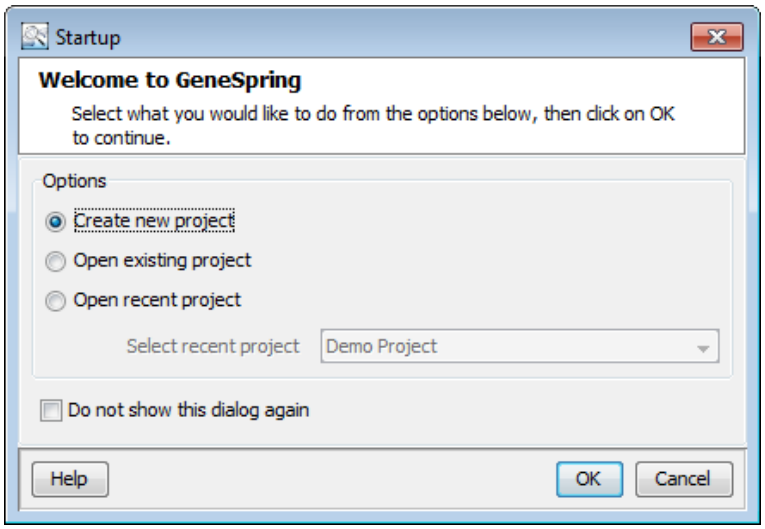


Figure 13.9: Welcome Screen

Either a new project can be created or else a previously generated project can be opened and re-analyzed. On selecting *Create New Project*, a window appears in which details (name of the project and notes) can be recorded. Press *OK* to proceed.

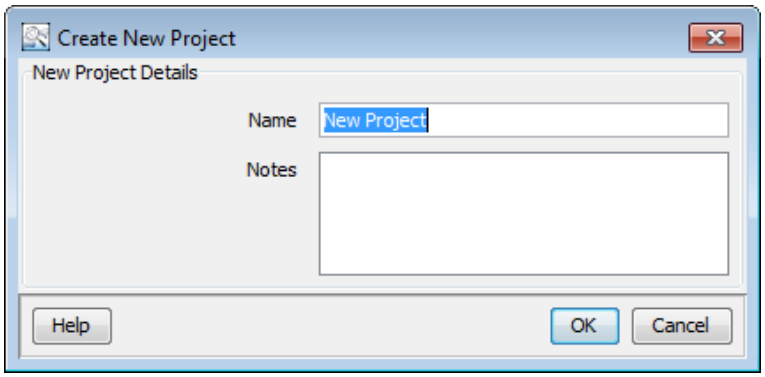


Figure 13.10: Create New Project

An Experiment Selection Dialog window then appears with two options.

1. **Create new experiment**
2. **Open existing experiment**

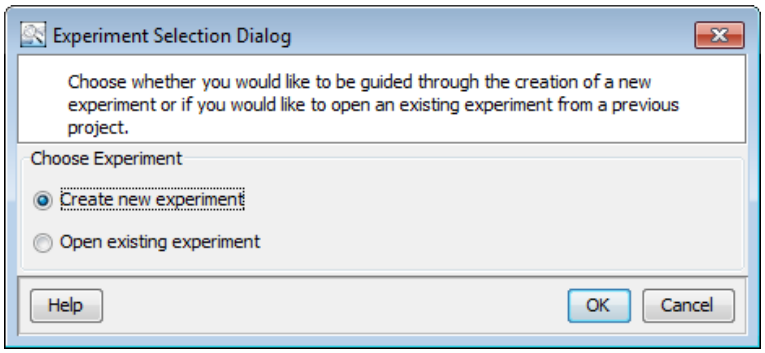


Figure 13.11: Experiment Selection

Selecting *Create new experiment* allows the user to create a new experiment (steps described below). *Open existing experiment* allows the user to use existing experiments from any previous projects in the current project. Choosing *Create new experiment* opens up a New Experiment dialog in which *Experiment name* can be assigned. Once the *Analysis Type* is selected (Expression), the *Experiment type* should then be specified (Generic Single Color), using the drop down button. The *Workflow Type* can be used to choose whether the workflow will be *Analysis: Biological Significance* or *Data Import Wizard*. Unlike the other technologies where *Analysis: Biological Significance* and *Data Import Wizard* analysis workflows are available, in case of Generic Single Color, only the *Data Import Wizard Workflow* is supported . Click *OK* will open a new experiment wizard. See Figure 13.12

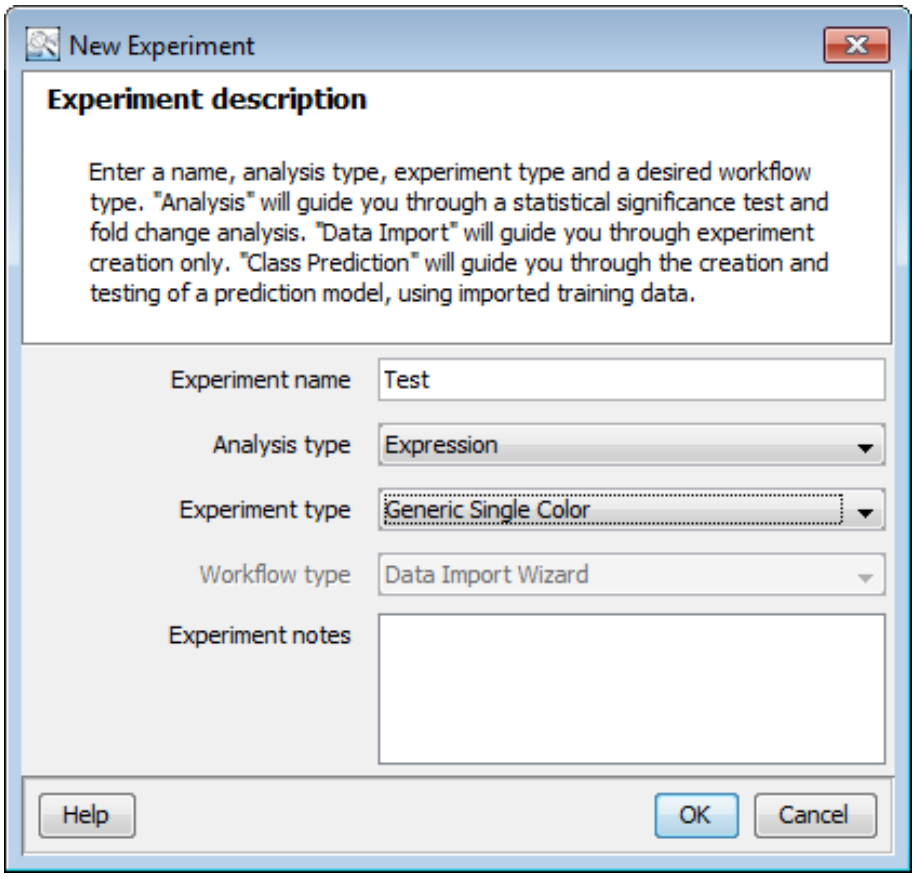


Figure 13.12: Experiment Description

13.2 Data Processing for Generic Single Color Experiment

1. **File formats:** The files should be tabular in nature. For example, .csv, .tsv, .txt etc. can be used.
2. **Raw:** The term ''raw'' signal values refer to the linear data after thresholding and summarization. Summarization is performed by computing the geometric mean.
3. **Normalized:**''Normalized'' value is the value generated after log transformation and normalization (Percentile Shift, Scale, Normalize to control genes or Quantile) and Baseline Transformation.
4. **Treatment of on-chip replicates:** The signal value of a probeset is the geometric mean of all its probes.
5. **Flag values:** The values for the probes are configured by the user during the creation of technology as either Present, Marginal or Absent. Based on the values of the probes, the probeset is assigned a flag value. The order of importance for flag values for probes in a probeset is *Present>Marginal>Absent*.
6. **Treatment of Control probes:** The control probes are included while performing normalization.

7. **Empty Cells:** Empty cells might be present in the intensity values column for certain genes in the data file. These genes are brought in **GeneSpring** . But an entity list containing these genes cannot be used for running Clustering and Class Prediction analyses. The user can choose to remove the missing values from an entity list using the option ***Remove Entities with missing signal values*** from the ***Results Interpretations*** section of the workflow.

8. **Sequence of events:** The sequence of events involved in the processing of the data files is: thresholding, summarization, log transformation and Normalization followed by Baseline Transformation.

9. **Merging of files:** Multiple files in Generic experiment creation are combined based on the Identifier column using the following rules. The very first file among the various files chosen server as a master reference (you can determine which file serves as the first file using the *Reorder* button on Page 1 of the New Experiment Creation page). The number of rows in this master must exceed the number of rows in all subsequent files, for extra rows in these subsequent files are dropped. Next, all identifiers in the Idenfier column of this first file are considered and missing values in these, if any, are discarded. This results in a set of valid identifier values; all rows in all other files whose identifier values are outside of this set are discarded. Next, on-chip replicates are determined by counting the number of occurrences of each valid identifier in the first file. Consider for example an identifier Id1 which appears 3 times in file 1. Then rows corresponding to the first 3 occurrences of Id1 are taken in each of the other files; if there are fewer than 3 rows, then as many rows that are present are taken; and if there are more than 3 rows, then the first 3 are taken. The summarized value for Id1 in each file is determined by taking a geometric mean over these chosen rows.

13.3 Data Import Wizard

The **Data Import Wizard** offers a variety of choices to the user for the analysis. Raw signal thresholding can be altered. Based upon the technology, Quantile or Median Shift normalization can be performed. Additionally there are options for baseline transformation of the data and for creating different interpretations. To create and analyze an experiment using the **Data Import Wizard**, choose the **Workflow Type** as Data Import Wizard. Clicking ***OK*** will open a New Experiment Wizard, which then proceeds as follows:

1. New Experiment (Step 1 of 4): The technology (created as mentioned above) can be selected and the new data files or previously used data files in **GeneSpring** can be imported in to create the experiment. A window appears containing the following options:

   (a) **Choose Files(s)**
   (b) **Choose Samples**
   (c) **Choose Raw Files**
   (d) **Reorder**
   (e) **Remove**

   An experiment can be created using either the data files or else using samples. Upon loading data files, **GeneSpring** associates the files with the technology (see below) and creates samples. These samples are stored in the system and can be used to create another experiment via the **Choose Samples** option through a sample search wizard. If the user has imported any custom experiments from **GeneSpring GX 7** and wants to recreate the

experiment in **GeneSpring**, then the user can create a new technology in the tool with an original raw file and later utilize the **Choose Raw Files** option to choose the raw files associated with the migrated custom experiment. For selecting data files and creating an experiment, click on the **Choose File(s)** button, navigate to the appropriate folder and select the files of interest. The files can be either tab separated (.txt or .tsv) or could be comma separated (.csv). Select ***OK*** to proceed.

The sample search wizard that comes up via the option *Choose Samples* has the following search conditions:

(a) **Search field** (which searches using any of the 6 following parameters- (Creation date, Modified date, Name, Owner, Technology, Type).

(b) **Condition** (which requires any of the 4 parameters- (equals, starts with, ends with and includes Search value).

(c) **Value**

Multiple search queries can be executed and combined using either *AND* or *OR*.

Samples obtained from the search wizard can be selected and added to the experiment using *Add* button, similarly can be removed using *Remove* button.

After selecting the files, clicking on the *Reorder* button opens a window in which the particular sample or file can be selected and can be moved either up or down by pressing on the buttons. Click on *OK* to enable the reordering or on *Cancel* to revert to the old order. See Figure 13.13

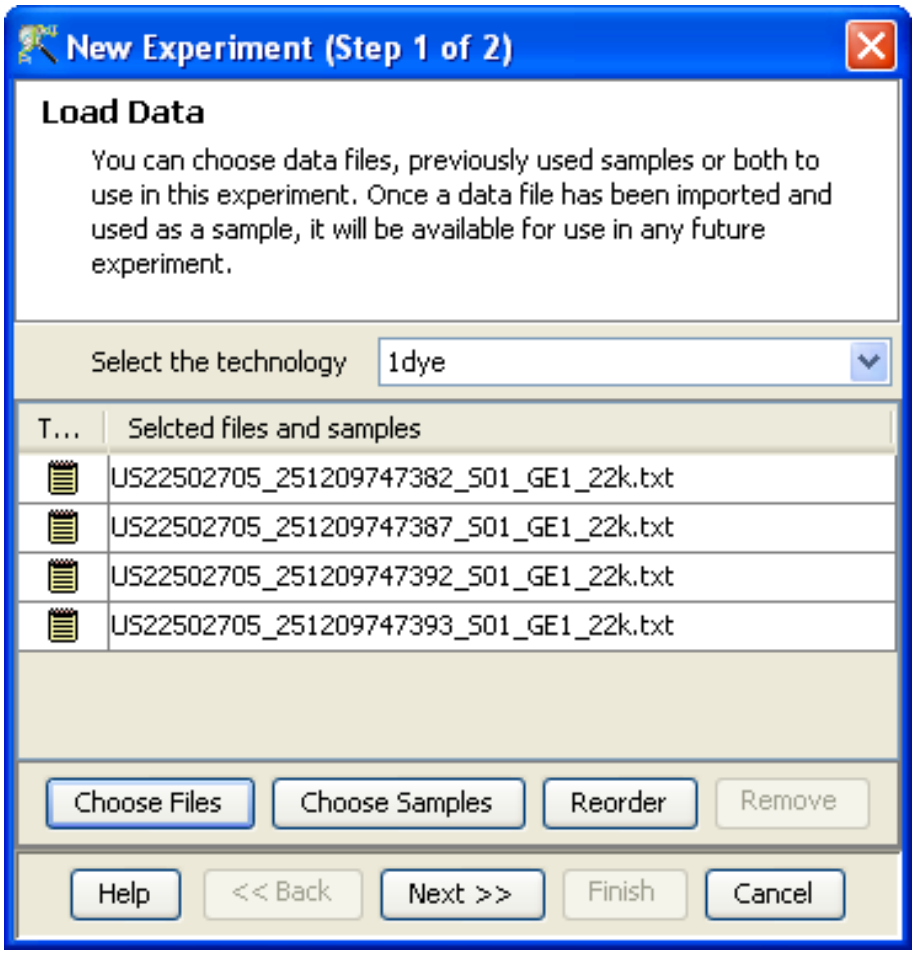


Figure 13.13: Load Data

2. New Experiment (Step 2 of 4): This gives the options for preprocessing of input data. It allows the user to threshold raw signals to chosen values and to select normalization algorithms(Quantile, Percentile Shift, Scale and Normalize to control genes).

   - **None:** No normalization is done. The threshold value for raw signals is set to 1.0 and can be changed by the user and all other options are disabled.
   - **Percentile Shift:** On selecting this normalization method, the **Percentile Target** box gets enabled allowing the user to enter a specific percentile value. The default value is set at 75^th percentile.
   - **Scale:** On selecting this normalization method, the user is presented with an option to scale it to the median of percentile target/mean of means of 'all' or 'control' samples. On choosing the latter, the user has to select the control samples from the available samples in the **Choose Samples** box. The **Percentile Target** box is set at a default value of 75.
   - **Normalize to control genes:** After selecting this option, the user has to specify the control genes in the next wizard. The **Percentile Target** box is disabled.
   - **Normalize to External Value:** This option will bring up a table listing all samples and a default scaling factor of '1.0' against each of them. The user can use the *'Assign Value'* button at the bottom to assign a different scaling factor to each of the sample;

multiple samples can be chosen simultaneously and assigned a value. The **Percentile Target** box is disabled.

For details on the above normalization methods, refer to section Normalization Algorithms.

In case, the data is already log transformed, the user can select the checkbox stating that their signal values are already in log scale. This will disable the thresholding option also.

See Figure 13.14.

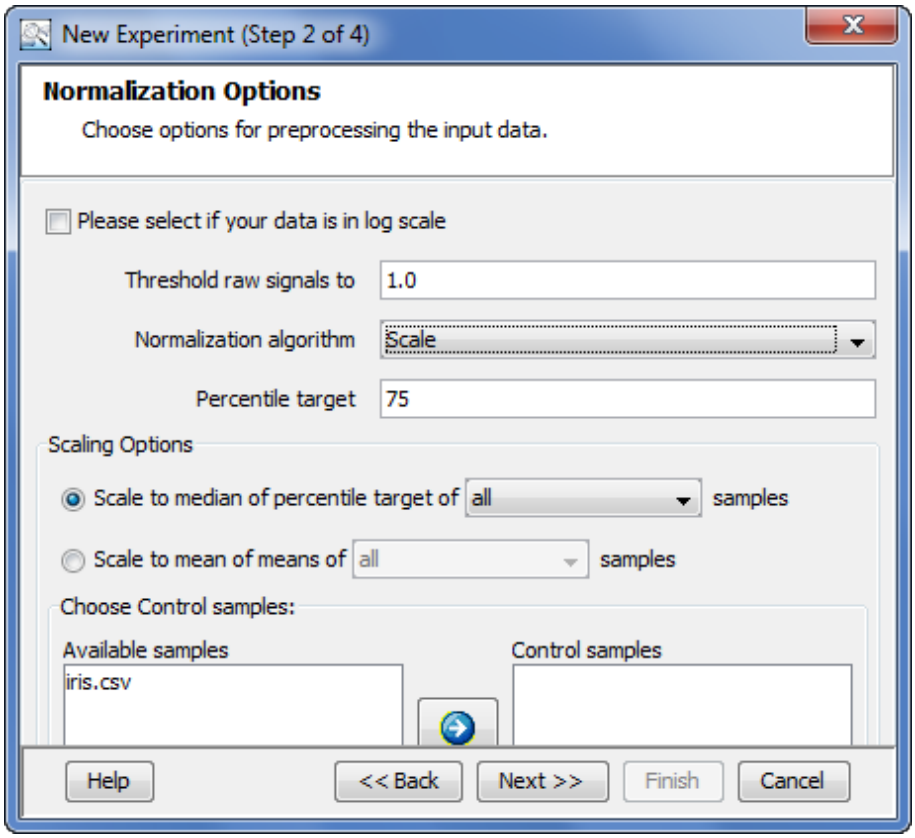


Figure 13.14: Preprocess Options

3. New Experiment (Step 3 of 4):

   If the **Normalize to control genes** option is chosen, then the list of control entities can be specified in the following ways in this wizard:

   - By choosing a file(s) (txt, csv or tsv) which contains the control entities of choice denoted by their probe id. Any other annotation will not be suitable.
   - By searching for a particular entity by using the ***Choose Entities*** option. This leads to a search wizard in which the entities can be selected. All the annotation columns present in the technology are provided and the user can search using terms from any of the columns. The user has to select the entities that he/she wants to use as controls when they appear in the **Output Views** page and then click ***Finish***. This will result in the entities getting selected as control entities and will appear in the wizard.

   You can choose either one or both the options to select control genes. The chosen genes can also be removed after selection is over. See Figure 13.15.

In case the entities chosen are not present in the technology or sample, they will not be taken into account during experiment creation. The entities which are present in the process of experiment creation will appear under matched probe ids whereas the entities not present will appear under unmatched probe ids in the experiment notes in the experiment inspector.

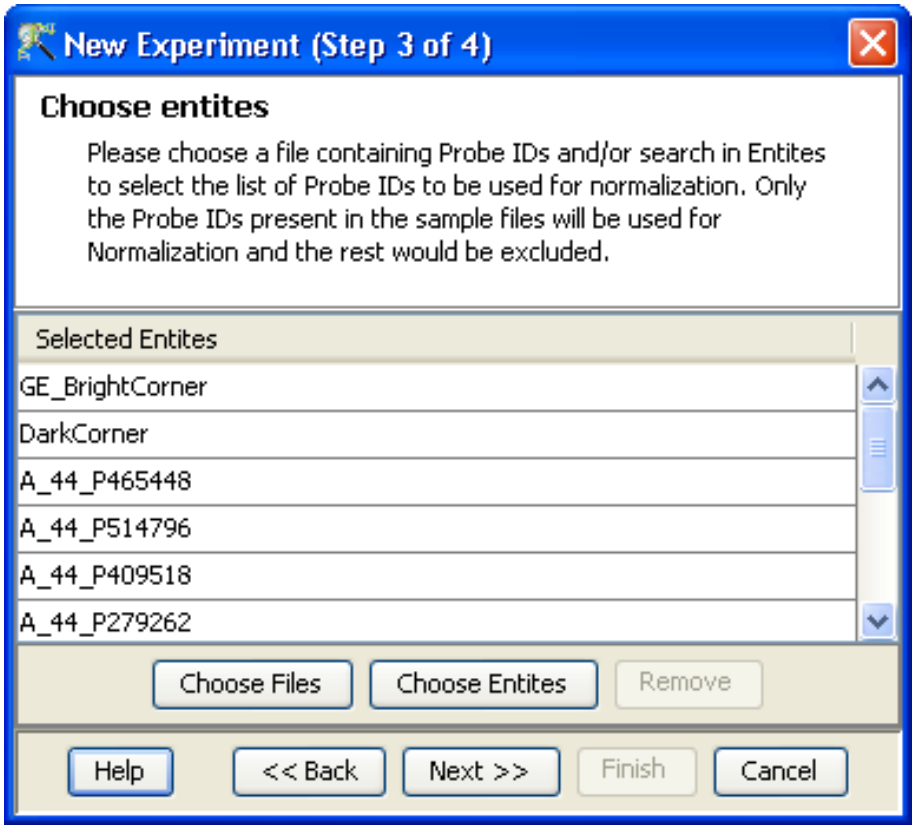


Figure 13.15: Choose Entities

4. New Experiment (Step 4 of 4):

   This step allows the user to perform baseline transformation. See Figure 13.16. The baseline options include

   - ***Do not perform baseline***
   - ***Baseline to median of all samples:*** For each probe the median of the log summarized values from all the samples is calculated and subtracted from each of the samples.
   - ***Baseline to median of control samples:*** For each sample, an individual control or a set of controls can be assigned. Alternatively, a set of samples designated as controls can be used for all samples. For specifying the control for a sample, select the sample and click on ***Assign value***. This opens up the ***Choose Control Samples*** window. The samples designated as Controls should be moved from the *Available Items* box to the *Selected Items* box. Click on ***Ok***. This will show the control samples for each of the samples.

     In *Baseline to median of control samples*, for each probe the median of the log summarized values from the control samples is first computed and then this is subtracted from the sample. If a single sample is chosen as the control sample, then the probe values of the control sample are subtracted from its corresponding sample.

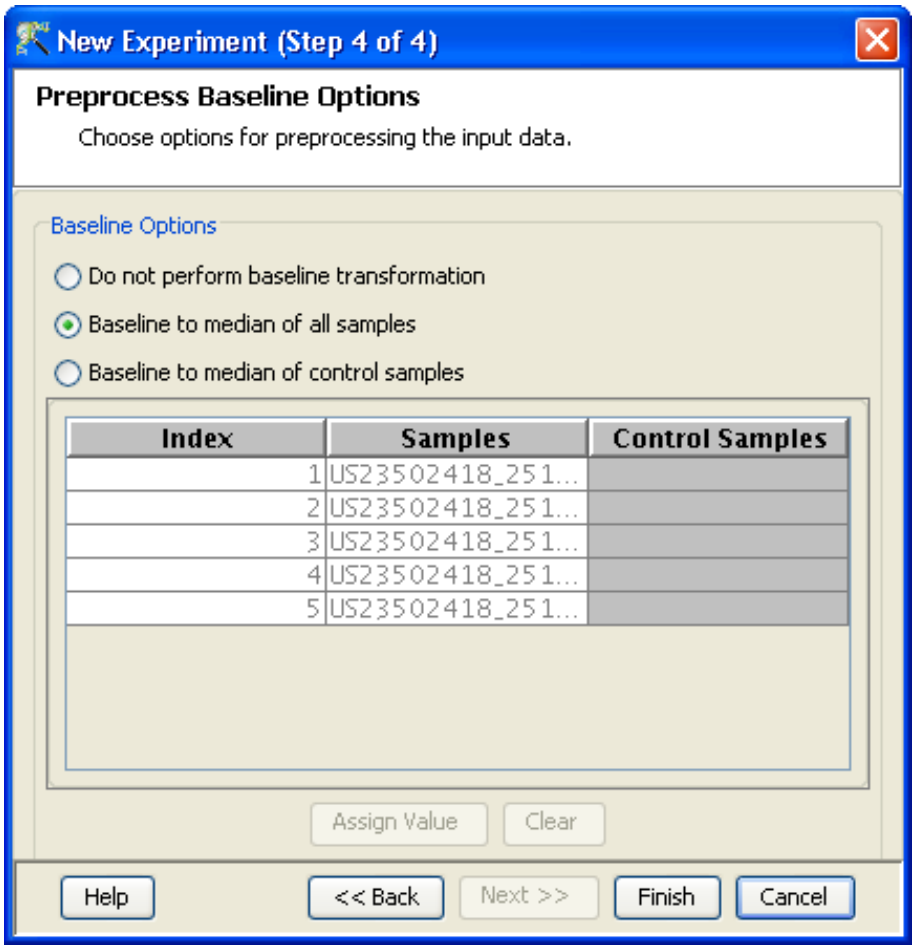


Figure 13.16: Preprocess Baseline Options

Clicking ***Finish*** creates an experiment, which is displayed as a Box Whisker plot in the active view. Alternative views can be chosen for display by navigating to ***View*** in Toolbar.

13.3.1 Experiment Setup

Once an experiment is created, the *Data Import Wizard* steps appear on the right hand side. This gives options to view quick start guide, create experiment grouping and interpretation and the utility to create gene level experiment from probe level experiment. See section Experiment Setup for complete details.

13.3.2 Quality Control

- **Quality Control on Samples:**

  Quality Control (QC) allows you decide which samples are ambiguous and which are passing the quality criteria. Based on the QC results, unreliable you can remove samples from further

downstream analysis. The QC dialog displays three tiled windows:

- 3D PCA scores, Correlation Analysis - Samples, and a Correlation Coefficients tab.
- Experiment Grouping.
- Legend.

Figure 13.17 illustrates the tiled windows for QC on samples.

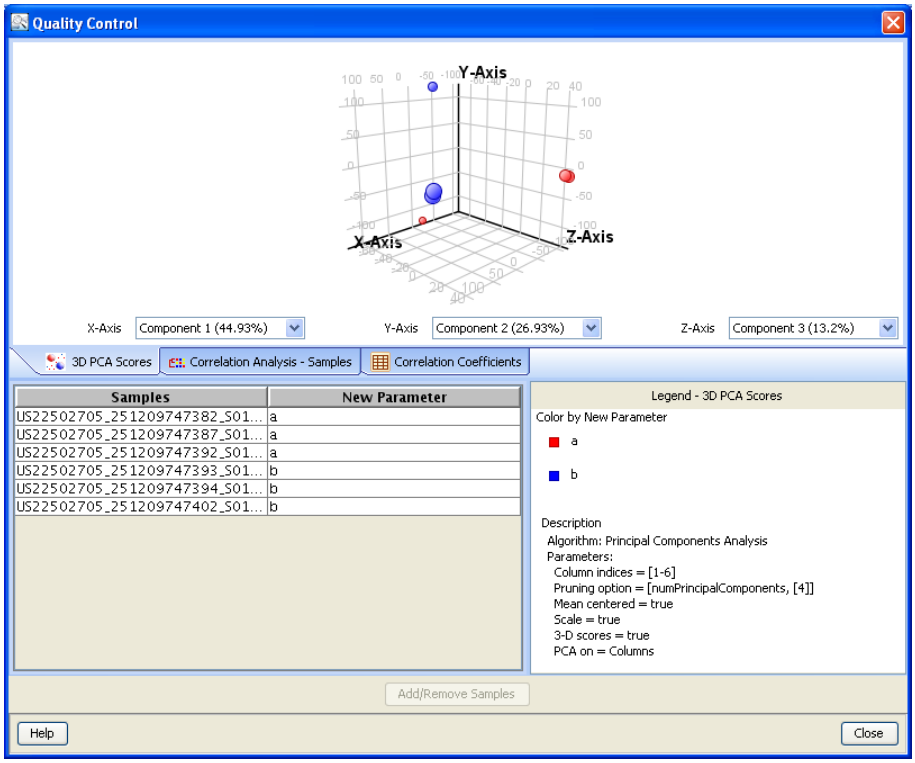


Figure 13.17: Quality Control Dialog

In the **Principal Component Analysis (PCA)** tab, PCA scores are visually represented in a 3D scatter plot. These scores are used to check data quality. It shows one point per array and is colored by the *Experiment Factors* provided earlier in the *Experiment Groupings* view. This allows viewing of separations between groups of replicates. Ideally, replicates within a group should cluster together and separately from arrays in other groups. The PCA components, represented in the X, Y, and Z axes are numbered 1, 2, 3... according to their decreasing significance. The 3D PCA scores plot can be customized via ***Right-Click*** $\longrightarrow$ ***Properties***. To zoom into a 3D Scatter plot, press the Shift key and simultaneously hold down the left mouse button and move the mouse upwards. To zoom out, move the mouse downwards instead. To rotate, press the Ctrl key, simultaneously hold down the left mouse button and move the mouse around the plot.

The **Correlation Analysis - Samples** tab shows the correlation analysis heatmap across arrays (samples) (See How to Perform Correlation Analysis on Samples on page 750). The correlation coefficient for each pair of arrays are displayed in the **Correlation Coefficients** tab. Correlation coefficients are calculated using Pearson Similarity Metric (see Pearson on page 799).

Pearson Similarity Metric: Calculates the mean of all elements in vector **a**. Then it subtracts that value from each element in **a** and calls the resulting vector **A**. It does the same for **b** to make a vector **B**. Correlation Coefficient is then computed as **A.B**/(—**A**——**B**—).

In a Correlation Analysis heatmap:

- Sample Order- In a correlation heatmap, the initial order of samples is the same as the input sample order at the time of experiment creation.
- Clustering - Perform clustering analysis on the correlation heatmap using the **Run clustering** icon from the correlation analysis heatmap toolbar (See To run clustering on a correlation heatmap on page 760). Samples are clustered based on correlation coefficients using hierarchical clustering algorithm and plotted in the order they cluster together. The linkage rule used is average and the distance metrics used is Euclidean.

For more information on the available toolbar icons in a correlation analysis heatmap, see Correlation Heatmap View Toolbar on page 776. See Properties on 792 to change the properties of the correlation analysis heatmap.

NOTE Correlation coefficients are computed on raw, non-normalized data and in linear scale.

Experiment Grouping shows the parameters and parameter values for each sample.

The third window shows the legend of the active QC tab.

Unsatisfactory samples or those that have not passed the QC criteria can be removed from further analysis, at this stage, using *Add/Remove Samples* button. Once a few samples are removed, re-normalization and baseline transformation of the remaining samples is carried out again. The samples removed earlier can also be added back. Click on **OK** to proceed.

- **Filter Probe Set by Expression:**

  Entities are filtered based on their signal intensity values. For details refer to the section on Filter Probesets by Expression

- **Filter Probe Set by Flags:**

  In this step, the entities are filtered based on their flag values P(present), M(marginal) and A(absent). Users can set what proportion of conditions must meet a certain threshold. The flag values that are defined at the creation of the new technology (Step 4 of 9) are taken into consideration while filtering the entities. The filtration is done in 4 steps:

  1. Step 1 of 4 : *Entity list and interpretation* window opens up. Select an entity list by clicking on *Choose Entity List* button. Likewise by clicking on *Choose Interpretation* button, select the required interpretation from the navigator window.

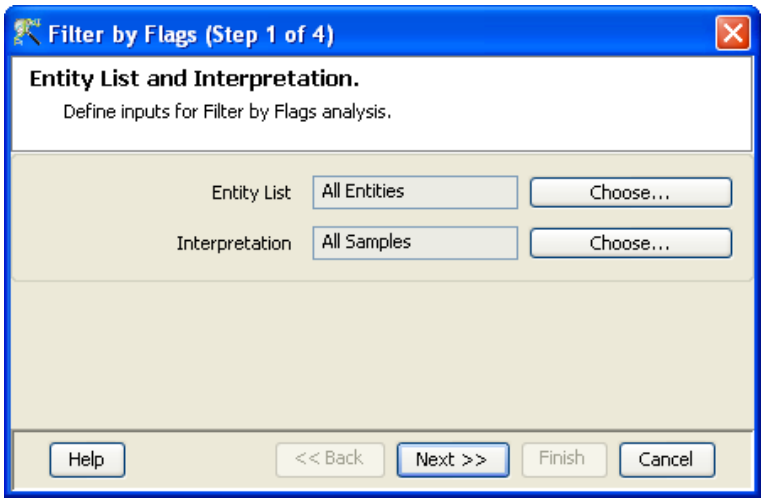


Figure 13.18: Entity List and Interpretation

2. Step 2 of 4: This step is used to set the Filtering criteria and the stringency of the filter. Select the flag values that an entity must satisfy to pass the filter. By default, the Present and Marginal flags are selected. Stringency of the filter can be set in *Retain Entities* box.

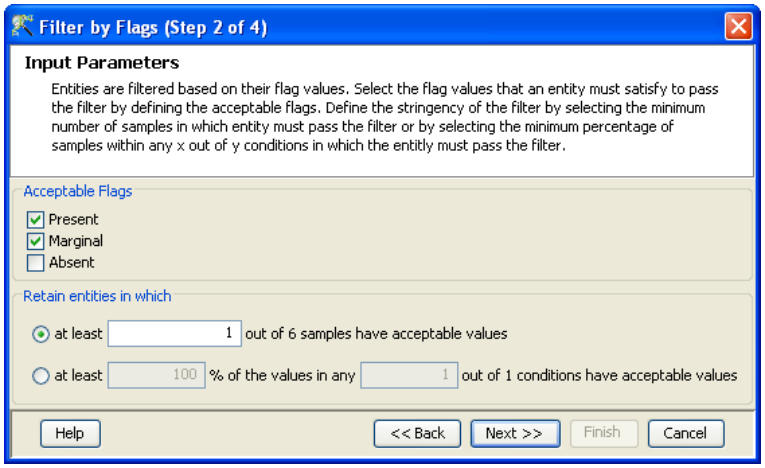


Figure 13.19: Input Parameters

3. Step 3 of 4: A spreadsheet and a profile plot appear as 2 tabs, displaying those probes which have passed the filter conditions. Baseline transformed data is shown here. Total number of probes and number of probes passing the filter are displayed on the top of the navigator window. (See Figure 13.20).

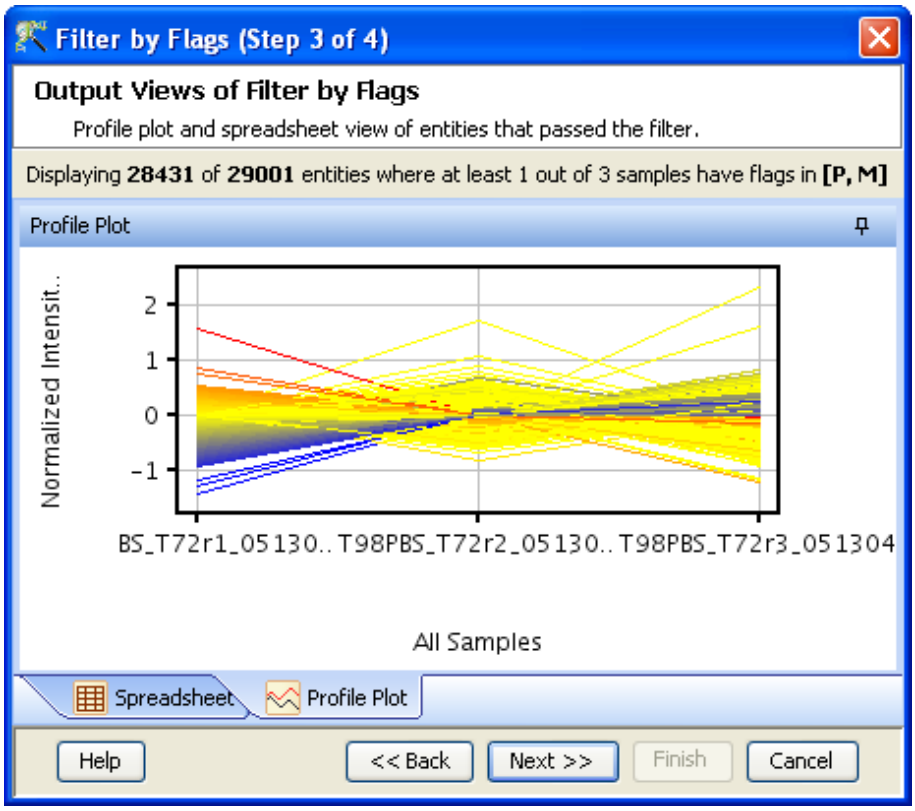


Figure 13.20: Output Views of Filter by Flags

4. Step 4 of 4: Click *Next* to annotate and save the entity list.(See Figure 13.21).

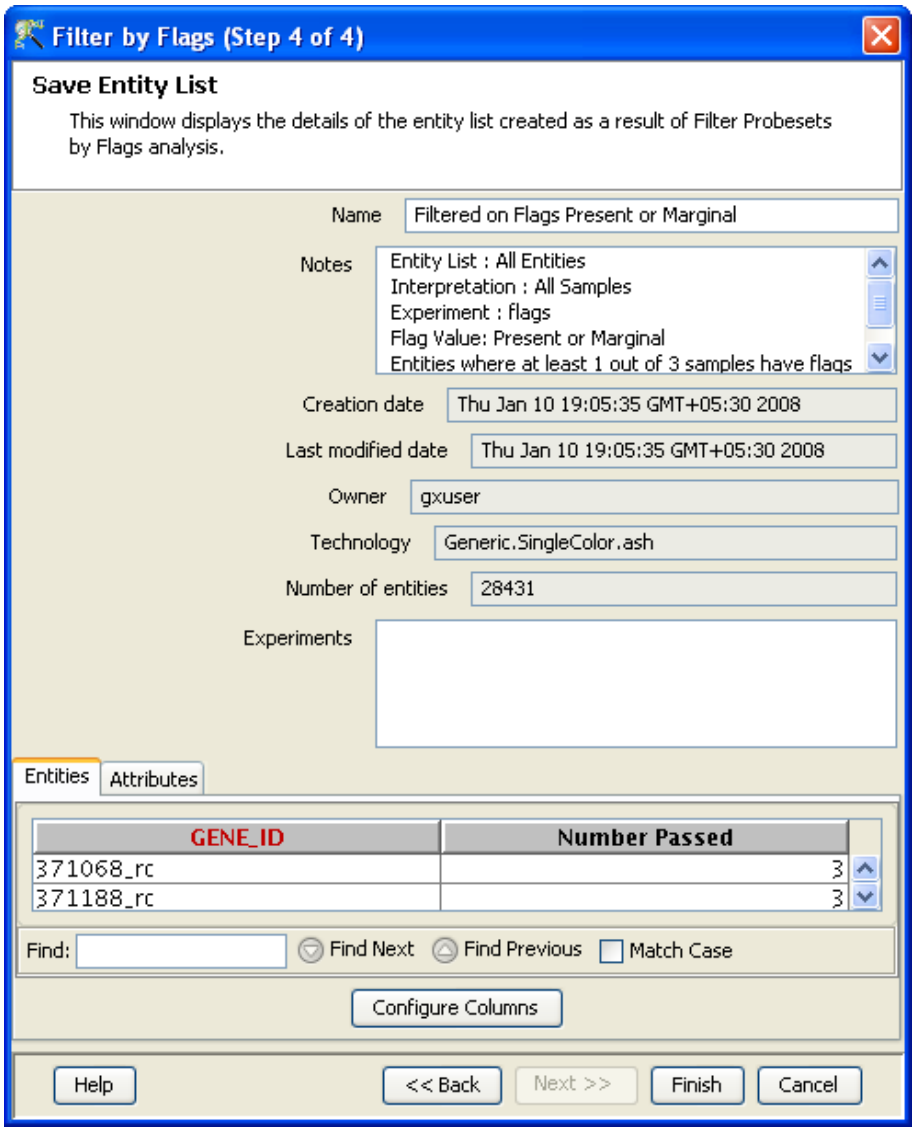


Figure 13.21: Save Entity List

- **Filter Probesets on Data Files:** Entities can be filtered based on values in a specific column of the original data files. For details refer to the section on Filter Probesets on Data Files
- **Filter Probesets by Error:** Entities can be filtered based on the standard deviation or coefficient of variation using this option. For details refer to the section on Filter Probesets by Error

13.3.3 Analysis

- **Statistical Analysis**

  For details refer to section Statistical Analysis in the Data Import Wizard section.
- **Filter on Volcano Plot**

For details refer to section Filter on Volcano Plot.

- **Fold Change**
  For details refer to section Fold change.
- **Clustering**
  For details refer to chapter Clustering Analysis.
- **Correlation**
  For details refer to chapter Correlation Analysis.
- **Find Similar Entities**
  For details refer to section Find similar entities.
- **Filter on Parameters**
  For details refer to section Filter on Parameters.
- **Principal Component Analysis**
  For details refer to section PCA.

13.3.4 Class Prediction

- **Build Prediction Model** For details refer to section Build Prediction Model.
- **Run Prediction** For details refer to section Run Prediction.

13.3.5 Results

- **Gene Ontology (GO) analysis**
  For more details see GO Analysis.
- **Gene Set Enrichment Analysis (GSEA)**
  For more details see Performing GSEA in **GeneSpring**.
- **Gene Set Analysis (GSA)**
  For more details see Gene Set Analysis.
- **Find Similar Entity Lists**
  This feature is discussed in a separate section called Find Similar Entity Lists.
- **Single Experiment Analysis or Multi-Omic Analysis**
  This feature is discussed in detail in the section Single Experiment and Multi-Omic Analysis.
- **Export to IPA**
  This feature is discussed in detail in the section Ingenuity Pathways Analysis (IPA) Connector.
- **Import IPA Entity List**
  This feature is discussed in detail in the section Import IPA Entity List.

- **NLP Network Discovery**

  This feature is discussed in detail in the section NLP Network Discovery.

- **MeSH Network Builder**

  This feature is discussed in detail in the section MeSH Network Builder.

- **Extract Relations via NLP**

  This feature is discussed in detail in the section Extract Relations via NLP.

13.3.6 Utilities

- **Import Entity list from File**

  For details refer to section Import Entity list from File.

- **Analysis: Biological Significance**

  For details refer to section Analysis: Biological Significance.

- **Filter On Entity List**

  For further details refer to section Filter On Entity List.

- **Remove Entities with missing signal values**

  For details refer to section Remove Entities with missing signal values.

Chapter 14

Analyzing Generic Two Color Expression Data

GeneSpring supports Generic Two color experiments, such as spotted cDNA arrays. However, a technology first needs to be created, based upon the file format being imported.

14.1 Creating Technology

Technology creation is a step common to both Generic Single Color and Two color experiments. Technology creation enables the user to specify the columns (Signals, Flags, Annotations etc.) in the data file and their configurations which are to be imported. Different technologies need to be created for different file formats. Custom technology can be created by navigating to ***Annotations*** in the toolbar and selecting ***Create Technology*** $\longrightarrow$ ***Custom from file***. **GeneSpring** also allows the user to create a technology specifically for GPR files via ***Annotations*** $\longrightarrow$ ***Create Technology*** $\longrightarrow$ ***From .gpr files***. This technology can later be used for creating a Generic Two Color experiment.

The process of creating a technology uses one data file as a sample file to mark the columns. Therefore, it is important that all the data files being used to create an experiment should have identical formats.

Technology creation using both the methods is detailed below:

14.1.1 Creation of Custom Technology-Non gpr files

The *Create Custom Technology* wizard has multiple steps. While steps 1, 2, 3 and 9 are common to both the Single color and Two Color, the remaining steps are specific to either of the two

technologies.

- **Technology Name (Step 1 of 9):** User input details, i.e., Technology type, Technology name, Organism, Sample data file location, Number of samples in a single data file and particulars of the annotation file are specified here. Click *Next*. See Figure 14.1.

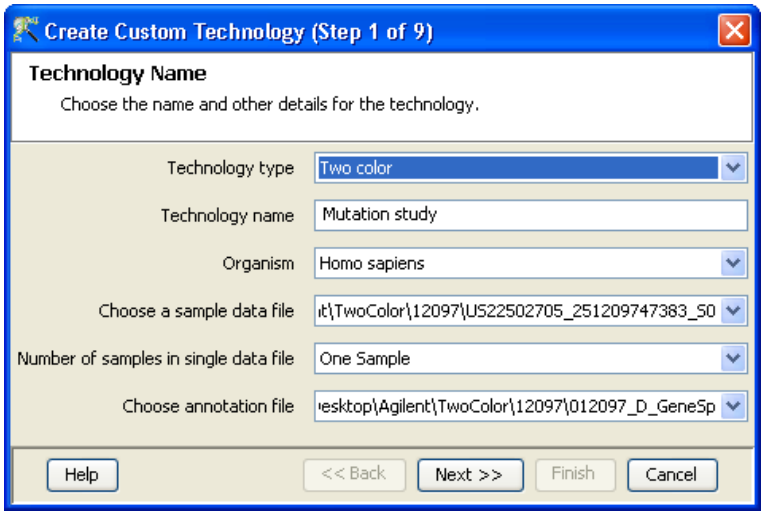


Figure 14.1: Technology Name

- **Format data set (Step 2 of 9):** This allows the user to specify the data file format. For this operation, four options are provided, namely, the *Separator*, the *Text qualifier*, the *Missing Value Indicator* and the *Comment Indicator*. The *Separator* option specifies if the fields in the file to be imported are separated by a tab, comma, hyphen, space etc. New separators can be defined by scrolling down to **Enter New** and providing the appropriate symbol in the textbox. *Text qualifier* is used for indicating characters used to delineate full text strings. This is typically a single or double quote character. The *Missing Value Indicator* is for declaring a string that is used whenever a value is missing. This applies only to cases where the value is represented explicitly by a symbol such as N/A or NA. The *Comment Indicator* specifies a symbol or string that indicates a comment section in the input file. Comment Indicators are markers at the beginning of the line which indicate that the line should be skipped (typical examples is the # symbol). See Figure 14.2.

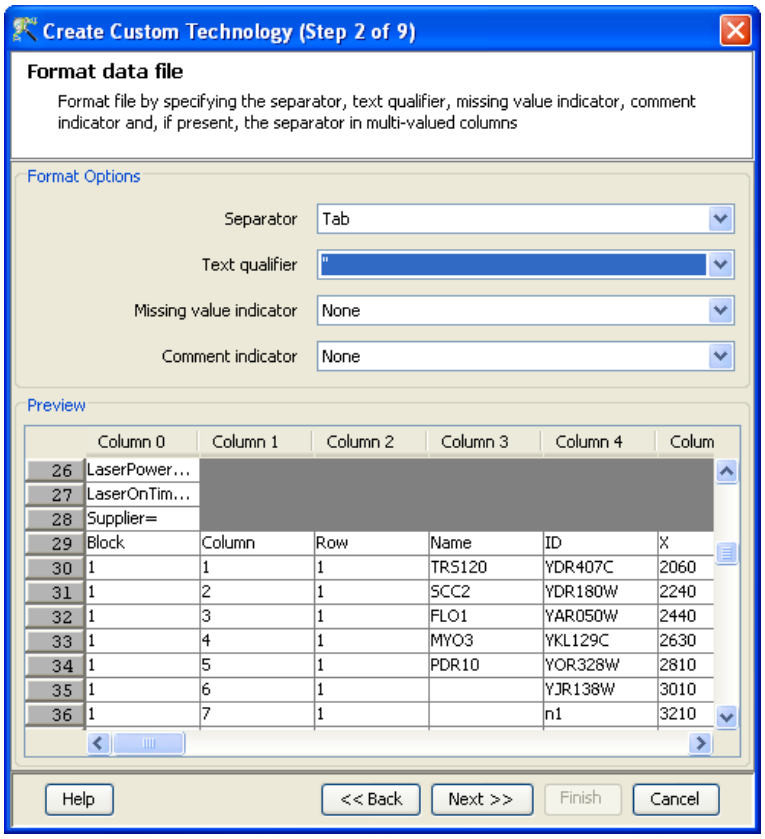


Figure 14.2: Format data file

- **Select Row Scope for Import (Step 3 of 9):** The data files typically contains headers which are descriptive of the chip type and are not needed for the analysis. Only those rows containing the data values are required. The purpose of this step is to identify which rows need to be imported. The rows to be imported must be contiguous in the file. The rules defined for importing rows from this file will then apply to all other files to be imported using this technology. Three options are provided for selecting rows:

  The default option is to select all rows in the file. Alternatively, one can choose to take a block of rows between specific row numbers (use the preview window to identify row numbers) by entering the row numbers in the appropriate textboxes. Remember to press the Enter key before proceeding. In addition, for situations where the data of interest lies between specific text markers, those text markers can be indicated. Note also that instead of choosing one of the options from the radio buttons, one can choose to select specific contiguous rows from the preview window itself by using Left-Click and Shift-Left-Click on the row header. The preview shows only the first 100 rows of the file by default. The user can change the default settings from ***Tools*** $\longrightarrow$ ***Options*** $\longrightarrow$ ***Miscellaneous*** $\longrightarrow$ ***Custom Data Library Creation*** $\longrightarrow$ ***Number of preview lines***. The panel at the bottom should be used to indicate whether or not there is a header row; in the latter case, dummy column names will be assigned. See Figure 14.3.

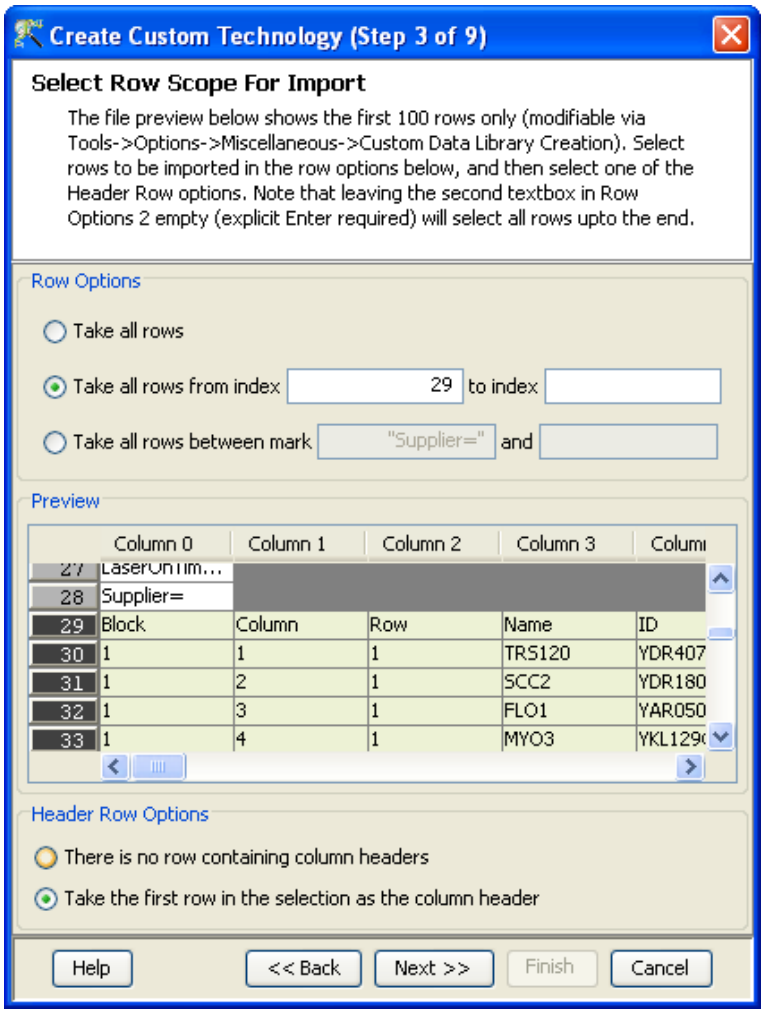


Figure 14.3: Select Row Scope for Import

- Steps 4 and 5 are used while creating custom technology for a single color experiment.
- **Create Custom technology (Step 6 of 9):** After the rows to be imported have been identified, columns for the identifier, background (BG) corrected signals and flag values for Cy5 and Cy3 channels in the data file have to be indicated. Please note that the identifier column should not be numeric. In case of a file containing a single flag column either the flag Cy3 or flag Cy5 can be used to mark the same. Categories within the flag columns can be configured to designate Present (P), Absent (A) or Marginal (M) values. Grid column can be specified to enable block by block normalization. See Figure 14.4.

Lowess sub-grid normalization can be performed by choosing the grid column.

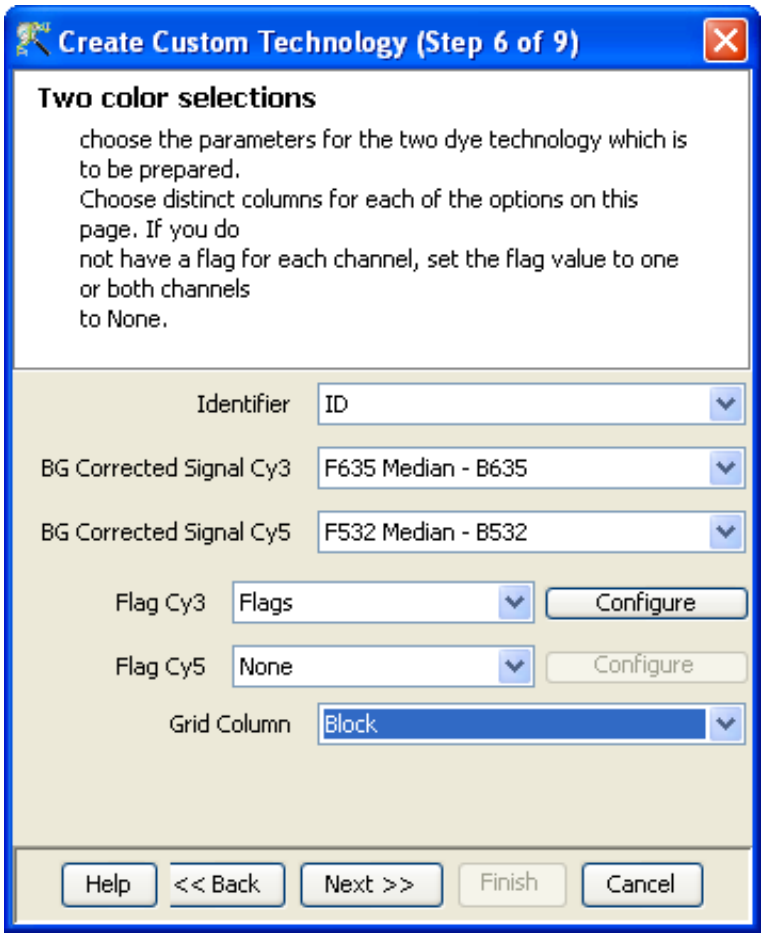


Figure 14.4: Two Color Selections

Annotation column options have to be specified from steps 7 to 9.

- **(Step 7 and 8 of 9):** These steps are similar to the step 2 of 9 and is used to format the annotation file. If a separate annotation file does not exist, then the same data file can be used as an annotation file, provided it has the annotation columns.
- **(Step 8 of 9):**Identical to step 3 of 9, this allows the user to select row scope for import in the annotation file.
- **(Step 9 of 9):** The Step 9 of technology creation is an extremely important step which allows the user to mark the columns appropriately. Proper marking of the various columns will enable the various functionalities like GO, GSEA, Genome Browser, Pathway Analysis to proceed smoothly. The markings to be given for all these functions is elaborated below:

- GSEA: The annotation file should contain a column containing Gene Symbol. This column should be marked as Gene Symbol from the drop-down menu.
- GSA: The annotation file should contain a column containing the gene Symbol. This column should be marked as Gene Symbol from the drop-down menu.
- GO: For carrying out GO analysis, the annotation file can either contain a single column with all the GO IDs in it, separated by a separator or it can contain separate columns for the

different GO processes. The single column with multiple GO IDs should be marked as *Gene Ontology accession* from the drop-down menu. Instead if columns containing individual GO processes(Biological Process, Cellular Component and Molecular Function) are present, they should be marked accordingly in the dropdown menu.

- Genome Browser: In order to view the data in Genome Browser, the annotation file should contain a Chromosome Start Index, Chromosome End Index, Strand and Chromosome Number columns. Provide the column mark for Chromosome Start index, Chromosome End index, Strand, Chromosome number respectively, from the drop-down menu.

  Note: The Chromosome Start Index < Chromosome End Index. For viewing *Profile* track only, in the Genome Browser, chromosome start index and chromosome number are needed. The labelling of the chromosome numbers should follow this convention-chr1, chr2 .....i.e. the word starts with chr followed by the chromosome number (without any space). For viewing data track, all four Chromosome Start Index, Chromosome End Index, Strand, Chromosome Number are needed.

  - If a custom technology is being created using an **Illumina** data and annotation file, then for the Genome Browser functionality, the column markings have to be handled as follows:

    For viewing using the Genome Browser, the annotation files has three columns which have values for all four (Chromosome Start Index, Chromosome End Index and Chromosome Number and Strand) Therefore before creating the custom experiment the user needs to parse these columns and create three new columns as follows :

    Probe_Chr_Orientation-- This column can be taken as it is. It should be marked as Strand.

    Chromosome -- A new column must be created wherein a 'chr' should be appended to each entry in the Chromosome column and this new column should be marked as Chromosome Number.

    Probe_Coordinates-- This column has each entry in the format a-b where a < b. Two new columns need to be created. one which has only the a values, (it should be marked as Chromosome Start Index) one which has only the b values (it should be marked as Chromosome End Index).

  - If a custom technology is being created using an **Agilent** data and annotation file, then for the Genome Browser functionality, the column markings have to be handled as follows:

    The annotation files have a single column 'Map' which has values for all four Chromosome Start Index, Chromosome End Index and Chromosome Number and Strand. Therefore before creating the custom experiment the user needs to parse the file and separate the four columns as Chromosome Start Index, Chromosome End Index Chromosome Number and Strand.

    Each entry in the Map column is typically in the format chrQ:a..b

    if a < b, the corresponding Chromosome Number is chrQ; the corresponding Chromosome Start Index is a; the corresponding Chromosome End Index is b; the corresponding Strand is +.

    if a > b the corresponding Chromosome Number is chrQ; the corresponding Chromosome Start Index is b; the corresponding Chromosome End Index is a; the corresponding Strand is - .

    For example, a Map value of chr14:34101457..34101398 corresponds to a Chromosome Start Index of 34101398, a Chromosome End Index of 34101457, a Chromosome Number of chr14 and a Strand of - (because in chrX:a..b a>b)

    For example, a Map value of chr6:46222041..46222100 corresponds to a Chromosome Start Index of 46222041, a Chromosome End Index of 46222100, a Chromosome Number of chr6 and a Strand of +(because in chrX:a..b a<b)

- Import BioPAX pathways: Pathways being imported should be in .owl format. During custom technology creation, provide the column mark for Entrez Gene ID/SwissProt from the drop-down menu. Only after this mark is provided can the proteins involved in a particular pathway be highlighted.
- Pathway Analysis: The annotation file should contain an Entrez Gene ID/SwissProt column, which have to be marked appropriately as Entrez Gene ID/SwissProt.
- Translation: This operation can be performed between organisms listed in the **Homologene table** in section Translation. Entrez Gene ID column has to be marked for performing translation.

See Figure 14.5.

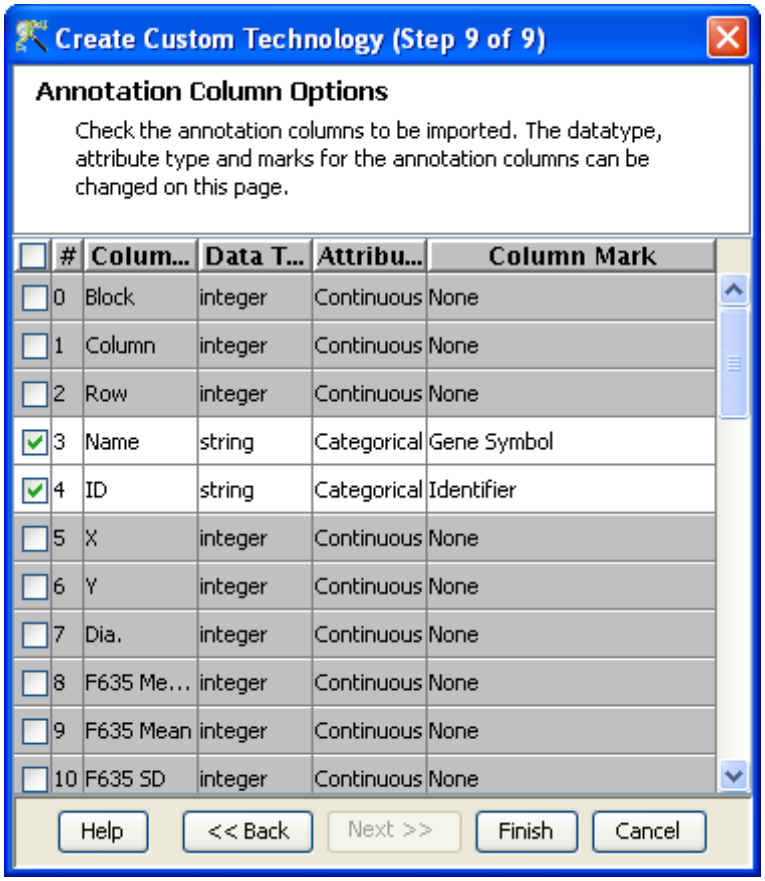


Figure 14.5: Annotation Column Options

The types of Data and Attribute marks available for the annotation columns are

- **Categorical:** A column marked as a ‘‘categorical” column means that the values in this column fall into certain finite distinct categories.
- **Continuous:** A column marked as a ‘‘continuous” column means that the values in this column can vary, potentially, over any large range.

- **String:** A continuous sequence of symbols or digits, not including a space.
- **Float:** A real number, i.e a number which can be given by a decimal representation.

Please note that the Entrez ID annotation column needs to be first specified as categorical and then as string in the dropdown.

The annotation marks are colored on the basis of their functionality in the tool. The meaning of the various colors are provided in the Figure 14.6.

Note:This figure is provided only for illustration purpose and is not available in the tool.

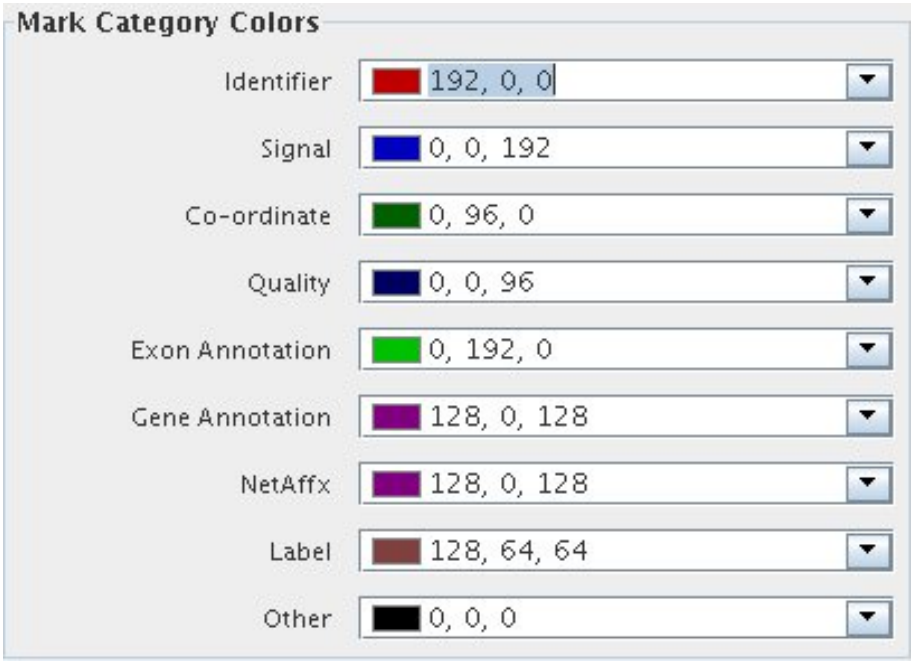


Figure 14.6: Annotation Mark Colors

Click ***Finish*** to exit the wizard.

14.1.2 GenePix Result Technology creation

This option allows the user to create a technology for files (.gpr) that have been generated using the GenePix Pro software. This feature is compatible for different versions of the gpr file. The gpr file used to create the technology should contain the following columns - ID, F635 Median - B635, F532 Median - B532, and Flags. Please note that the identifier column should not be numeric. This technology creation option is accessible from ***Annotations*** $\longrightarrow$ ***Create Technology*** $\longrightarrow$ ***From .gpr file***. On selecting this option, the user has to go through the following step for custom technology creation:

- **Input Data(Step 1 of 1):**

  This step allows the user to input data required for technology creation. The user has to provide the technology name, organism and the sample data file. See Figure 14.7. The organism name is optional, but in the event of Biological Genome creation, the organism name is a must. Biological Genome contains most of the annotations using which additional analysis like GO, GSEA etc can be performed. For further details, refer to the section on Biological Genome.

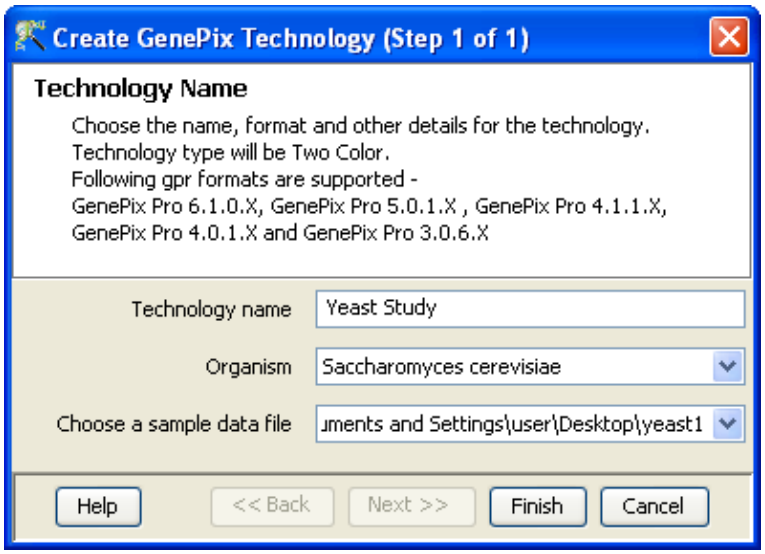


Figure 14.7: Technology Creation

The technology created does not have any annotations associated with it. The user can add annotations via ***Annotations*** $\longrightarrow$ ***Update Technology Annotations*** $\longrightarrow$ ***From file or Biological Genome***. For more details on the same, refer to Update Technology Annotations

14.1.3 Project and Experiment Creation

After technology creation, data files satisfying the file format can be used to create an experiment. The following steps will guide you through the process of experiment creation.

Upon launching **GeneSpring**, the startup is displayed with 3 options. See Figure 15.1.

1. **Create new project**
2. **Open existing project**
3. **Open recent project**

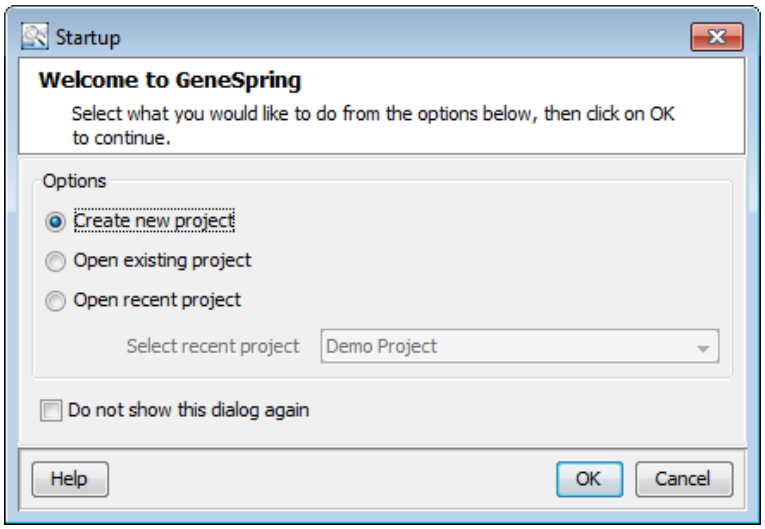


Figure 14.8: Welcome Screen

Either a new project can be created or else a previously generated project can be opened and re-analyzed. On selecting *Create New Project*, a window appears in which details (name of the project and notes) can be recorded. Press *OK* to proceed. See Figure 15.2.

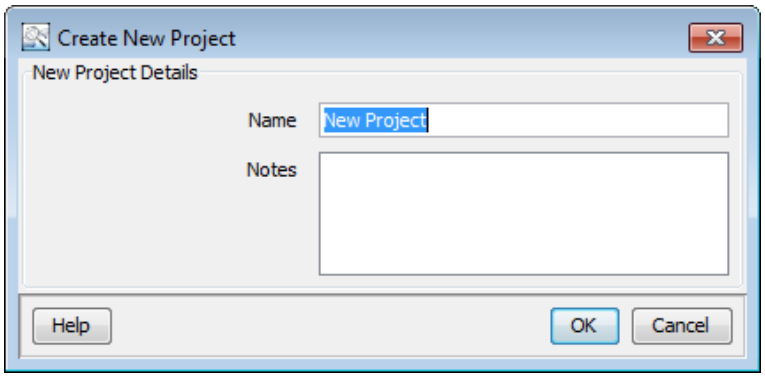


Figure 14.9: Create New Project

An Experiment Selection Dialog window then appears with two options

1. **Create new experiment**
2. **Open existing experiment**

See Figure 15.3.

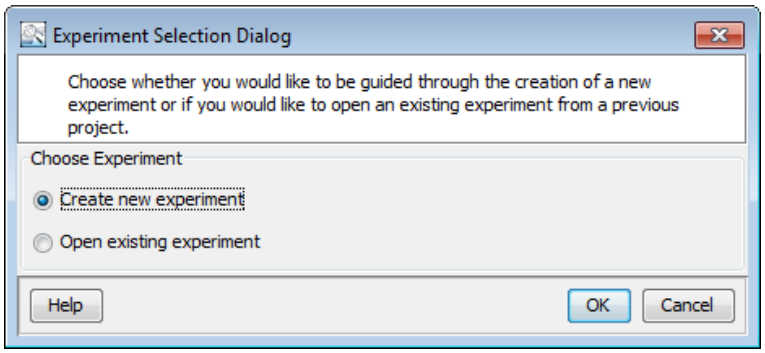


Figure 14.10: Experiment Selection

Selecting *Create new experiment* allows the user to create a new experiment (steps described below). *Open existing experiment* allows the user to use existing experiments from any previous projects in the current project. Choosing *Create new experiment* opens up a New Experiment dialog in which *Experiment name* can be assigned. Once the *Analysis Type* is selected (Expression), the *Experiment type* should then be specified (Generic Two Color), using the drop down button. The *Workflow Type* that appears is the *Data Import Wizard* type. Unlike the other technologies where *Analysis: Biological Significance* and *Data Import Wizard* analysis workflows are available, in case of Generic Two-color, only the *Data Import Wizard* is supported . Click *OK* will open a new experiment wizard. See Figure 14.11

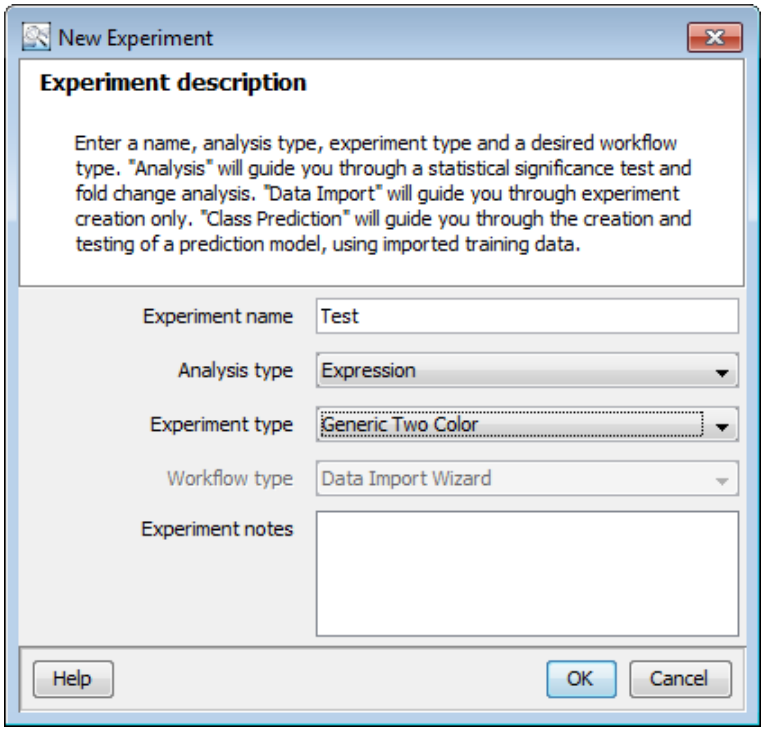


Figure 14.11: Experiment Description

14.2 Data Import Wizard

The *Data Import Wizard* offers a variety of choices to the user for the analysis. Thresholding can be performed. Based upon the technology, Lowess or sub-grid Lowess normalization can be performed. Additionally there are options for baseline transformation of the data and for creating different interpretations.

The *New Experiment Wizard* has the following steps:

1. **New Experiment (Step 1 of 4):** The technology (created as mentioned above) can be selected and the new data files or previously used data files in **GeneSpring** can be imported in to create the experiment. A window appears containing the following options:

   (a) **Choose Files(s)**
   (b) **Choose Samples**
   (c) **Choose Raw Files**
   (d) **Reorder**
   (e) **Remove**

   An experiment can be created using either the data files or else using samples. Upon loading data files, **GeneSpring** associates the files with the technology (see below) and creates samples. These samples are stored in the system and can be used to create another experiment via the *Choose Samples* option through a sample search wizard. If the user has imported any custom experiments from **GeneSpring GX 7** and wants to recreate the experiment in **GeneSpring**, then the user can create a new technology in the tool with an original raw file and later utilize the **Choose Raw Files** option to choose the raw files associated with the migrated custom experiment. For selecting data files and creating an experiment, click on the *Choose File(s)* button, navigate to the appropriate folder and select the files of interest. Select *OK* to proceed.

   The sample search wizard that comes up via the option *Choose Samples* has the following search conditions:

   (a) **Search field** (which searches using any of the 6 following parameters- (Creation date, Modified date, Name, Owner, Technology, Type).
   (b) **Condition** (which requires any of the 4 parameters-Equals, Starts with, Ends with and includes Search value).
   (c) **Value**

   Multiple search queries can be executed and combined using either *AND* or *OR*.

   Samples obtained from the search wizard can be selected and added to the experiment using *Add* button, similarly can be removed using *Remove* button.

   After selecting the files, clicking on the *Reorder* button opens a window in which the particular sample or file can be selected and can be moved either up or down by pressing on the buttons. Click on *OK* to enable the reordering or on *Cancel* to revert to the old order. See Figure 14.12.

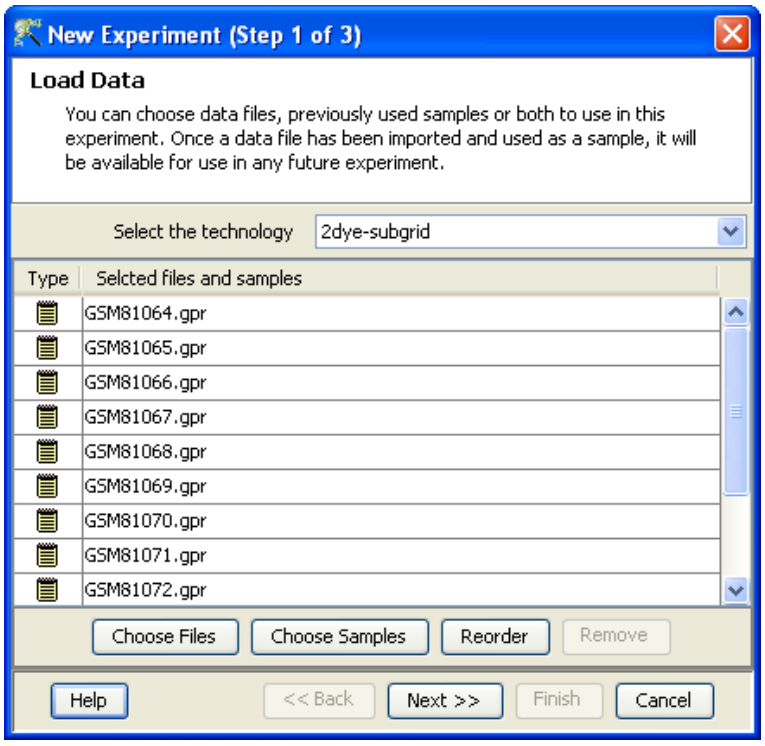


Figure 14.12: Load Data

2. **New experiment (Step 2 of 4):** Dye-swapping accounts for dye related bias (as both dyes bind DNA with different affinities) that cannot be removed by standard normalization methods.

   In **GeneSpring**, during the creation of two color experiments, users can specify sample(s) that have been dye swapped. The experiment creation wizard has an additional step (step 3 of 5) that lists all the sample(s) chosen by the user for experiment creation. At this step users can select sample(s) that have been dye swapped. For sample(s) that have been marked as dye-swapped, **GeneSpring** treats Cy3 as‘‘signal’’ and Cy5 as ‘‘control” and computes the signal as Cy3/Cy5. For non-dye swapped sample(s), **GeneSpring** treats Cy5 as ‘‘signal’’ and Cy3 as ‘‘control” and computes the signal as Cy5/Cy3. See Figure 14.13.

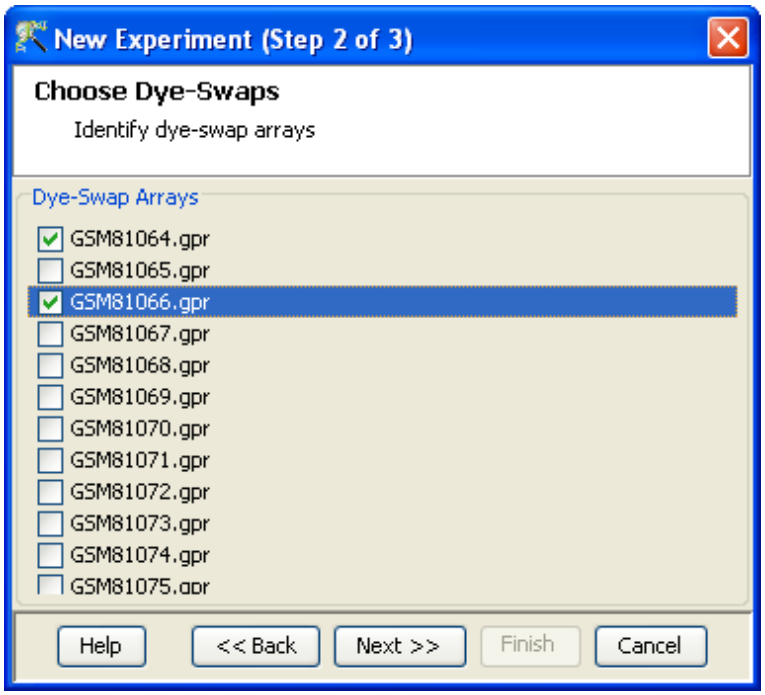


Figure 14.13: Choose Dye-Swaps

3. **New experiment (Step 3 of 4):** This gives the options for preprocessing of input data. It allows the user to threshold raw signals to chosen values and the selection of Lowess normalization. In case of experiment creation using .gpr files, the option to perform sub-grid Lowess is not present.

   In case, the data is already log transformed, the user can select the checkbox stating that their signal values are already in log scale. This will disable the thresholding option also.

   See Figure 14.14.

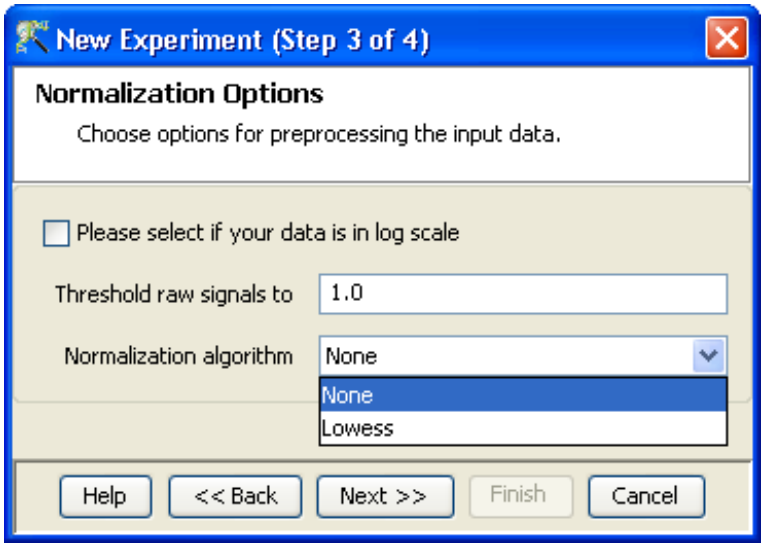


Figure 14.14: Preprocess Options

4. **New experiment (Step 4 of 4):**

   This step provides the baseline options which include:

   - ***Do not perform baseline***
   - ***Baseline to median of all samples:*** For each probe the median of the log summarized values from all the samples is calculated and subtracted from each of the samples.
   - ***Baseline to median of control samples:*** For each sample, an individual control or a set of controls can be assigned. Alternatively, a set of samples designated as controls can be used for all samples. For specifying the control for a sample, select the sample and click on ***Assign value***. This opens up the ***Choose Control Samples*** window. The samples designated as Controls should be moved from the *Available Items* box to the *Selected Items* box. Click on ***Ok***. This will show the control samples for each of the samples.

     In *Baseline to median of control samples*, for each probe the median of the log summarized values from the control samples is first computed and then this is subtracted from the sample. If a single sample is chosen as the control sample, then the probe values of the control sample are subtracted from its corresponding sample.

   See Figure 14.15.

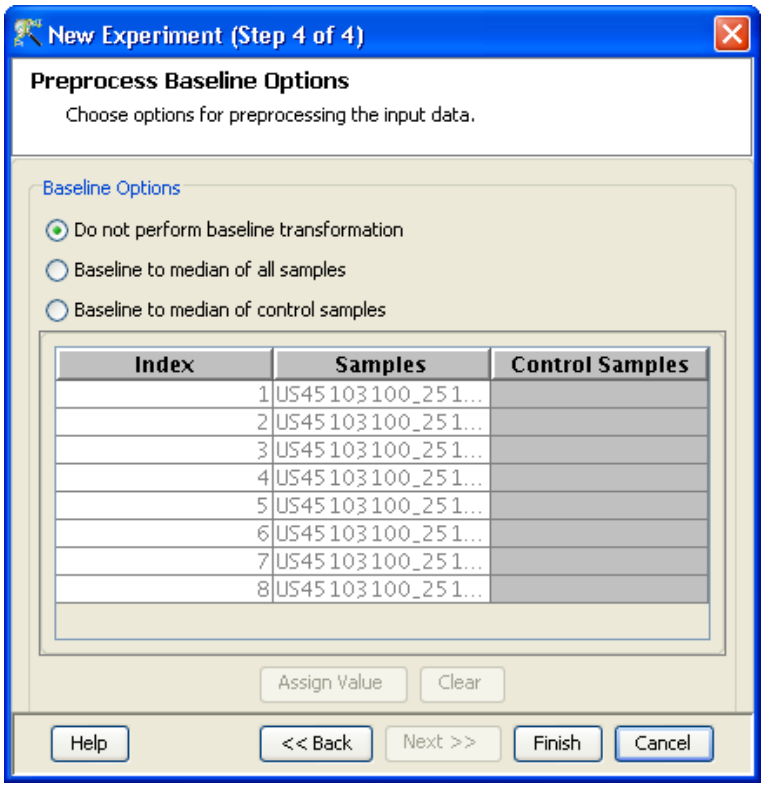


Figure 14.15: Preprocess Baseline Options

Clicking *Finish* creates an experiment, which is displayed as a Box Whisker plot in the active view. Alternative views can be chosen for display by navigating to *View* in Toolbar.

14.2.1 Data Processing for Generic Two Color Data

1. **File formats:** The files should be tabular in nature. For example, .csv, .tsv , .gpr etc.
2. **Signal Columns**: When custom technology is created via ***Annotations*** $\longrightarrow$ ***Create Technology*** $\longrightarrow$ ***From .gpr file***, the signal columns taken from files are F532 Median-B532 for cy3 and F635 Median-B635 for cy5.
3. **Raw:** The term ''raw'' signal values refer to the linear data after thresholding to 1.0 and summarization for the individual channels (cy3 and cy5). Summarization is performed by computing the geometric mean.
4. **Normalized:** The term Normalized signal value refers to the raw data after normalization of cy5 channel, ratio computation (cy5/cy3), log transformation and Baseline Transformation.
5. **Treatment of on-chip replicates:** The signal value of a probeset is the geometric mean of all its probes.
6. **Flag values:** The values for the probes are configured by the user during the creation of technology as either present, marginal or absent. Based on the values of the probes, the probeset is assigned a flag value. The order of importance for flag values for probes in a probeset is *Present>Marginal>Absent*. When custom technology is created via ***Annotations*** $\longrightarrow$ ***Create Technology*** $\longrightarrow$ ***From .gpr file***, flags are configured by the tool. A value of -50 is designated as Marginal(M) and anything below is considered Absent(A) and anything above is considered as Present(P).
7. **Treatment of Control probes:** The control probes are included while performing normalization.
8. **Empty Cells:** Empty cells might be present in the intensity values column for certain genes in the data file. These genes are brought in **GeneSpring** . But an entity list containing these genes cannot be used for running clustering and class prediction.
9. **Sequence of events:** The sequence of events involved in the processing of the data files is: thresholding$\longrightarrow$ summarization$\longrightarrow$ normalization$\longrightarrow$ ratio computation$\longrightarrow$ log transformation$\longrightarrow$ Baseline Transformation.
10. **Merging of files:** Multiple files in Generic experiment creation are combined based on the Identifier column using the following rules. The very first file among the various files chosen server as a master reference (you can determine which file serves as the first file using the *Reorder* button on Page 1 of the New Experiment Creation page). The number of rows in this master must exceed the number of rows in all subsequent files, for extra rows in these subsequent files are dropped. Next, all identifiers in the Idenfier column of this first file are considered and missing values in these, if any, are discarded. This results in a set of valid identifier values; all rows in all other files whose identifier values are outside of this set are discarded. Next, on-chip replicates are determined by counting the number of occurrences of each valid identifier in the first file. Consider for example an identifier Id1 which appears 3 times in file 1. Then rows corresponding to the first 3 occurrences of Id1 are taken in each of the other files; if there are fewer than 3 rows, then as many rows that are present are taken; and if there are more than 3 rows, then the first 3 are taken. The summarized value for Id1 in each file is determined by taking a geometric mean over these chosen rows.

14.2.2 Experiment Setup

Once an experiment is created, the *Data Import Wizard* steps appear on the right hand side. This gives options to view quick start guide, create experiment grouping and interpretation and the utility to create gene level experiment from probe level experiment. See section Experiment Setup for complete details.

14.2.3 Quality Control

- **Quality Control on Samples:**

  Quality Control (QC) allows you decide which samples are ambiguous and which are passing the quality criteria. Based on the QC results, unreliable you can remove samples from further downstream analysis. The QC dialog displays three tiled windows:

  1. 3D PCA scores
  2. Experiment grouping
  3. Legend

  See Figure 14.16.

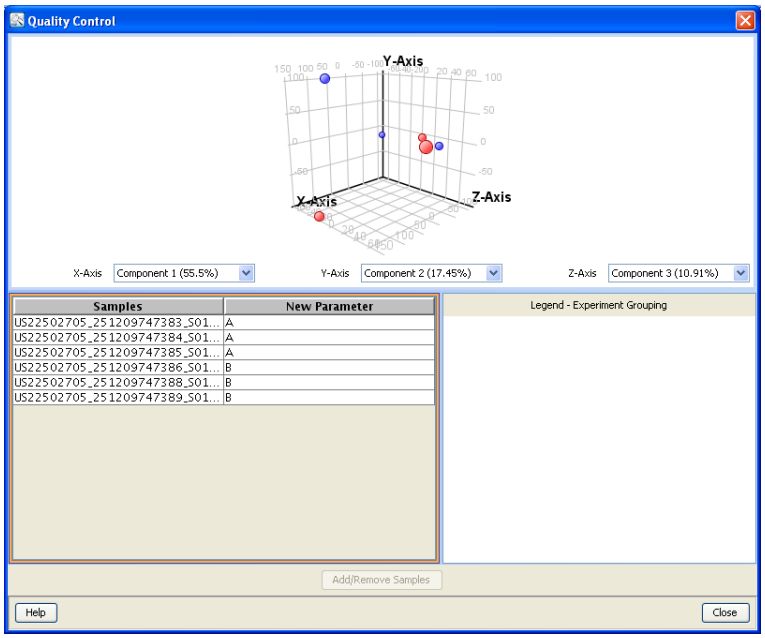


Figure 14.16: Quality Control

In the **Principal Component Analysis (PCA)** tab, PCA scores are visually represented in a 3D scatter plot. These scores are used to check data quality. It shows one point per array and is colored by the *Experiment Factors* provided earlier in the *Experiment Groupings* view. This allows viewing of separations between groups of replicates. Ideally, replicates within a group should cluster together and separately from arrays in other groups. The PCA components, represented in the X, Y, and Z axes are numbered 1, 2, 3... according to their

decreasing significance. The 3D PCA scores plot can be customized via ***Right-Click*** $\longrightarrow$ ***Properties***. To zoom into a 3D Scatter plot, press the Shift key and simultaneously hold down the left mouse button and move the mouse upwards. To zoom out, move the mouse downwards instead. To rotate, press the Ctrl key, simultaneously hold down the left mouse button and move the mouse around the plot.

Experiment Grouping shows the parameters and parameter values for each sample.

The third window shows the legend of the active QC tab.

Click on **OK** to proceed.

- **Filter Probe Set by Expression:** Entities are filtered based on their signal intensity values. for details refer to the section on Filter Probesets by Expression
- **Filter Probe Set by Flags:**

  In this step, the entities are filtered based on their flag values, the P(present), M(marginal) and A(absent). Users can set what proportion of conditions must meet a certain threshold. The flag values that are defined at the creation of the new technology (Step 2 of 3) are taken into consideration while filtering the entities. The filtration is done in 4 steps:

  1. Step 1 of 4 : *Entity list and interpretation* window opens up. Select an entity list by clicking on *Choose Entity List* button. Likewise by clicking on *Choose Interpretation* button, select the required interpretation from the navigator window. This is illustrated in Figure 14.17.

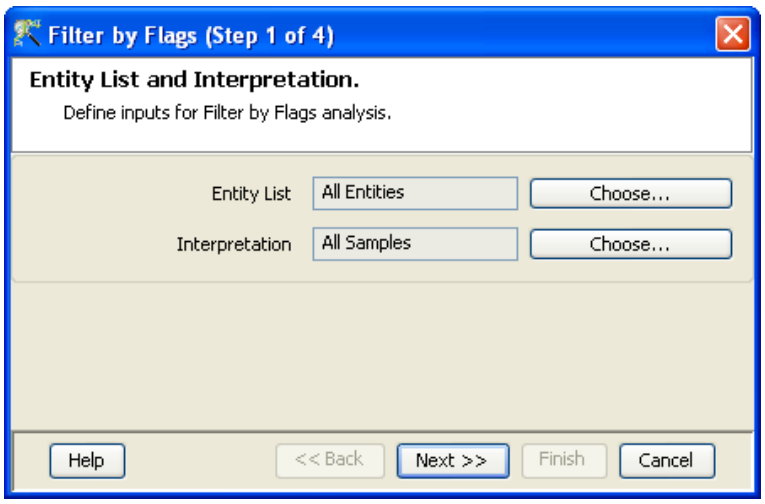


Figure 14.17: Entity List and Interpretation

  2. Step 2 of 4: This step is used to set the Filtering criteria and the stringency of the filter. Select the flag values that an entity must satisfy to pass the filter. By default, the Present and Marginal flags are selected. Stringency of the filter can be set in *Retain Entities* box.(See Figure 14.18) .

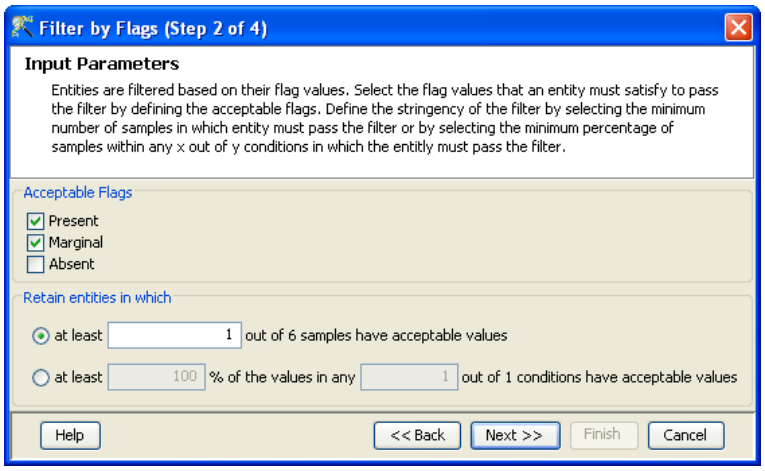


Figure 14.18: Input Parameters

3. Step 3 of 4: A spreadsheet and a profile plot appear as 2 tabs, displaying those probes which have passed the filter conditions. Baseline transformed data is shown here. Total number of probes and number of probes passing the filter are displayed on the top of the navigator window.(See Figure 14.19).

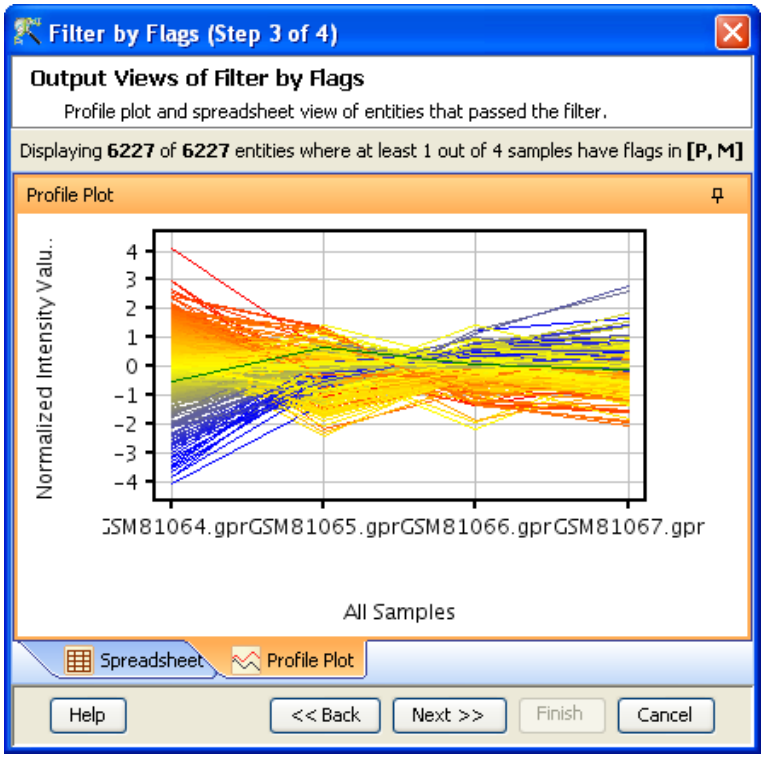


Figure 14.19: Output Views of Filter by Flags

4. Step 4 of 4: Click *Next* to annotate and save the entity list. (See Figure 14.20).

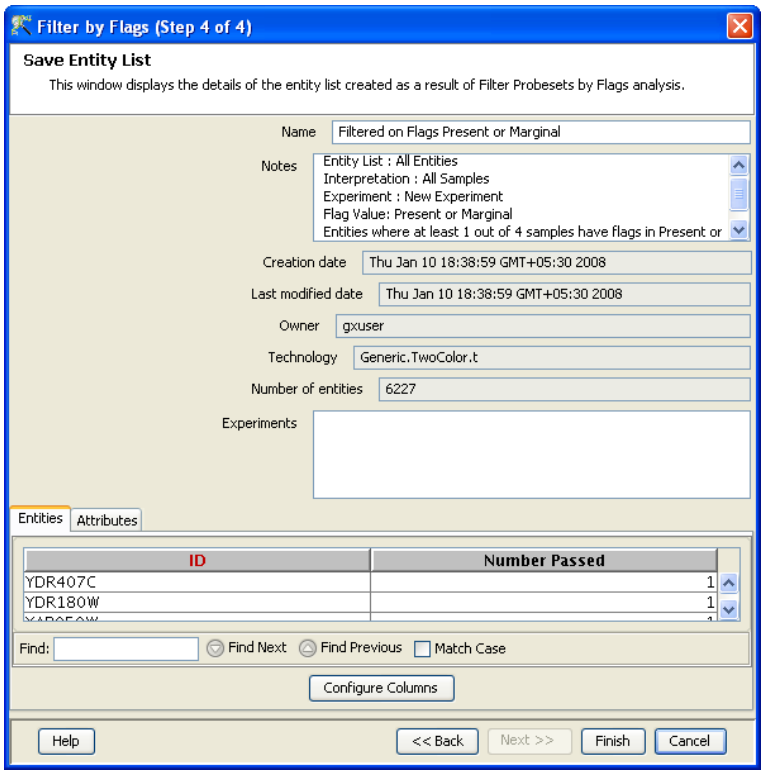


Figure 14.20: Save Entity List

- **Filter Probesets on Data Files:** Entities can be filtered based on values in a specific column of the original data files. For details refer to the section on Filter Probesets on Data Files
- **Filter Probesets by Error:** Entities can be filtered based on the standard deviation or coefficient of variation using this option. For details refer to the section on Filter Probesets by Error

14.2.4 Analysis

- **Statistical Analysis**

  For details refer to section Statistical Analysis in the Data Import Wizard section.

- **Filter on Volcano Plot**

  For details refer to section Filter on Volcano Plot.

- **Fold Change**

  For details refer to section Fold change.

- **Clustering**

  For details refer to chapter Clustering Analysis.

- **Correlation**

  For details refer to chapter Correlation Analysis.

- **Find Similar Entities**
  For details refer to section Find similar entities.
- **Filter on Parameters**
  For details refer to section Filter on Parameters.
- **Principal Component Analysis**
  For details refer to section PCA.

14.2.5 Class Prediction

- **Build Prediction Model** For details refer to section Build Prediction Model.
- **Run Prediction** For details refer to section Run Prediction.

14.2.6 Results

- **Gene Ontology (GO) analysis**
  For more details see GO Analysis.
- **Gene Set Enrichment Analysis (GSEA)**
  For more details see Performing GSEA in **GeneSpring**.
- **Gene Set Analysis (GSA)**
  For more details see Gene Set Analysis.
- **Find Similar Entity Lists**
  This feature is discussed in a separate section called Find Similar Entity Lists.
- **Single Experiment Analysis or Multi-Omic Analysis**
  This feature is discussed in detail in the section Single Experiment and Multi-Omic Analysis.
- **Export to IPA**
  This feature is discussed in detail in the section Ingenuity Pathways Analysis (IPA) Connector.
- **Import IPA Entity List**
  This feature is discussed in detail in the section Import IPA Entity List.
- **NLP Network Discovery**
  This feature is discussed in detail in the section NLP Network Discovery.
- **MeSH Network Builder**
  This feature is discussed in detail in the section MeSH Network Builder.
- **Extract Relations via NLP**
  This feature is discussed in detail in the section Extract Relations via NLP.

14.2.7 Utilities

- **Import Entity list from File**

  For details refer to section Import Entity list from File.

- **Analysis: Biological Significance**

  For details refer to section Analysis: Biological Significance.

- **Filter On Entity List**

  For further details refer to section Filter On Entity List.

- **Remove Entities with missing signal values**

  For details refer to section Remove Entities with missing signal values.

Chapter 15

Analyzing Agilent miRNA Data

Micro RNAs or the miRNAs are small (22-25 nucleotides long), single-stranded, regulatory RNA molecules identified first in 1993. They are found in almost all of the life forms. Regulation is achieved by binding to regions of mRNA which share sequence complementarity with the miRNA. miRNAs affect the expression of genes involved in several physiological, developmental and pathological processes. Hence, expression studies of miRNA became important to understand their role in controlling biological and pathological processes. Advanced high throughput technologies like expression arrays enable us to study expression patterns of miRNA under given conditions. These studies can be correlated with their target gene expression studies.

GeneSpring supports all the Agilent miRNA microarray chip types. It supports data files obtained in text (.txt) format from Agilent Feature Extraction (FE) version 10.7.3 or above. GeneView files are not supported.

GeneSpring creates standard technologies for running miRNA experiments based on the miRBase version ID as described in the section Selecting technology for Agilent miRNA.

15.1 Running the Agilent miRNA Workflow

Upon launching **GeneSpring**, the startup is displayed with 3 options.

- **Create new project**
- **Open existing project**
- **Open recent project**

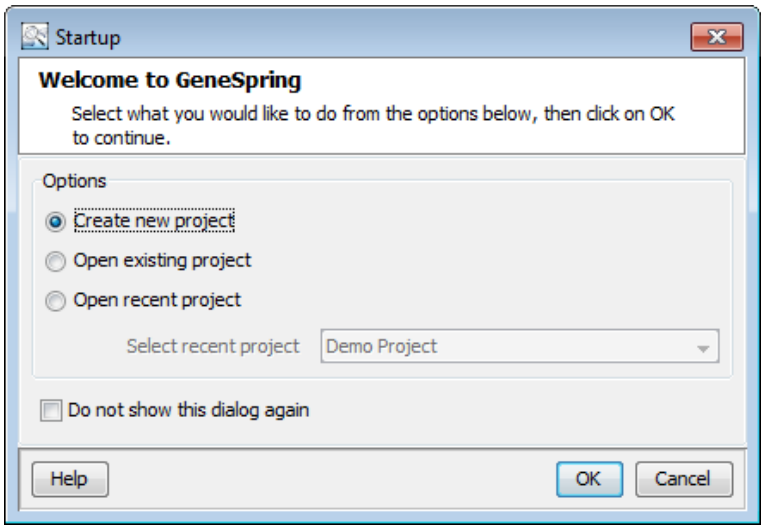


Figure 15.1: Welcome Screen

Either a new project can be created or a project created previously can be opened and re-analyzed. On selecting **Create new project**, a window appears in which details (Name of the project and Notes) can be recorded. **Open recent project** lists all the projects that were recently worked on and allows you to select a project. After selecting any of the above 3 options, click on ***OK*** to proceed.

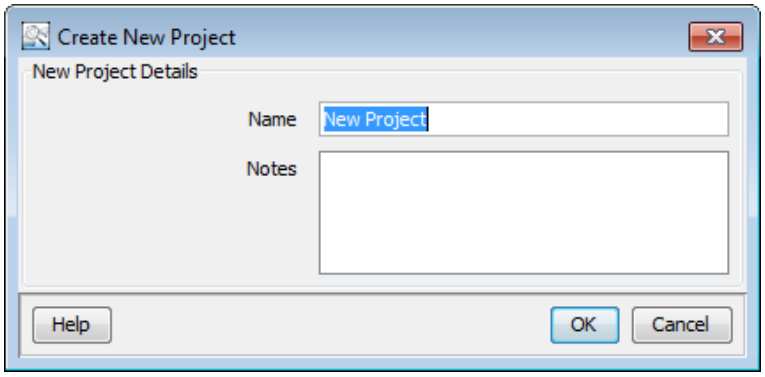


Figure 15.2: Create New Project

If **Create new project** is chosen, then an Experiment Selection dialog window appears with two options

1. **Create new experiment**: This allows the user to create a new experiment (Refer 15.1).
2. **Open existing experiment**: This allows you to use existing experiments from previous projects for further analysis.

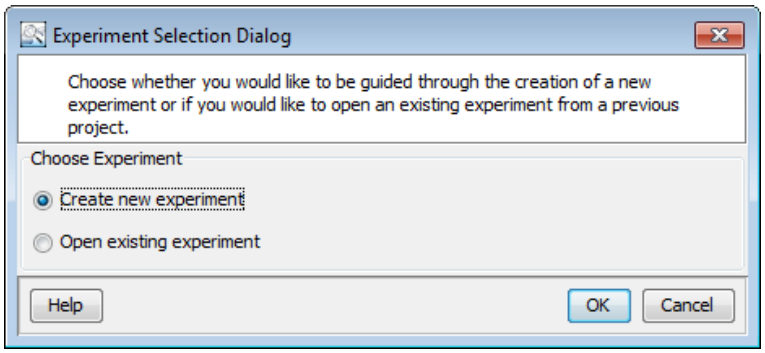


Figure 15.3: Experiment Selection

On selecting the **Create new experiment** option, a New Experiment dialog opens up where a **Experiment name** can be assigned. The drop-down menu for the analysis type gives the user the option to choose between multiple analysis types namely, Association, Copy Number, Exon, Expression, Mass Profiler Professional, Pathway, RTPCR, and miRNA. After selecting the desired analysis type, you need to select an experiment type of interest (whenever available).

Next, select a workflow type from the options based on your familiarity with the analysis type and **GeneSpring**.

1. **Analysis: Biological Significance** (Equivalent to Guided Workflow)
2. **Data Import Wizard** (Equivalent to Advanced Analysis)

Analysis: Biological Significance workflow is primarily meant for a new user and is designed to assist you through the creation and basic analysis of an experiment. Analysis involves default parameters which are not configurable in majority of the cases. However, in **Data Import Wizard**, the parameters can be changed to suit your individual requirements.

Upon selecting a workflow, a window opens with the following options:

1. Choose Files(s)
2. Choose Samples
3. Reorder
4. Remove

An experiment can be created using either the data files or else using samples. **GeneSpring** differentiates between a data file and a sample. A data file refers to the hybridization data obtained from a scanner. On the other hand, a sample is created within **GeneSpring**, when it associates the data files with its appropriate technology (See the section on Technology). Thus a sample created with one technology cannot be used in an experiment of another technology. These samples are stored in the system and can be used to create another experiment of the same technology via

the ***Choose Samples*** option. For selecting data files and creating an experiment, click on the ***Choose File(s)*** button, navigate to the appropriate folder and select the files of interest. Click on ***OK*** to proceed.

Selecting the ***Choose Samples*** option opens a sample search wizard, with the following search conditions:

1. **Search Field**: Requires one of these parameters; Build, Creation Date, Library ID, Library Layout, Modified Date, Name, Organism, Platform, Sample Type, Taxonomy ID, Technology, Type, chunk.range, object.class, sm.id, and sm.version. Select any one of these to perform the search.
2. **Condition**: Requires one of the these parameters; equals, starts with, ends with, and includes. Select any one of these to perform the search.
3. **Search Value**

Note: The default options for these fields are, *Technology* (Search Field), *starts with* (Condition), and *Agilent.miRNA* (Search Value).

Multiple search queries can be executed and combined using either *OR* or *AND*.

Select samples obtained from the search wizard and add to the experiment by selecting the ***Add*** option, or remove them from the list using the ***Remove*** option.

Either remove, or reorder files during the data loading step by using the ***Remove*** or ***Reorder*** button.

Figures 15.4 and 15.5, illustrate the process of choosing analysis type and loading data.

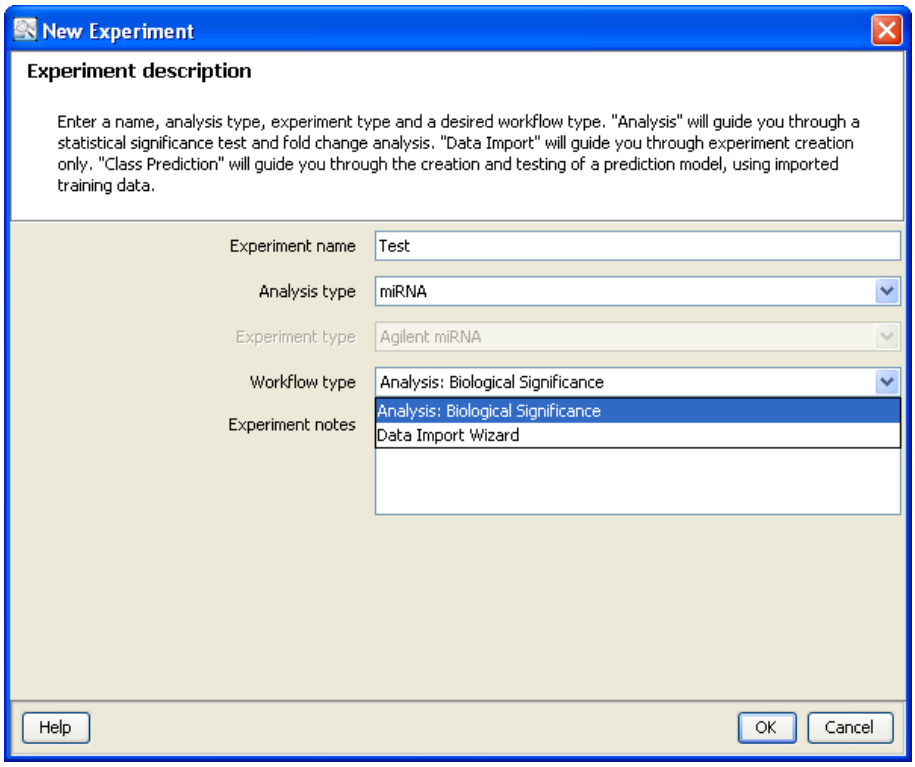


Figure 15.4: Experiment Selection

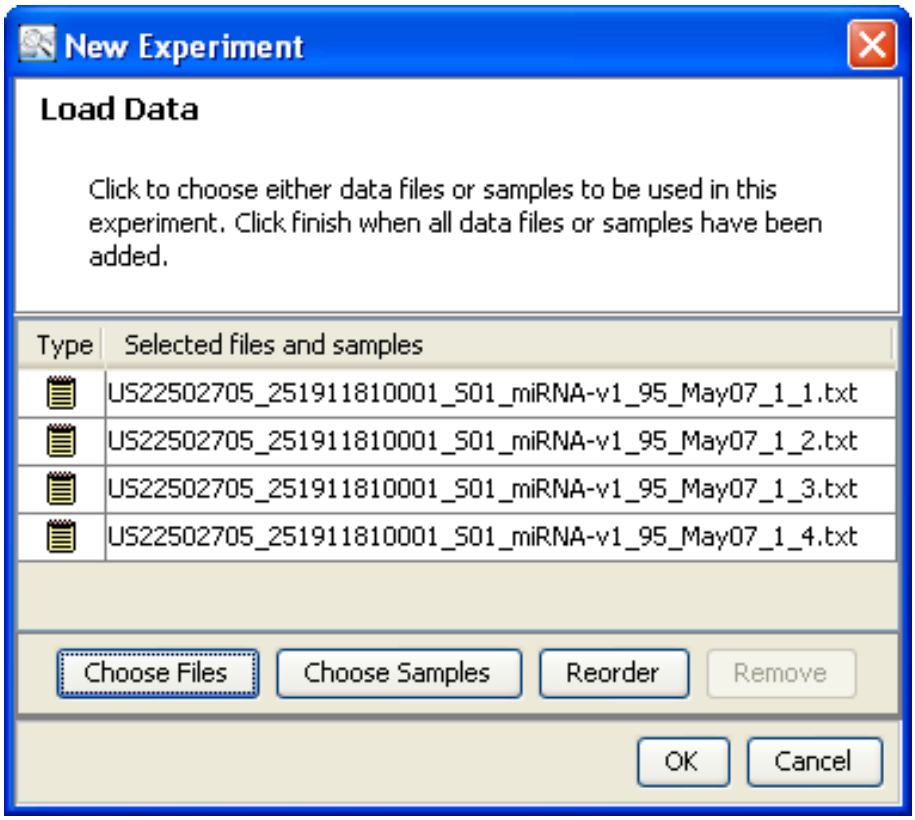


Figure 15.5: Load Data

15.1.1 Selecting technology for Agilent miRNA

Prior to the creation of a miRNA experiment, you should download the technology or technologies from the menu **Annotations → Create Technology → From Agilent Server or From Agilent eArray**. For each AMADID (Design ID) for catalog arrays, multiple technologies have been packaged on the **GeneSpring** update server. Each technology corresponds to a particular version of miRBase. You can view and update these from the list in **GeneSpring** UI by choosing: **Annotations ⟶ Update Technology Annotations ⟶ From Agilent Server**.
For more information refer to section on: Update Technology Annotations. After loading data, **GeneSpring** will prompt you to select a technology for creating the miRNA experiment. See Figure 15.6.

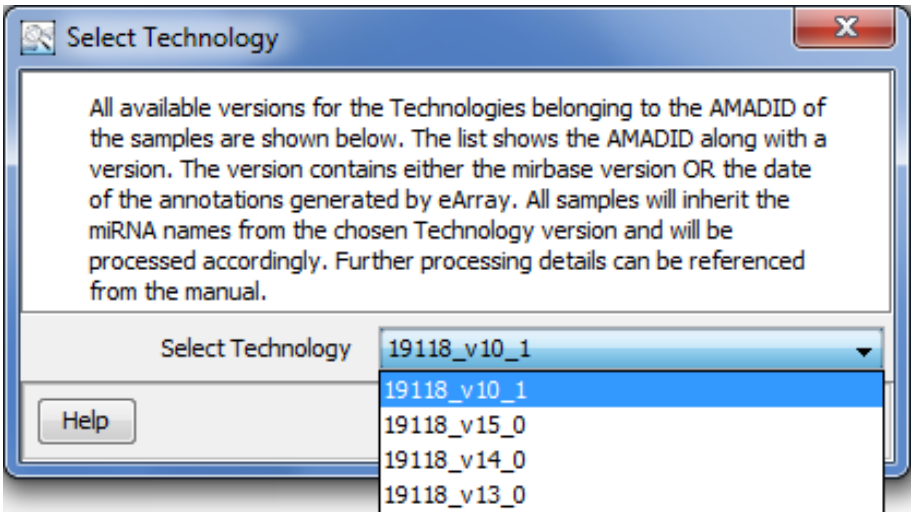


Figure 15.6: Select Technology in miRNA Experiment

The drop down menu in 'Select Technology' step has a list of all available technologies which have been downloaded. Upon selecting one and clicking OK, the Agilent miRNA workflow appears. If the **Analysis: Biological Significance** option is chosen, the Analysis: Biological Significance wizard appears with the sequence of steps on the left side highlighting the current step. The workflow allows you to proceed in a schematic fashion and does not allow you to skip steps. If the **Data Import Wizard** has been chosen, then the step 2 of 4 of miRNA experiment creation wizard appears. For more details on experiment creation via **Data Import Wizard**, refer to the section on Data Import Wizard.

15.1.1.1 Technology Creation

Technology for miRNA is no longer created on the fly. It can be created in two ways as follows:

- From Agilent eArray
- From GEML File

From Agilent eArray:

Technology can be created from eArray portal of Agilent. This can be accessed from:
Annotations*⟶ *Create Technology*⟶ *From Agilent eArray
An account for eArray is required to create technology using this feature. Details such as user name, password for your eArray account, AMADID Number (Agilent's microarray design identification numbers), organism, and technology type are required for creating an eArray technology. See Figure 15.7. Once the details are provided, a technology is created along with annotation columns like Entrez-ID, GO etc (if available) from the tab delimited text (TDT) file of the specified eArray. Using this AMADID number, **GeneSpring** fetches the GEML file from Agilent eArray that was used to generate the technology.

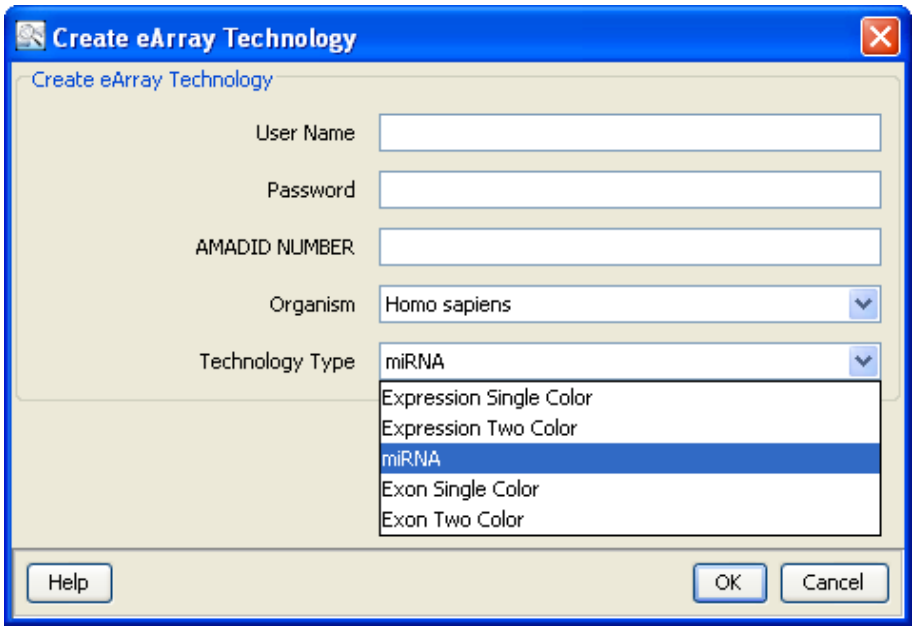


Figure 15.7: Creating Technology: From Agilent eArray

From GEML File:

Technology can also be created from a GEML file. This can be accessed from: ***Annotations⟶ Create Technology⟶ From Custom Agilent miRNA***

In this case, technology will be created after parsing the GEML file. You have to provide a path for the GEML file and also select an organism from the drop down list. See Figure 15.8.

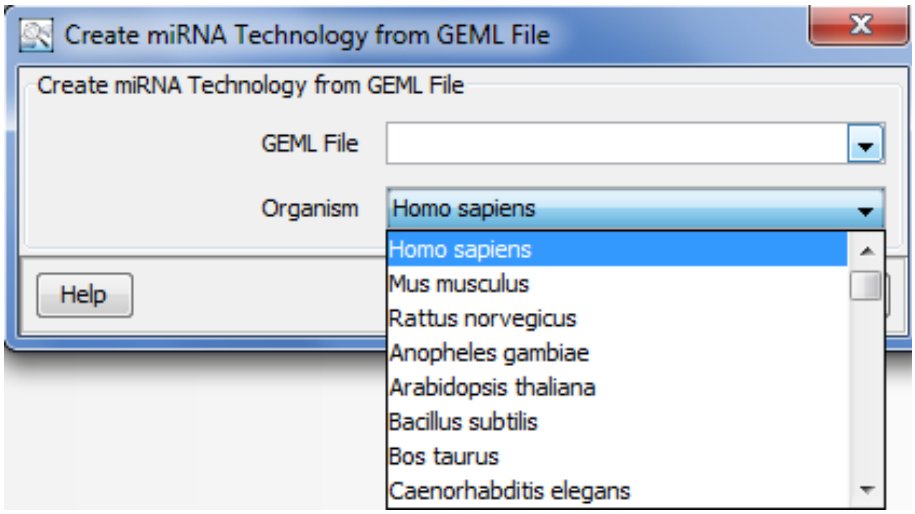


Figure 15.8: Creating miRNA Technology from GEML File

A technology name is created of the form: `Agilent.miRNA.<AMADID>-<miRBaseId>` when **GeneSpring** can find (from the Lookup table) a corresponding miRBase version for the grid date extracted from its GEML file. However, if the grid date in the GEML file

is not present in the Lookup table, the technology name will be created of the form: `Agilent.miRNA.<AMADID>-<GRID\_DATE>`.

Table 15.1 shows how the fields in the (GEML) XML file are extracted and mapped to their corresponding fields in the technology during technology creation.

Technology Annotation Column	**Field from GEML file**
AMADID	AMADID:_
miRBase ID	annotate_*to*_mirbaseversion
Grid date	date_ in $< Project >$ tag
miRNA name	systematic_name in reporter tag
Control_type	control_type in reporter tag
miRBase Accession no.	'mirbase accession No' tag
Chromosome	'chr' in 'probe-mappings' tag
Chr Start	'chr' in 'probe-mappings' tag
Chr End	'chr' in 'probe-mappings' tag
Strand	generated separately using 'start' and 'end' information
Sequence of Probe	'active_sequence' tag
MIMAT ID	'accessions' tag

Table 15.1: Probe Mappings from tags in GEML(XML) file

The technology will store the following two maps that will be later used for downstream analysis:

1. Agilent Probe IDs to its corresponding miRNA names (Systematic names).
2. miRNA names to its corresponding MIMAT accession numbers. This will be used to find targets for miRNAs across miRBase versions during TargetScan.

15.1.2 Sample validation

GeneSpring will attempt to match the set of samples loaded to a standard technology from the drop down list of technologies using a look up table. Please refer to section 15.2 for the Lookup table. This look up table maps ranges of grid dates to miRBase versions based on which the technologies are created. If it can find an exact match, then **GeneSpring** will highlight this technology by default during the 'Select Technology' step; and you can still override this selection by choosing another technology from the drop down list. But if there is no match (the range of grid dates in the sample file may point to more than one miRBase version ID or they may not find a matching miRBase ID in the table), then **GeneSpring** will prompt you to choose from the drop down list showing miRBase version IDs parsed from the available technologies. The drop down list in the 'Select Technology' step will display all the technologies available for the given Design ID across mirBase versions which have been updated in **GeneSpring** from the update server. It

is recommended that you download all the technologies available for the given Design ID used for experiment creation.

15.1.3 Backward Compatibility with old miRNA experiments

miRNA experiments created prior to **GeneSpring** version 11.5 can be successfully imported into this version using the menu **Project** → **Import Project**.
Note that the following will not work as expected:

- Add/remove samples will not be supported for older miRNA experiments since it requires rerunning experiment creation process. Experiment creation from version 11.5 requires additional mappings in the technology which are not present in the older technologies for miRNA.
- TargetScan will not work on old experiments unless you update annotations of the technology with the MIMAT (miRBase accession) IDs. Refer section Running TargetScan for older miRNA experiments for further details.
- Translation across the old and new miRNA experiments will not work unless the MIMAT accession IDs are updated in the old experiment as explained.

15.2 Data Processing for miRNA experiments

- **File formats**: The data files should be in text (*.txt*) format and obtained from Agilent Feature Extraction (FE) versions 10.7.3 or higher. **GeneSpring** supports the full file format and does not support the GeneView format files. For different FE versions, the probe name remains constant but the systematic name varies due to miRBase updates.
- **Sample Processing based on miRBase version**: Once the technology is created, a lookup table 15.2 is used to map the Grid dates to a particular miRBase version. If all samples belong to a particular grid date, then the technology is autoselected as shown in Figure 15.9 with the miRBase version corresponding to the Grid date. However, if the grid dates in the sample files are not all the same, then the user is asked to 'Select' the version from a drop down menu of miRBase versions as shown in Figure 15.10. Depending on the version selected, the technology will map the probe IDs to the corresponding miRNA names of the chosen technology.

  The technology uses the lookup table 15.2 to map the grid dates to the miRBase versions.

Grid date Range	miRBase version
$04/26/2007 <= griddate < 02/14/2008$	9.1
$02/14/2008 <= griddate < 11/29/2008$	10.1
$11/29/2008 <= griddate < 05/19/2009$	12
$05/19/2009 <= griddate < 10/30/2009$	13
$10/30/2009 <= griddate < 06/04/2010$	14
$06/04/2010 <= griddate < 10/22/2010$	15
$10/22/2010 <= griddate < 07/07/2011$	16
$07/07/2011 <= griddate < 11/04/2011$	17
$11/04/2011 <= griddate < 12/05/2012$	18
$12/05/2012 <= griddate < 03/20/2013$	19
$03/20/2013 <= griddate <$	20

Table 15.2: Lookup Table

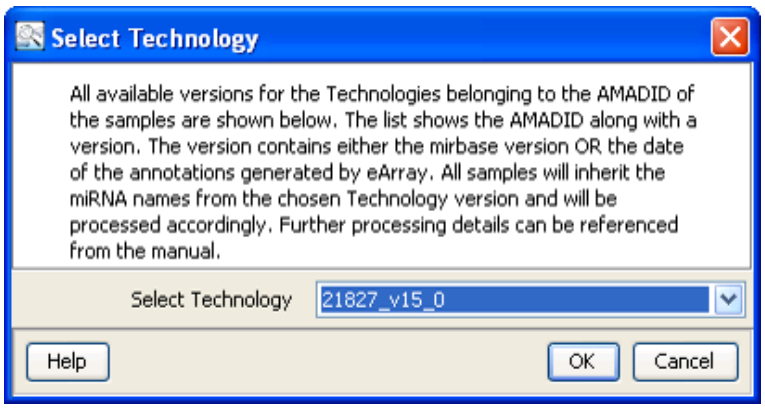


Figure 15.9: Sample processing based on miRBase version - Autoselected

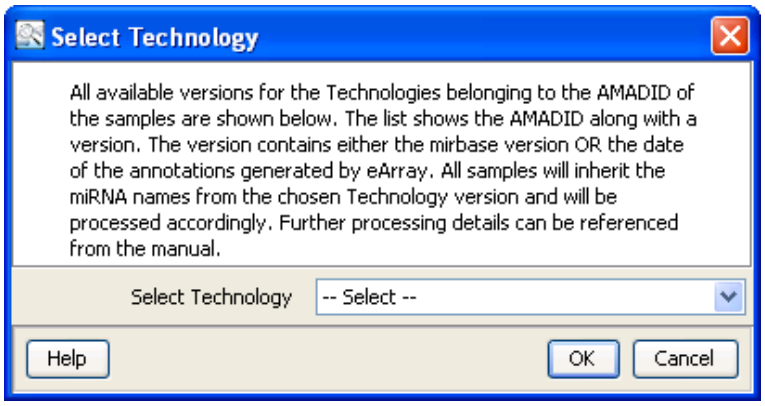


Figure 15.10: Sample processing based on miRBase version - Select from list

Note: There can be multiple technologies for a given AMADID number, corresponding to different miRBase versions.

- **Raw Signal Values**: The term ''raw'' signal values refer to the linear data after summarization and thresholding. Summarization is done by taking the sum of all the probes for a given miRNA and not the average or mean as is normally done for expression experiments.

- **Treatment of Probes**: Each miRNA is represented by atleast two different probes to give information about specificity. Different miRNA probes have different lengths and specificity. For each ProbeID in the FE sample file, three columns are considered for data analysis; gTotalProbeSignal, gTotalProbeError and gIsGeneDetected . There can be multiple probe names for a given systematic name and a given probe is present in replicates. If probes are in replicates, the first gTotalProbeSignal value is picked. Using the unique Probe IDs and a map to the systematic name provided in an annotation file (different for each miRBase version), each ProbeID is mapped to a given Systematic name.

- **Filtering by Flags**: a filtering step is done wherein both flag (IsGeneDetected column) and error is taken into account. The flag value of a particular probeset is dependant on the flag values of the probes in it. In filtering by flag values, only those rows are used for summarizations that have a flag value of 1, ignoring values that are Not Detected. The 'gIsGeneDetected' is taken as flag column and a value of 0 is considered as 'Not Detected or Marginal' and 1 is considered as 'Detected'. For each miRNA, if all are 0 (Marginal) then the signal value is taken as 0.1. If some values are 0 and some are 1, then **GeneSpring** uses only those rows for summarization which have value of 1. Unlike expression arrays, there is no 'Compromised' flag.

- **Filtering by Error**: Rows that passed the Flag filter are now filtered for Error (gTotalProbeError column). If the signal value is greater than 3 times the error value, only then the signal value passes the filter and a value of 0 gets assigned for error. If not, an error value of 1 gets assigned and those rows are disregarded. Error for such is the Root Mean Square of all values.

- **Summarization**: Following filtering and before we do the thresholding, the multiple probes that map to a given systematic name are first summed up. The signal value is computed as the SUM of all values and not average. Summarization is based on the systematic name as explained above and is done before thresholding of signal values.

  Table 15.3 shows how the processing is done (For example: for miRNA 'hsa-miR-216') during probe summarization and error handling:

Systematic-Name	gTotalProbeSignal	gTotalProbeError	gIsGeneDetected	Comments
hsa-miR-216	10	3	1	Pass
hsa-miR-216	15	3	1	Pass
hsa-miR-216	12	3	0	Fail because though Signal > 3 times error but flag is 0
hsa-miR-216	11	4	1	Fail because Signal < 3 times error even though flag is 1

Table 15.3: Example: Probe Summarization

So total signal is calculated as: $10 + 15 = 25$

- **Thresholding**: Following filtering and summarization, thresholding is done only for all the valid signals.
- **Normalization**: Various criteria for preprocessing of input data is set here. It allows you to threshold raw signals to chosen values and select normalization algorithms (None, Percentile Shift, Scale, Quantile, Normalize to control genes, or Normalize to External Value).
- **Treatment of Control probes**: All probes named NC* and Negative Control are control probes which report signals near background levels. The control probes are included while performing Normalization.
- **Missing Values**: The missing values in the raw dataset are now being assigned a value of 0.1. Though these will not be considered for thresholding, they will be used in all further downstream analysis.
- **Sequence of events**: The sequence of events involved in the processing of the data files is: Summarization $\longrightarrow$ Thresholding $\longrightarrow$ Log Transformation $\longrightarrow$ Normalization $\longrightarrow$ Baseline Transformation.
- **Changes from GeneSpring GX 11.0.2 to 11.5**: There is a change in the way data is processed before and after Agilent Feature Extraction version 10.7.3. Files extracted in FE versions less than 10.7.3 were handled differently. Earlier technology was created on the fly so that no annotations were associated with each miRNA. Thresholding is now done after adding up all the probe signals for a particular miRNA.

15.3 Analysis: Biological Significance

15.3.1 Summary Report (Step 1 of 8)

The Summary report displays the summary view of the created experiment. It shows a Box Whisker plot, with the samples on the X-axis and the Normalized Intensity Values on the Y axis. An information message on the top of the wizard shows the number of samples in the file and the sample processing details. By default, the *Analysis: Biological Significance* does a thresholding of the signal values to 1. It then normalizes the data to 90^th percentile and does not perform baseline transformation. If the number of samples are more than 30, they are only represented in a tabular column. Select the ***Next*** option to proceed to the next step and on clicking ***Finish***, an entity list will be created on which analysis can be done. By placing the cursor on the screen and selecting by dragging on a particular probe, the probe in the selected sample as well as those present in the other samples are displayed in green. Figure 15.11 illustrates the Summary Report with a box-whisker plot.

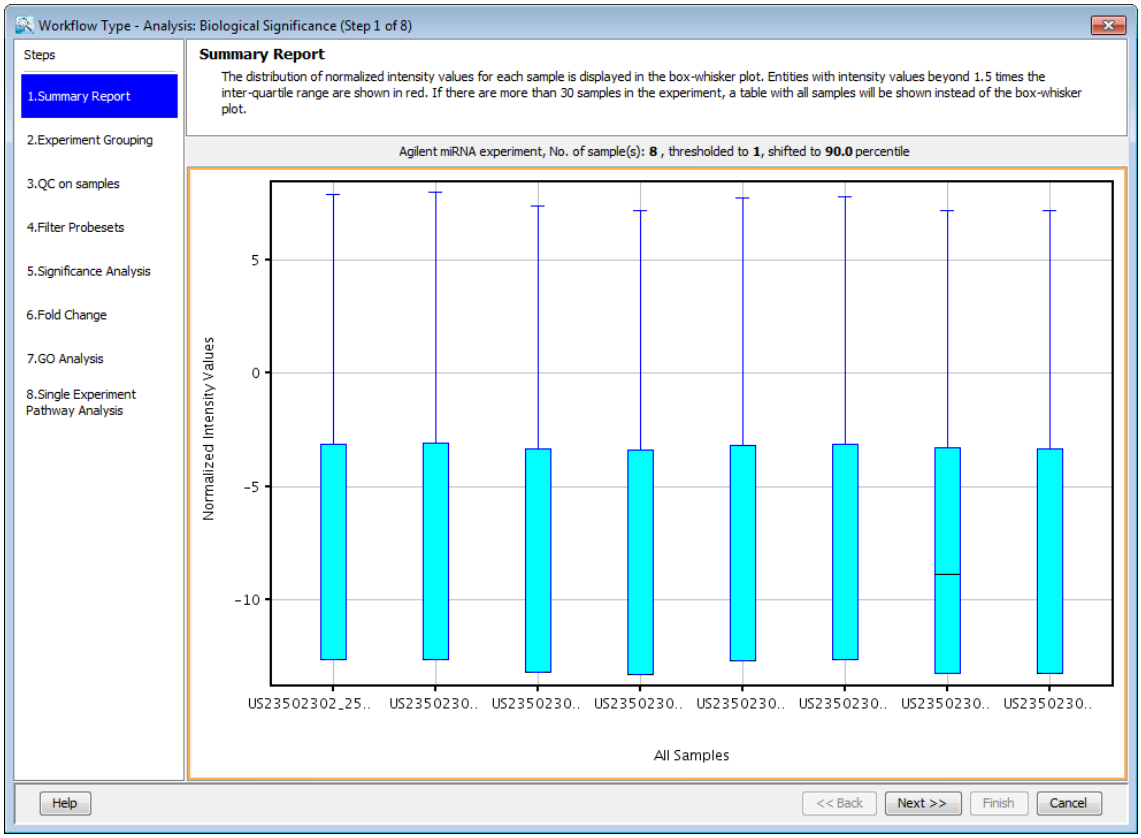


Figure 15.11: Summary Report

Note: In *Analysis: Biological Significance*, these default parameters cannot be changed. To choose different parameters use *Data Import Wizard*.

15.3.2 Experiment Grouping (Step 2 of 8)

On clicking ***Next***, the *Experiment Grouping* window appears which is the 2nd step in the **Analysis: Biological Significance** workflow. It requires parameter values to be defined to group samples. Samples with same parameter values are treated as replicates. To assign parameter values, click on the ***Add parameter*** button. Parameter values can be assigned by first selecting the desired samples and assigning the corresponding parameter value. For removing any value, select the sample and click on ***Clear***. Press ***OK*** to proceed. Although any number of parameters can be added, only the first two will be used for analysis in the **Analysis: Biological Significance** workflow. The other parameters can be used in the **Data Import Wizard**.

Note: The *Analysis: Biological Significance* workflow does not proceed further without grouping information.

Experimental parameters can also be loaded externally by clicking on Load experiment parameters from file icon button. The file containing the *Experiment Grouping* information should be a tab or comma separated text file. The experimental parameters can also be imported from previously used samples, by clicking on Import parameters from samples icon. In case of file import, the file should contain a column containing sample names; in addition, it should have one column per factor containing the grouping information for that factor. Here is an example of a tab separated text file.

Sample genotype dosage

A1.txt NT 20
A2.txt T 0
A3.txt NT 20
A4.txt T 20
A5.txt NT 50
A6.txt T 50

Reading this tab file generates new columns corresponding to each factor.

The current set of experiment parameters can also be saved to a local directory as a tab separated or comma separated text file by clicking on the Save experiment parameters to file icon button. These saved parameters can then be imported and used for future analysis. In case of multiple parameters, the individual parameters can be re-arranged and moved left or right. This can be done by first selecting a column by clicking on it and using the Move parameter left icon to move it left and Move parameter right icon to move it right. This can also be accomplished using the Right click$\longrightarrow$ *Properties*$\longrightarrow$ *Columns* option. Similarly, parameter values, in a selected parameter column, can be sorted and re-ordered, by clicking on Re-order parameter values icon. Sorting of parameter values can also be done by clicking on the specific column header.

Unwanted parameter columns can be removed by using the Right-click$\longrightarrow$ *Properties* option. The *Delete parameter* button allows the deletion of the selected column. Multiple parameters can be

deleted at the same time. Similarly, by clicking on the *Edit parameter* button the parameter name as well as the values assigned to it can be edited.

> Note: *Analysis: Biological Significance* workflow by default creates averaged and unaveraged interpretations based on parameters and conditions. It takes average interpretation for analysis in the Analysis: Biological Significance wizard.

Windows for Experiment Grouping and Parameter Editing are shown in Figures 15.12 and 15.13 respectively.

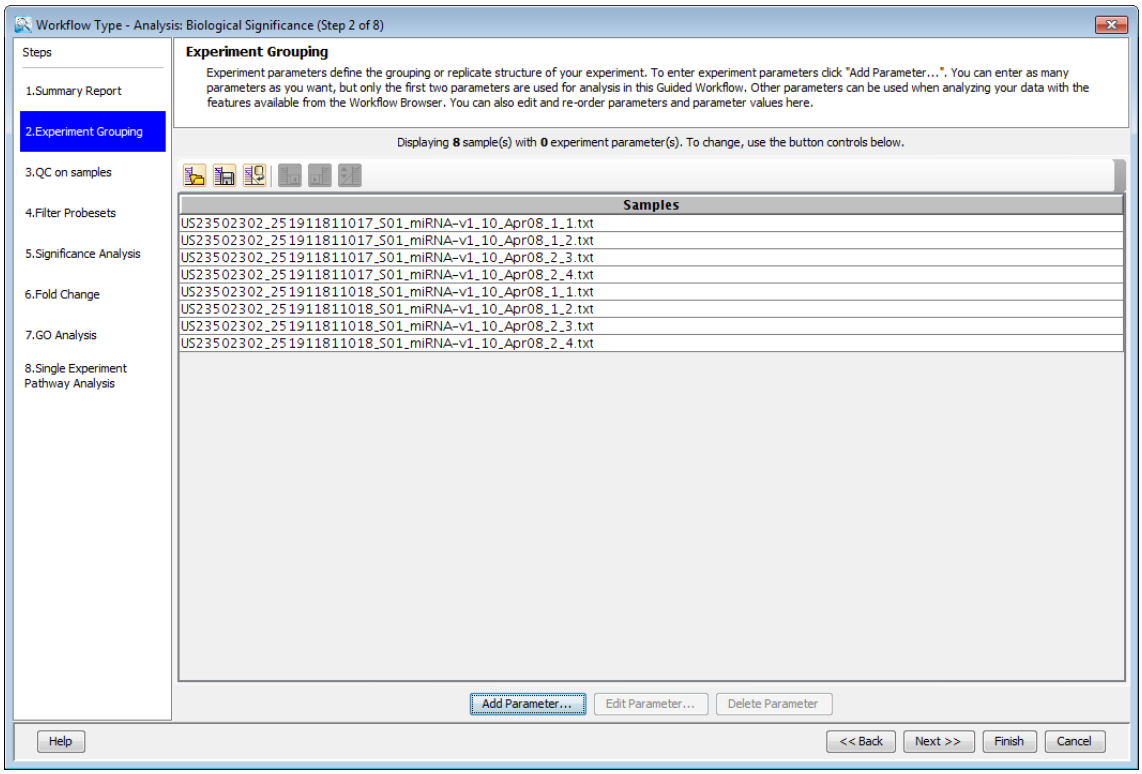


Figure 15.12: Experiment Grouping

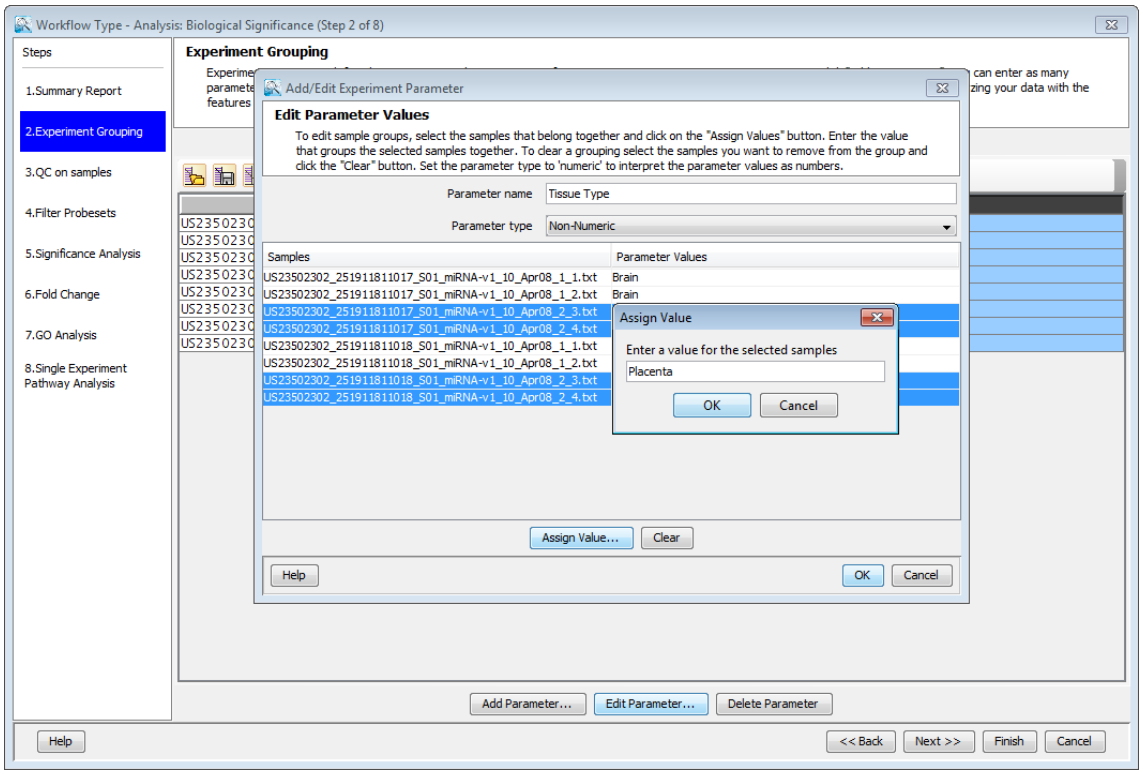


Figure 15.13: Add/Edit Parameters

15.3.3 Quality Control (QC) on samples (Step 3 of 8)

The 3rd step in Analysis: Biological Significance is the QC on samples which is displayed in the form of four tiled windows. They are as follows:

- Quality Controls Metrics- Report and Experiment grouping tabs
- Quality Controls Metrics- Plot
- 3D PCA Scores.
- Legend

These four tiled windows are illustrated in Figure 15.14.

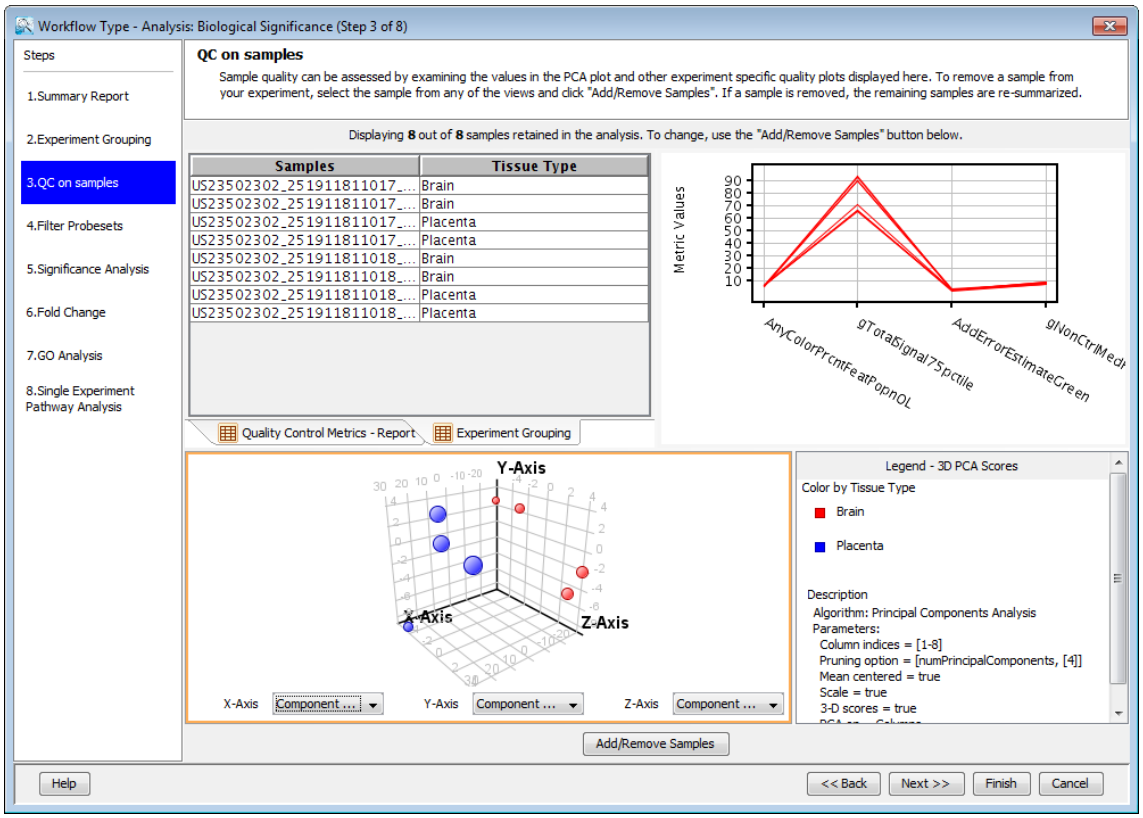


Figure 15.14: Quality Control on Samples

The *Experiment Grouping* tab has the grouping information provided by you in the Step 2.

The QC metrics report aids you in evaluating the reproducibility and reliability of the microarray data. The quality metrics scores are obtained directly from the sample file. Please refer to Agilent Feature Extraction Software Reference Guide (v10.7.3 or higher) for detailed descriptions of the QC Metrics.

Principal Component Analysis (PCA) calculates the PCA scores and visually represents them in a 3D scatter plot. These scores can also be used to check data quality. It displays one point per array and is colored by the *Experiment Factors* provided earlier in the *Experiment Groupings* view. This allows viewing of separations between groups of replicates. Ideally, replicates within a group should cluster together and separately from arrays in other groups. The PCA components, represented in the X, Y and Z axes are numbered 1, 2, 3... according to their decreasing significance. The 3D PCA scores plot can be customized via ***Right-Click*** $\longrightarrow$ ***Properties***. To zoom into a 3D Scatter plot, press the Shift key and simultaneously hold down the left mouse button and move the mouse upwards. To zoom out, move the mouse downwards instead. To rotate, press the Ctrl key, simultaneously hold down the left mouse button and move the mouse around the plot.

The *Add/Remove* samples option can be used to remove unsatisfactory samples and to add the samples back if required. Whenever samples are removed or added back, normalization is performed again. Click on ***OK*** to proceed.

Note:Add/remove samples will not be supported for older miRNA experiments since Add/Remove requires rerunning experiment creation process and now the experiment creation process requires additional mappings in the technology which are not there in the older technologies for miRNA.

The fourth window shows the legend of the active QC tab.

15.3.4 Filter Probesets (Step 4 of 8)

In this step, the entities are filtered based on their flag values *Detected* and *Not Detected*. Information pertaining to the flags is present in the data file. **GeneSpring** considers the "gIsGeneDetected" as the flag column and marks entities having '0' as *Not Detected* and '1' as *Detected*. Only entities having the Detected flag in at least 1 sample are displayed in the profile plot. The selection can be changed using *Rerun Filter* option. The plot is generated using the normalized signal values and samples grouped by the active interpretation. Options to customize the plot can be accessed via the Right-click menu. An *Entity List*, corresponding to this filtered list, will be generated and saved in the Navigator window. The Navigator window can be viewed after exiting from *Analysis: Biological Significance*. Double clicking on an entity in the Profile Plot opens up an *Entity Inspector* giving the information corresponding to the selected entity. Newer annotations can be added and existing ones removed using the *Configure Columns* button. An additional tab in the *Entity Inspector* shows the raw and normalized values for that entity. A plot which shows the distribution of the normalized intensity values of that entity over the current interpretation is present as a tab in the same window. The cutoff for filtering can be changed using the *Rerun Filter* button. Newer Entity lists will be generated with each run of the filter and saved in the Navigator. The information message on the top shows the number of entities satisfying the flag values.

Figures 15.15 and 15.16 illustrate the profile plot obtained with one and two parameters.

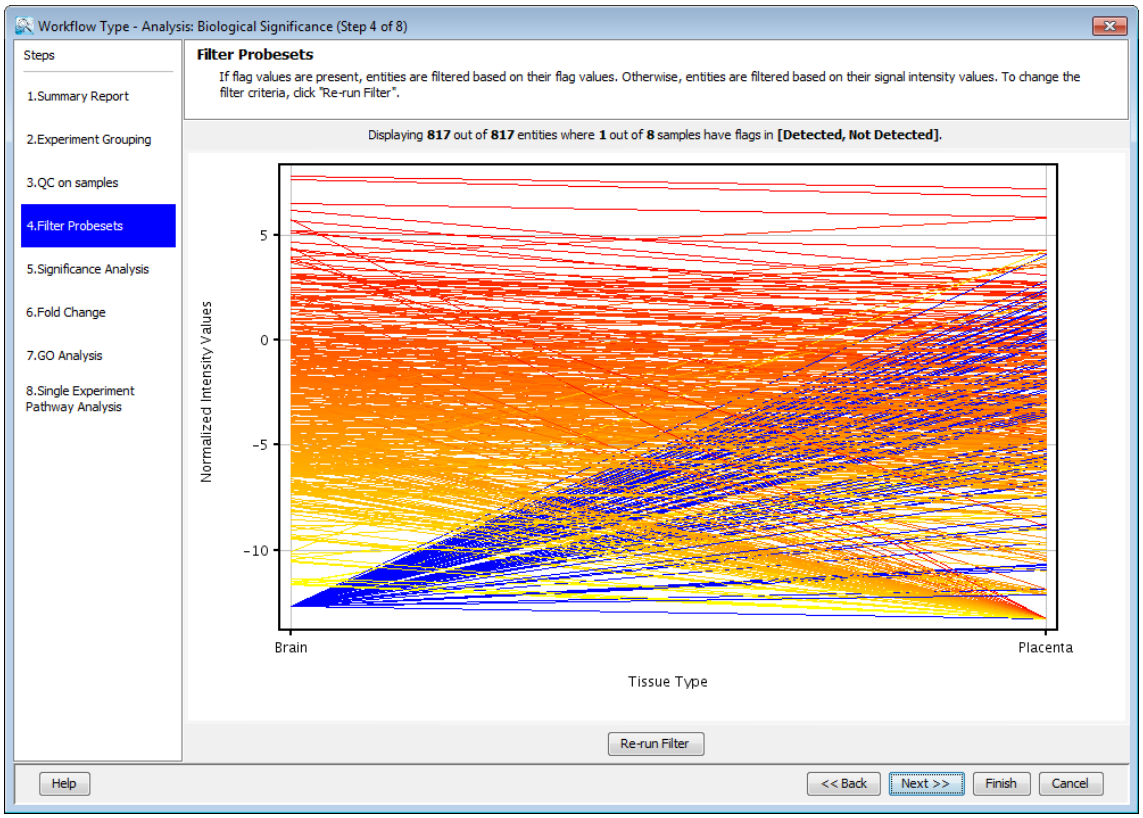


Figure 15.15: Filter Probes-Single Parameter

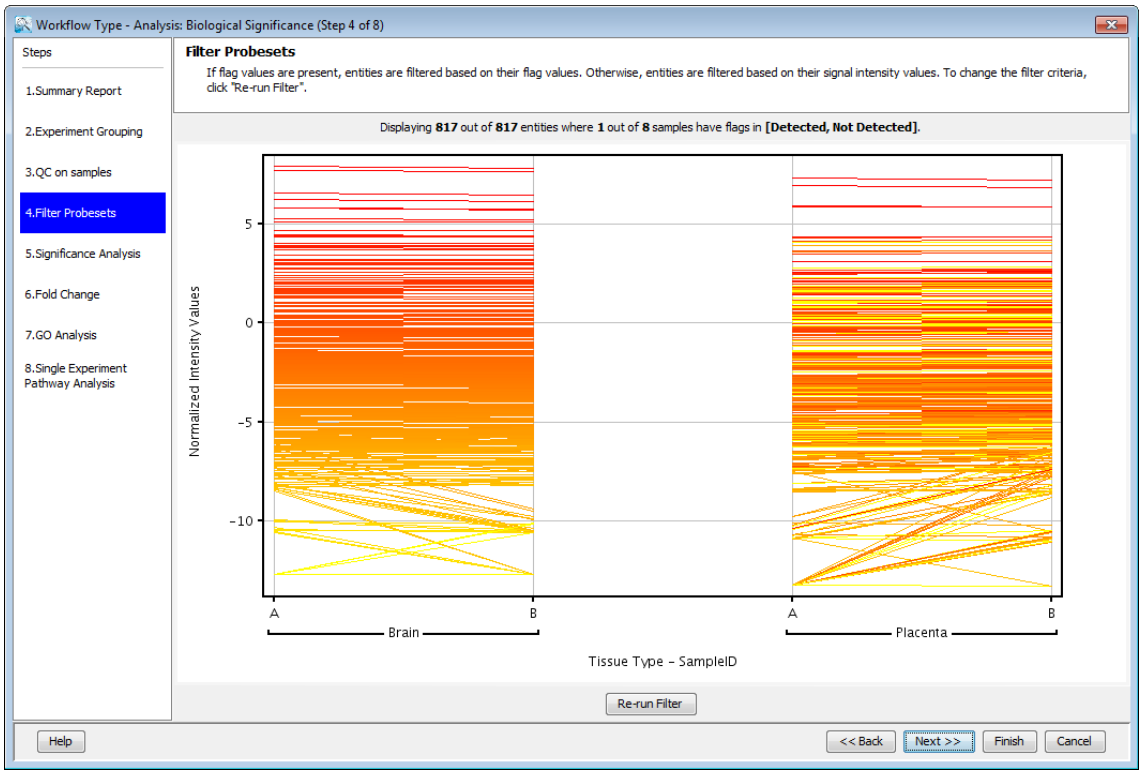


Figure 15.16: Filter Probes-Two Parameters

15.3.5 Significance Analysis (Step 5 of 8)

miRNAs are filtered based on p-values calculated from statistical analysis. To apply a new p-value cut-off, use the **p-value cut-off** slider at the bottom of the window. If none of the miRNAs pass the p-value cut-off, the workflow is not able to proceed. Depending upon the sample grouping of the experiment, **GeneSpring** performs either t-test or Analysis of Variance (ANOVA). The tables below illustrate the types of statistical test performed given a specific experimental grouping with examples:

- **Example Sample Grouping I:** The example outlined in Table 15.4 has two groups, Normal and Tumor, each with replicates. In such a situation, **GeneSpring** performs a Filter on Volcano Plot and an unpaired t-test.

Samples	Grouping
S1	Normal
S2	Normal
S3	Normal
S4	Tumor
S5	Tumor
S6	Tumor

Table 15.4: Sample Grouping and Significance Tests I

- **Example Sample Grouping II:** In this example (see Table 15.5), only one group, Tumor, is present. The Filter on Volcano Plot and *t*-test against zero is performed here.

Samples	Grouping
S1	Tumor
S2	Tumor
S3	Tumor
S4	Tumor
S5	Tumor
S6	Tumor

Table 15.5: Sample Grouping and Significance Tests II

- **Example Sample Grouping III:** When three groups are present (in this example Normal, Tumor1 and Tumor2) and one of the groups (Tumor2 in this case) does not have replicates, statistical analysis cannot be performed.

NOTE

In cases where some conditions do not have replicates, statistical analysis can only be performed on conditions with replicates if the samples representing the conditions with out replicates (Sample 6 in this example) are removed from the **Interpretation** used for analysis. Therefore, we recommend that you create a new **Interpretation** (excluding the condition without replicate samples) once experiment creation is complete and re-run statistical analysis using the option available from the **Workflow Browser**.

Samples	Grouping
S1	Normal
S2	Normal
S3	Normal
S4	Tumor1
S5	Tumor1
S6	Tumor2

Table 15.6: Sample Grouping and Significance Tests III

- **Example Sample Grouping IV:** When there are three groups within an Interpretation and each group has replicates (see Table 15.7), **GeneSpring** performs a one-way ANOVA.

Samples	Grouping
S1	Normal
S2	Normal
S3	Tumor1
S4	Tumor1
S5	Tumor2
S6	Tumor2

Table 15.7: Sample Grouping and Significance Tests IV

- **Example Sample Grouping V:** The Table 15.8 shows an example of an experiment where two parameters are present. The presence of two parameters means that **GeneSpring** looks for samples in all the possible parameter value pairings (in this example: Normal/10 min, Normal/50 min, Tumor/10 min, Tumor/50 min). However, there are no samples for the conditions Normal/50 min and Tumor/10 min. Due to the absence of such samples, no statistical significance test is performed.

Samples	Grouping A	Grouping B
S1	Normal	10 min
S2	Normal	10 min
S3	Normal	10 min
S4	Tumor	50 min
S5	Tumor	50 min
S6	Tumor	50 min

Table 15.8: Sample Grouping and Significance Tests V

- **Example Sample Grouping VI:** For experiments with the kind of experiment grouping illustrated in Table 15.9, a two-way ANOVA is performed.

Samples	Grouping A	Grouping B
S1	Normal	10 min
S2	Normal	10 min
S3	Normal	50 min
S4	Normal	50 min
S5	Tumor	50 min
S6	Tumor	50 min
S7	Tumor	10 min
S8	Tumor	10 min

Table 15.9: Sample Grouping and Significance Tests VI

- **Example Sample Grouping VII:** In the example shown in Table 15.10, a two-way ANOVA is performed. However, the result only contains a p-value for each parameter, *i.e.* for Grouping A and Grouping B. The p-value for the combined parameters, Grouping A- Grouping B is not computed. In this particular example, there are six conditions (Normal/10min, Normal/30min, Normal/50min, Tumor/10min, Tumor/30min, Tumor/50min), which is the same as the number of samples. The p-value for the combined parameters can only be computed when the number of samples exceeds the number of possible groupings.

Samples	Grouping A	Grouping B
S1	Normal	10 min
S2	Normal	30 min
S3	Normal	50 min
S4	Tumor	10 min
S5	Tumor	30 min
S6	Tumor	50 min

Table 15.10: Sample Grouping and Significance Tests VII

Results are displayed in the following ways:

- ***t*-test**: *t*-test unpaired is chosen as a test of choice for experimental samples as shown in Table 15.4. Upon completion of the *t*-test the results are displayed as tiled windows.
  - A ***p*-value table** consisting of Probe Names, *p*-values, corrected *p*-values, and Fold change. If an entity is 2-fold up or 2-fold down, it will be called +2.0 or -2.0 respectively. The fold change value is colored in Red or Blue by default for up-regulated and down-regulated entities, respectively.

– The **Differential expression analysis report** containing the Test Description, *i.e.* the test that has been used for computing p-values, as well as the selected type of p-value computation (**Asymptotic** or **Permutative**) and multiple testing correction. This report also provides a result summary that indicates the number of entities satisfying different p-value cut-offs.

– A **Volcano plot** is only present if there are two groups provided in Experiment Grouping. In this plot, the entities are colored depending on their passing the default cut-offs:

  * **Red**: Entities which pass both **Corrected p-value** and **Fold change** cut-offs and are up-regulated.
  * **Blue**: Entities which pass both **Corrected p-value** and **Fold change** cut-offs and are down-regulated.
  * **Grey**: Entities which neither pass the **Corrected p-value** cut-off nor the **Fold change** cut-off.

  The Volcano plot shows the negative log_{10} of p-value vs log_2 of fold change. Entities with large fold-change and low p-value are easily identifiable on this view.

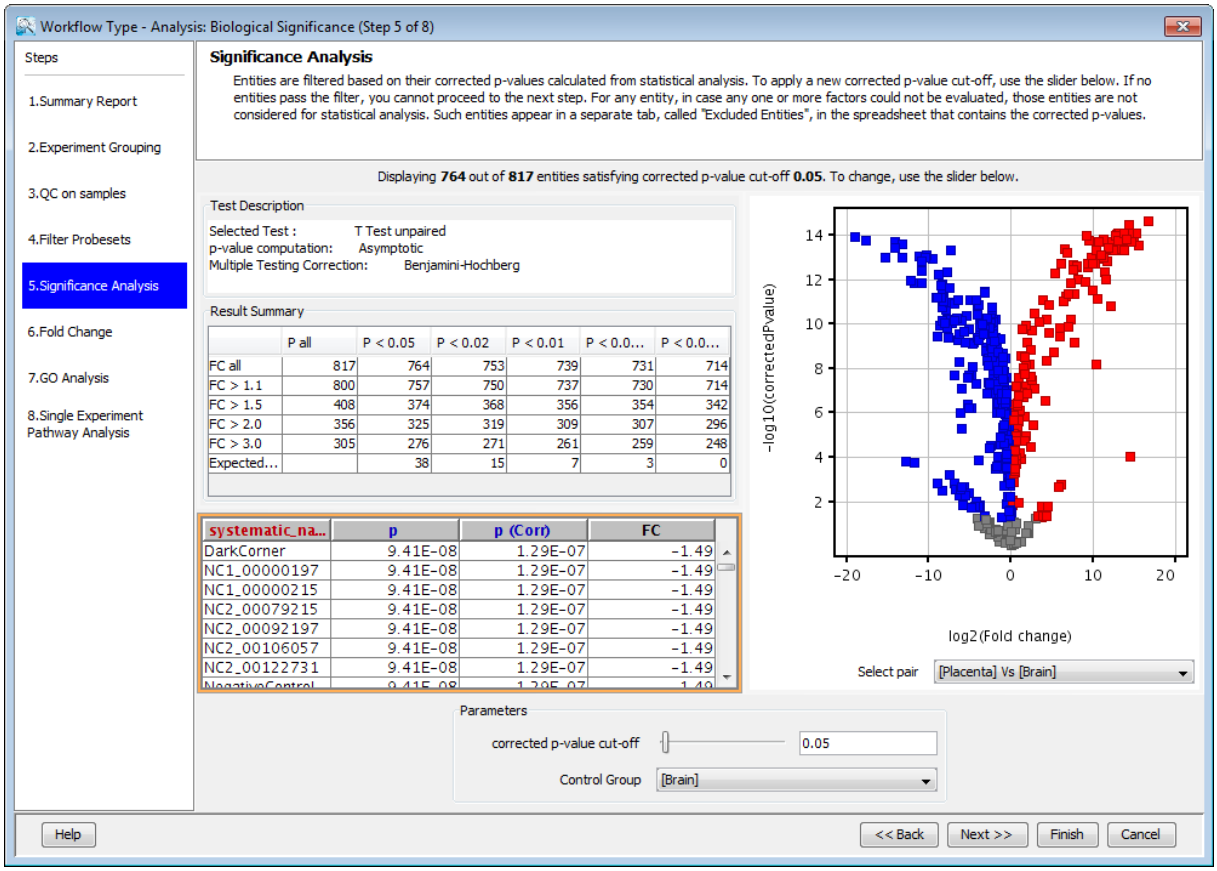


Figure 15.17: Significance Analysis T-Test

- **ANOVA**: For 2-way ANOVAs, the p-value table contains SS ratio columns in addition to the Fold change columns and the Volcano plot is replaced by a Venn Diagram:

– A *p*-**value table** consisting of Probe Names, *p*-values, corrected *p*-values, SS ratio and Fold change. The SS ratio is the mean of the sum of squared deviates (SSD) as an aggregate measure of variability between and within groups.

– The **Differential expression analysis report** containing the Test Description, *i.e.* the test that has been used for computing *p*-values, as well as the selected type of *p*-value computation (**Asymptotic** or **Permutative**) and multiple testing correction. This report also provides a result summary that indicates the number of entities satisfying different *p*-value cut-offs.

– A Venn Diagram that reflects the union and intersection of entities passing the *p*-value cut-off. You can save any section of the Venn Diagram as a separate entity list by selecting it and clicking the **Save custom list...** button.

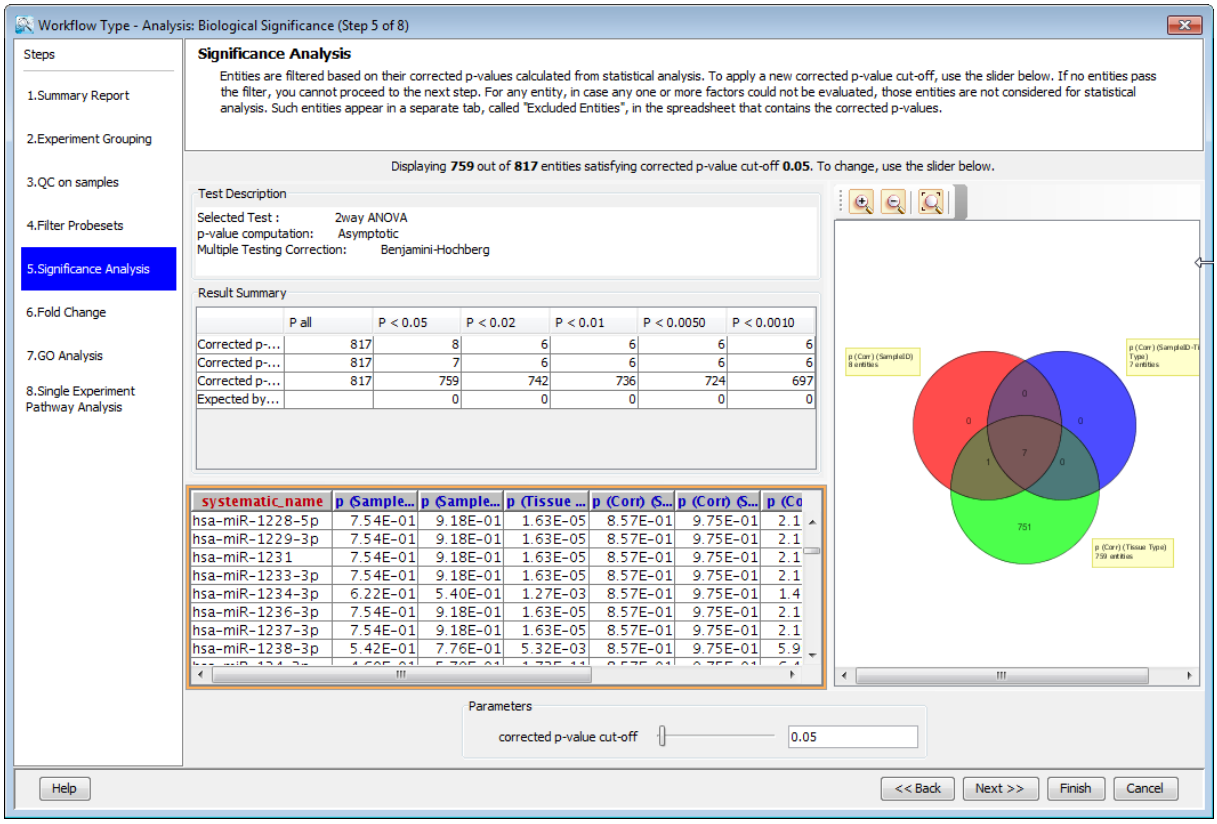


Figure 15.18: Significance Analysis ANOVA

Note: If the experiment grouping has two parameters, and one of the parameters has only one condition, significance analysis will be carried out using the other parameter. If both parameters, have only one condition, significance analysis is carried out on the first parameter only.

15.3.6 Fold Change (Step 6 of 8)

Fold change analysis is used to identify genes with expression ratios or differences between a treatment and a control that are outside of a given cut-off or threshold. Fold change is calculated between any 2 conditions, Condition 1 and Condition 2. The ratio between Condition 2 and Condition 1 is calculated (Fold change = Condition 1/Condition 2). Fold change gives the absolute ratio of normalized intensities (no log scale) between the average intensities of the samples grouped. The entities satisfying the significance analysis are passed on for the fold change analysis. The wizard shows a table containing the Probe ID, Fold change value in normal and log scale; the Regulation column and the Absolute fold change value columns can also be shown here by changing the defaults from the menu **Tools** → **Options** → **Miscellaneous** → **Fold Change Settings**. The fold change and log fold change values are colour coded (red for up regulation and blue for down regulation, by default; can be changed from the configuration menu). The absolute fold change values are neither colour coded nor carry a sign. The regulation column depicts which one of the groups has greater or lower intensity values with respect to the other group. The cut off can be changed using ***Re-run Filter***. The default cut off is set at 2.0 fold. So it shows all the entities which have fold change values greater than or equal to 2. The fold change value can be manipulated by either using the sliding bar (goes up to a maximum of 10.0) or by typing in the value and pressing Enter. Fold change values cannot be less than 1.

Along with this table, a profile plot is also generated and shown. Up-regulated entities are shown in red while down regulation ones are shown in blue. The color can be changed using the Right-click⟶ *Properties* option. Double click on any entity in the plot shows the *Entity Inspector* giving the annotations corresponding to the selected entity. An entity list will be created corresponding to entities which satisfied the cut-off in the experiment Navigator.

Fold Change view with the spreadsheet and the profile plot is illustrated in Figure 15.19.

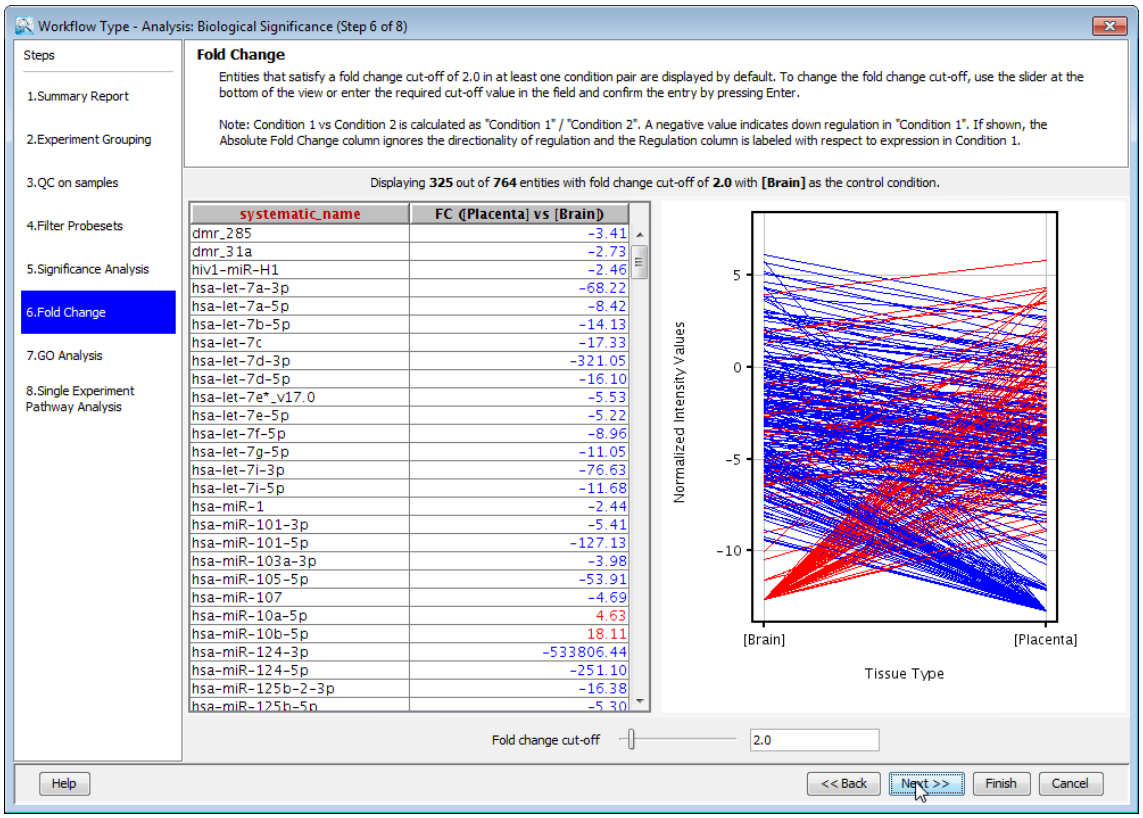


Figure 15.19: Fold Change

The Analysis: Biological Significance guided workflow in miRNA experiments ends after the **Fold Change** step. Click **Next** or **Finish** to exit the workflow and see the results in the Experiment Navigator.

The default parameters for miRNA experiment creation from the option of *Analysis: Biological Significance* are summarized in Table 15.11.

15.4 Data Import Wizard Workflow

The *Data Import Wizard* offers a variety of choices for the analysis. Flag options can be changed and raw signal thresholding can be altered. Additionally, there are options for baseline transformation of the data and for creating different interpretations. To create and analyze an experiment using the *Data Import Wizard*, load the data as described earlier. In the **New Experiment Dialog**, choose the **Workflow Type** as Data Import Wizard. Clicking on ***OK*** will open a new experiment wizard and proceeds as follows:

1. **Load Data (Step 1 of 4):**

	Parameters	**Parameter values**
Expression Data Transformation	Thresholding	1.0
	Normalization	Shift to 90th Percentile
	Baseline Transformation	Not Applicable
	Summarization	Not Applicable
Filter by		
1.Flags	Flags Detected	Retained
2.Expression Values	(i) Upper Percentile cutoff	Not Applicable
	(ii) Lower Percentile cutoff	
Significance Analysis	p-value computation	Asymptotic
	Correction	Benjamini-Hochberg
	Test	Depends on Grouping
	p-value cutoff	0.05
Fold change	Fold change cutoff	2.0

Table 15.11: Table of Default parameters for Analysis: Biological Significance

Similar to *Analysis: Biological Significance*, either data files can be imported or pre-created samples can be used.

- For loading new data files, use ***Choose Files***.
- If the data files have been previously used in **GeneSpring** experiments, select ***Choose Samples***.

Experiment Creation Step 1 of 4, the 'Load Data' window, is illustrated in Figure 15.20.

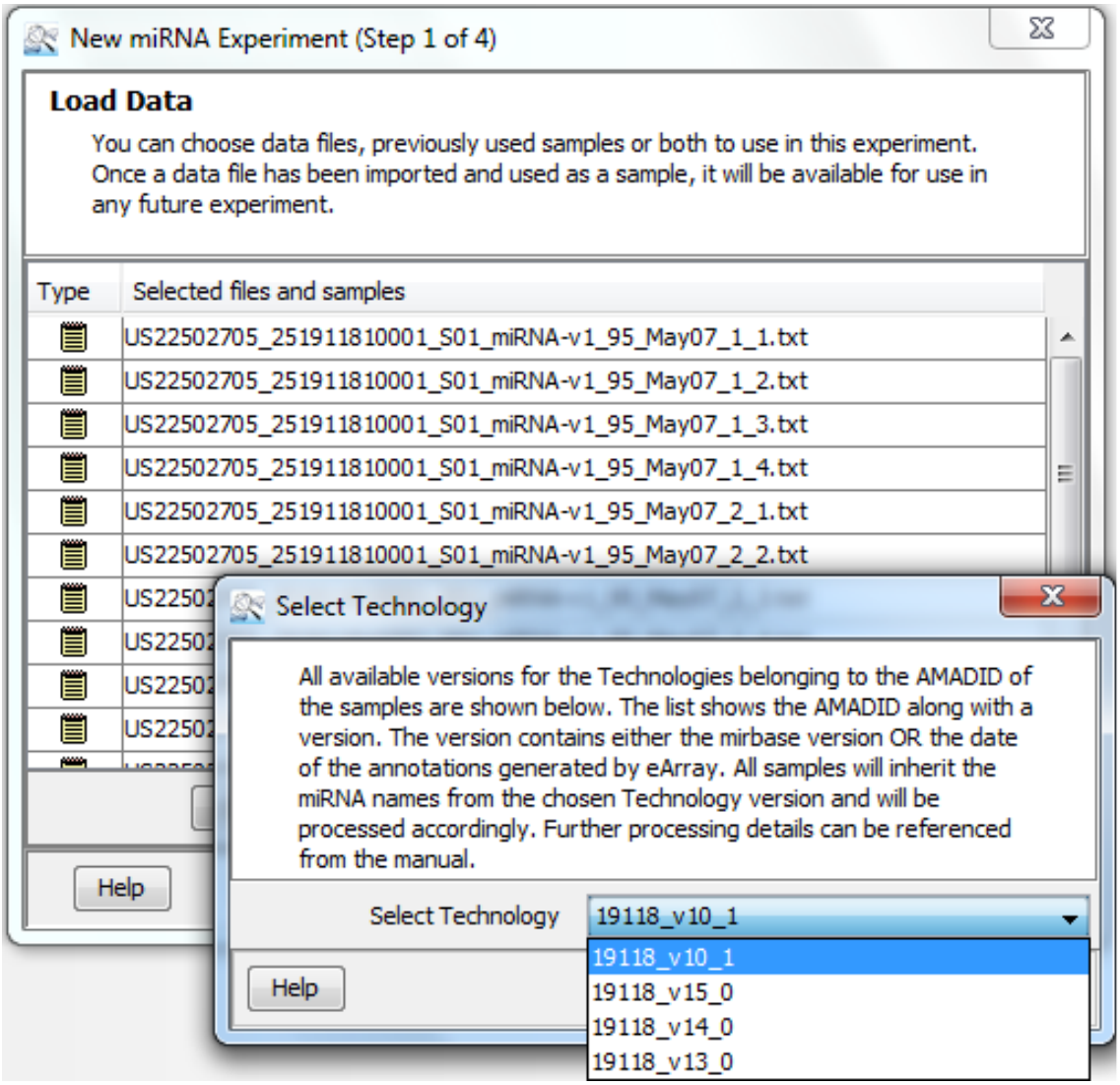


Figure 15.20: Load Data

After loading data, **GeneSpring** will prompt you to select a standard technology for creating the miRNA experiment. Refer section Selecting technology for miRNA.

2. **Normalization Options (Step 2 of 4):**

   Criteria for preprocessing of input data should be set here. It allows you to threshold raw signals to chosen values and select normalization algorithms . The gTotalGeneSignal from FE output which is already background subtracted is brought in. All additional processing steps are performed on this column.

   Normalization algorithm:

   - **None:** No normalization is done. The threshold value for raw signals is set to 1.0 and can be altered and all other options are disabled.
   - **Percentile Shift:** On selecting this normalization method, the **Percentile target** option gets enabled allowing you to enter a specific percentile value using which

normalization is performed. The default is set at 90.

- **Scale:** On selecting this normalization method, you will have an an option to either scale it to the percentile/mean of all samples or to scale it to the percentile/mean of control samples. The Threshold raw signals is set to 1.0. On choosing the latter, the user has to select the control samples from the available samples in the **Choose Samples** option. The **Percentile target** option is enabled and the percentile is set at a default value of 90. The default is set as scale to percentile of all samples.
- **Quantile:** Makes all statistical parameters of the sample like, mean, median, and percentile identical. The Percentile target box is disabled. The user does not have to enter any specifications for this normalization.
- **Normalize to control genes:** After selecting this option, you have to specify the control genes in Step 3. The **Percentile target** option is disabled and the percentile is set at a default value of 50.
- **Normalize to External Value:** This option will open a table listing all samples and a default scaling factor of '1.0' for each of them. You can use the *'Assign Value'* option to assign a different scaling factor to each of the sample; multiple samples can be chosen simultaneously and assigned a value. The **Percentile target** option is disabled and the percentile is set to a default value of 50.

For further details on the above normalization methods, refer to section Normalization Algorithms.

Figure 15.21 illustrated the Step 2 of 4 of experiment creation.

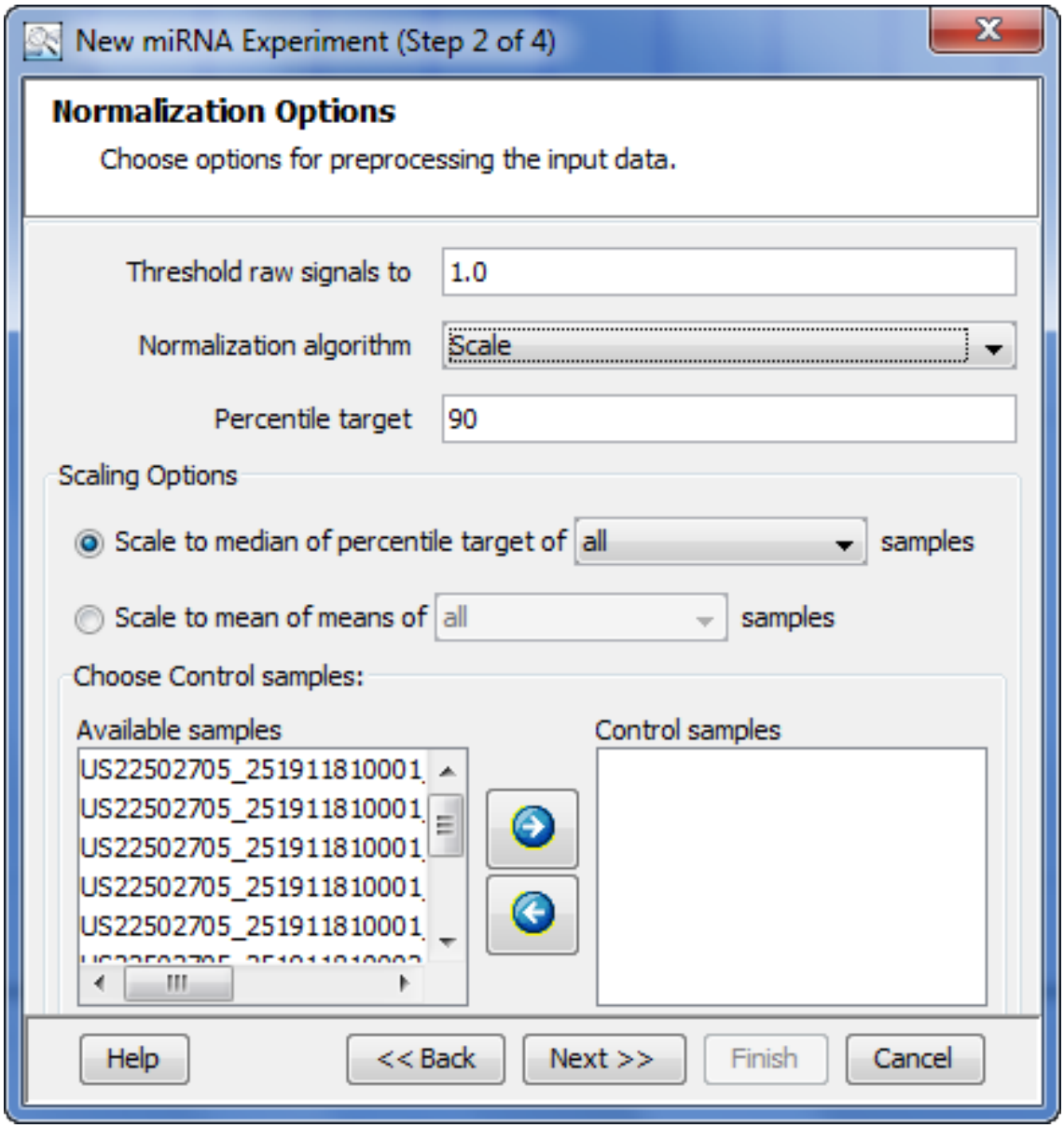


Figure 15.21: Normalization Options

Note: If no normalization is required, the option ***None*** can be selected.

3. **Choose entities (Step 3 of 4):**

   If the **Normalize to control genes** option is selected, then the list of control entities can be specified in the following ways:

   - By choosing a file(s) (txt, csv, or tsv) which contains the control entities of choice denoted by their probe ID. Any other annotation will not be suitable.
   - By searching for a particular entity by using the ***Choose Entities*** option. This opens a search wizard in which the entities can be selected. All the annotation columns present in the technology are provided and you can search using terms from any of the columns. You also have to select the entities as controls when they appear in the **Output Views**

page and then click ***Finish***. This will result in the entities getting selected as control entities and will appear in the wizard.

You can choose either one or both the options to select control genes. The chosen genes can also be removed after selecting the same. See Figure 15.22

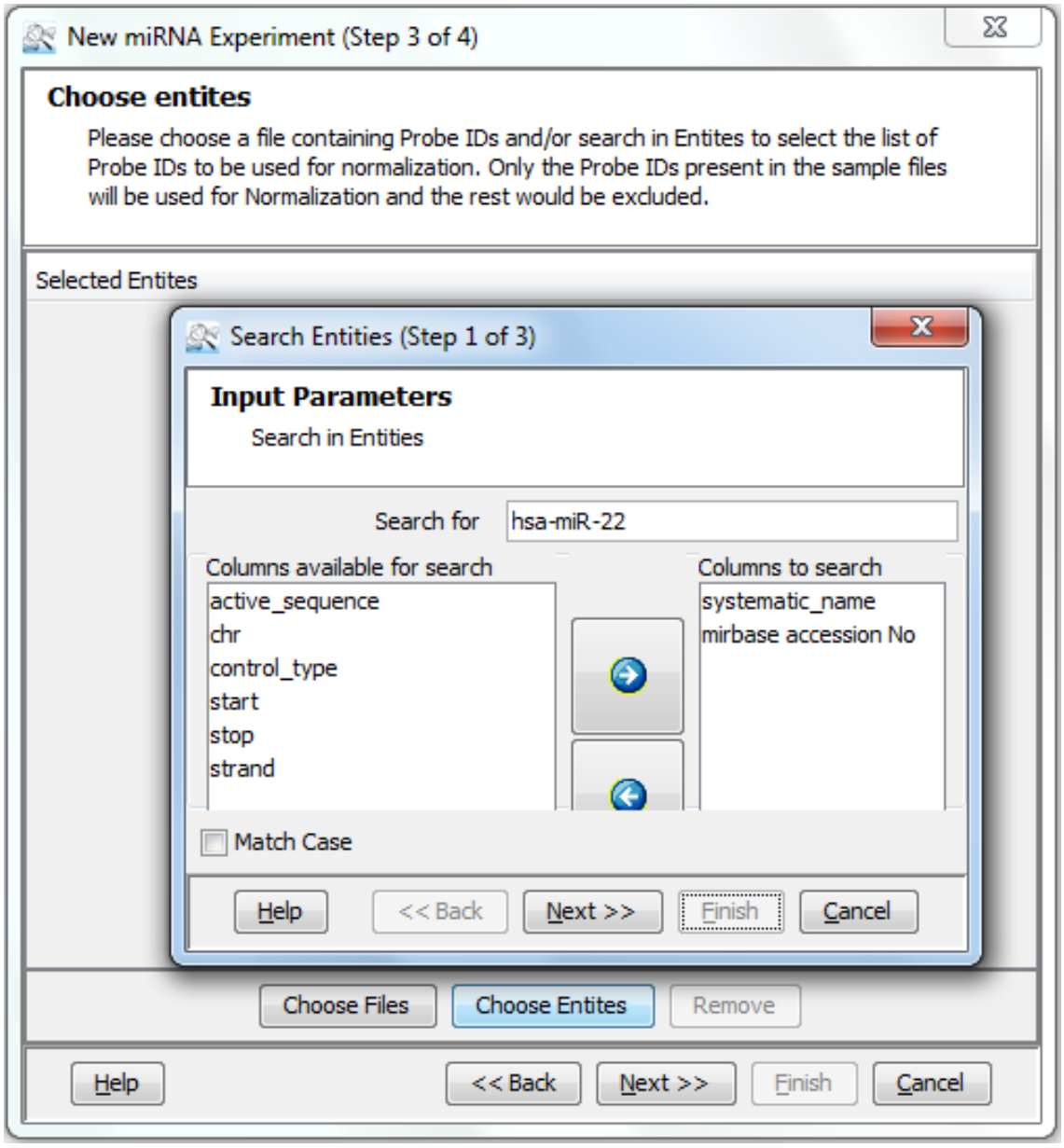


Figure 15.22: Choose entities

If the chosen entities are not present in the technology or sample, they will not be taken into account during experiment creation. Such entities will appear under unmatched probe IDs in the experiment notes of the experiment inspector.

4. **Preprocess Baseline Options (Step 4 of 4):**

Baseline Transformation is carried out row-wise across all samples. This data processing step is particularly useful when visualizing the results in a profile plot or heat map. The baseline transformation options (Refer Figure 15.23), available in **GeneSpring** are:

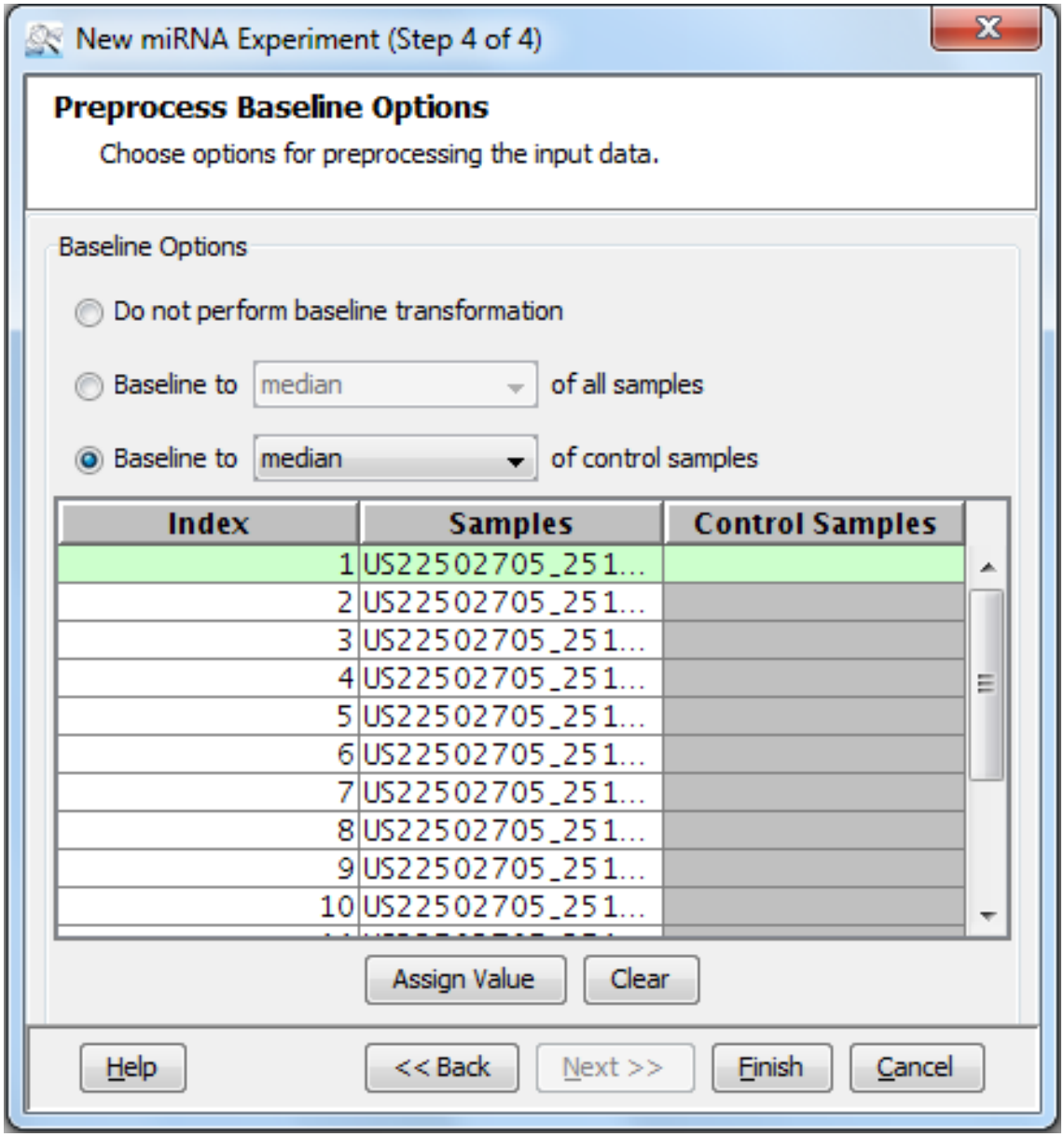


Figure 15.23: Baseline Transformation

- **Do not perform baseline**
- **Baseline to median of all samples:** For each row (probe), the median of the log summarized values across all the samples is calculated. This value is then subtracted from the probe value for all samples.
- **Baseline to median of control samples:** Here control samples are used to calculate the median value for each probe. This value is then subtracted from the probe value for all samples. The controls could be an individual control for each sample or it could

be a set of controls. Alternatively, a set of samples can be used as controls for all samples. For specifying the control for a sample, select the sample and click on ***Assign value***. This opens up the ***Choose Control Samples*** window from where the samples designated as *Controls* should be moved from the *Available Items* box to the *Selected Items* box. Refer Figure 15.24. Click on ***Ok***. This will show the control samples for each of the samples.

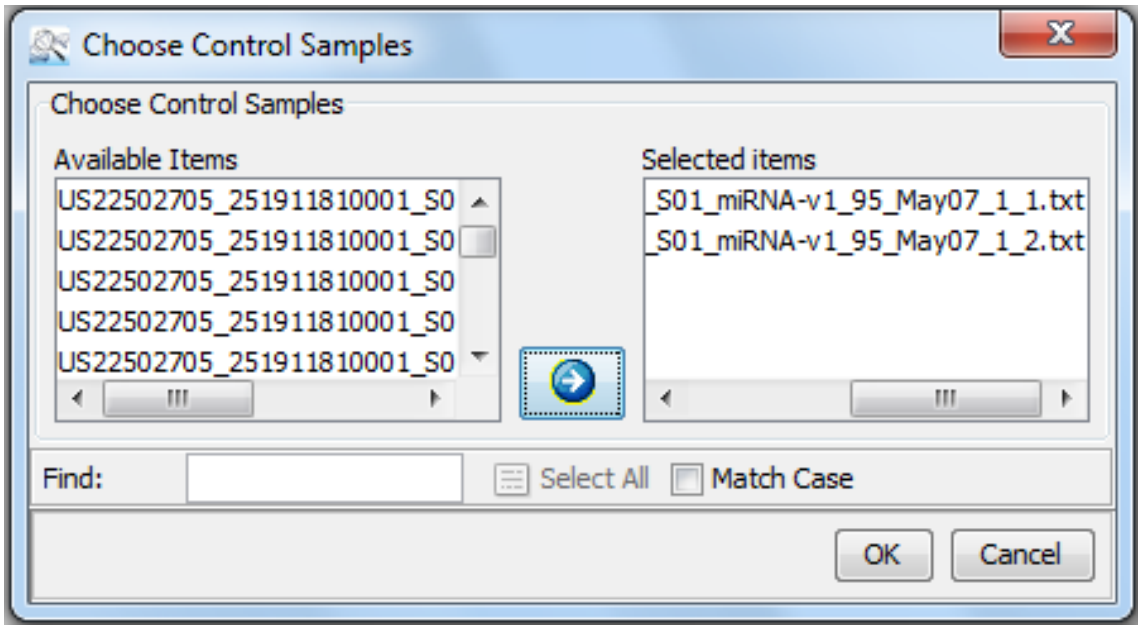


Figure 15.24: Selection of Controls

In *Baseline to median of control samples*, for each probe the median of the log summarized values from the control samples is first computed and then this is subtracted from the sample. If a single sample is chosen as the control sample, then the probe values of the control sample are subtracted from its corresponding sample.

Selecting ***Finish*** creates an experiment, which is displayed as a Box Whisker plot in the active view. Alternative views can be chosen for display by navigating to ***View*** in Toolbar.

Once an experiment is created, the *Data Import Wizard* steps appear on the right hand side. Following is an explanation of the various workflow links:

15.5 Workflow Options for a miRNA experiment

miRNa experiments created either through Analysis: Biological Significance or Data Import Wizard have the options detailed in this section.

15.5.1 Experiment Setup

- **Quick Start Guide:** Selecting this link will direct you to the appropriate chapter in the manual giving details of loading files into **GeneSpring**, Data Import Wizard, various analysis methods, the details of the algorithms used and the interpretation of results
- **Experiment Grouping:** *Experiment Parameters* defines the grouping or the replicate structure of the experiment. For details refer to the section on Experiment Grouping.
- **Create Interpretation:** An interpretation specifies how the samples would be grouped into experimental conditions for display and used for analysis. For details refer to the section on Create Interpretation

15.5.2 Quality Control

- **Quality Control on Samples:** Quality Control (QC) or the Sample QC allows you to decide which samples are ambiguous and which are passing the quality criteria. Based upon the QC results, the unreliable samples can be removed from the analysis.
  - 3D PCA scores, Correlation Analysis - Samples, and a Correlation Coefficients tab.
  - Quality Control Metrics - Report, Quality Control Metrics - Plot, and Experiment Grouping.
  - Legend.

  Figure 15.25 has tiled windows which shows the QC on Samples.

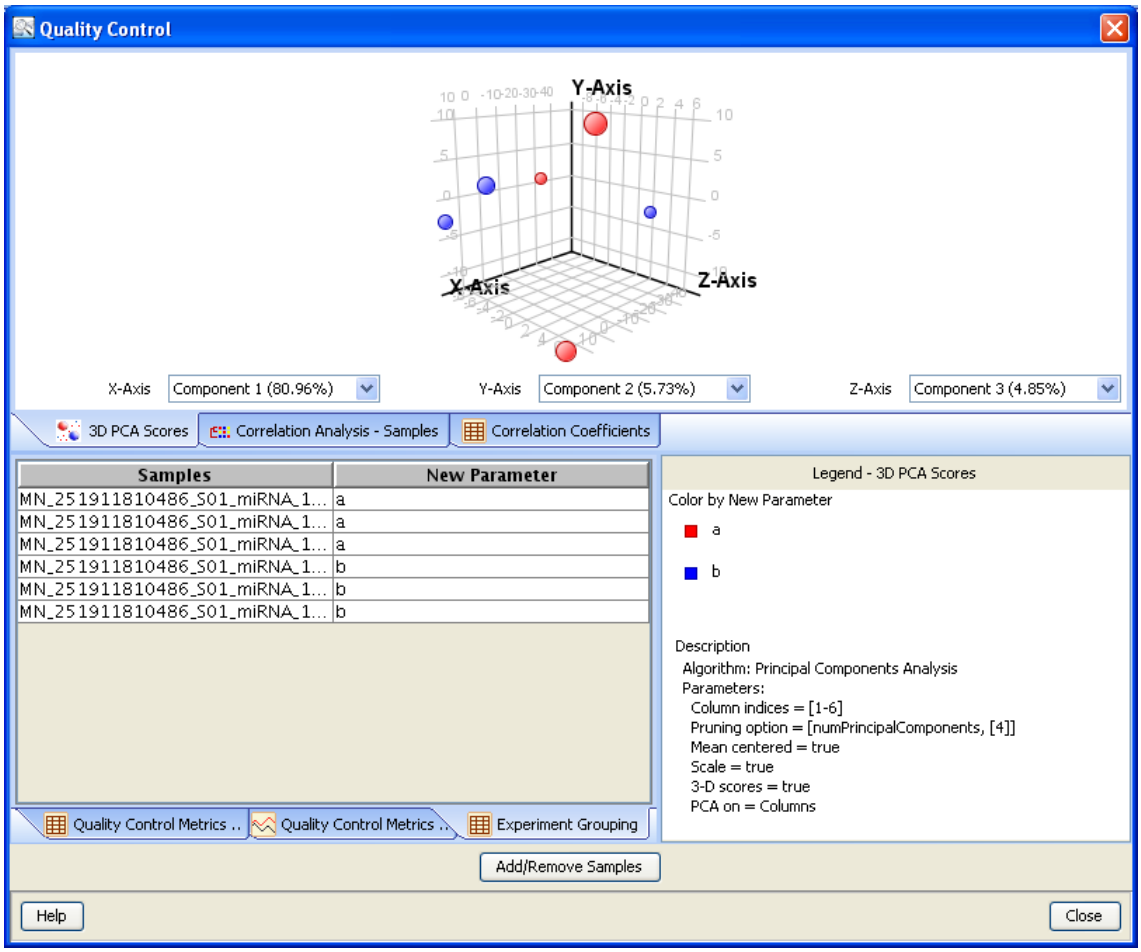


Figure 15.25: Quality Control

In the **Principal Component Analysis (PCA)** tab, PCA scores are visually represented in a 3D scatter plot. These scores are used to check data quality. It shows one point per array and is colored by the *Experiment Factors* provided earlier in the *Experiment Groupings* view. This allows viewing of separations between groups of replicates. Ideally, replicates within a group should cluster together and separately from arrays in other groups. The PCA components, represented in the X, Y, and Z axes are numbered 1, 2, 3... according to their decreasing significance. The 3D PCA scores plot can be customized via ***Right-Click*** $\longrightarrow$ ***Properties***. To zoom into a 3D Scatter plot, press the Shift key and simultaneously hold down the left mouse button and move the mouse upwards. To zoom out, move the mouse downwards instead. To rotate, press the Ctrl key, simultaneously hold down the left mouse button and move the mouse around the plot.

The **Correlation Analysis - Samples** tab shows the correlation analysis heatmap across arrays (samples) (See How to Perform Correlation Analysis on Samples on page 750). The correlation coefficient for each pair of arrays are displayed in the **Correlation Coefficients** tab. Correlation coefficients are calculated using Pearson Similarity Metric (see Pearson on page 799).

Pearson Similarity Metric: Calculates the mean of all elements in vector **a**. Then it subtracts that value from each element in **a** and calls the resulting vector **A**. It does the same for **b** to make a vector **B**. Correlation Coefficient is then computed as **A.B**/(—**A**——**B**—).

In a Correlation Analysis heatmap:

- Sample Order- In a correlation heatmap, the initial order of samples is the same as the input sample order at the time of experiment creation.
- Clustering - Perform clustering analysis on the correlation heatmap using the **Run clustering** icon from the correlation analysis heatmap toolbar (See To run clustering on a correlation heatmap on page 760). Samples are clustered based on correlation coefficients using hierarchical clustering algorithm and plotted in the order they cluster together. The linkage rule used is average and the distance metrics used is Euclidean.

For more information on the available toolbar icons in a correlation analysis heatmap, see Correlation Heatmap View Toolbar on page 776. See Properties on 792 to change the properties of the correlation analysis heatmap.

NOTE Correlation coefficients are computed on raw, non-normalized data and in linear scale.

The **Quality Control Metrics - Report** includes the statistical results to help you evaluate the reproducibility and reliability of your microarray data. The quality metrics scores are obtained directly from the sample file.

Quality controls Metrics - Plot shows the QC metrics present in the QC report in the form of a plot. Please refer to Agilent Feature Extraction Software Reference Guide (v10.7.3 or higher) for detailed descriptions of the QC Metrics.

Experiment Grouping shows the parameters and parameter values for each sample.

The third window shows the legend of the active QC tab.

Unsatisfactory samples or those that have not passed the QC criteria can be removed from further analysis at this stage using *Add/Remove Samples* button. Once a few samples are removed, re-normalization and baseline transformation(if chosen) of the remaining samples is carried out again. The samples removed earlier can also be added back. Click on **OK** to proceed.

NOTE Add/remove samples will not be supported for older miRNA experiments since Add/Remove requires rerunning experiment creation process and now the experiment creation process requires additional mappings in the technology which are not present in the older technologies for miRNA.

- **Filter Probesets by Expression:**

  Entities are filtered based on their signal intensity values. For details refer to the section on Filter Probe by Expression

- **Filter Probesets by Flags:**

  In this step, the entities are filtered based on their flag values as either Detected or Not Detected. Information pertaining to the flags is present in the data file.**GeneSpring** considers the ‘‘gIsGeneDetected” as the flag column and marks entities having ‘0’ as *Not Detected* and ‘1’ as *Detected*. This process is done in four steps:

  1. Step 1 of 4 : *Entity list and Interpretation* window opens up. Select an entity list by clicking on *Choose Entity List* button. Likewise by clicking on *Choose Interpretation* button, select the required interpretation from the navigator window. This is illustrated in Figure 15.26

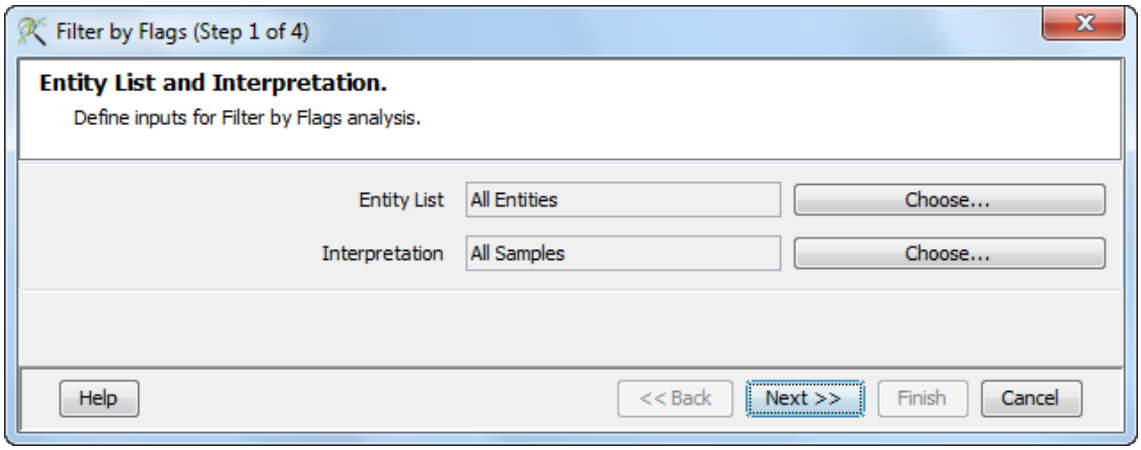


Figure 15.26: Entity list and Interpretation

  2. Step 2 of 4: This step is used to set the filtering criteria and the stringency of the filter. Select the flag values that an entity must satisfy to pass the filter. By default, the Detected and Not Detected flags are selected. Stringency of the filter can be set in ***Retain Entities*** option. Refer Figure 15.27.

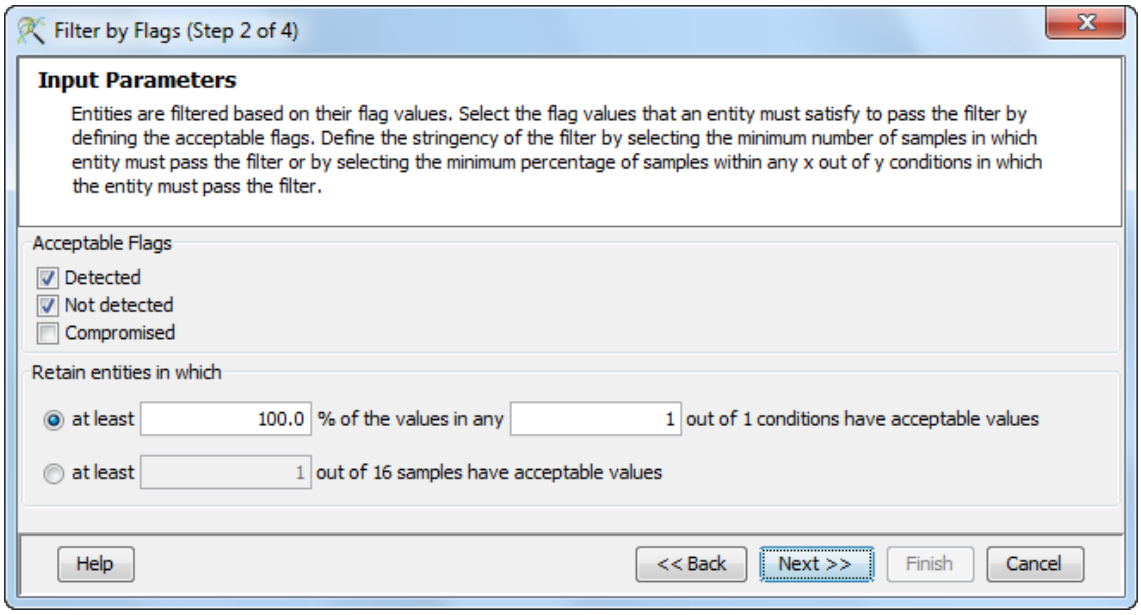


Figure 15.27: Input Parameters

3. Step 3 of 4: A profile plot and a spreadsheet are displayed as 2 tabs, displaying those probes which have passed the filter conditions. Total number of probes and number of probes passing the filter are displayed on the top of the navigator window. Refer Figure 15.28.

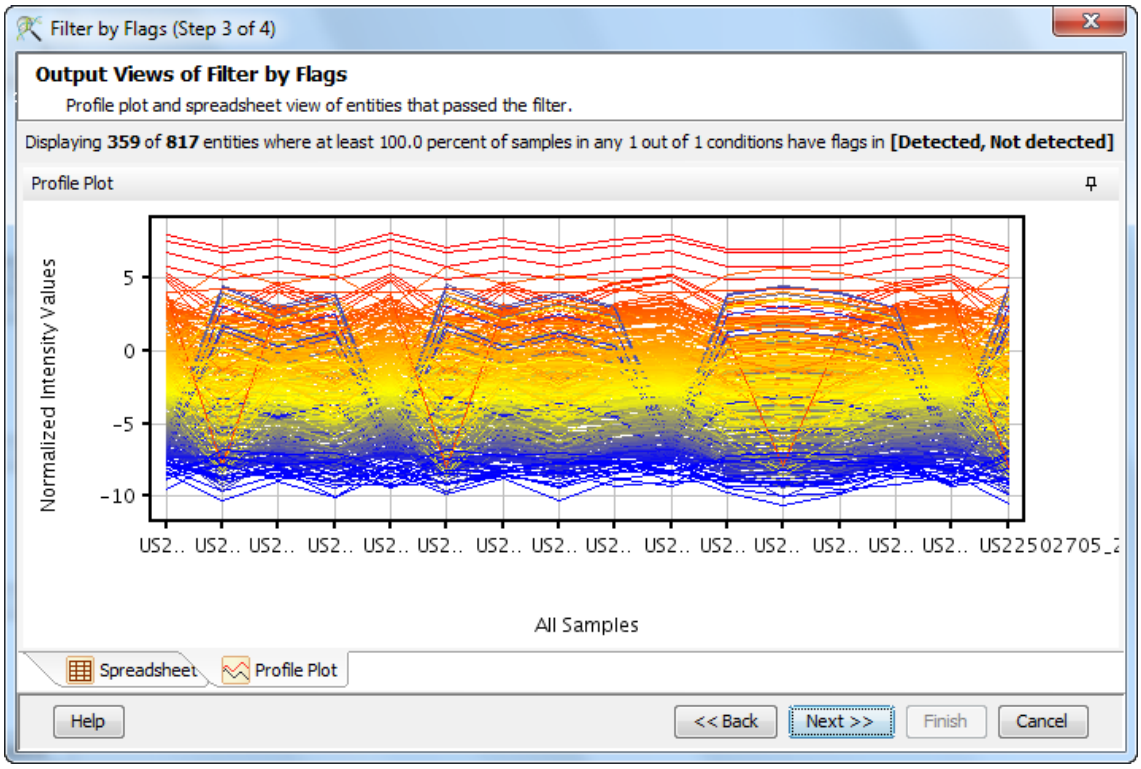


Figure 15.28: Output Views of Filter by Flags

4. Step 4 of 4: Click ***Next*** to save the entity list created as a result of this analysis. Refer Figure 15.29.

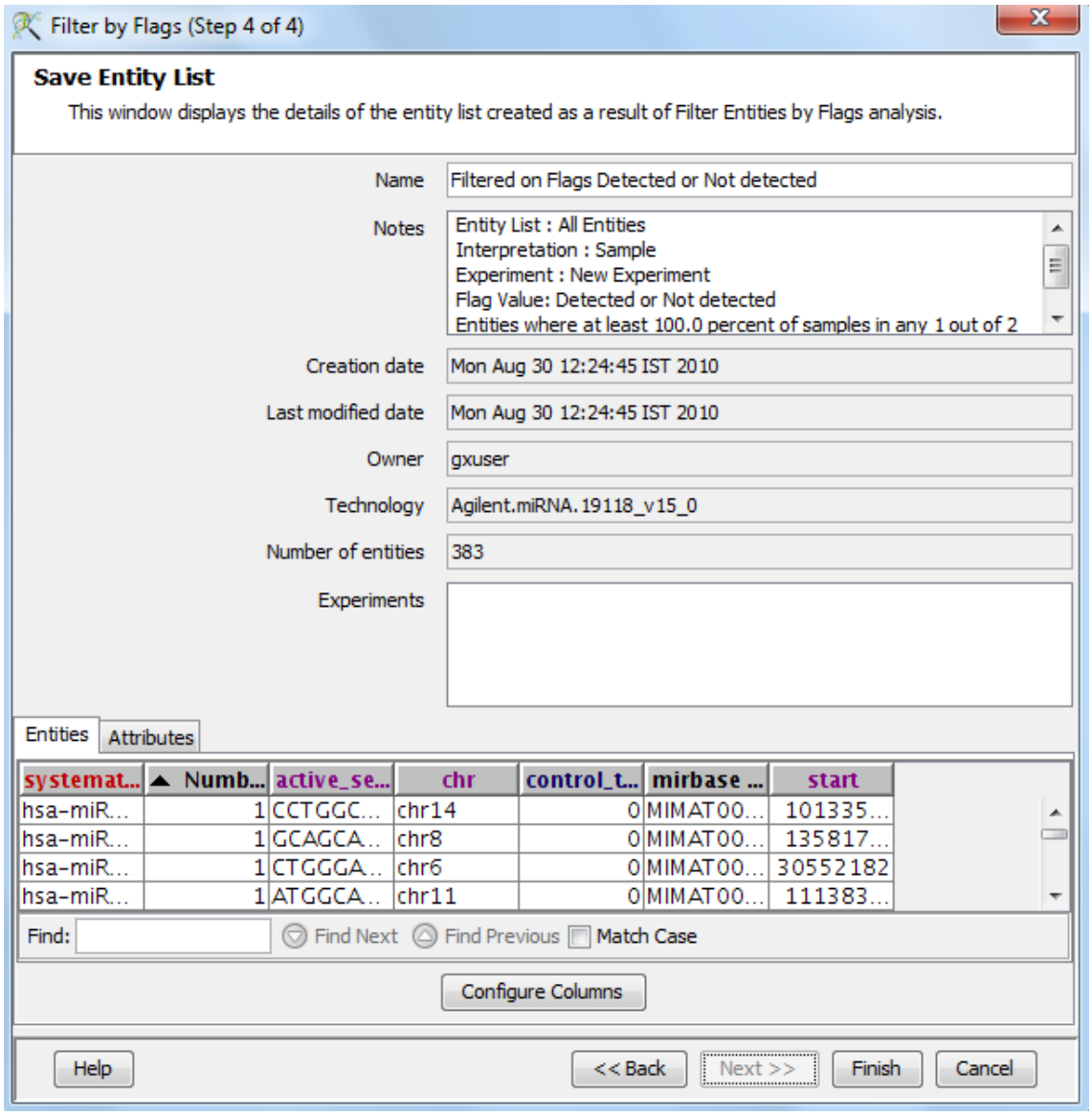


Figure 15.29: Save Entity List

- **Filter Probesets on Data Files:** Entities can be filtered based on values in a specific column of the original data files. For details refer to the section on Filter Probesets on Data Files.
- **Filter Probesets by Error:** Entities can be filtered based on the standard deviation or coefficient of variation using this option. For details refer to the section on Filter Probesets by Error.

15.5.3 Analysis

- **Statistical Analysis**

For details refer to section Statistical Analysis in the Data Import Wizard section.

- **Filter on Volcano Plot**

  For details refer to section Filter on Volcano Plot.

- **Fold Change**

  For details refer to section Fold change.

- **Clustering**

  For details refer to chapter Clustering Analysis.

- **Correlation**

  For details refer to chapter Correlation Analysis.

- **Find Similar Entities**

  For details refer to section Find similar entities.

- **Filter on Parameters**

  For details refer to section Filter on Parameters.

- **Principal Component Analysis**

  For details refer to section PCA.

15.5.4 Class Prediction

- **Build Prediction Model** For details refer to section Build Prediction Model.
- **Run Prediction** For details refer to section Run Prediction.

15.5.5 Results Interpretation

- **Find Targeted Genes**

  GeneSpring miRNA workflow identifies significant miRNAs and facilitates identification of the target genes regulated by the miRNAs. The 'Find Targeted Genes' option under the 'Results Interpretation' requires an entity list as the input, and, outputs an entity list of all the mRNA genes targeted by the miRNA.

  Note: Find Targeted Genes process in **GeneSpring** has altered, and so, you have to download the new Target Prediction databases available from Annotation Manager.

  The mapping of the miRNA active regions to the target genes is accomplished from one or more of the target prediction databases which are selected by you. **GeneSpring** supports PITA, PicTar, TarBase, TargetScan, and microRNA.org databases (Refer Target prediction databases for more details). These databases can be downloaded from the the Annotations $\longrightarrow$Annotations Manager.

  Note:

  – Download at least one target prediction database before running Find Targeted Genes or else, **GeneSpring** enables you to download databases during Find Targeted Genes step.

- If you have an organism build downloaded previously, running Find Targeted Genes will prompt you for a download, and, all the target databases for that particular build will be downloaded. Otherwise, go to Annotations Manager to download a specific target database for a specific build/ organism.
- If you don't have a build downloaded, you are prompted to download the same from the Annotations Manager.

Filtering criteria such as, context score in the case of TargetScan to be applied on the target mapping can be specified in the Configuration Dialog in Tools ⟶Options ⟶Miscellaneous ⟶Target Detection.

Find Targeted Genes workflow creates a list of the targets for the entities under the original entity list. Analysis methods such as **GO Analysis** accept only target entity lists as input. Since the miRNA technology is not associated with the target gene annotations, it is necessary to have the appropriate organism's **Biological Genome** created. Using this feature, the GO terms are obtained for the TargetScan list. Similarly, for Pathway Analysis, the Entrez IDs are obtained from **Biological Genome**. For more information on creating a genome refer to the section on Biological Genome.

How to Find Targeted Genes

Note: If you don't have a Biological Genome and run Find Targeted Genes, you will be prompted to download the same. Download the Biological genome at this step or go to Annotations ⟶Create technology ⟶From Agilent Server, and download the required Biological Genome.

1. **Select Inputs**

   - In this wizard, choose an entity list, and specify the target prediction databases to be used for target prediction. All Entities list is selected by default.
   - If multiple databases are selected, you can also specify if only those entities that are supported by all the selected databases should be considered for target prediction. This option is selected by default.
   - If you have downloaded two different builds of the same organism, you will be prompted to select a build and only the databases specific to the selected build will be displayed in this window.
   - Refer Figure 15.30 for an illustration.

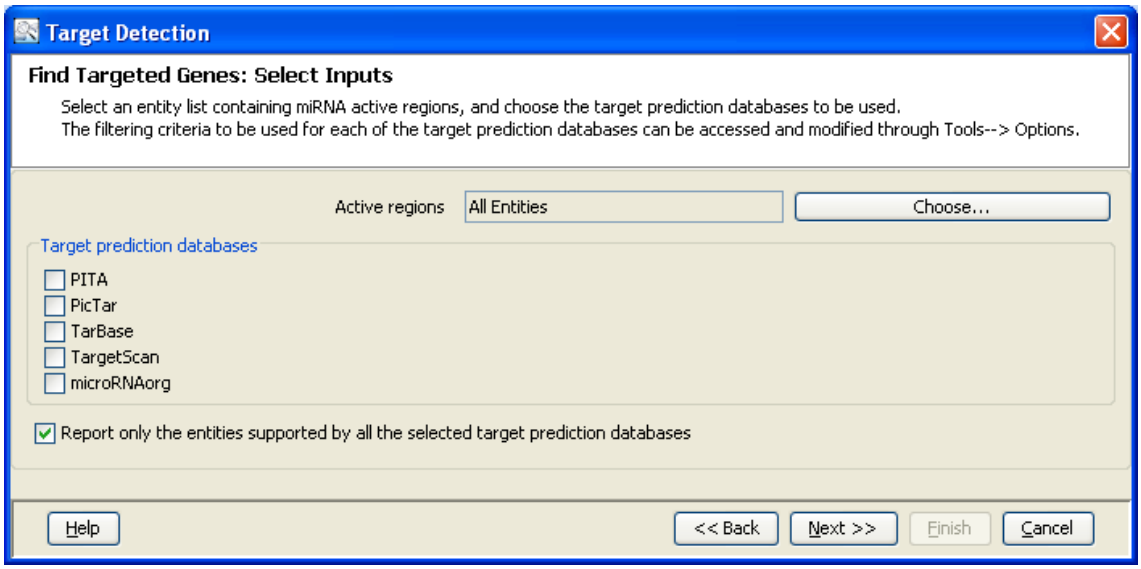


Figure 15.30: Find Targeted Genes: Select Inputs

Note: **GeneSpring** uses MIMAT IDs for all Target Prediction databases. Only if miRNA data was imported using the Generic workflow to create the experiment **GeneSpring** uses systematic names for Target Prediction Analysis. In this case, make sure you have provided updated systematic names.

2. **Output Views**

   - Depending on the target databases selected, a table is generated with the miRNA genes mapped to the active regions (miRNAs).
   - The columns comprise of Gene ID, systematic name (for Generic miRNA experiments) or miRBase accession number (for Agilent miRNA experiments), and a p-value.
   - For the p-value computation method, refer Section 7.3.5.3. The default p-value cut-off for filtering the genes is 0.05 which can be altered by you.

3. **Object Details**

   - A final entity list is created with the targeted genes. All the parameters on which the targeted genes were created are present in the notes.
   - A table is displayed with all the entities that satisfied the filters/ cut-offs with their attributes.
   - Depending on the databases selected in the previous steps, the table has multiple columns with attributes for the genes and can be sorted/ configured by you.
   - Refer Figure 15.31 for an illustration.

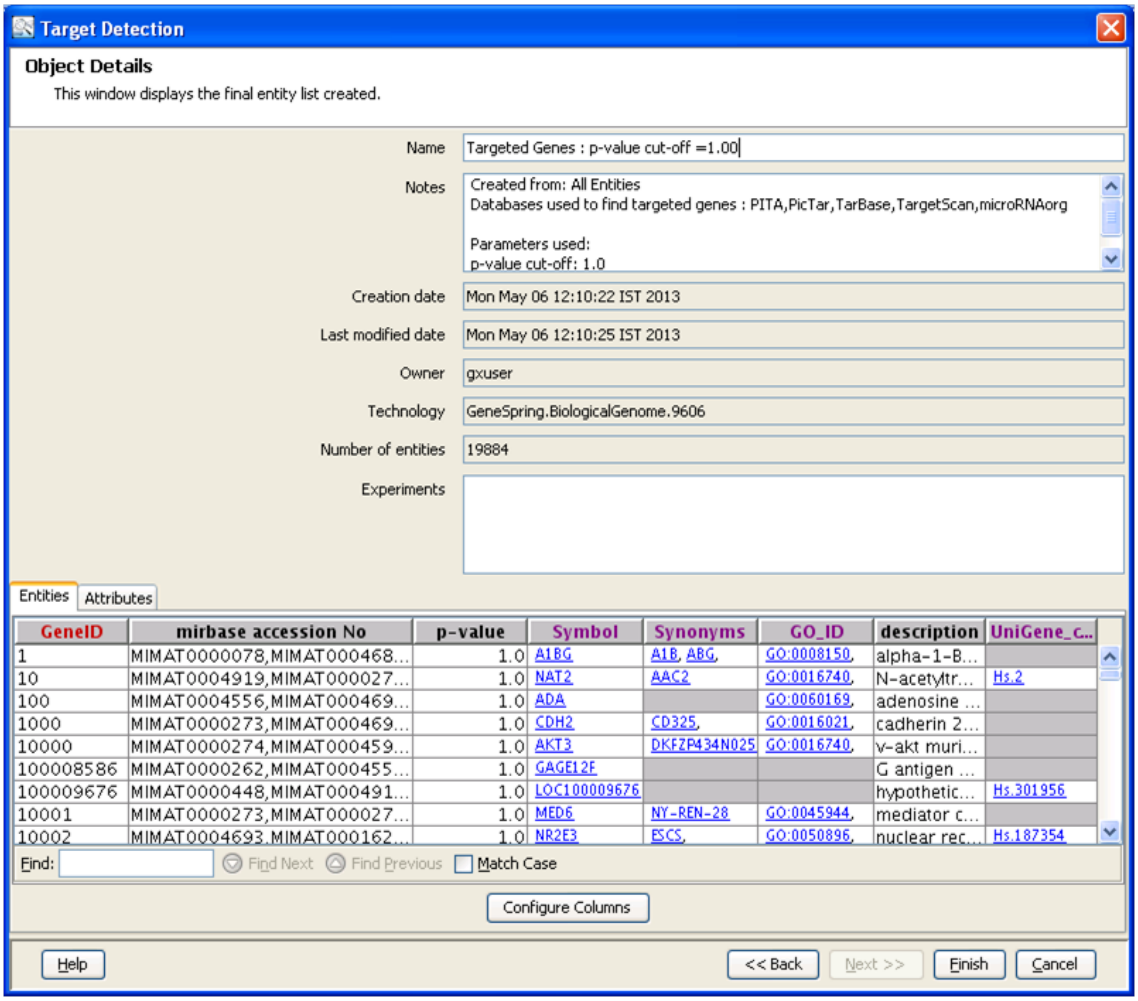


Figure 15.31: Find Targeted Genes: Object Details

Target Prediction Databases in GeneSpring

1. **PITA**

   PITA target prediction database is created by downloading files from the Segal Lab portal. The portal has files with individual binding site locations and since the same miRNA can have multiple sites on the same UTR, PITA database also has a file with this combined information.

   In **GeneSpring**, these files with site-level details are used to visualize the binding sites in Genome Browser and a scaled score (from the combined interaction score of all the miRNAs) is used for target Prediction Analysis.

2. **PicTar**

   PicTar target prediction database for hg19/ hg18 build is based on the pic- TarMiRNA5Way table at UCSC. The UCSC track was converted from hg17 build to hg19 build using the hg liftOver utility.

   Predictions with obsolete RefSeq transcripts or transcripts for which Entrez ID and

Ensembl IDs cannot be associated were removed from the database as these entries cannot be used for Target Prediction.

Each predicted target us accompanied with a **score**. This score is not the original PicTar score, but a renormalized score on a 0-1000 scale by the UCSC.

3. **TarBase**

   This database has been downladed from the TarBase portal.

4. **Target Scan**

   TargetScan identifies mRNA targets for any specific miRNA, based on the context percentile and the database which are user defined. Context percentile is derived from context score and has been described as:

   "Sites within 15 nt of a stop codon are flagged because these are typically not effective. The context of each of the remaining sites has been evaluated and scored considering the following four features:

   - site-type contribution: reflects the type of seed match (8mer, 7mer-m8, and 7mer-1A)
   - 3' pairing contribution: reflects consequential miRNA-target complementarity outside the seed region
   - local AU contribution: reflects transcript AU content 30 nt upstream and downstream of predicted site
   - position contribution: reflects distance to nearest end of annotated UTR of target

   With all the four features, a more negative score is associated with a more favorable site. The context score is the sum of the above scores, and the context score percentile is the percentile rank of each site compared to all sites for this miRNA family. Thus a high context score percentile (between 50 and 100) shows that a specific site is more favorable than most other sites of this miRNA." (Refer http://www.TargetScan.org/docs/help.html for more details)

   The other criteria which determines target selection is the database. There are two databases, conserved and non-conserved. For miRNA target sites, conservation is defined using the conserved branch length which is based on the sum of phylogenetic branch lengths between species that contain a site and also dependent on site type and UTR conservation.

   The conserved branch length score (Friedman et al., 2008) is the sum of phylogenetic branch lengths between species that contain a site. To help control for individual UTR conservation, 3' UTRs were separated by conservation rate into ten equally sized bins, and a unique set of branch lengths based on 3' UTR sequence alignments was constructed for each bin. Site conservation is defined by conserved branch length, with each site type having a different threshold for conservation:

   - 8mer: 0.8
   - 7mer-m8: 1.3
   - 7mer-1A: 1.6

   (Refer http://www.targetscan.org/cgi-bin/targetscan/vert_50/view_gene.cgi?taxid=9606&gs=TNKS2&members=miR-1/206 for more details)

   Note that in TargetScan, definitions of conservation can apply to

   (a) miRNA families and

   (b) miRNA target sites. In the context of **GeneSpring**, it is only the miRNA target sites and their relevant databases described above are relevant.

The TargetScan prediction database is created by downloading files from the TargetScan portal. The files provide the target site information in the form of an offset with respect to the 3'UTR of a RefSeq transcript. Using the RefSeq transcript annotations from NCBI, these coordinates are translated into build hg19/ hg18 genomic coordinates.

Note: **GeneSpring** utilizes the liftOver utility to translate hg18 co-ordinates for target prediction.

Entrez and Ensembl IDs are also added to the database to make it suitable for Target Prediction Analysis.

TargenScan analysis is based on MIMAT accession nos. The MIMAT accession numbers are packaged along with TargetScan data `http://www.targetscan.org//vert_50//vert_50_data_download/miR_Family_Info.txt.zip`. The technology maps each systematic name to MIMAT accession IDs. The MIMAT accession numbers present in the technology are then matched against these in TargetScan to retrieve the targets as Gene Symbols. Conversion of Gene Symbol to Gene ID is done using the Biological Genome. The miRBase Accession numbers for which no match was found in Targetscan will be reported missing and would be collected in the Entity list notes.

Running TargetScan for older miRNA experiments

Targetscan will not work on miRNA experiments created in **GeneSpring** prior to version 11.5 unless the user updates annotations of the technology with the MIMAT accession numbers. The MIMAT accession numbers are present in the Feature Extraction (FE) files which could be processed and put in an appropriate format. Create a tab separated file with the appropriate MIMAT accession numbers for the miRNA names and use this file to update the technology annotation as described below.

Go to the menu **Annotations** → **Update Technology Annotations** → **From File or Biological Genome**; a 3 step wizard will come up.

- **Step 1: Input parameters** Choose the technology you want to update and click on 'Update from file'. Choose the appropriate tab separated file (txt, tsv, csv or xls) containing the MIMAT accession numbers and click 'Next'.
- **Step 2: Format Data File** Format the input file if required.
- **Step 3: Choose Annotation Columns** This step gives options for matching fields while updating the technology annotation. Choose the appropriate technology column and genome/file column. **GeneSpring** allows appending, overwriting or filling in missing values and user can choose the required option. Remember to assign a column mark to the miRBase accession number column in this step. Check the column name showing mirbase Accession No and in the last column 'Column Mark', choose 'Enter New' from the drop down list and enter 'mirbase.accession'. This mark is required to assign the MIMAT accession numbers and to run the targetscan analysis. See Figure 15.32.

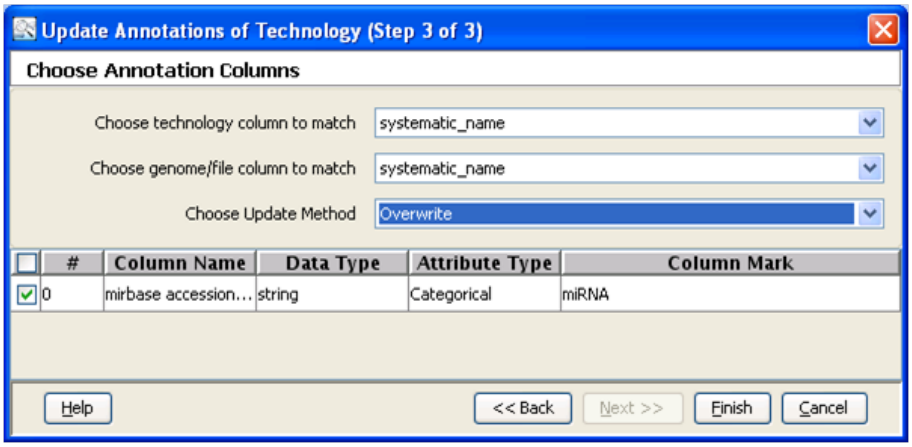


Figure 15.32: Step 3 of Update Annotations

TargetScan analysis can be carried out on older miRNA experiments as well.

Translation

- **Translation between miRNA and non-miRNA experiments:** Translation of miRNA entity lists into non-miRNA experiments is not allowed because of absence of Entrez GeneIDs in miRNA entity list. However, the gene lists created following TargetScan analysis can be translated into Gene Expression experiments.
- **Translation between miRNA experiments:** Translation of miRNA entity lists between miRNA experiments is allowed and if the technologies are same, it is done using miRNA (Systematic) names. However, if the technologies between miRNA experiments are different, then it uses the miRBase accession numbers(MIMAT IDs) for translation.

5. **microRNA.org**

   The microRNA.org prediction database is created by downloading the files from microRNA.org portal.

   Each target qualifies for prediction if it satisfies these four attributes;

   - mirSVR score
   - conservation of the target site
   - whether the miRNA is conserved or not
   - seed class: 8-mer, 7-mer(m8), 7-mer(1A) and 6-mer

Note: The number of databases available for Find Targeted Genes are organism and build specific. All the databases have been mapped to miRBase version 18 and will be updated accordingly.

- **GO Analysis**

  Refer Gene Ontology Analysis chapter for a detailed description.

- **Find Similar Entity Lists** Refer Find Similar Entity Lists section for a detailed description.

15.5.6 Pathway Analysis

- **Single Experiment Analysis**

  This feature is discussed in detail in the section Single Experiment Analysis.

- **Export to IPA**

  This feature is discussed in detail in the section Ingenuity Pathways Analysis (IPA) Connector.

- **Import IPA Entity List**

  This feature is discussed in detail in the section Ingenuity Pathways Analysis (IPA) Connector.

15.5.7 NLP Networks

- **NLP Network Discovery** This feature is discussed in detail in the section NLP Network Discovery.

- **MeSH Network Builder**

  This feature is discussed in detail in the section MeSH Network Builder.

- **Extract Relations via NLP**

  This feature is discussed in detail in the section Extract Relations via NLP.

15.5.8 Utilities

- **Import Entity list from File**

  For details refer to section Import Entity list from File.

- **Analysis: Biological Significance**

  For details refer to section Analysis: Biological Significance.

- **Filter On Entity List**

  For further details refer to section Filter On Entity List.

- **Remove Entities with missing signal values**

  For details refer to section Remove Entities with missing signal values.

Chapter 16

Analyzing Agilent Exon Data

Splice variants are variable sequences of RNA produced from the same gene in DNA, resulting in creation of different proteins potentially affecting cellular regulation. **GeneSpring** supports Agilent Exon Splicing analysis using Agilent Exon arrays. Both single and two color technologies are supported. Gene-level and exon-level splicing analysis can be performed to investigate global gene expression and to uncover novel alternative splicing events. The data files are in *.txt format and obtained from Agilent Feature Extraction Software 10.7.3 or higher.

An Agilent exon experiment in **GeneSpring** aims to answer the following questions:

- Which known genes are differentially expressed between the experiment groups?
- Which known genes are differentially spliced between the experiment groups?
- Identify the exons that are differentially spliced.

What can the user do in GeneSpring with Agilent Exon Technology?

GeneSpring supports both expression and splicing analysis for Agilent Exon experiments. Various capabilities within **GeneSpring** for Agilent Exon Experiments are listed here:

- Create Single Color / Two Color Experiments
- Add/Remove samples through Quality Control
- Gene Level Analysis
- Exon Level Analysis
- Visualization of Splicing

See section Detailed Descriptions for in-depth explanations on various technical aspects underlying the analysis.

16.1 Running the Agilent Exon Workflow

16.1.1 Overview

Experiment creation broadly involves the following steps. See additional sections for further details.

Technology Download: Before creating an Agilent Exon experiment, the technology needs to be downloaded. Go to the menu **Annotations** → **Create Technology** → **From Agilent Server**; all the available technologies are listed here for Agilent Exon Single and Two Color. The technology type would be of the form: 'Agilent.Exon< $SingleorTwo$ >Color.< $AMADID$ >'. The user can recognize a specific technology type based on the AMADID number.

See section Agilent Exon Technology for details.

Choose the relevant technology(ies) and download.

Experiment Creation:

- **Loading Samples**: Create a new experiment by choosing experiment type as Agilent Exon Single Color or Agilent Exon Two Color. Choose sample files (*.txt format) from FE files ensuring that all the sample files are of same technology (as deciphered from the AMADID number); if not, **GeneSpring** will display an error message.
- **Choosing Technology**: Choose the genome build of the technology for the set of samples. Note that while several technologies can co-exist, **GeneSpring** will automatically choose the latest version of technology for that build available in **GeneSpring** .
- **Setting experiment parameters**: Input the flag settings, normalization and baseline transformation options for single color experiment; for two color, give the dye swap information (no normalization option exists for two color arrays since these are LOWESS normalized in FE). The experiment is now created for the chosen samples using the selected technology.

Choose the analysis and experiment type in the 'New Experiment Window' as shown in Figure 16.1.

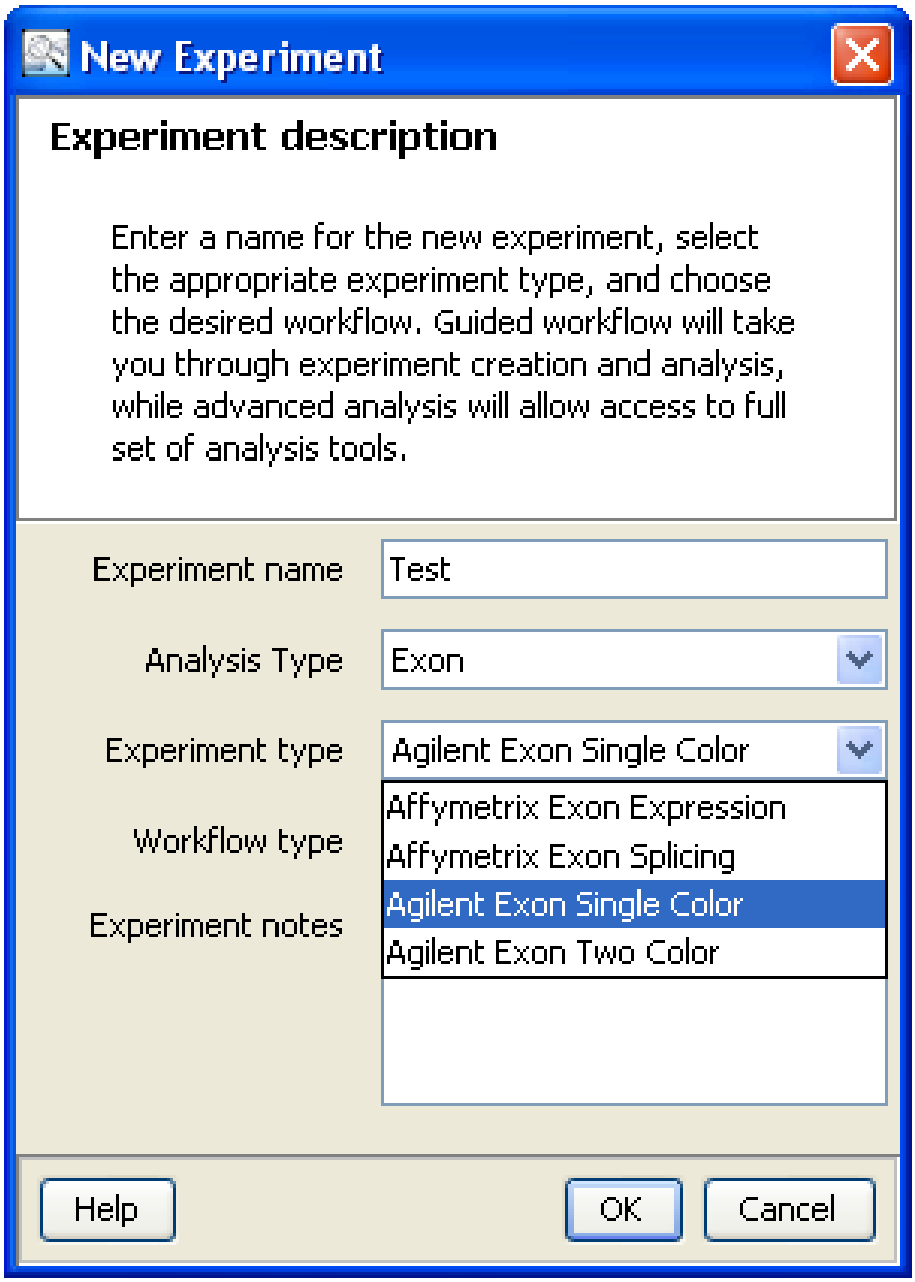


Figure 16.1: New Experiment for Agilent Exon

Note: There is no *Analysis: Biological Significance* for Agilent Exon Single Color and Two Color Experiments. The only option is to run *Data Import Wizard*.

16.1.2 Data Processing for Agilent Exon Experiments

Agilent Flag Logic:
Probes are first considered horizontally across the 5 columns to assign a resultant flag value to each

probe. There are five criteria which determine the flag value of each probe. The spot information in the data file is used to flag the data as Detected, Not Detected or Compromised based on the five feature criteria, namely: not positive or significant, not uniform, not above background, is saturated or a population outlier. These criteria are decided in Step 3 (Figure 16.6) of experiment creation. For assigning overall flag for a given probe P1, the resultant flags are then considered vertically across all replicates. In Example 1, we exclude or ignore the 'Compromised' signal(1) and assign the majority flag which is: Not Detected. Hence the overall signal = $(2+3+4+5)/4 = 3.5$

The following two examples illustrate this.

Replicate Probes	Signal	F1	F2	F3	F4	F5	Resultant Flag
P1	1	Detected	NotDetected	Compromised	NotDetected	Detected	Compromised
P1	2	Detected	NotDetected	NotDetected	NotDetected	Detected	NotDetected
P1	3	Detected	Detected	Detected	Detected	Detected	Detected
P1	4	NotDetected	NotDetected	Detected	Detected	NotDetected	NotDetected
P1	5	NotDetected	Detected	NotDetected	NotDetected	Detected	NotDetected

Table 16.1: Agilent Flag Logic: Example 1

In Example 2, even if there is one 'Compromised' flag, the resultant flag becomes 'Compromised'. To get the overall flag across replicates, if there is a compromised flag, we normally exclude or ignore it. But in this example, since there are only 'Compromised' flags, the overall flag is assigned 'Compromised' for this probe. Overall signal = $(1+2+3+4+5)/5 = 3$

Replicate Probes	Signal	F1	F2	F3	F4	F5	Resultant Flag
P1	1	Detected	NotDetected	Compromised	NotDetected	Detected	Compromised
P1	2	Detected	NotDetected	NotDetected	Compromised	Detected	Compromised
P1	3	Detected	Detected	Compromised	Detected	Detected	Compromised
P1	4	NotDetected	NotDetected	Detected	Compromised	NotDetected	Compromised
P1	5	NotDetected	Detected	Compromised	NotDetected	Detected	Compromised

Table 16.2: Agilent Flag Logic: Example 2

The Experiment Navigator for Agilent Exon Experiments shows only gene-level data. Ways to obtain the probe-level data are explained in section Probe-level Data.

Sequence of events during Agilent Exon Experiment Creation: The sequence of steps done to create Agilent Exon Experiment is depicted pictorially in Figure 16.2. All the terms are explained in this section.

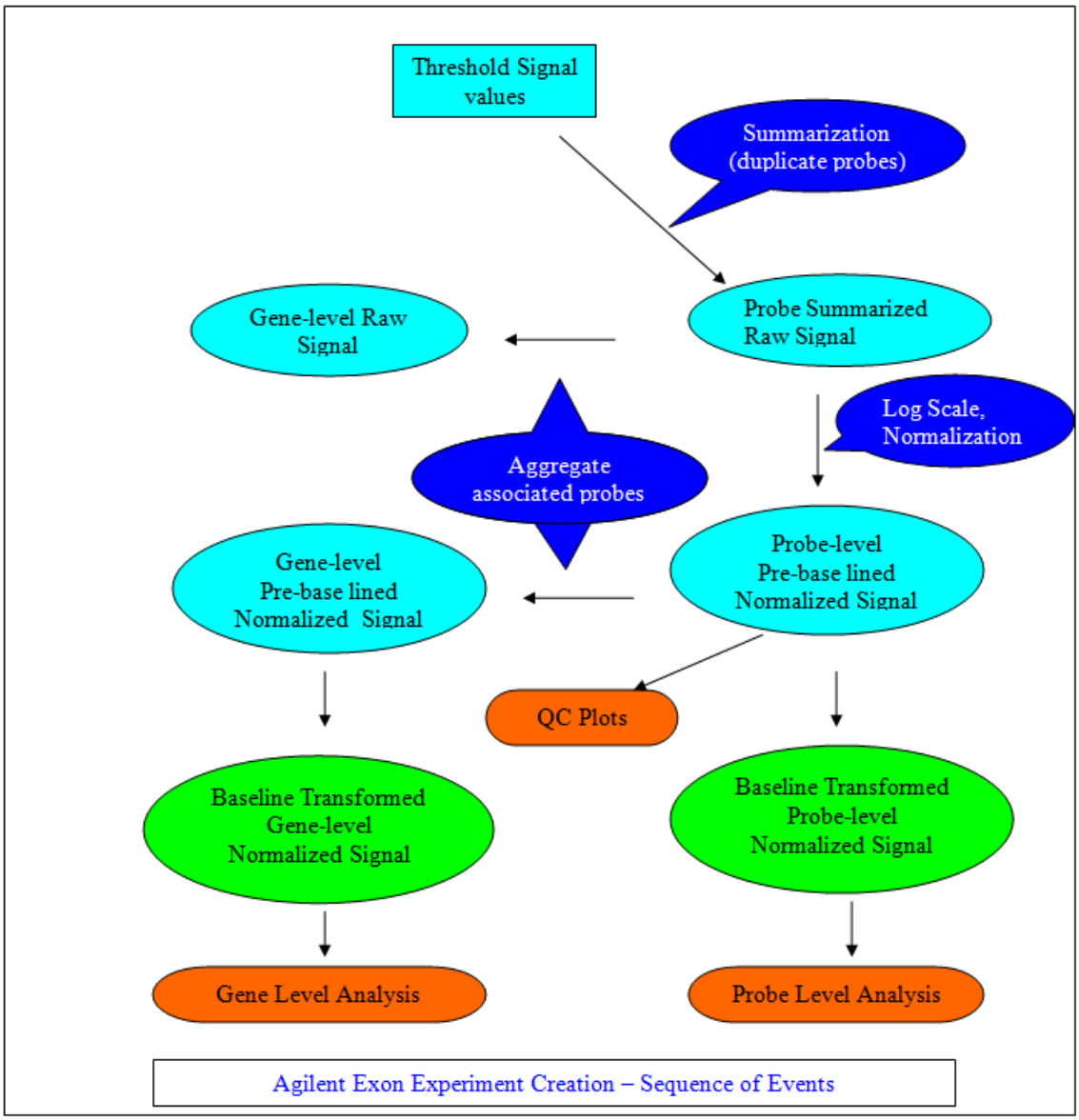


Figure 16.2: Sequence of events while creating Agilent Exon Experiment

File formats: The data file should be present as a text file (*.txt), obtained from Agilent FE 10.7.3 or higher.

Probe Summarized (Raw) Signal: Probe-summarized data refers to the probe-level data obtained after threshold. In case of on-chip replicates, a summarization step is carried out. The geometric mean of signals of all replicates of a particular probe is referred to as the probe-summarized value. For ease of computation, the signal values (which are in linear scale) are converted to log scale; arithmetic mean is calculated and re-transformed to linear scale.

Let v_1j, v_2j, ..., v_nj be the signal values for the jth probe.
Then the log summarized value, LS_j= (log(v_{1j}) + log(v_{2j}) + ... + log(v_{nj}))/n.
The summarized value, S_j= Antilog(LS_j).

Raw Gene Level Signal: For computing Raw Gene Level (Summarized) Signal, all the associated probes for that gene are identified using a map stored in the standard technology and a geometric mean of all these probe summarized (raw) signal values is taken after applying the flag logic. For more details on Agilent Flag Logic, please refer to section 16.1.2.

Let v_1, v_2, ..., v_k be the summarized raw signal intensities of the exon probes e_1, e_2, ..., e_k respectively. Raw Gene Level Signal = Antilog$(1/k \times (log(v_1) + log(v_2) + \ldots + log(v_k)))$

Normalized Signal Values: GeneSpring generates normalized signal values for the probe and gene level as explained below.

Probe Level Normalized Signal Computation: The Probe Level Pre-Baseline Normalized value is generated by performing log transformation and normalization on the probe summarized (raw) values. For two color experiments an additional step of $Cy5/Cy3$ ratio computation is carried out before log transformation. The probe-level pre-baseline normalized signal is baseline transformed using the baseline choice specified in the step 5 of experiment creation. Refer to Figure 16.8 for more details.

Gene Normalized Signal Computation: GeneSpring computes gene-level normalized pre-baseline transformed signal values by taking geometric mean of all the associated probe-level normalized pre baseline transformed values after applying the flag logic. For more details on Agilent Flag Logic, please refer to section 16.1.2.

Let $v_1, v_2, \ldots, v_k$ be the normalized, pre baseline transformed signals of the exon probes $e_1, e_2, \ldots, e_k$ respectively. Normalized gene signal (Pre baseline transformed) value= $1/k * (v_1 + v_2 + \ldots + v_k)$

Following this, Baseline transformation is explicitly carried out to obtain Gene Level Normalized Signal Values.

Treatment of on chip replicates: Probes which are present multiple times on the chip are termed as on-chip replicates. The overall flag is assigned to such probes after applying the Agilent Flag logic. The final signal value will be the geometric mean of all the signal values of replicates that were not excluded after applying the flag logic. For more details on Agilent Flag Logic, please refer to section 16.1.2. Please see Figure 16.6 showing the criteria for flag import (F1-F5) in the examples shown below.

> Note: The data processing steps are the same for both single and two-color experiments except for computation of Normalized signal values. As no normalization is done during two-color experiment creation, the ratio of the log values of signal and control columns is used for subsequent analyses.

16.1.3 Creating an Agilent Exon Single Color Experiment

The steps to create an Agilent Exon Single Color Experiment are explained here.

Loading Data (Step 1 of 5)

The user can select the data files, previously used samples, or both to create an Agilent exon experiment. See Figure 16.3.

- The user can select the data files using the *Choose Files* button, and navigate to the folder containing the data files.

- The user can select already used samples using the *Choose Samples* button. This opens the *Sample/RawFile Search Wizard*. The *Sample/RawFile Search Wizard* has two steps.
  1. Allows the user to Add, Remove, and Combine Search Conditions (by "OR" or "AND"). The user can select "Show User Attributes" check box to open the search results with the user attributes, and enter an appropriate value for the "Max Results per page" (default value: 100).
  2. Launches the search results of samples (select and double click to launch the Sample Inspector). Select the Samples the user wants to use in the Experiment, and click "Finish".

  The user can click on the *Reorder* button to change the ordering of the samples.

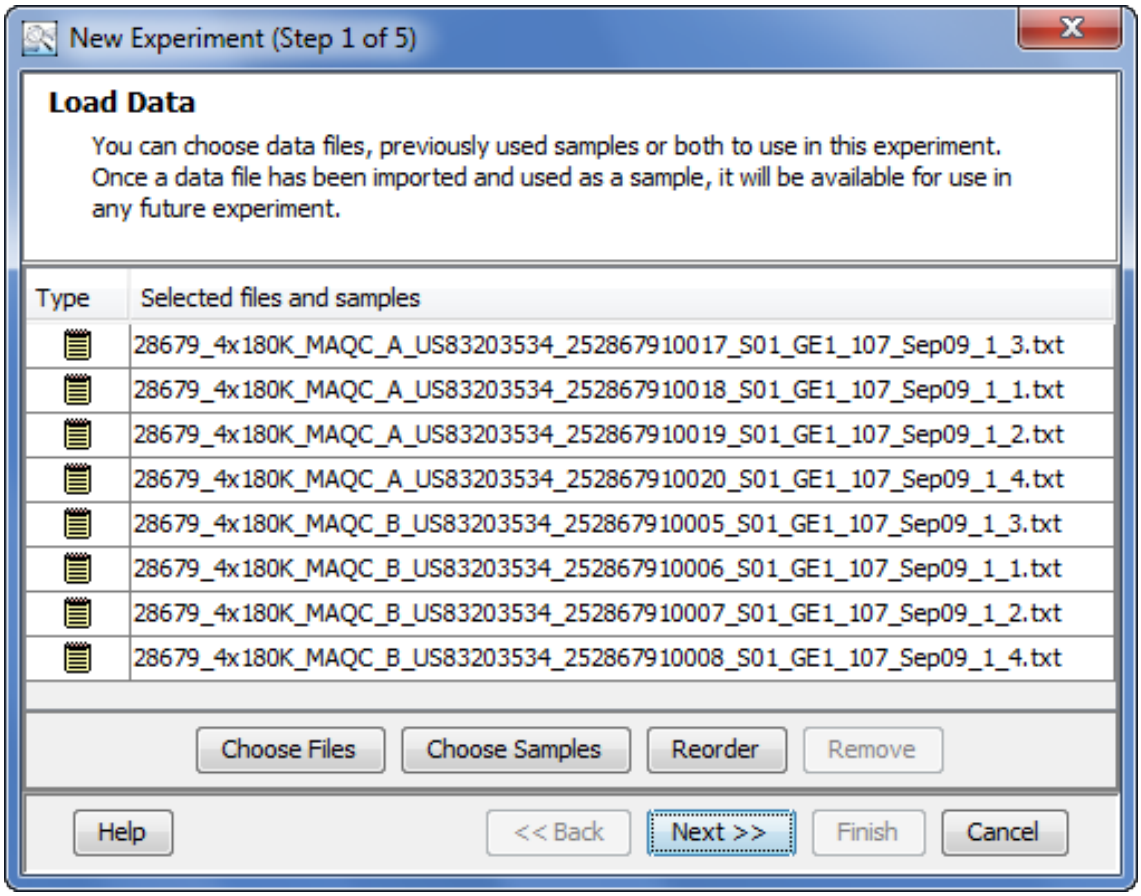


Figure 16.3: Load Data

If there is no Agilent Exon technology detected, then **GeneSpring** displays an error message (as shown in Figure 16.4) prompting you to download the technology before proceeding.

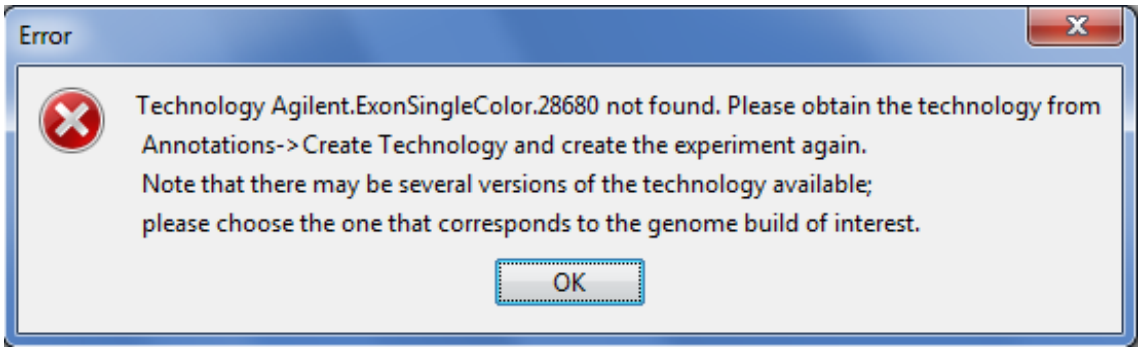


Figure 16.4: Error message if technology is not found

If this occurs, cancel the experiment creation and go to the menu **Annotations** → **Create Technology** → **From Agilent Server** and download the relevant technology and then proceed with experiment creation.

Selecting Build for the Technology (Step 2 of 5)

Currently Agilent Exon uses hg19 build. Fig 16.5 shows the second step **Chose Corresponding Build for the Technology**.

- Select the Technology Build hg19 from the drop-down menu. The summary box will show only the latest version of the technologies available for the chosen build. For instance, if there are two technologies available for hg19 build, only the one carrying the latest date will be shown. The versions for any one build differ in annotations.

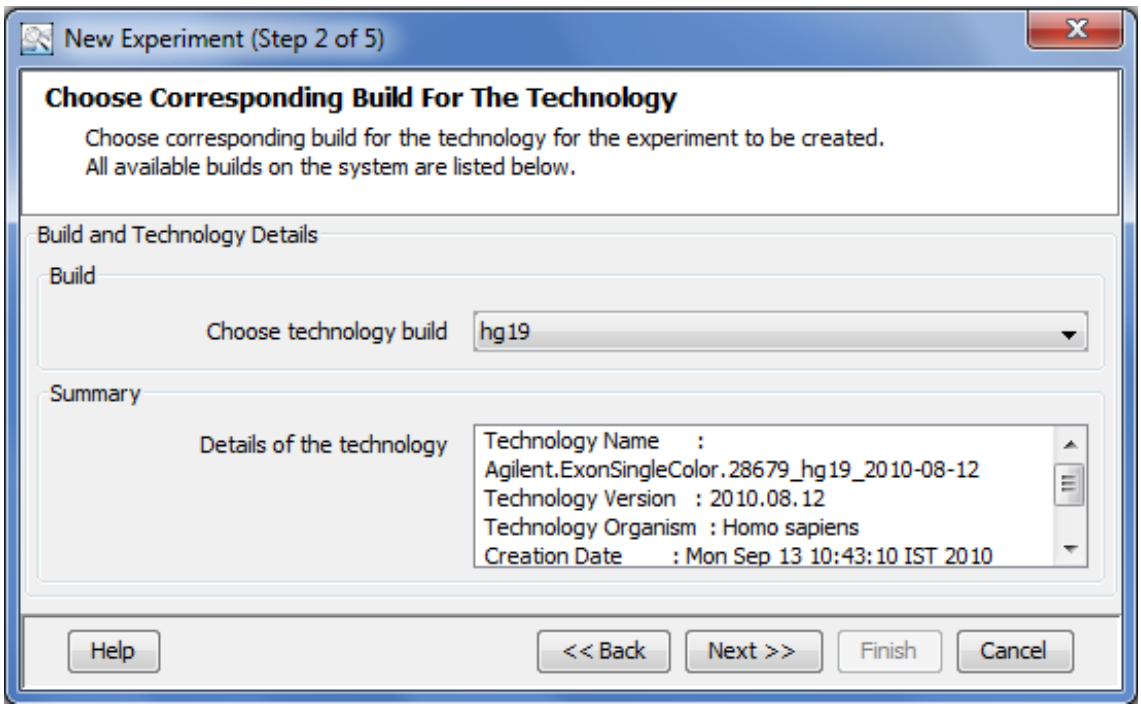


Figure 16.5: Chose Corresponding Build for the Technology

Advanced Flag Import (Step 3 of 5)

Flag information derived from the Feature columns in data file is imported at this step. The recommended flag settings appear as default. These can be changed and re-set as default using the "Save as default" box. See Fig 16.6

Alternatively, these settings can be changed from Tools → Options → Miscellaneous → Agilent Flag Settings.

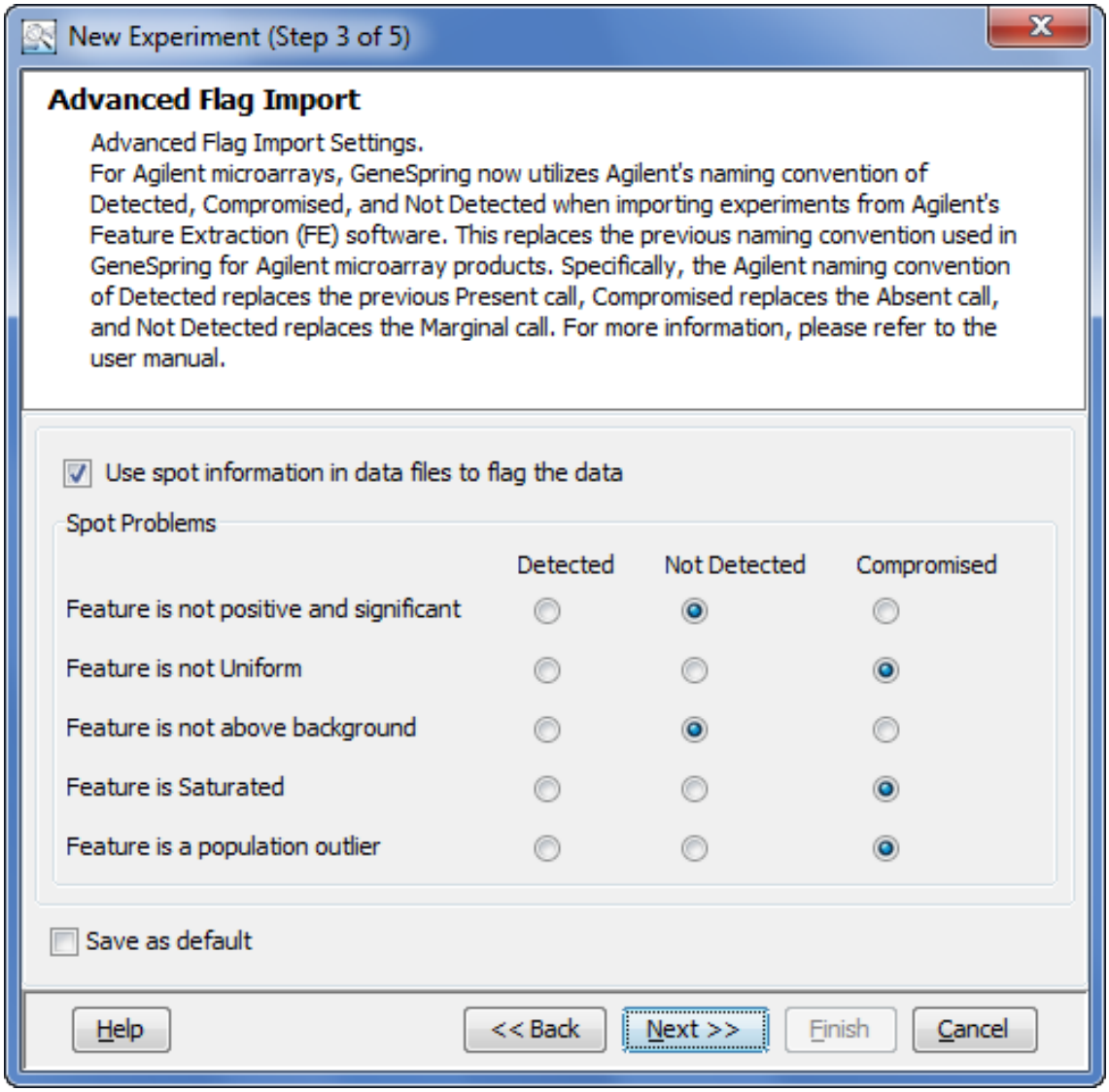


Figure 16.6: Advanced Flag Import

Normalization Options (Step 4 of 5)

GeneSpring offers various data pre-processing criteria. This step allows the user to threshold raw signals to chosen values and select normalization algorithms (None, Percentile Shift, Scale, Quantile, or Normalize to External Value). See Fig 16.7. For detailed description of each normalization methods, please refer to section Normalization Algorithms.

- **None:** No normalization is done.

- **Percentile Shift:** Percentile shift normalization is a per-sample normalization. The data of each sample is shifted such that it is now centred around the defined percentile value. The recommended percentile value is 75^{th} percentile.
- **Scale:** Scaling helps the user in overcoming the inter array differences. Scaling can be done to either the 'Mean of Means' or to 'Median of target percentiles'. For more details refer to section: 24.2.2. If control samples are used for scaling, the user must select the control samples from the available samples in the **Choose Samples** box. The **Percentile target** box is set to a default value of 75 when *Scale to median of percentile target* is chosen and is disabled if *Scale to mean of means* option is chosen.
- **Quantile:** Quantile normalization helps to make the distribution of expression values of all samples in an experiment the same. Thus after normalization, all statistical parameters of the sample, i.e., mean, median and percentiles of all samples will be identical. This works quite well at reducing variance between arrays.
- **Normalize to External Value:** It allows the user to configure the *Scaling Factor (default value: 1)* for each sample using the *'Assign Value'* button. The user can select one or multiple samples simultaneously and assign a value.

Note: 'Normalize to Control genes' is not available for Agilent Exon

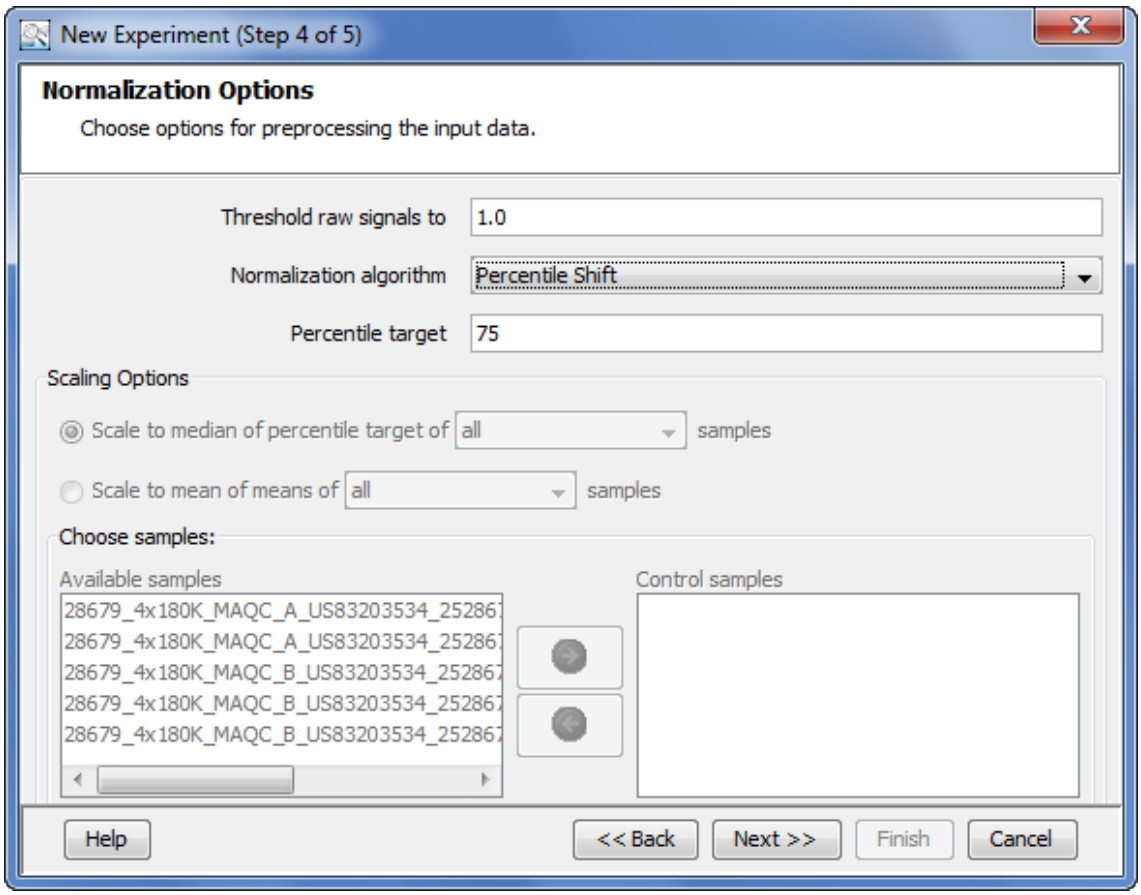


Figure 16.7: Normalization Options

Preprocess Baseline Options (Step 5 of 5)

The baseline transformation option chosen at this stage is carried out at both gene-level and

probe-level; Median/Mean of the log normalized values is first computed and then this value is subtracted row-wise from each of the sample values. See Fig 16.8. The baselining can be done using either Mean/Median of 'All' samples or the 'Control' samples.

Baseline to median/mean of control samples allows the user to assign a control or a set of controls for each sample. Alternatively, the user can assign a set of samples as controls for all samples. If a single sample is chosen as the control sample, then the probe values of the control sample is used.

In order to specify the control for a sample, select the sample and click on ***Assign value***. This opens up the ***Choose Control Samples*** window. The samples assigned as *Controls* should be moved from the *Available Items* box to the *Selected Items* box. Click on ***Ok***. This shows the controls for each sample.

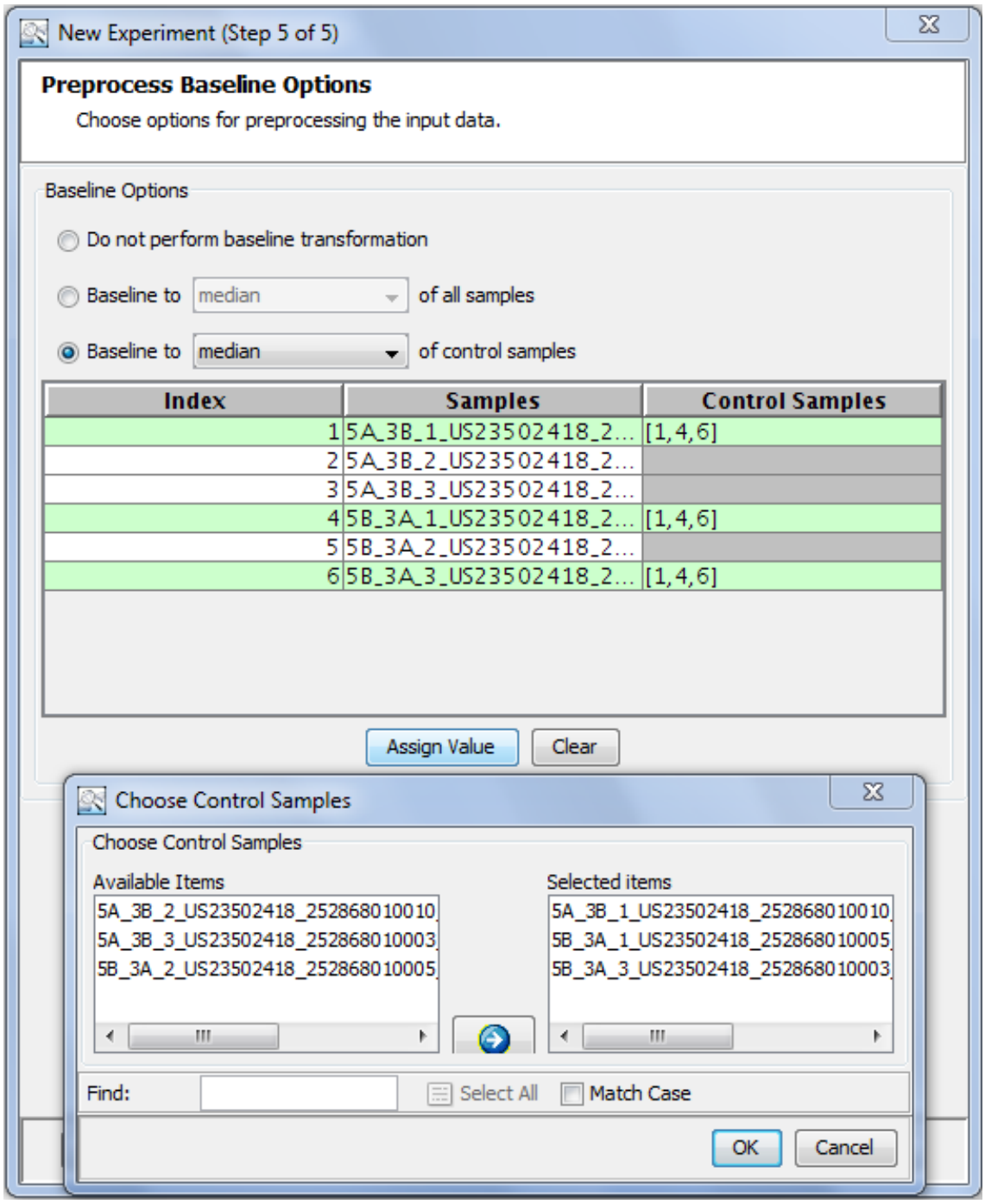


Figure 16.8: Preprocess Baseline Options

16.1.4 Creating an Agilent Exon Two Color Experiment

Choose the experiment type as 'Agilent Exon Two Color' as the first step. The steps to create an Agilent Exon Two Color Experiment remains almost the same as in single color except for the additional step to perform dye swap and the absence of normalization step (In Two Color Experiments, Lowess normalization is done by default in FE files).

Loading Data, Selecting Build for the Technology (Steps 1,2 of 5)

These are exactly as in Agilent Exon Single Color Experiment explained above in section Creating Agilent Exon Single Color Experiment

Choose Dye-Swaps (Step 3 of 5)

Dye-swapping accounts for dye related bias (as both dyes bind DNA with different affinities) that cannot be removed by standard normalization methods.

In **GeneSpring**, during the creation of two color experiments, users can specify sample(s) that have been dye swapped. The experiment creation wizard has an additional step (step 3 of 5) that lists all the sample(s) chosen by the user for experiment creation. At this step users can select sample(s) that have been dye swapped. For sample(s) that have been marked as dye-swapped, **GeneSpring** treats Cy3 as"signal" and Cy5 as "control" and computes the signal as Cy3/Cy5. For non-dye swapped sample(s), **GeneSpring** treats Cy5 as "signal" and Cy3 as "control" and computes the signal as Cy5/Cy3.

The following dialog allows the user to specify the dye-swap arrays: Fig 16.9.

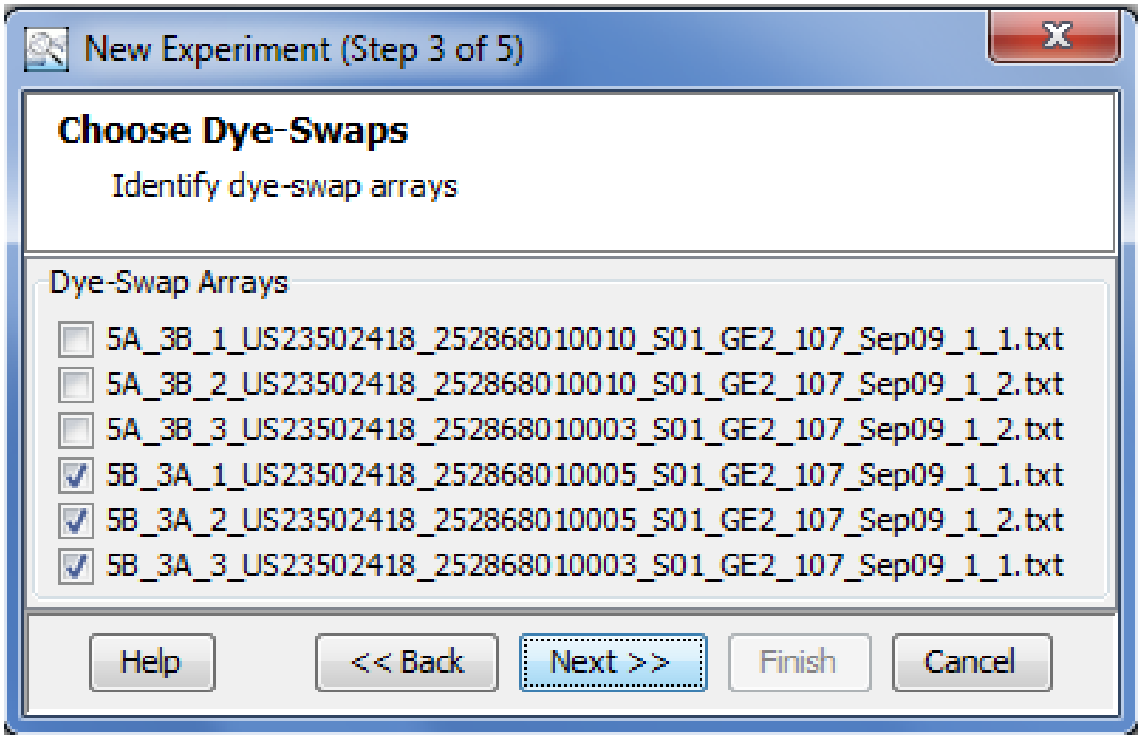


Figure 16.9: Choose Dye Swaps

Advanced Flag Import, Preprocess Baseline Options (Step 4,5 of 5)

These are exactly as in Agilent Exon Single Color Experiment explained above in section Creating Agilent Exon Single Color Experiment

Once the experiment is created **Data Import Wizard** steps appear on the right side of the desktop under the heading 'Workflow'.

16.1.5 Experiment Setup

This gives options to view quick start guide, create experiment grouping and interpretation. See section Experiment Setup for complete details.

16.1.6 Quality Control

16.1.6.1 Quality Control on Samples

Quality Control(QC) on Samples lets the user decide which samples are ambiguous and which are passing the quality criteria. Based upon the QC results, the unreliable samples can be removed from the analysis. The QC view for Agilent Exon shows three tiled windows as shown in Figure 16.10.

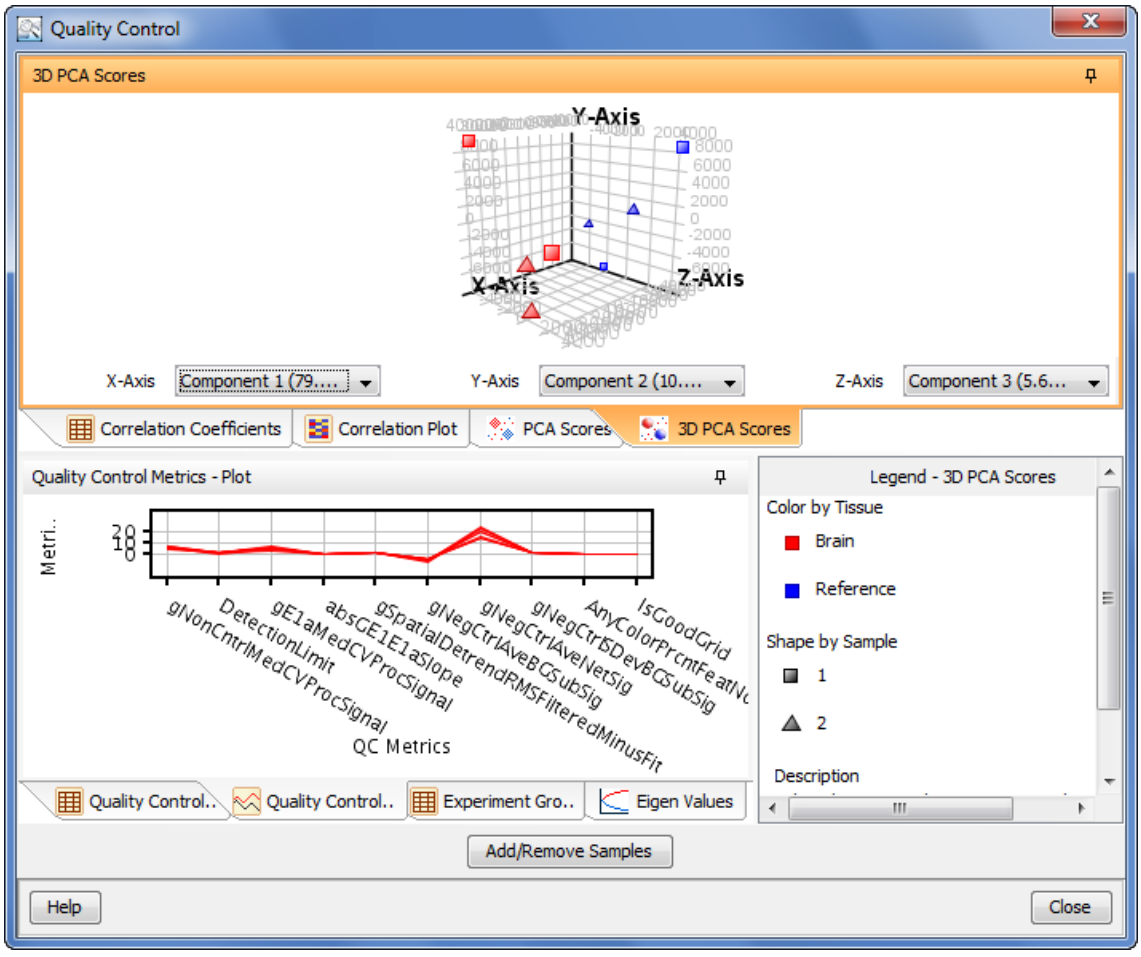


Figure 16.10: Quality Control on Samples for Agilent Exon Experiment

- PCA Scores and 3D PCA scores
- Eigen values, Quality Metrics Report, Quality Metrics Plot and Experiment Grouping tabs.
- Legend

Principal Component Analysis (PCA) calculates the PCA scores and visually represents them in a 3D scatter plot. The PCA scores are calculated from the Covariance matrix which is computed and stored during experiment creation. It shows one point per array and is colored by the *Experiment Factors* provided earlier in the *Experiment Groupings* view. This allows viewing of separations between groups of replicates. Ideally, replicates within a group should cluster together and separately from arrays in other groups. The PCA components, represented in the X, Y and Z axes are numbered 1, 2, 3... according to their decreasing significance. The 3D PCA scores plot can be customized via ***Right-Click*** $\longrightarrow$ ***Properties***. To zoom into a 3D Scatter plot, press the Shift key and simultaneously hold down the left mouse button and move the mouse upwards. To zoom out, move the mouse downwards instead. To rotate, press the Ctrl key, simultaneously hold down the left mouse button and move the mouse around the plot.

Adding / removing samples will essentially re-do all the steps associated with experiment creation as the covariance matrix has to be recomputed with the new set of samples.

For both single and two color experiments, **GeneSpring** uses normalized pre-baseline transformed probe level data to compute correlation or co-variance matrices. However, no Correlation Coefficient and QC plots are displayed for two color experiments. Please refer to Data Processing for more details.

Quality control Metrics Plot shows the QC metrics present in the QC report in the form of a plot. The QC Metrics report is different for single and two color experiments.

Please refer to Agilent Feature Extraction Software Reference Guide (v10.7.3 or higher) for detailed descriptions of the QC Metrics.

16.1.6.2 Filter Genes by Expression, by Flags, by Error

Genes are filtered based on their signal intensity values, flags or error. All these filters work on gene level values (see section Data Processing for details on how these values are generated from probe level data). The steps remain the same for expression, flags or error except for the filtering conditions and are explained here. At the end, an entity list containing the filtered genes is created.

Step 1: Entity List and Interpretation: An entity list is chosen and an interpretation is selected to begin the filtering.

Step 2: Input Parameters: The filtering conditions and the stringency of the filter is defined in this step.

1. For **Filter by Expression**, choose either raw or normalized data to be taken as input for filtering. The range of interest for filtering can be absolute value or in terms of percentile of the signal intensity values. The stringency is defined as either the number of samples in which the entities (genes) must pass the filtering condition or entities within a defined number of conditions. The step is shown in Figure 16.11.

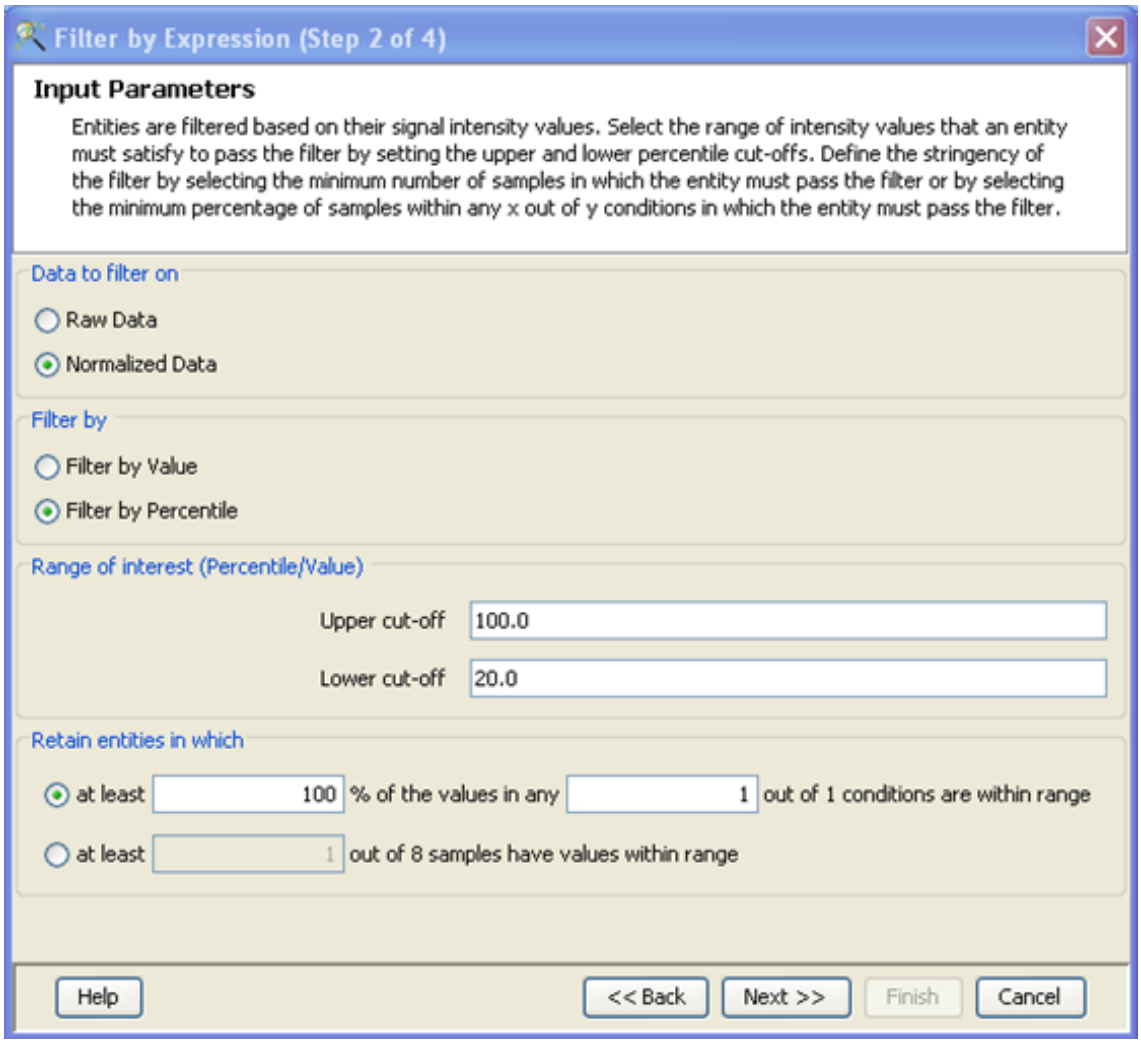


Figure 16.11: Filter by Expression - Input Parameters

2. For **Filter By Flags**, the filtering condition is the acceptable flag(s) out of Detected, Not Detected or Compromised flags. The stringency is defined as either the number of samples in which the entities(genes) must pass the filtering condition or the percentage of samples in which the entities (genes) must pass the filtering condition. The step is shown in Fig 16.12. For more information on Flag logic, please refer to examples in section on: Agilent Flag Logic

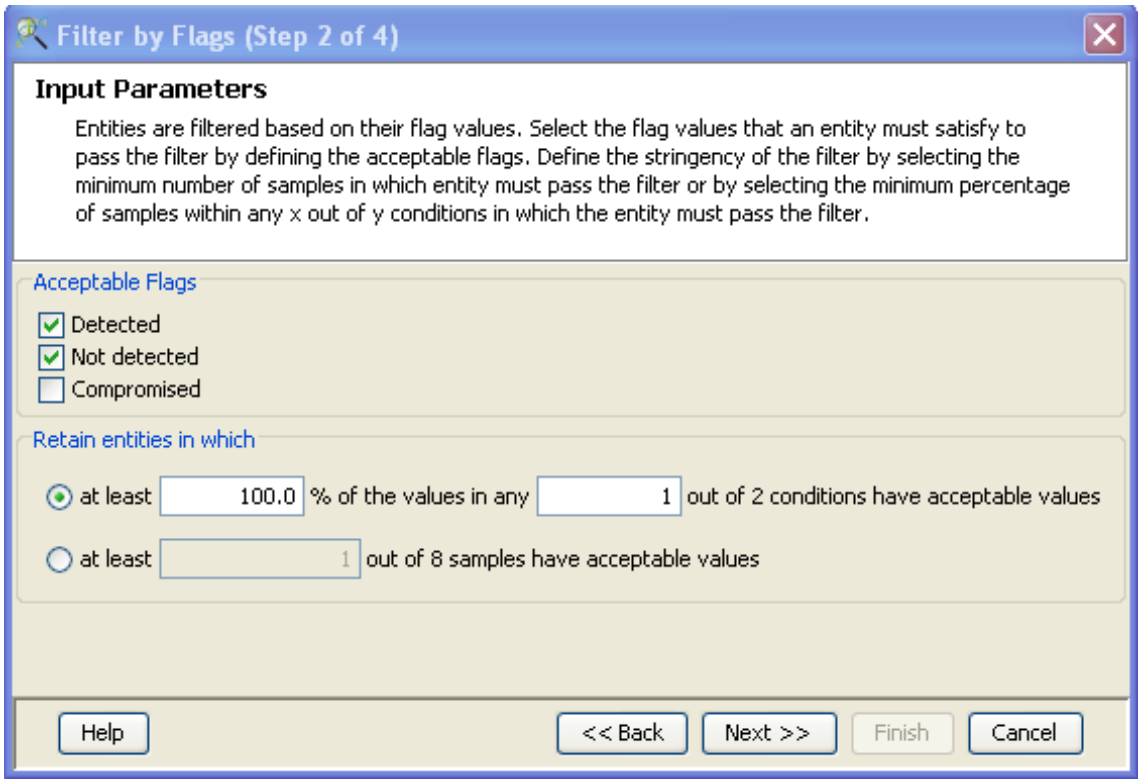


Figure 16.12: Filter by Flags - Input Parameters

3. For **Filter by Error**, the filtering condition is the deviation of the expression values within the condition. The deviation can be in terms of Coefficient of Variation or Standard Deviation. Standard Deviation is a measure of variation or dispersion from the average or mean. Coefficient of Variation is a normalized measure of this dispersion. It is defined as the ratio of Standard Deviation to the Mean. Choose one of them and define a range of interest by choosing a lower threshold or an upper threshold. The stringency is defined by selecting the number of conditions in which the entities (genes) must pass the filtering condition. The step is shown in Fig 16.13. This filter runs on Gene-normalized values.

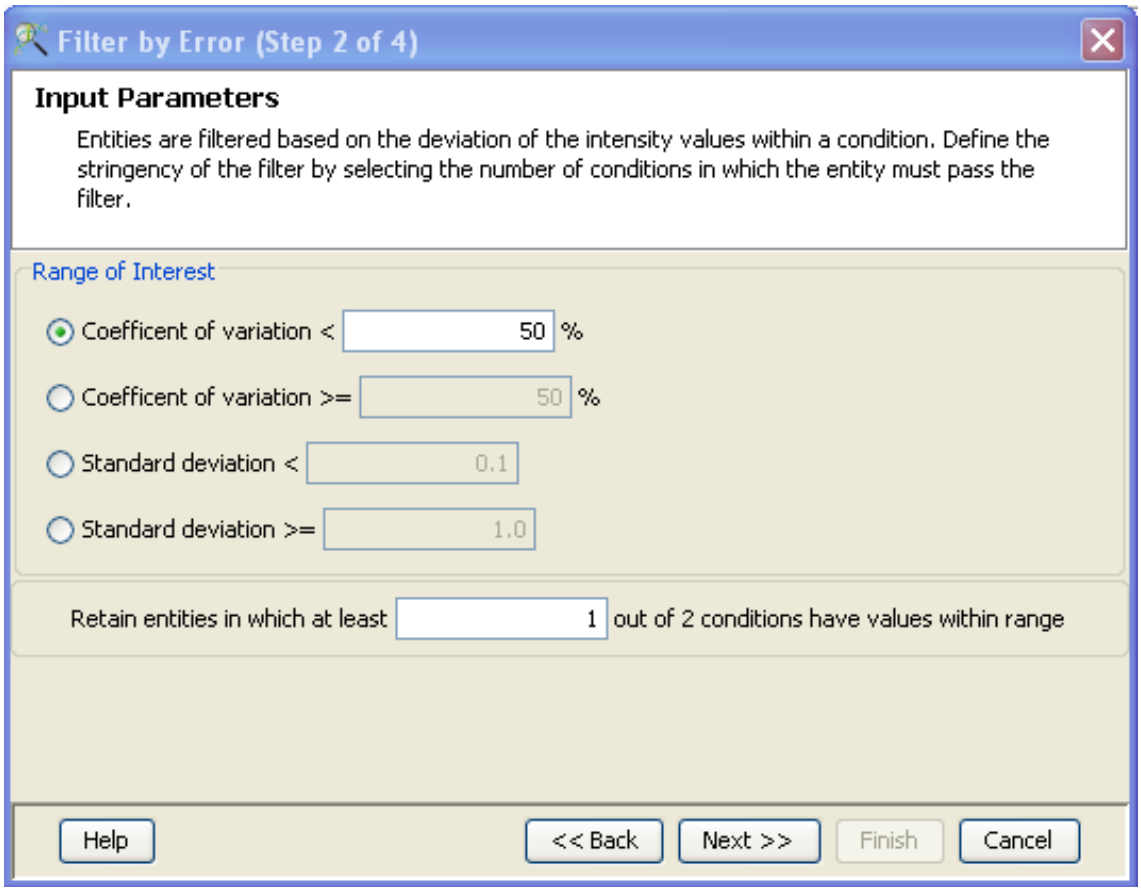


Figure 16.13: Filter by Error - Input Parameters

Step 3: Output Views: Output view is shown in Fig 16.14.

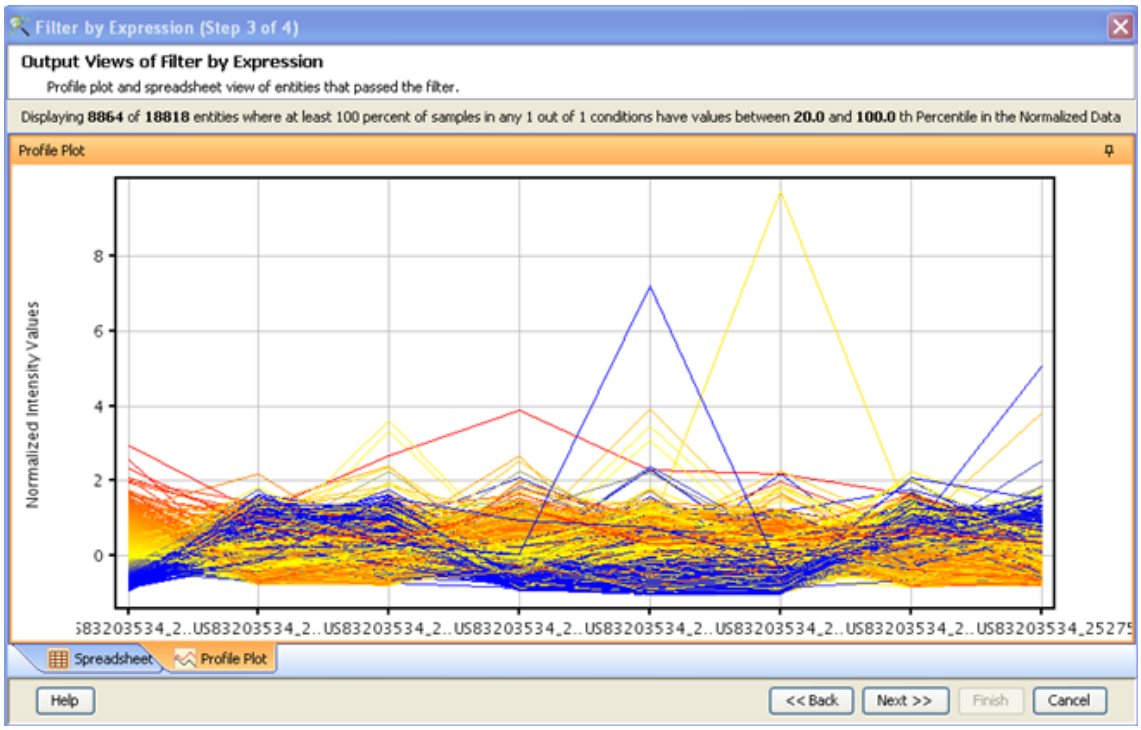


Figure 16.14: Filter - Output Views

The 'Output Views' dialog shows two tabs: Profile plot and Spreadsheet view of entities that passed the filter criteria specified in Step 2. The 'Number Passed' column in the spreadsheet view displays the number of samples or conditions that satisfied the filter criteria.

Step 4: Save Entity List: The last step allows configuring and saving the entity list created by the filtering process. This entity list contains only the gene level information.

16.1.7 Gene Level Analysis

This section describes the expression analysis for an Agilent Exon Experiment. The process is similar to expression experiments e.g 44k, which usually have a single probe per gene. In Exon arrays the gene level values are obtained by summarizing the signal intensity values of probes mapping to various exons in the gene.

- **Statistical Analysis**

  For details refer to section Statistical Analysis in the Data Import Wizard section.

- **Filter on Volcano Plot**

  For details refer to section Filter on Volcano Plot.

- **Fold Change**

  For details refer to section Fold change.

- **Clustering**

  For details refer to chapter Clustering Analysis.

- **Correlation**

  For details refer to chapter Correlation Analysis.

- **Find Similar Entities**

  For details refer to section Find similar entities.

- **Filter on Parameters**

  For details refer to section Filter on Parameters.

- **Principal Component Analysis**

  For details refer to section PCA.

16.1.8 Exon Level Analysis

Before we discuss the UI for Exon Level Analysis, it is important to understand the theory behind Splicing Analysis and how **GeneSpring** computes the Splicing Index and p-value for a given exon.

16.1.8.1 Splicing Analysis

Splicing analysis helps to find out genes that are differentially spliced between the experiment groups as well as to identify the exons that are differentially spliced within that gene. Essentially, the analysis happens at the probe (exon) level but the results are summarized and shown at the gene level. However, please note that a probe may not be representative of an exon. Some exons with alternative splice sites may have two or more probes mapping to them. The parameters useful in identifying the splicing events are Splicing Index and the p-value associated with the splicing. The following sections explain how it is done in **GeneSpring** .

- **Splicing Index**
- **Statistical Splicing Analysis**

Both analyses take an entity list and an interpretation as input; Note that it is possible for certain genes to be excluded from the analysis based on the permitted flags used to create the interpretation.

16.1.8.2 Splicing Index

Splicing Index is the difference between the Gene Normalized Probe Normalized (GNPN) values between any two conditions. The GNPN value is the relative intensity resulting from the division of the probe-normalized value by the geometric mean of probe -normalized values of all probes mapping to that gene. The GNPN value of a probe is reflective of the differences in its probe signal intensity across conditions due to splicing alone unlike the probe-normalized intensities which may vary due to many factors like expression level of the gene, transcript stability and differential splicing. Hence **GeneSpring** uses the GNPN values for splicing index computation.

The Splicing Index is calculated for every probe in every gene and finally one value is assigned to each gene. This value is then used to filter the genes in the 'Filter on Splicing Index' workflow. The steps involved in computing the Splicing Index and filtering of genes on the basis of Splicing Index are explained here.

- Get all associated probes for each gene.
- Take one gene; let P1, P2, P3, P4 and P5 be the associated probes for this Gene G1.
- Assume there are 2 conditions (C1 and C2) and 2 samples in each condition (S1 and S2; S3 and S4). For the sake of this example, we will assume that the splicing index is calculated for 'pairs of conditions - C1 vs C2'.
- Let e11 be the probe normalized value for P1 for S1, e12 for P1 for S2, e21 for P2 for S1, e22 for P2 for S2, etc.
- Let g1 be the gene normalized value for Gene G1 in sample 1, g2 for G1 in sample 2, etc.
- Divide each of the 'e' values for this probe by g1 ; this ratio of Probe Normalized Value to the Gene Normalized Value is the **Gene Normalized Probe Normalized Value**.
- Take an average of this ratio for each condition.
- For each probe, we get the Averaged Gene Normalized Probe Normalized Value for Condition 1 and subtract the value from that of Condition 2 to obtain the Splicing Index for that probe.

  Table showing Splicing Index Calculation

  Please note that since the values are log transformed, subtraction is carried out instead of division.
- Now, Gene G1 has 5 Probes and 5 Splicing Indices; Choose the **Maximum value** from this list and associate that value to the Gene. In this example, let us assume that a5c1/a5c2 is the highest value and Gene G1 would be assigned this Splicing Index Value. For example, if a gene had 5 associated exons and their splicing indices were 2.67, 2.01, 2.99, 1.50 and 2.22, the gene would get a splicing index value of 2.99.

Probes	Sample 1	Sample 2	Average for C1	Sample 3	Sample 4	Average for C2	Splicing Index
P1	e11/g1	e12/g2	a1c1	e13/g3	e14/g4	a1c2	a1c1/a1c2
P2	e21/g1	e22/g2	a2c1	e23/g3	e24/g4	a2c2	a2c1/a2c2
P3	e31/g1	e32/g2	a3c1	e33/g3	e34/g4	a3c2	a3c1/a3c2
P4	e41/g1	e42/g2	a4c1	e43/g3	e44/g4	a4c2	a4c1/a4c2
P5	e51/g1	e52/g2	a5c1	e53/g3	e54/g4	a5c2	a5c1/a5c2

- The process is repeated for every gene and a Splicing index value is obtained for all genes; apply cut-off value (default value is 2) and filter the genes.
- Default Splicing Index cut off value of 2.0 is in log scale (base 2).

16.1.8.3 Filter On Splicing Index

Input Parameters (Step 1 of 4)
Choose the entity list and the interpretation for filtering on Splicing Index.

Pairing Options (Step 2 of 4)
This step lets the user decide the pairing option for computing the Splicing Index. The Splicing Index can be computed 'pair-wise' or 'all conditions against a single condition'. For each of them, choose the condition(s) in the drop down box as shown in Figure 16.15.

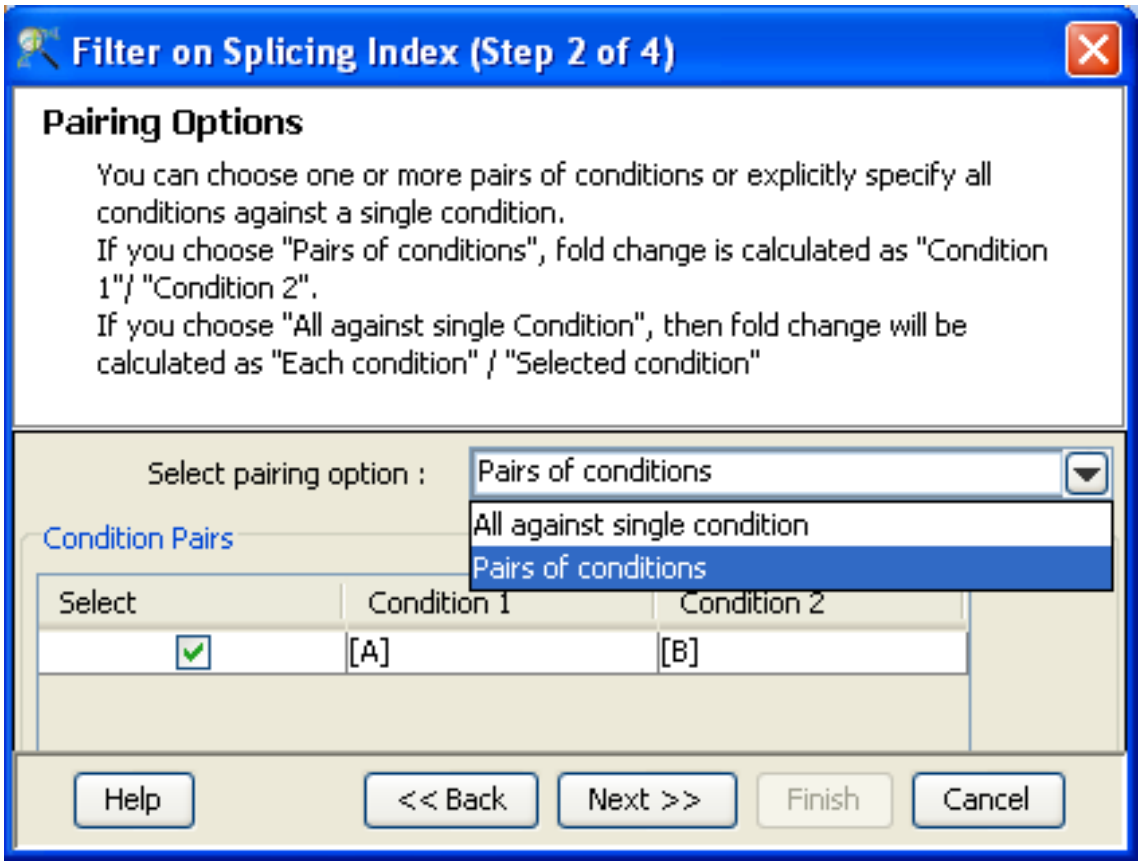


Figure 16.15: Filter On Splicing Index - Pairing Options

Filter on Splicing Index (Step 3 of 4)

This step shows the result of the filtering using the default cut-off value of 2.0. Entities that passed the cut-off and show an upregulated Splicing Index (greater than or equal to 2.0) are displayed in red color and down regulated entities that passed the cut-off are displayed in blue color in the spreadsheet by default. The color for up-regulation and down-regulation can be configured using: Tools ⟶ Options ⟶ Agilent Exon ⟶ Splicing Index Settings. The Splicing Index cut-off can be changed by clicking on the 'Change cut-off' button. The window will dynamically refresh to show the results as per the new cut-off. User can choose gene(s) from this list and click on 'Save Custom List' to create a customized list. Please note that the Minimum number of pairs cut-off signifies the minimum number of condition pairs which should satisfy the Splicing Index cut-off for the gene to pass the filter. See Figure 16.16.

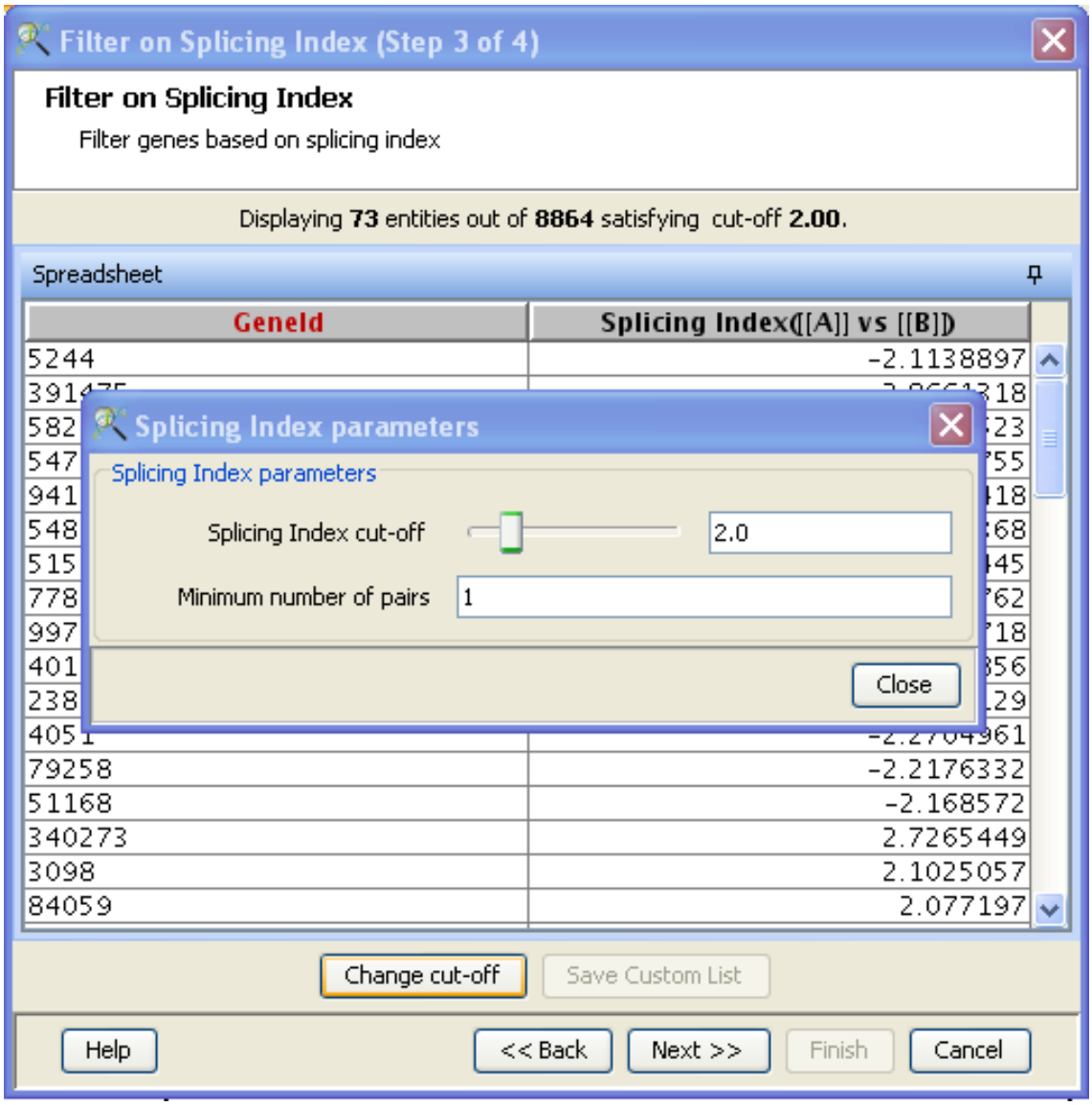


Figure 16.16: Filter On Splicing Index - Filter On Splicing Index

Object Details (Step 4 of 4)

The last step shown in Figure 16.17 displays the results of the filter. The splicing index value and the regulation type (up or down) are shown here along with the other details about the genes. Upon completion of filtering, a separate entity list for both UP and DOWN regulated Splicing Indices are created for all the pairs of conditions.

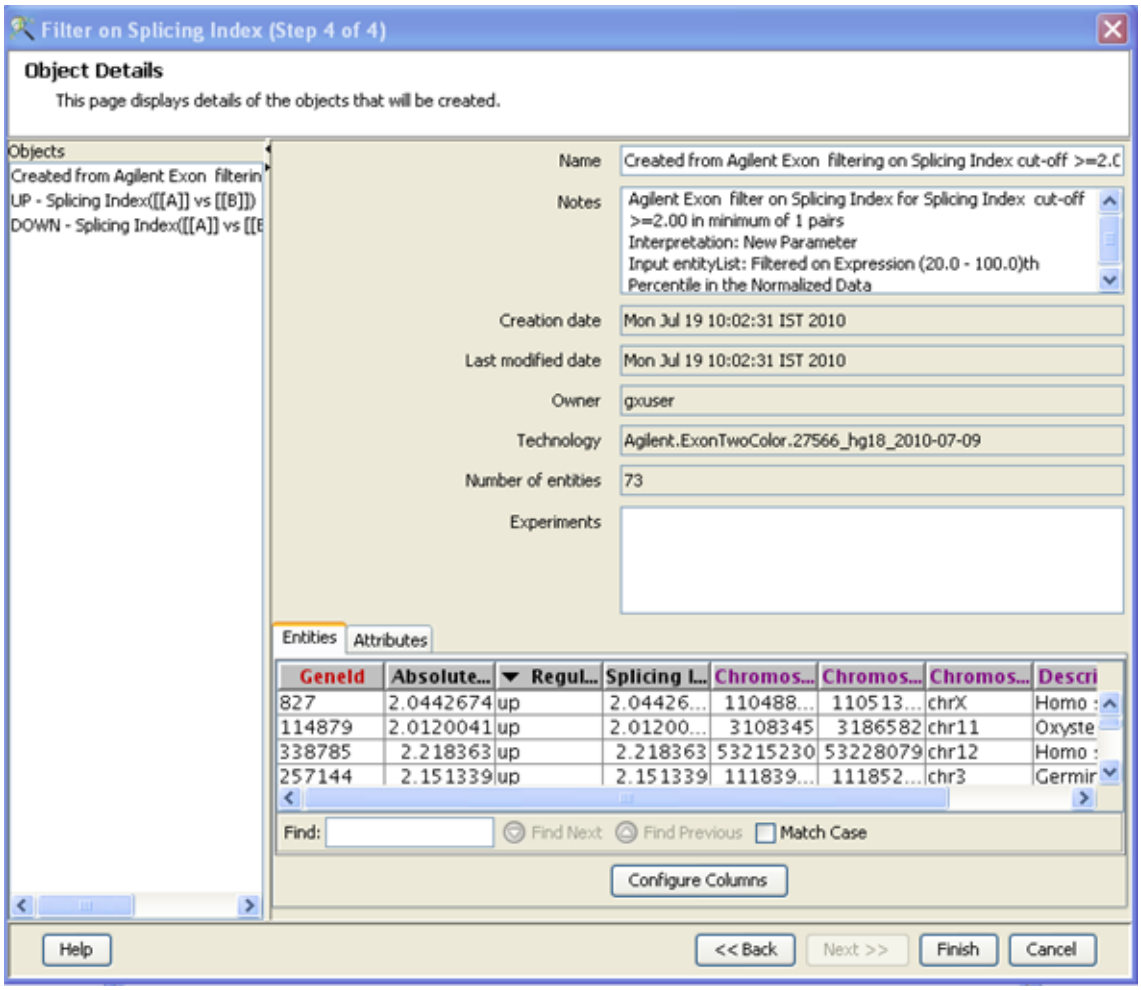


Figure 16.17: Filter On Splicing Index - Object Details

Note: In case of genes with a single probe, the Splicing Index will always be 0. This is because for such genes the GNPN values will be 0.

See section Probe-level Data for Splicing on how probe-level data can be extracted from here.

16.1.8.4 Statistical Splicing Analysis

Statistical Splicing Analysis computes the p-value for all probes for each gene in the experiment. A gene gets assigned a p-value which is the lowest p-value amongst the list of probes for that gene. It can take any entity list (gene list) as input and allows the user to filter out probes based on flag information as well as filter probes based on expression levels. The result of the statistical splicing analysis is a special node called **Results Node** that carries information on all the parameters that were used to run the statistical splicing analysis. In **GeneSpring**, the Statistical Splicing Analysis is done for each probe as explained below.

- Get all associated probes for each gene.

- Take one gene; let P1, P2, P3, P4 and P5 be the associated probes for this Gene G1.
- If their p-values were 0.0008, 0.007, 0.01, 0.0009 and 0.005, the gene would get a p-value of 0.0009.
- The process is repeated for all genes and a gene list with p-values (and corrected p-values) is created.

Input Parameters (Step 1 of 9)

It allows the user to select an entity list and an interpretation with at least one parameter.

Select Test (Step 2 of 9)

It allows the user to select a statistical test for computing the p-value.

Agilent Exon experiment in **GeneSpring** provides the following statistical tests when the user has an interpretation with 1 parameter and 2 conditions:

- T-test unpaired
- T-test paired
- T-test unpaired unequal variance
- MannWhitney unpaired
- MannWhitney paired

For interpretations with 1 parameter and more than 2 groups, the following tests are available:

- ANOVA
- ANOVA Unequal Variance
- Kruskal Wallis
- Repeated Measures
- Friedman

Please note that post-hoc is not performed and ANOVA is at Exon level. For interpretations with 2 parameters and multiple groups, 2-way ANOVA is the only option.

Refer to Details of Statistical Tests for more information on these statistical tests.

Figure 16.18 shows step 2 of Statistical Splicing Analysis.

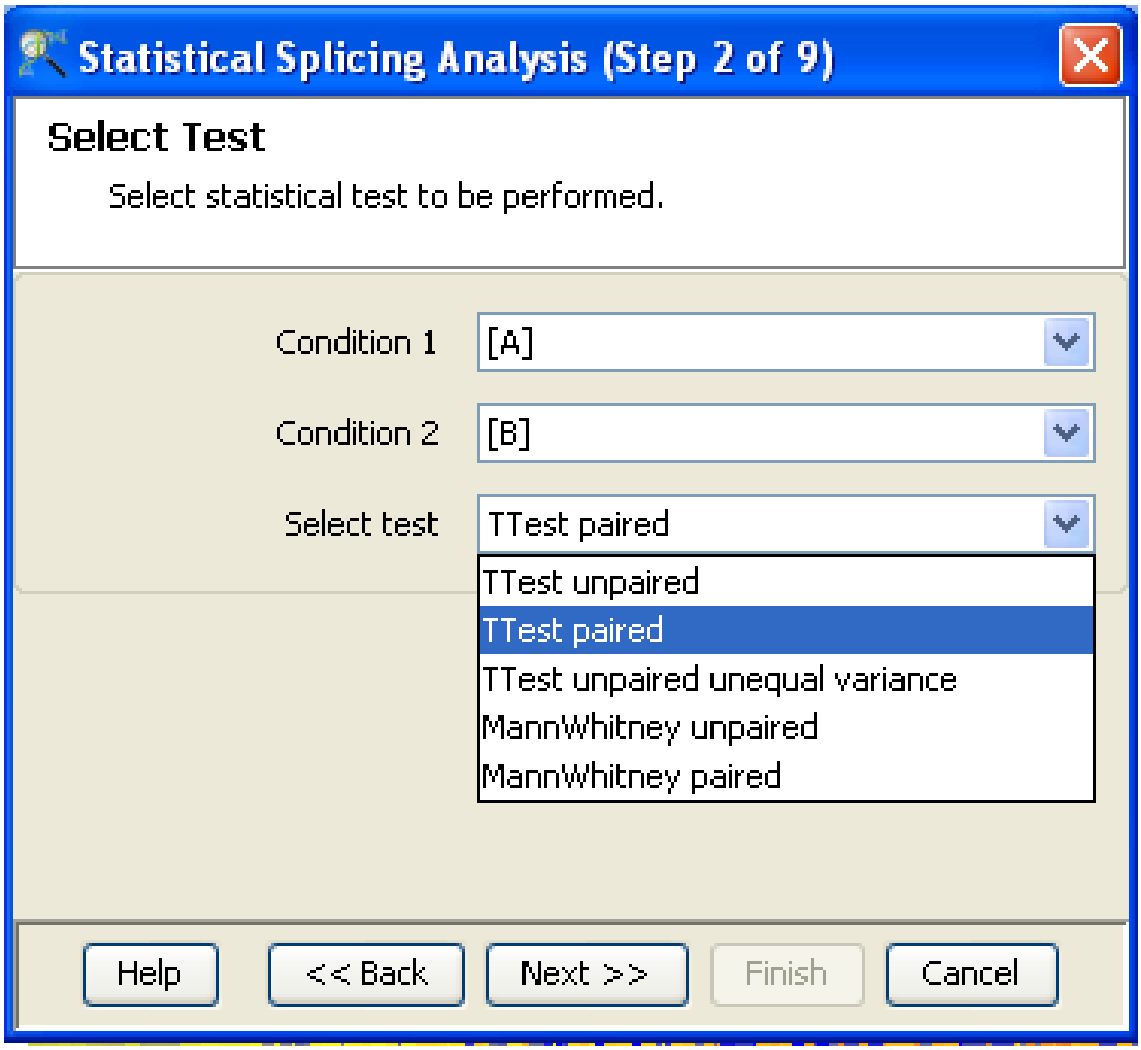


Figure 16.18: Statistical Splicing Analysis - Select Test

For unpaired tests, steps 3 and 4 are skipped. In step 5, user can select a p-value computation method: Asymptotic or Permutative. For Permutative option enter the Number of Permutations in the text box. The specified number of permutations defines the maximum number of permutations that will be performed. The user can also select a Multiple testing correction method as explained in step 4.

Column Reordering (Step 3 of 9)

This step comes up if a paired test (T test or MannWhitney) is chosen in step 2 and it allows reordering of columns.

p-value Computation (Step 4,5 of 9)

Step 4 appears for (T-test or MannWhitney) paired or 2-way ANOVA tests where the p-value computation method is Asymptotic. However, Step 5 appears for unpaired tests and ANOVA. It allows the user to choose either Asymptotic or Permutative method for p-value computation. When Permutative is chosen, then the number of permutations can be specified by the user. Step 4 is shown in Figure 16.19 and Step 5 is shown in Figure 16.20.

It allows user to choose a Multiple Testing Correction Method from the following options.

- No Correction
- Bonferroni FWER
- Bonferroni Holm FWER
- Benjamini Hochberg FDR (default option)

See section Adjusting for Multiple Comparisons for details.

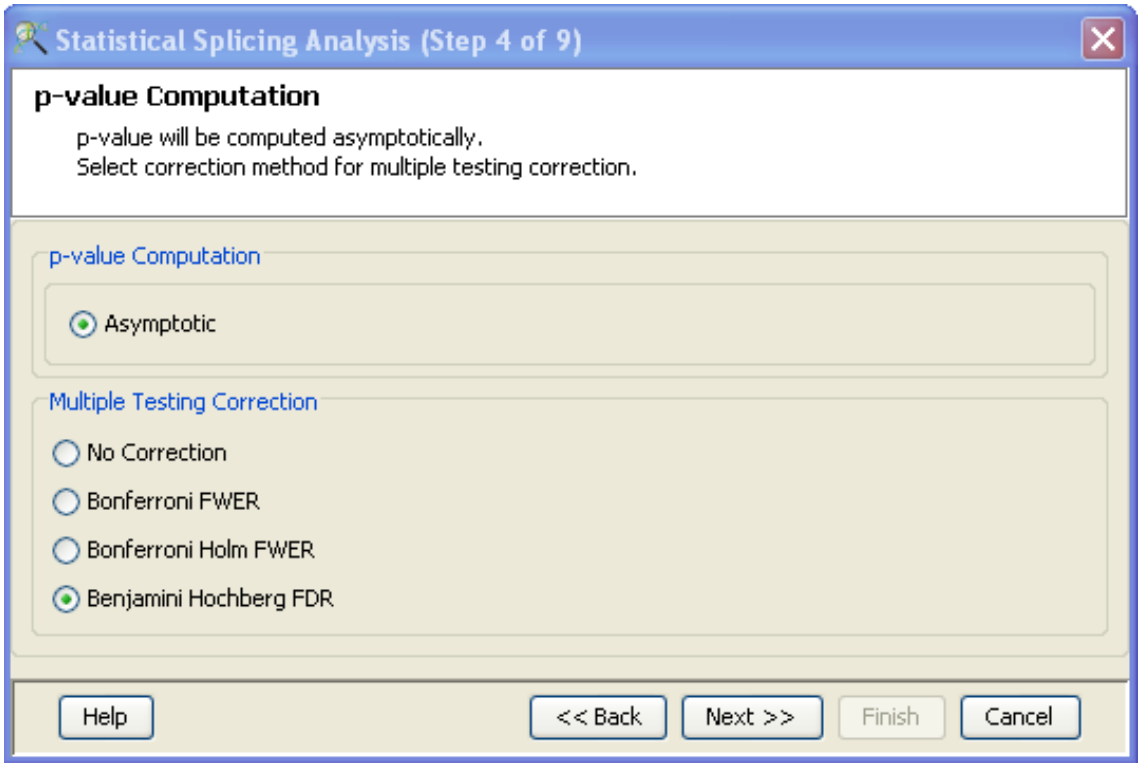


Figure 16.19: Statistical Splicing Analysis - p-value Computation

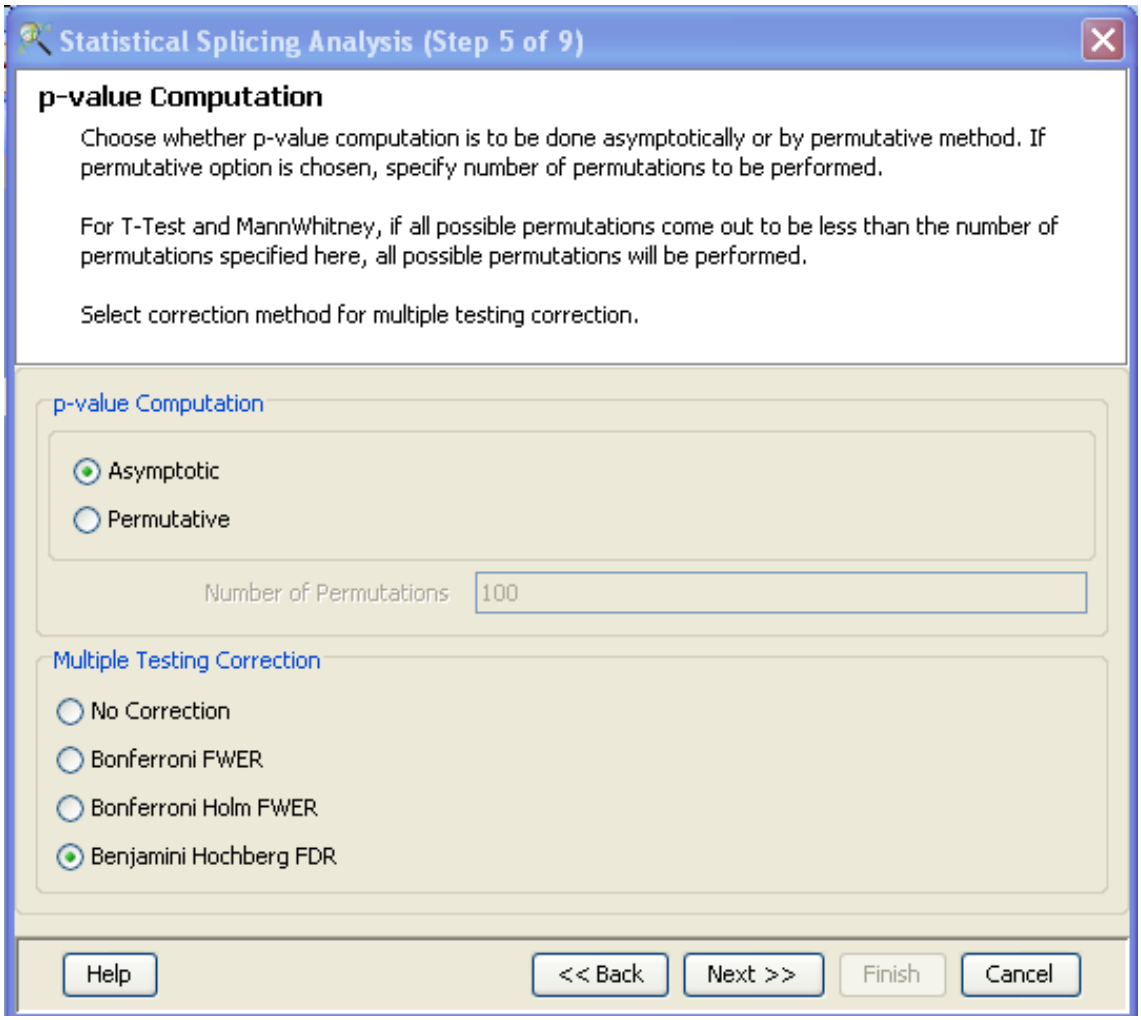


Figure 16.20: Statistical Splicing Analysis - p-value Computation

Filter probes by Flags (Step 6 of 9)

The methodology is exactly as in the QC Filter Genes by Flags step, but this filter runs at the probe level and not at the gene-level. Even if a gene is present in the input entity list, if all of its probes have been filtered out due to the flag criteria chosen, that gene will not be included in the analysis.

Filter Probes (Step 7 of 9)

This step allows the user to filter probes by expression. **GeneSpring** identifies probes to be filtered based on the criteria defined here. Note that the filter is 'off' by default.

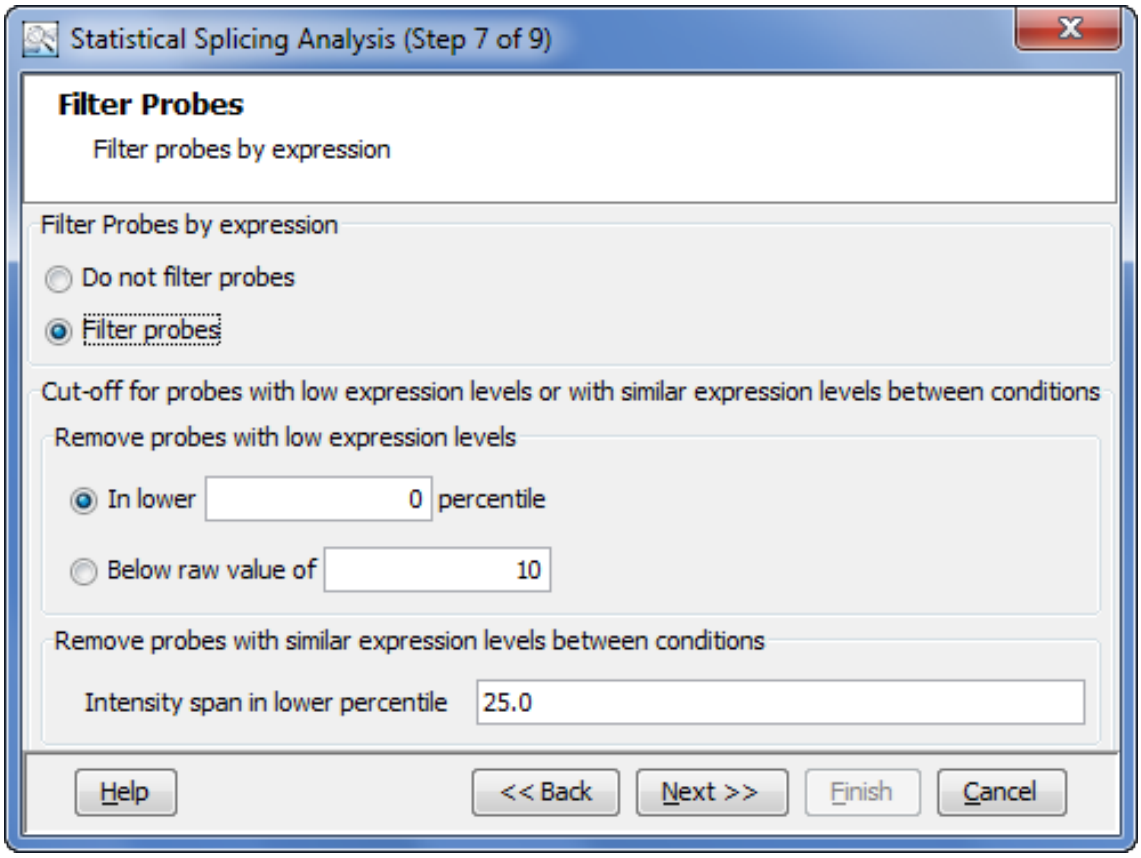


Figure 16.21: Statistical Splicing Analysis - Filter Probes

Probes with low expression levels

This step filters out the probes that show a very low signal value across all samples. Here raw data is used to remove the probes with low expression levels. **GeneSpring** identifies and removes probe(s) whose maximum intensity **across all the samples** in that interpretation is in the following range:

- In lower $< Defined >$ percentile; default is 0
- Below raw value of $< Defined >$; default is 10

Probes with similar expression levels

The probes for which the signal values do not change significantly across samples or conditions are considered to be having similar expression levels. A probe is removed if its Span (i.e., Maximum Intensity - Minimum Intensity) is less than the $< defined >$ percentile of the spans of **all probes in all samples in that interpretation.** The default value is 25. This filter runs on Probe Normalized values.

In case of percentile cut-off, the cut-off is applied to all the probes in the experiment, corresponding to all the genes and not just those corresponding to the input Entity List. For the 'Below Raw Value' option during filtering probes with low expression (default is 10), the cut-off applies only to probes in the input entity list.

Results of Splicing Analysis (Step 8 of 9)

Results of Splicing Analysis are displayed as a Result Summary, Spreadsheet and a Volcano plot.

The **Result Summary** is a matrix that displays the number of genes that would satisfy the given combination of Splicing Index and p-value cut-off.

The **Spreadsheet** gives the corrected p-value and Splicing Index assigned for each gene.

The **Volcano Plot** shows the negative log_{10} of p-value vs the Splicing Index. The entities are colored depending on their passing the default cut-offs:

- **Red**: Entities which pass both **Corrected p-value** and **Splicing Index** cut-offs and are up-regulated.
- **Blue**: Entities which pass both **Corrected p-value** and **Splicing Index** cut-offs and are down-regulated.
- **Green**: Entities which pass the **Corrected p-value** cut-off and fail to pass the**Splicing Index** cut-off.
- **Grey**: Entities which neither pass the **Corrected p-value** cut-off nor the **Splicing Index** cut-off.

The default cut-offs are set as 0.05 (p-value) and 2 (Splicing Index) which can be changed by clicking the **Change cut-off** button. The **Spreadsheet** and **Volcano Plot** will dynamically refresh to show the results as per the new cut-offs. Gene(s) from this list can be chosen to create a customized list which then appears along with the entity list and result node created at the end of the analysis.

This step is shown in Figure 16.22.

The p-value and Splicing Index value are combined using 'AND' condition. The genes passing this test could have the same probe as being the most significant and having the highest Splicing Index. Alternatively, for some genes, the lowest p-value and highest Splicing Index could belong to different probes.

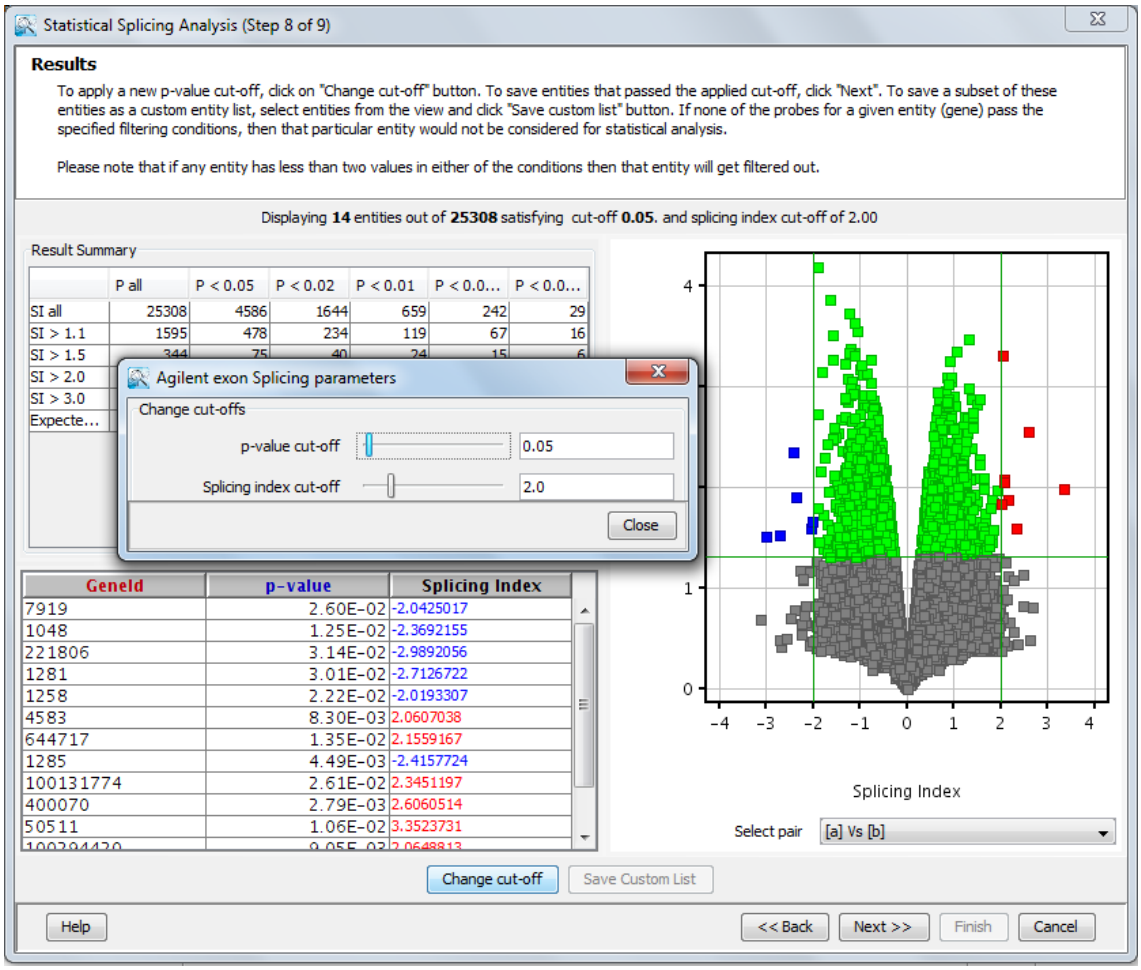


Figure 16.22: Statistical Splicing Analysis - Results

Save Objects (Step 9 of 9)

The gene-list with the p-values are shown here along with other annotation details for the gene. The columns can be configured here.

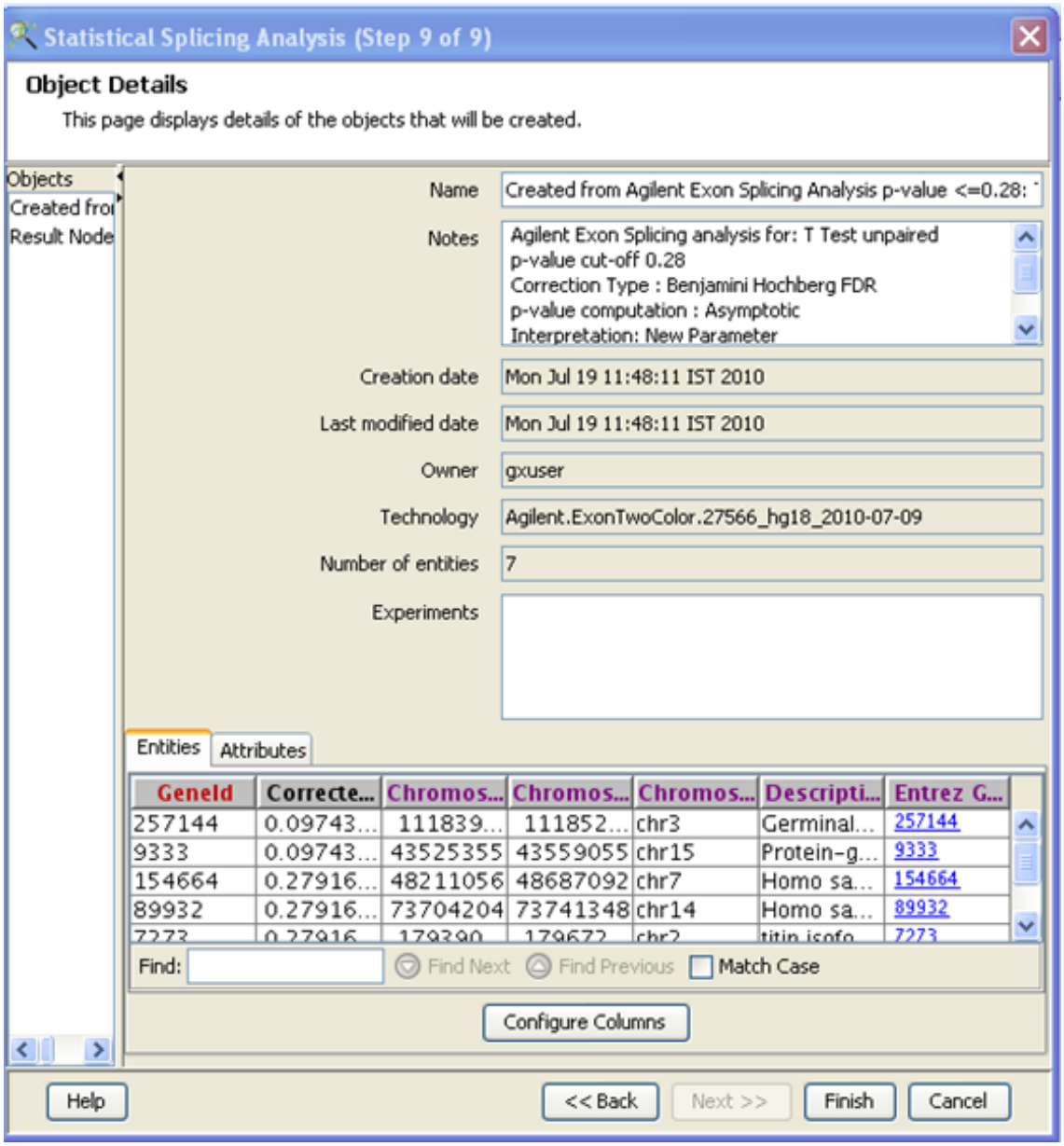


Figure 16.23: Statistical Splicing Analysis - Object Details

On finishing this process, two nodes/objects are created as a result of the Statistical Splicing Analysis, that appear in the experiment navigator section on the left side of the desktop.

1. An Entity List created from Agilent Exon Splicing Analysis p-value <> ; < *StatisticalTest* >
2. A Result node created from Agilent Exon Splicing Analysis p-value <> ; < *StatisticalTest* >

Result Node:
The Result Node is created in the experiment navigator of **GeneSpring** UI. The Result Node is created as a result of Statistical Splicing Analysis and it captures the probe level data generated during Splicing Analysis e.g probe level p-values, probes filtered out as a result of applying the Filter on Expression step of Statistical Splicing Analysis. The information

stored in the Result Node can be viewed by clicking on the Splicing Visualization tab in Exon Level Analysis (user can specify either a Interpretation or a Result Node to visualize the splicing analysis data. Right-clicking on the Result Node, gives three options:

- Inspect Stats Node: This option allows the user to view the test details, splicing status, the date the entity list was created and last modified etc.
- Export Data: User can export either the p-values or Filtered Probes, associated with the genes and/or probes as a tab separated text file. While exporting p-values, the user can choose between exporting Probe or Gene p-values. Though all entities are considered while filtering out the probes with low expression and probes with similar expression across conditions; while exporting the probes, only the probes corresponding to the input entity list are exported.
- Delete Node: This option allows the user to permanently delete the result node and all its descendants (if any).

These nodes are shown in Figure 16.24.

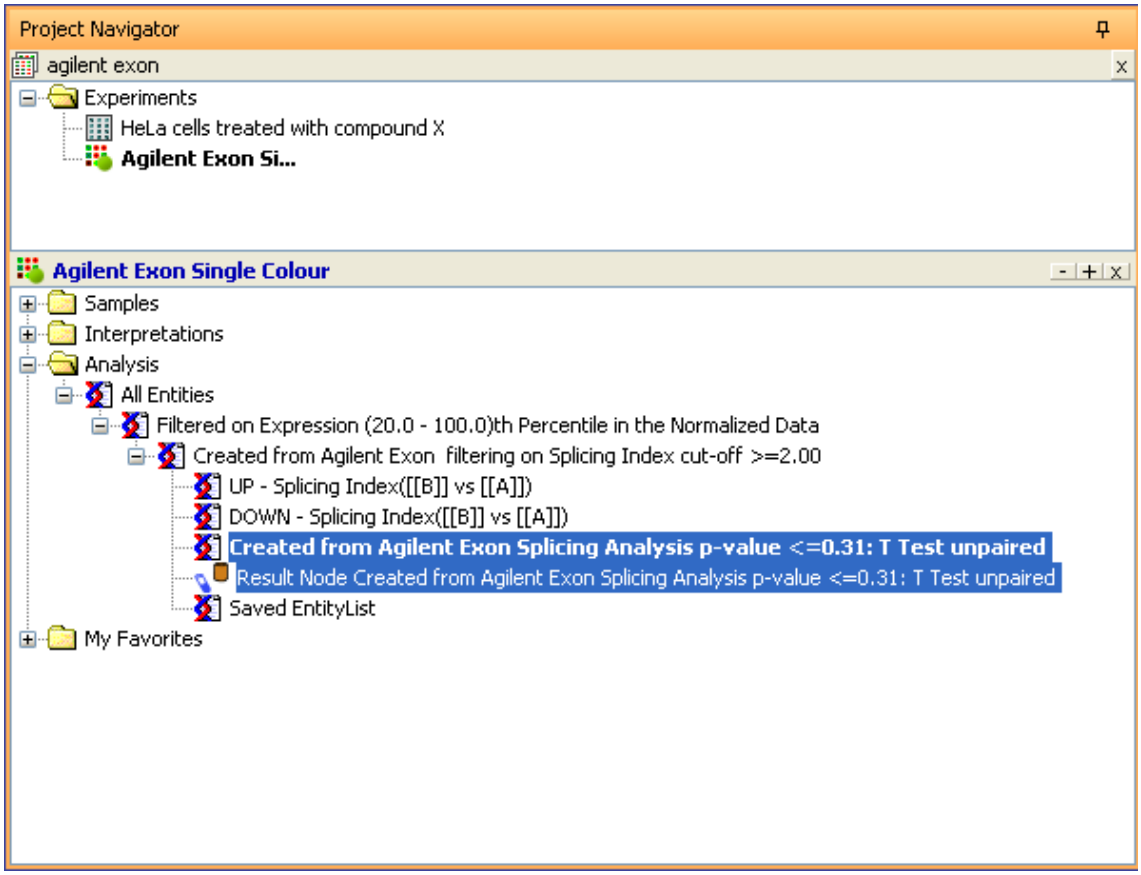


Figure 16.24: Results of Statistical Splicing Analysis

16.1.8.5 Splicing Visualizations

GeneSpring provides a special interface to visualize and analyze splicing events in Agilent Exon experiments. 'Splicing visualization' in **GeneSpring** is a composite view which allows visualizing transcript, exons and their properties in the same view.

Step 1: Input parameters: This step gives the user the option to choose an Entity List and an Interpretation or Result Node. If 'Result Node' is chosen, Statistical Splicing Analysis p-values will be shown in views that follow. The p-value will not be displayed if an interpretation is chosen. The step is shown in Figure 16.25.

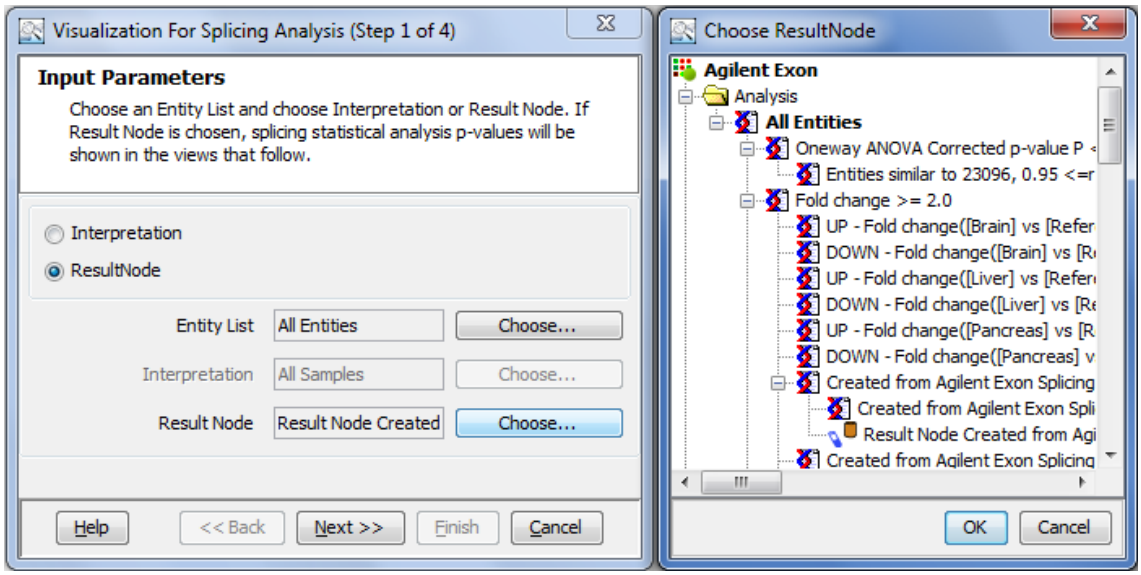


Figure 16.25: Splicing Visualization - Input Parameters

Step 2: Splicing Index: This step gives the user the option to choose one or more pairs of conditions or explicitly specify all conditions against a single condition for the purpose of displaying the Splicing Index. Please note that the p-value and Splicing Index for a gene may not necessarily come from the same probe. This step is shown in Figure 16.26

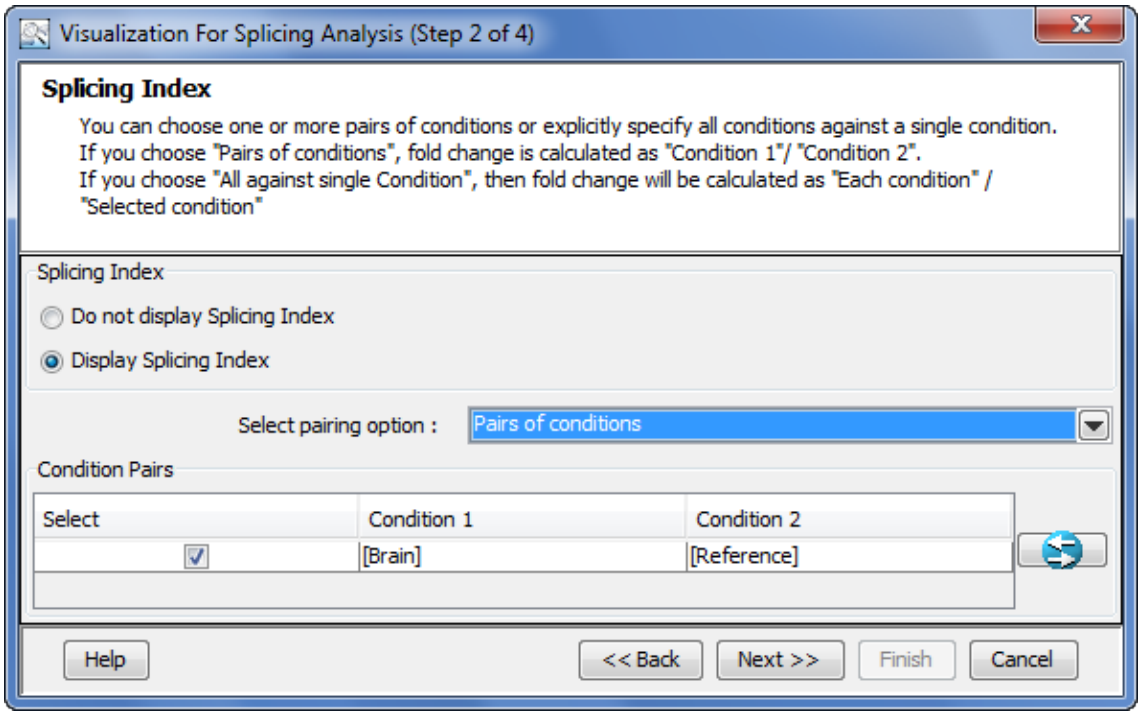


Figure 16.26: Splicing Visualization - Splicing Index

Step 3: Visualization: This window shows two panels: A selection spreadsheet panel on the left and a Gene Data, Probe Data and Gene Sub View panel on the right. Here we introduce a new term called the GeneBinID. It is of the form **geneid:Chr:+/-:Chromosome-Start:Chromosome-Stop**. This effectively contains all the relevant information for the gene entity and is stored in the technology as 'GeneBinID'. It appears in the 'Gene Data' tab on the right panel. GeneBinID can be used to distinguish different instances of genes where Entrez ID is the same but it's genomic cordinates are different.

- A Selection Panel on the left hand side of this screen gives a spreadsheet view of the chosen entities; it lists the GeneIDs, Gene Symbols, GeneBinIDs. In addition to that, it also displays any list associated values present in the chosen entity list. Double click on a gene to view its associated data on the other panel. The user can select gene(s) of interest by checking the check box in the 'Select' Column. The selected genes would then be carried forward in the wizard and an entity list will be created. The 'Search' tab at the bottom can be used to search for genes. While GeneID, GeneSymbol, ProbeID can be used as search terms, GeneBinID cannot be used. To run search, type in the search term and press 'Enter'. Double click on a gene in the search results to bring it to focus. Please note that the search term should start from the beginning of the annotation you are looking for e.g If you are searching for a gene with GeneId 55220, the search term can be 552 or 5522 but it cannot be 220.
- A tabbed view consisting of Gene Data, Probe Data and Gene Sub View on the right side.
  - Gene Data - Lists the annotation data for that gene; gives corrected p-value if launched from the Result Node. The p-value is not available for genes which contain probes all of which have been filtered.
  - Probe Data - Based upon the chosen interpretation, lists the annotation for all the probes (exons) that are associated with the chosen gene and their associated

Gene Normalized Probe Normalized (GNPN) values as determined by the chosen interpretation; If we choose a Result Node in the previous step, then it also gives the p-value and corrected p-value. The p-value cell for a probe is greyed if the probe was filtered out during analysis.

– Gene Sub View - This view shows annotation at the Gene, Transcript and Exon level and also probe level data in a single view.

The third step showing the gene sub view is shown in Figure 16.27

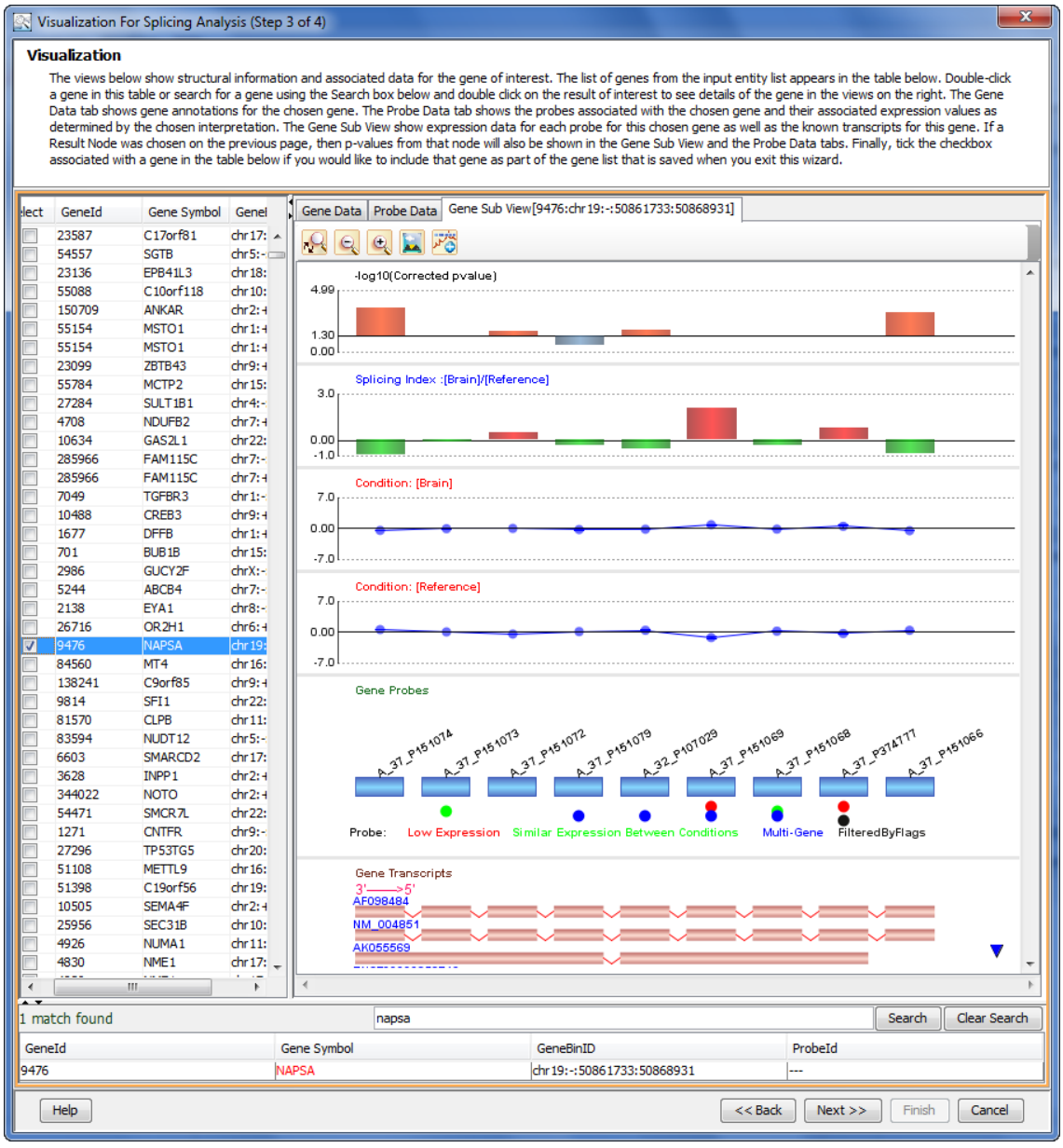


Figure 16.27: Splicing Visualization for Agilent Exon

Gene Sub View:
The Gene Sub View is a composite view consisting of independent plots as explained below. Properties of each plot can be changed by using options on right click menu on each plot.

1. **Corrected p-value track** is shown right on top, only if a result node is visualized. Value that is plotted is actually in (-log10) scale where the default p-value of 0.05 is equivalent to a value of 1.30. Values above and below this threshold are colored different; the thresholds, scales and colors can be changed from the right click option 'Edit Plot Properties → Ruler'.
2. **Splicing Index track** is shown below the p-value track, only if we choose to 'Display splicing index' in step2 of Splicing Visualization. The Splicing indices are shown as a histogram.
3. **Various Condition tracks** - The conditions of the interpretation are obtained from the Node (Result Node or Interpretation) and are shown here based on their Gene Normalized Probe Normalized Value.
4. **Gene Probes track** is the second one from the bottom. This shows the associated probes for the gene arranged in their increasing order of start location on the chromosome. This plot also obtains information on the nature of probes such as probes with low expression, probes with similar expression across conditions, probes which have been filtered by flags and Multi-Gene probes. The first three are generated during the analysis and obtained from the result node (ResultNode) while the last is available from the technology. Using right click 'Edit Plot Properties' option, one can change the legends for these probes and also choose to mark only some of these probes. In the Figure 16.27, all the 4 probe states are seen.
5. **Gene Transcripts track** is the bottom most track and depicts the Ensemble and RefSeq transcripts aligned with the Gene Probes track. In the Figure 16.27, Transcript AK055569 corresponding to Gene NAPSA contains two exon blocks. The 1st block has 4 probes mapping to it as seen in the Exon Probes Plot right on top; similarly 4 other probes map to the 2nd exon block for this transcript. Please note that the size of the exon blocks correspond to the number of probes mapping to it and does not reflect the actual biological size. Please note that the blue arrow at the bottom of this track indicates that there more transcripts which can be viewed by increasing the plot size using the right click menu on the plot. Clicking on the arrow will not increase the size of the plot.

Configuring Tracks in Gene Sub View:

In addition to the above mentioned features in Gene Sub View, various right click options for configuring data columns are available for the user as detailed below.

- **Configuring track properties**
  Right click options are track specific. In any track, on right click choose Edit Plot properties; go to the tab "Rendering" and choose a data column to be shown. For example, for a result node, it is possible to view either the p-value or the corrected p-value using this option.
- **Probes with low expression, Probes with similar expression across conditions, Filtered By Flags, Multi-Gene Probes**
  Right click on the Exon Probes Track and choose the tab 'Probe'. Here, using the check boxes, one can configure the probes that are to be marked along with their legend.
- **Viewing transcript, gene and exon**
  The transcript track has options to configure how the transcript, gene and exon are shown. Right click and choose 'Edit plot properties' and choose the gene, transcript or exon tab to change the way the view is shown.
- **Change Plot size**
  Change the plot size to make the track bigger. It may be useful to bring the bottom most transcript track into full view.

- **Changing references and scales of plot**
  For the data columns (Exon p-value plot and Exon Variance Plot), there is a tab called 'Ruler' in Edit Plot Properties; reference points, scales can be changed from here.
- **Export**
  Individual plots can be exported as images using the right click option 'Export'. The 'Export Multiple plots as image' tool bar icon on top will export all the plots as a composite view as is shown in the tool along with properties.

Select gene(s) of interest by ticking those of interest from the 'Select' Column on the left side to proceed to the next step.

Step 4: Save Entity List The chosen genes will be shown here along with the annotations of the gene. The user can configure the columns of annotations to save. The annotation columns chosen can also be saved as default for the next experiment by clicking on the check box. The list is saved on finishing the process. This step is shown in Figure 16.28

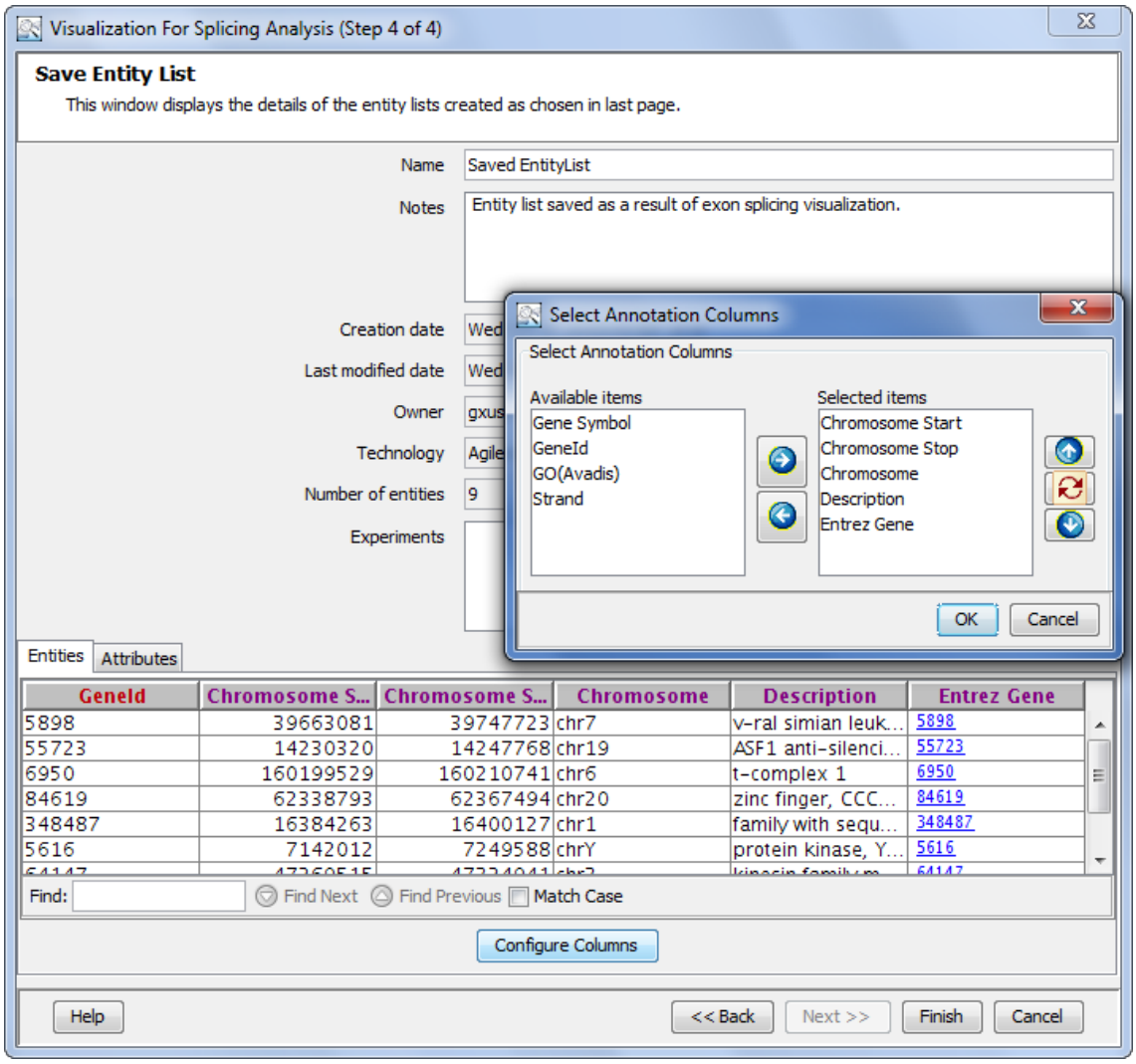


Figure 16.28: Splicing Visualization - Save Entity List

Please note the p-values in the left panel of 'Visualization' UI, is obtained from list associated values, while the p-value used in the right panel displayed in the 'Gene sub view' are obtained from the Result Node. If the Result Node chosen was generated for a different gene list or interpretation, then the values for SI or p-value in the left and the right panels would be different.

16.1.8.6 Probe-level Data

Right click on the entity list in Agilent Exon experiment, for instance, in the above example clicking on the object 'Created from Agilent Exon Splicing Analysis' - there is an option to **'Export Probe Data'**, as shown in Figure 16.29.

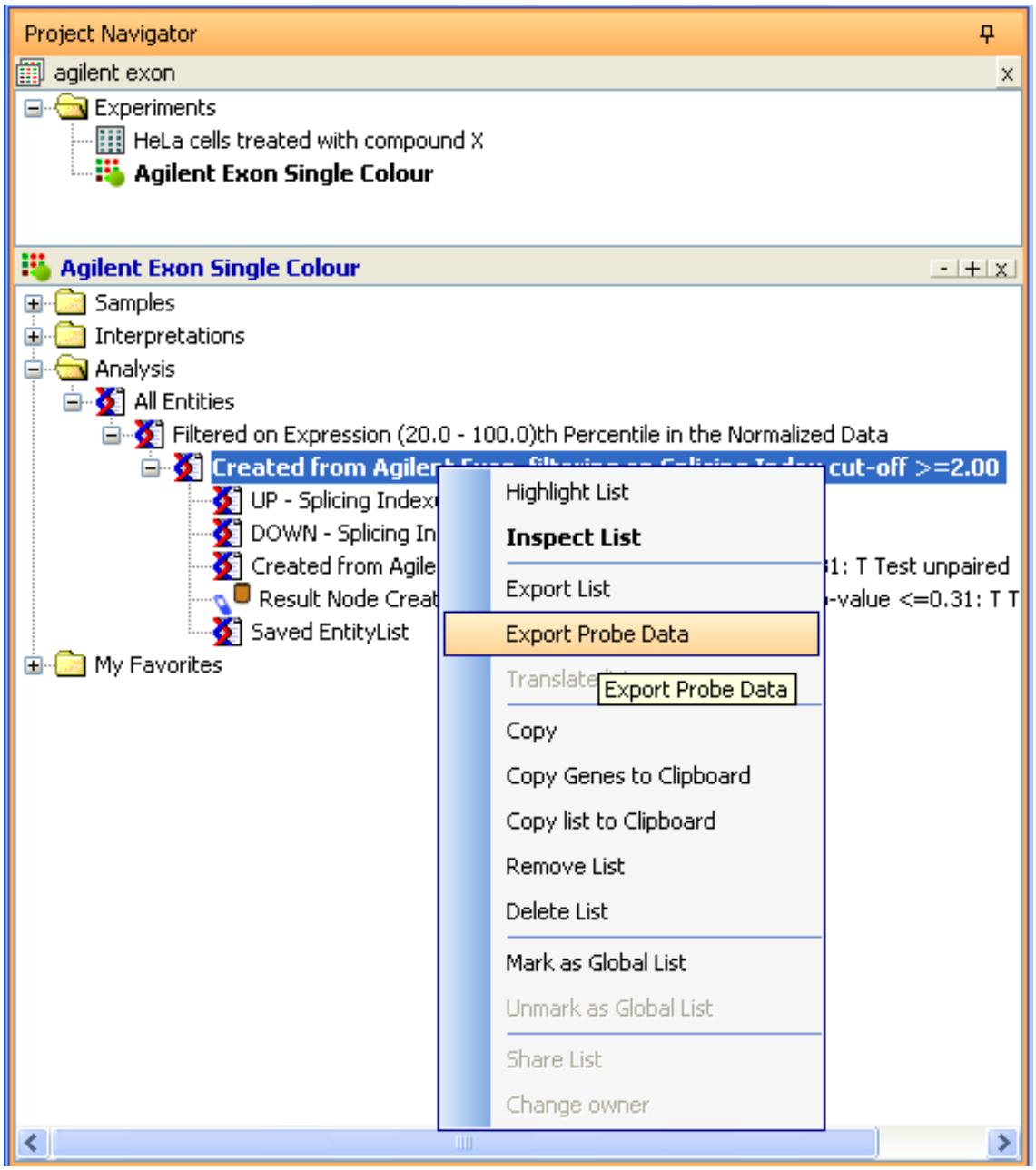


Figure 16.29: Export Probe Data

This opens a window to select an interpretation, the type of probe data to export and the output file name and path.

1. Probe Raw Data
2. Probe Normalized Data
3. Probe Error Data

4. Probe Flag Data
5. Gene Normalize Probe Normalize Data
6. Probe Annotations

The window is shown in Figure 16.30.

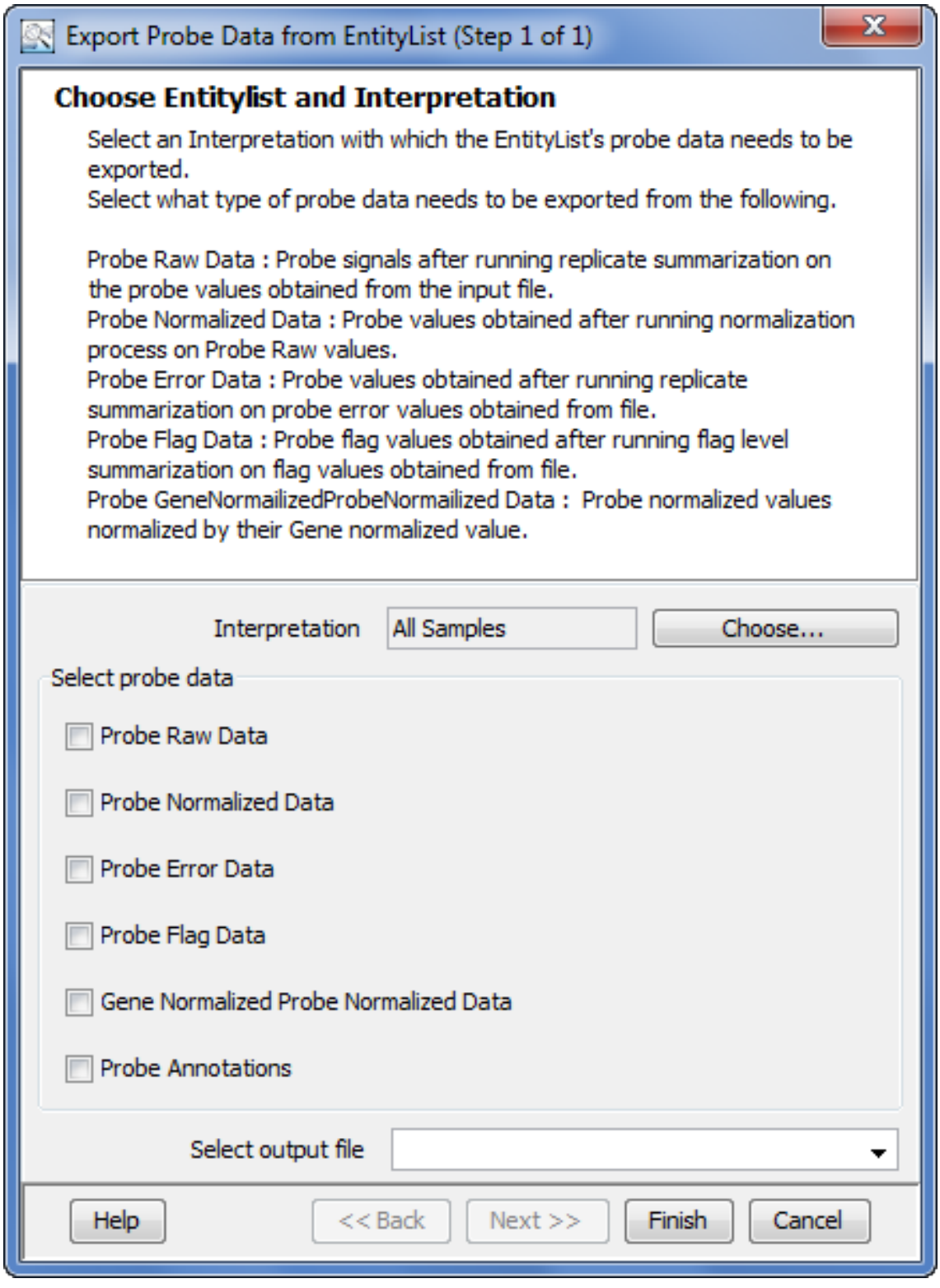


Figure 16.30: Export Probe Level Data - Choose Entity list and Interpretation

For the Result Node that is created as a result of the Statistical Splicing Analysis, there are additional right click options as shown in Figure 16.31. p-values can be exported at probe or gene level or both. **Export Filtered Probes** exports out probes low expression and probes with similar expression across conditions.

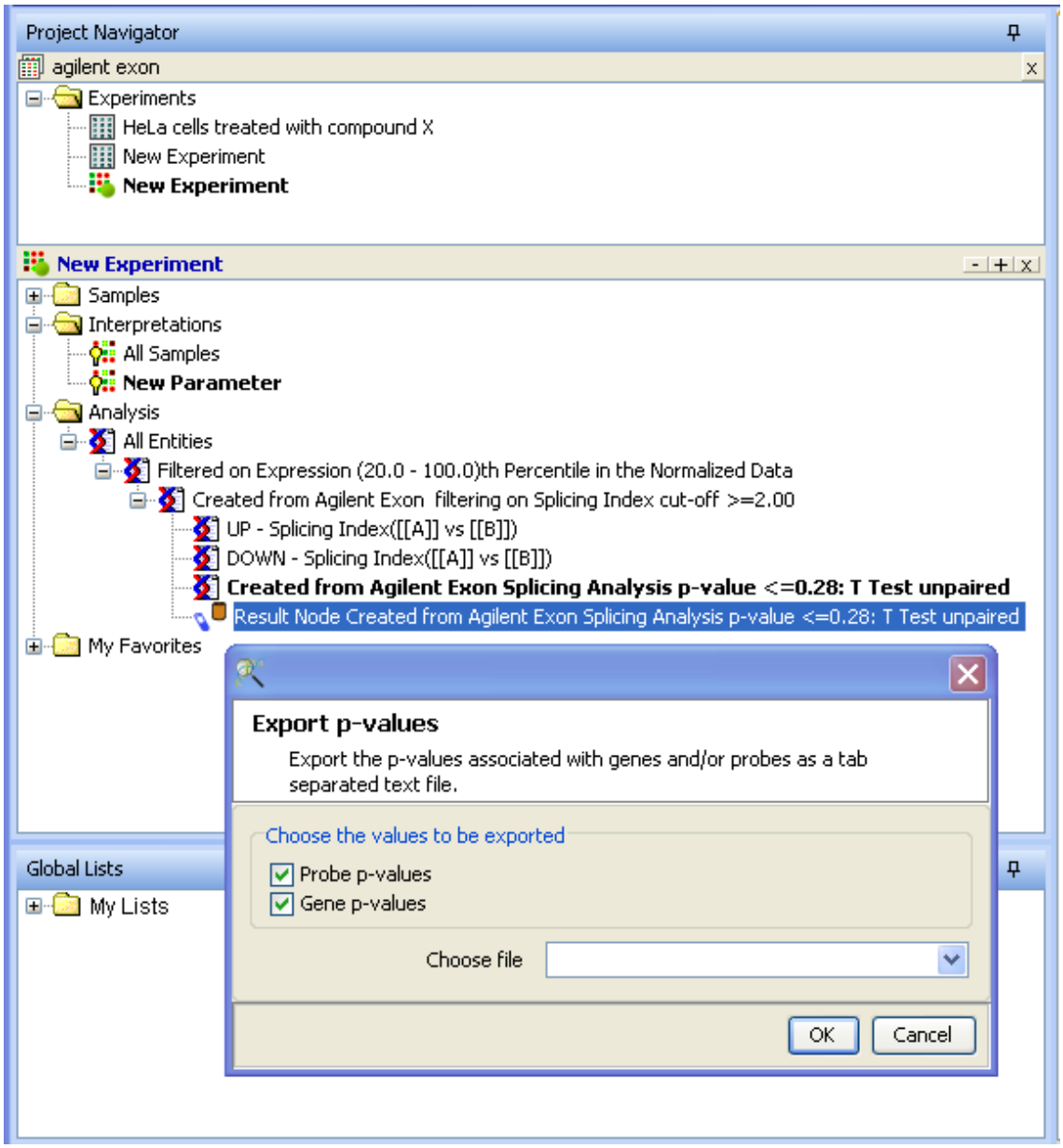


Figure 16.31: Export p-values

16.1.9 Class Prediction

- **Build Prediction Model** For details refer to section Build Prediction Model.

- **Run Prediction** For details refer to section Run Prediction.

16.1.10 Results Interpretation

- **Gene Ontology (GO) analysis**
  For more details see GO Analysis.
- **Gene Set Enrichment Analysis (GSEA)**
  For more details see Performing GSEA in **GeneSpring**.
- **Gene Set Analysis (GSA)**
  For more details see Gene Set Analysis.
- **Find Similar Entity Lists**
  This feature is discussed in a separate section called Find Similar Entity Lists.
- **Single Experiment Analysis or Multi-Omic Analysis**
  This feature is discussed in detail in the section Single Experiment and Multi-Omic Analysis.
- **Export to IPA**
  This feature is discussed in detail in the section Ingenuity Pathways Analysis (IPA) Connector.
- **Import IPA Entity List**
  This feature is discussed in detail in the section Import IPA Entity List.
- **NLP Network Discovery**
  This feature is discussed in detail in the section NLP Network Discovery.
- **MeSH Network Builder**
  This feature is discussed in detail in the section MeSH Network Builder.
- **Extract Relations via NLP**
  This feature is discussed in detail in the section Extract Relations via NLP.

16.1.11 Utilities

- **Import Entity list from File**
  For details refer to section Import Entity list from File.
- **Analysis: Biological Significance**
  For details refer to section Analysis: Biological Significance.
- **Filter On Entity List**
  For further details refer to section Filter On Entity List.
- **Remove Entities with missing signal values**
  For details refer to section Remove Entities with missing signal values.

16.2 Detailed Descriptions

16.2.1 Agilent Exon Technology

An exon array can have single or multiple probes for each exon. In an Agilent exon array, on an average there is only one probe for each exon, however there are instances of a single exon having multiple probes. In **GeneSpring**, data is computed at probe-level (exon-level) and at gene-level. Visualization is possible at transcript level also in addition to probe level and gene level.

16.2.1.1 Inputs for Technology Creation

Agilent Exon Technology is created from the following files:

- GEML file specific to the design.
- Text file on 'Transcript Alignments'
- Text file on 'Probe-to-Genome-To-Transcript'
- Text file on 'Gene to GO annotation'

The GEML file is specific to each chip and is used as the primary source of information for creating the technology. The three annotation files [.txt] remain the same for each organism irrespective of chip type. These arrays are designed using the hg19 Genome Build.

16.2.1.2 Technology creation

Briefly, the technology creation has the following steps:

- From GEML file, the probes specific for that technology are obtained; this is used to filter the probe annotation file and the transcript annotation file. The fields in the Transcript Annotation file include: Transcript ID, Strand information, GeneID, Chromosome, Chromosome Start, Chromosome Stop, Number of Exon blocks, Size of Exon blocks and Chromosomal Exon Starts.
- Create a GeneID to Transcript ID mapping. Recall that a gene is a collection of transcripts.
- Filter the Transcript Annotation File by removing duplicate transcripts. Duplicate transcripts are those that have the same ID but have different number of constituting exon blocks or different sizes of constituting exon blocks. If all transcripts of a gene are removed due to the above filtering, remove that Gene.

- Group all transcripts with same Entrez Gene ID, irrespective of their chromosome and Strand. For each gene, create sub groups of these transcripts based on chromosome and strand. For example, assume that Gene G1 is made up of Transcripts T1, T2, T3, T4, T5 and T6. Let T1, T2 and T3 belong to Chr1, '+' Strand while T4 belongs to Chr1 '-' Strand and T5 and T6 belongs to Chr4,'+' Strand. In this case, G1 will have 3 sub groups: (T1, T2, T3), (T4), (T5,T6). Each sub group will consist of transcripts from same chromosome and same strand.
- For each gene, order the Transcripts within the sub groups in ascending order with respect to the Chromosome Start location. Take the first Transcript from this ordered list and check if the second transcript's start location is within the stop location of the first transcript. The hypothetical example given below explains this step.

Transcript ID	Chromosome Start Position	Chromosome Stop Position
T11	65267587	65267645
T12	65267540	65267600
T13	65267501	65267590
T14	65267595	65267660
T15	65267600	65267730
T16	65267890	65267960

  The ordered list would be T13, T12, T11, T14, T15, T16.
- Create clusters of these sub groups looking at their overlaps. In the above example, consider T13 and see if the Start position of the next transcript T12 is within its Stop position. If so, group T13 and T12 together as a cluster. Repeat for all the other transcripts. If the condition is not satisfied, create a new cluster. In this example, 3 transcript clusters are created for G1, viz, (T13, T12, T11),(T14, T15),(T16). This way, each cluster would only consist of transcripts from same chromosome and same Strand.
- Create a Cluster ID of the form **geneid:Chr:+/-:Chromosome-Start:Chromosome-Stop**. This effectively contains all the relevant information for the gene and is stored in the technology as 'GeneBinID'.
- **Create a GeneBinID → Transcript Row Indices mapping by looking at all the constituting transcripts.**. This is stored in the technology.
- **Create a GeneBinID → Probe Row Indices mapping by getting all the constituting probes.** This is stored in the technology.
- Get the Entrez Gene ID → Probe Row Indices mapping from above and reverse the map to obtain all those probes that have multiple Gene IDs associated with them. **These are the 'Multi-Gene' Probes; this information is used in visualization.** This is stored in the technology.

16.2.1.3 Structure of Technology

The Agilent Exon Technology is given a name that is self-explanatory. The technology type would be Agilent.Exon< $SingleorTwo$ >Color and the generic name of the technology would be:

`Technology Type.<AMADID>_genomebuild_Versiondate in yyyy_mm_dd` format.
For example: `Agilent.ExonSingleColor.28679_hg19_2010-08-12`

Date of original creation of technology will be stored inside the technology and the version date will reflect the date of last modification.

16.2.1.4 Updating Technology Annotation

Before creating Agilent Exon experiments, download the technology from the Agilent Server. Go to the menu **Annotations** → **Create Technology** → **From Agilent Server**; all the available technologies are listed here for Agilent Exon Single and Two color giving the type and the name of the technology. Choose the relevant technology(ies) and download. Please note that **GeneSpring** will automatically choose the latest version of available technology for that particular genome build. For example, let us assume that a user has downloaded two technologies of names
`27566_hg19_2010_07_15` and
`27566_hg19_2010_07_20`

In this case, only the latter technology (which is the most recent version) will be shown.

Please note that the same AMADID number can be used for both Single Color and Two Color Experiments.

Chapter 17

Analyzing Affymetrix Exon Expression Data

Affymetrix Exon chips are being increasingly used for assessing the expression levels of transcripts. **GeneSpring** supports this Affymetrix Exon Expression Technology.

17.1 Running the Affymetrix Exon Workflow

Refer to the common section on Running the Expression Experiment Workflow for more details on project creation. Figures 17.1, 17.2, 17.3, 17.4 show the process of choosing analysis type, experiment type, loading data, choosing samples and re-ordering the data files, for Affymetrix Exon Expression experiment.

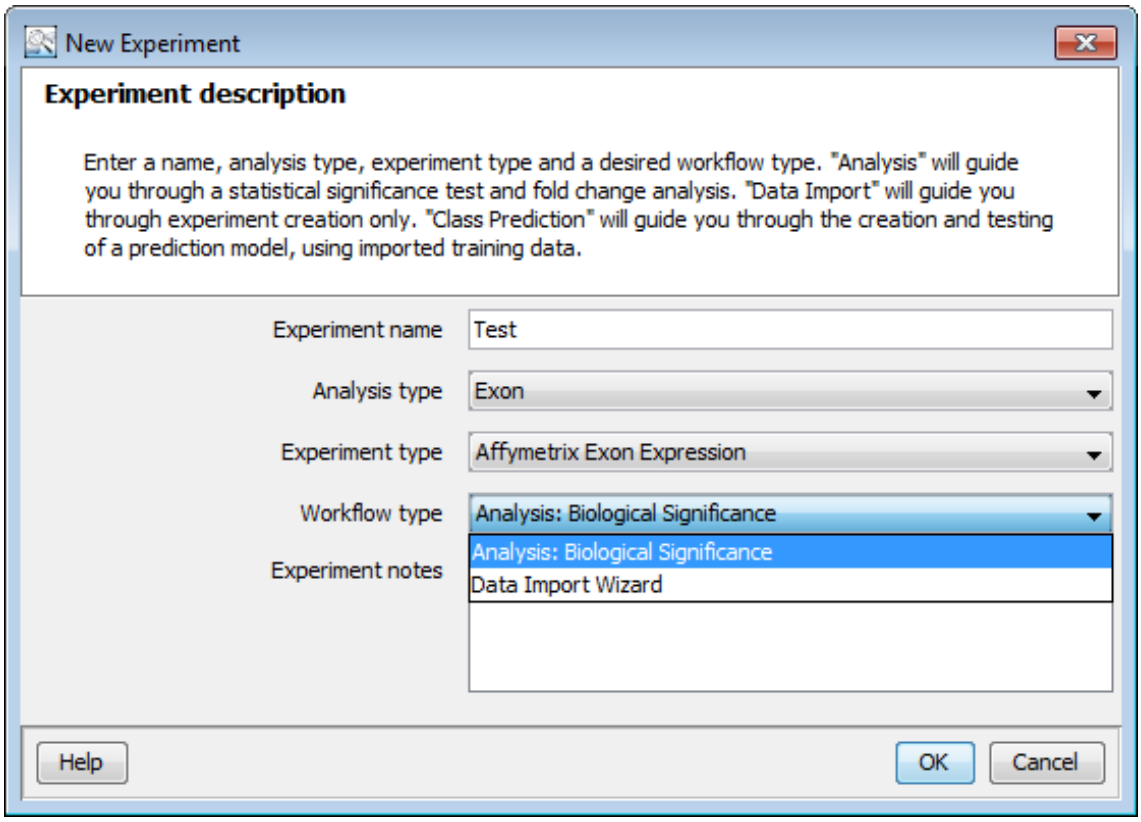


Figure 17.1: Experiment Description

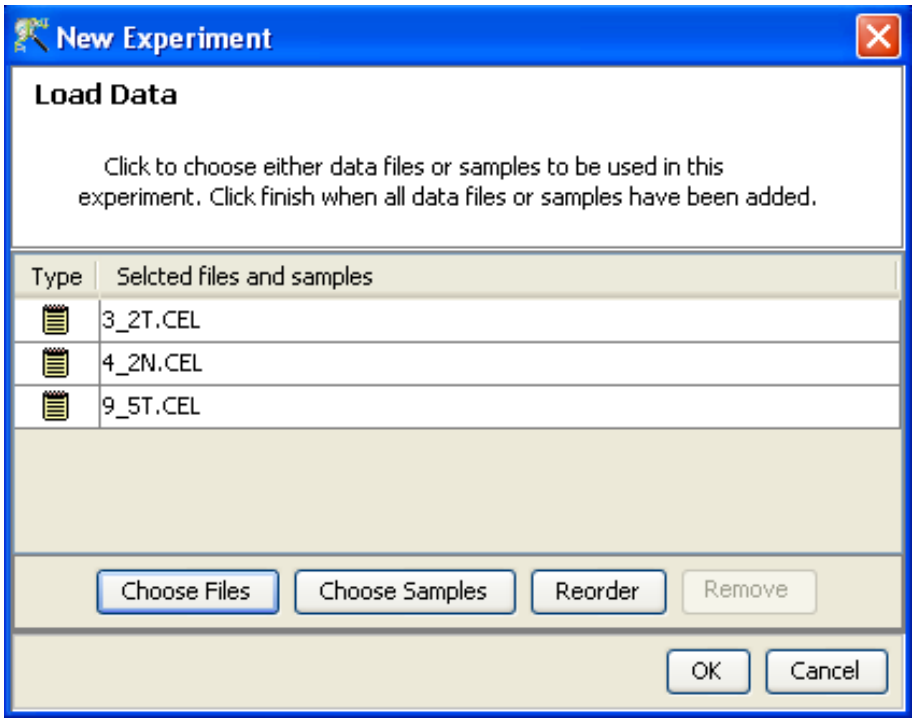


Figure 17.2: Load Data

Figure 17.3: Choose Samples

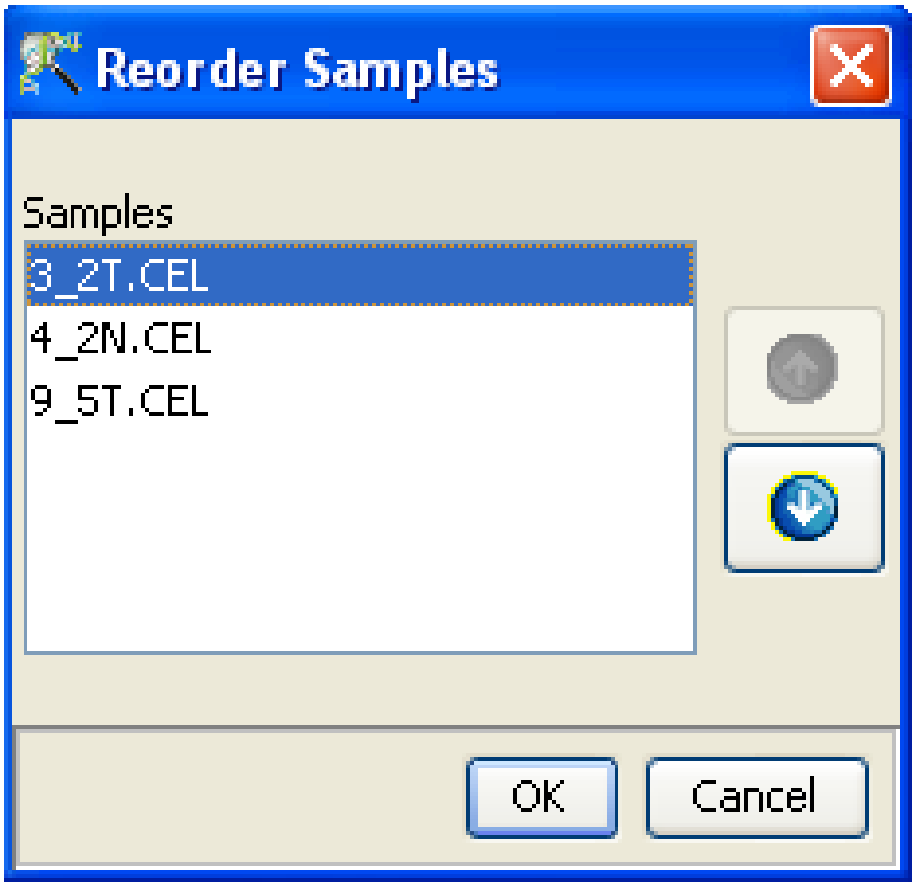


Figure 17.4: Reordering Samples

17.2 Data Processing

1. **File formats**: The data file should be present either as a CEL file or a CHP file. However while creating an experiment; only one type of file (CEL/CHP) can be used.
2. **Raw signal values (CEL files)**: In an Affymetrix Exon Expression experiment, the term ‘‘raw’’ signal values refers to the linear data which has been summarized using a summarization algorithm (RMA16, PLIER 16 and Iterative PLIER 16). All summarization algorithms also do variance stabilization by adding 16.
3. **Raw signal values (CHP files)**: In an Affymetrix Exon Expression experiment, the term ”raw” files refers to the linear data obtained from the CHP files.
4. **Normalized signal values (CEL files)**: ”Normalized’’ values are generated after the log transformation and baseline transformation step.
5. **Normalized signal values (CHP files)**: The term ”Normalized” refers to values generated after log transformation, normalization (Percentile Shift, Scale or Normalize to control genes) and baseline transformation.
6. **Treatment of on-chip replicates**: Not Applicable.

7. **Flag values**: Not Applicable.
8. **Treatment of Control probes**: Not Applicable.
9. **Empty Cells**: Not Applicable.
10. **Sequence of events (CEL files)**: The sequence of events involved in the processing of a CEL file is: Summarization⟶ log transformation⟶ Baseline Transformation.
11. **Sequence of events (CHP files)**: If the data in the CHP file is already log transformed, then **GeneSpring** detects it and proceeds with the normalization step.

17.3 Analysis: Biological Significance

Analysis: Biological Significance has 8 steps:

1. Summary Report
2. Experiment Grouping
3. QC on samples
4. Filter Probe Sets
5. Significance analysis
6. Fold Change
7. GO analysis
8. Pathway Analysis

These steps are explained in section Analysis: Biological Significance. Details specific to Affymetrix Exon Expression Data are explained here.

Quality Control (Step 3 of 8): The 3rd step in Analysis: Biological Significance is the QC on samples which displays three tiled windows when *CHP* files are used and four when *CEL* files are used as samples. They are as follows:

- Experiment grouping
- Hybridization Controls(only for CEL files)
- PCA scores
- Legend

See Figure 17.5 for more details.

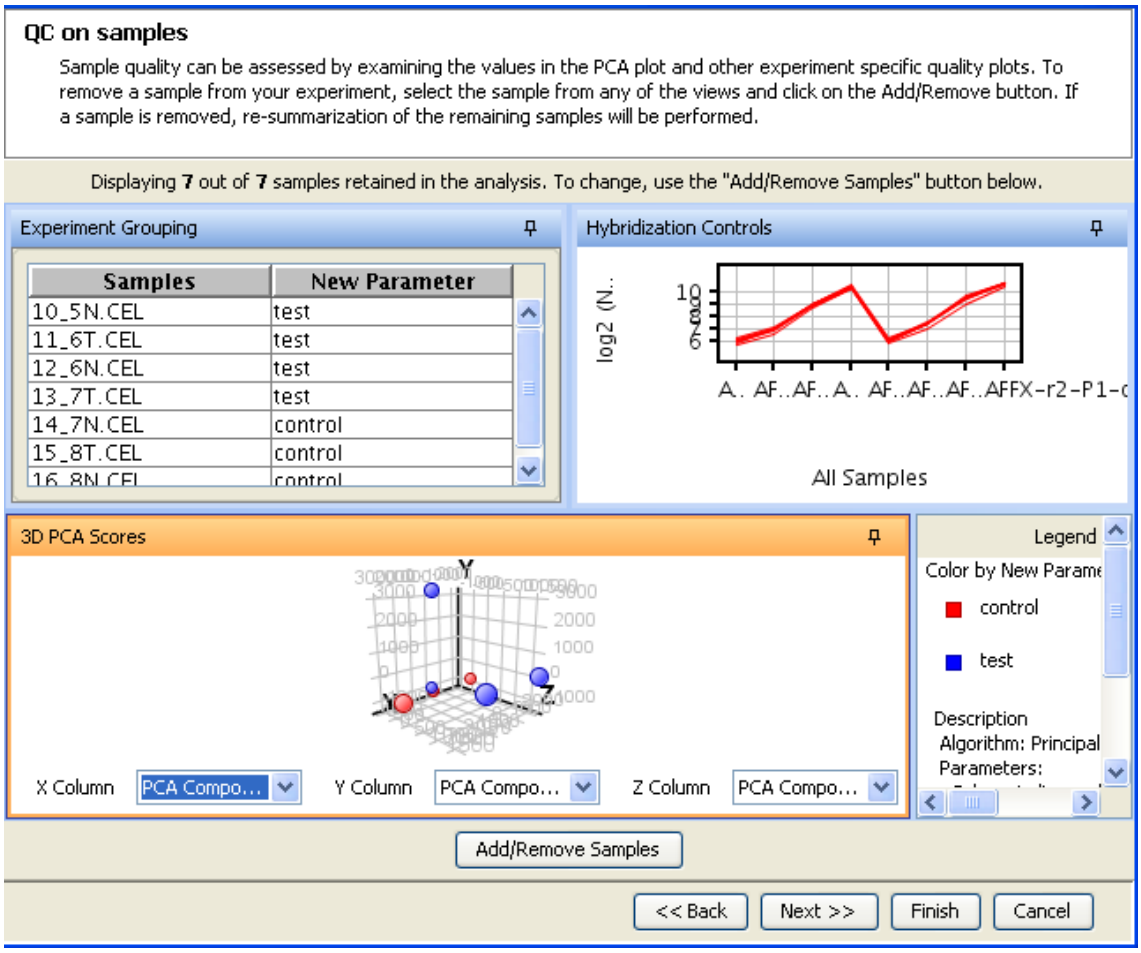


Figure 17.5: Quality Control on Samples

The views in these windows are lassoed i.e., selecting the sample in any of the view highlights the sample in all the views.

The *Experiment Grouping* view shows the samples and the parameters present.

The *Hybridization Controls* view depicts the hybridization quality. Hybridization controls are composed of a mixture of biotin-labelled cRNA transcripts of bioB, bioC, bioD, and cre prepared in staggered concentrations (1.5, 5, 25, and 100pm respectively). This mixture is spiked-in into the hybridization cocktail. bioB is at the level of assay sensitivity and should be called Present at least 50% of the time. bioC, bioD and cre must be Present all of the time and must appear in increasing concentrations. The X-axis in this graph represents the controls and the Y-axis, the log of the Normalized Signal Values.

Principal Component Analysis (PCA) calculates the PCA scores and visually represents them in a 3D scatter plot. The scores are used to check data quality. It shows one point per array and is colored by the *Experiment Factors* provided earlier in the *Experiment Groupings* view. This allows viewing of separations between groups of replicates. Ideally, replicates within a group should

cluster together and separately from arrays in other groups. The PCA components, represented in the X, Y and Z axes are numbered 1, 2, 3... according to their decreasing significance. The 3D PCA scores plot can be customized via ***Right-Click*** $\longrightarrow$ ***Properties***. To zoom into a 3D Scatter plot, press the Shift key and simultaneously hold down the left mouse button and move the mouse upwards. To zoom out, move the mouse downwards instead. To rotate, press the Ctrl key, simultaneously hold down the left mouse button and move the mouse around the plot.

The *Add/Remove* samples allows the user to remove the unsatisfactory samples and to add the samples back if required. Whenever samples are removed or added back, summarization as well as baseline transformation is performed again on the samples. Click on *OK* to proceed.

The fourth window shows the legend of the active QC tab.

17.4 Data Import Wizard

The *Data Import Wizard* offers a variety of choices to the user for the analysis. Several different summarization algorithms are available for probeset summarization. Additionally there are options for baseline transformation of the data and for creating different interpretations. To create and analyze an experiment using the *Data Import Wizard*, load the data as described earlier. In the **New Experiment Dialog**, choose the **Workflow Type** as Data Import Wizard. Clicking ***OK*** will open a New Experiment Wizard, which then proceeds as follows:

17.4.1 Creating an Affymetrix ExonExpression Experiment

Data Import Wizard Analysis can be done using either CEL or CHP files. However, a combination of both file types cannot be used. Only transcript summarized CHP files can be loaded in a project.

New Experiment (Step 1 of 7): Load data As in case of *Analysis: Biological Significance*, either data files can be imported or else pre-created samples can be used.

- For loading new CEL/CHP files, use *Choose Files*.
- If the CEL/CHP files have been previously used in experiments *Choose Samples* can be used.

Step 1 of 7 of Experiment Creation, the **Load Data** window, is shown in Figure 17.6.

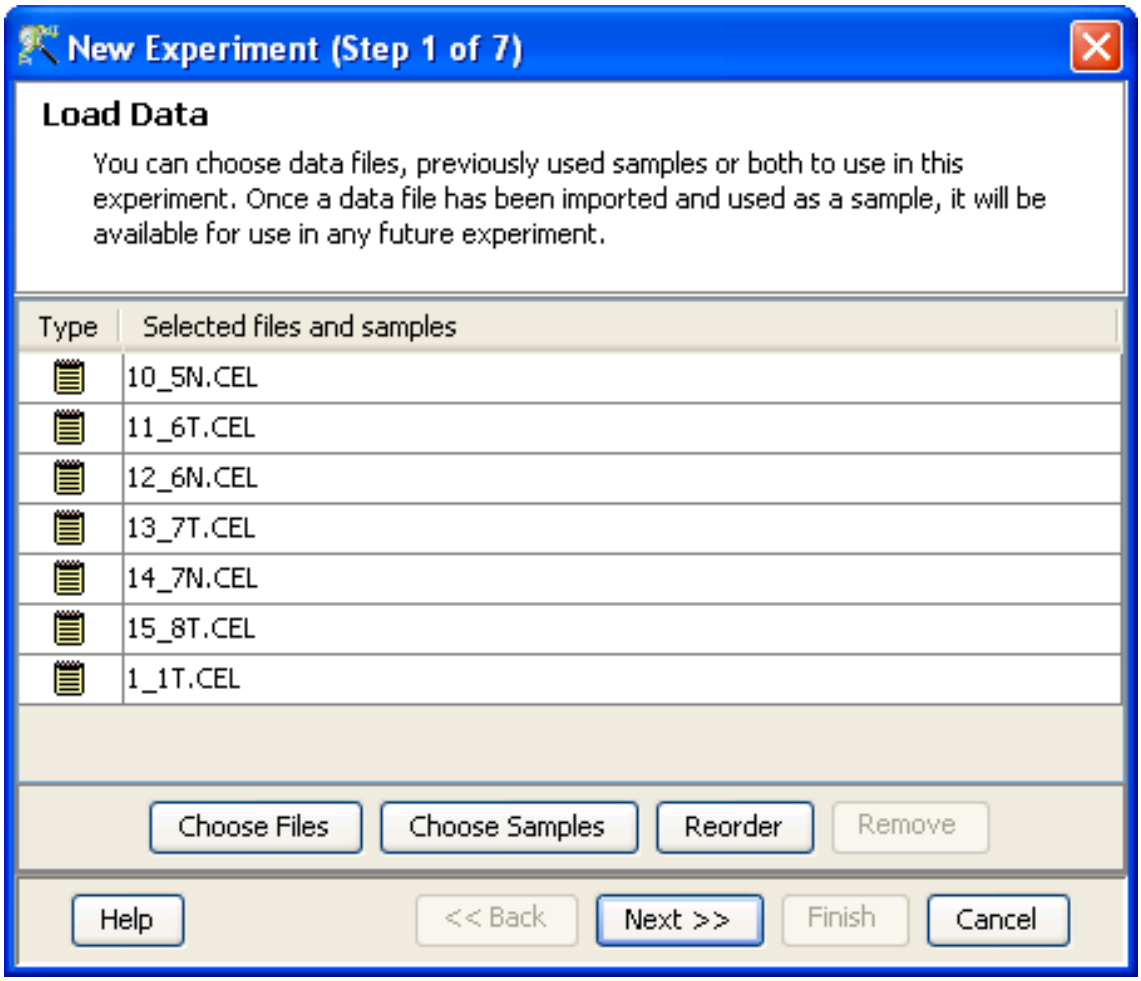


Figure 17.6: Load Data

New Experiment (Step 2 of 7): Selecting ARR files ARR files are Affymetrix files that hold annotation information for each sample CEL and CHP file and are associated with the sample based on the sample name. These are imported as annotations to the sample. Click on ***Next*** to proceed to the next step.

Step 2 of 7 of Experiment Creation, the Select ARR files window, is depicted in the Figure 17.7.

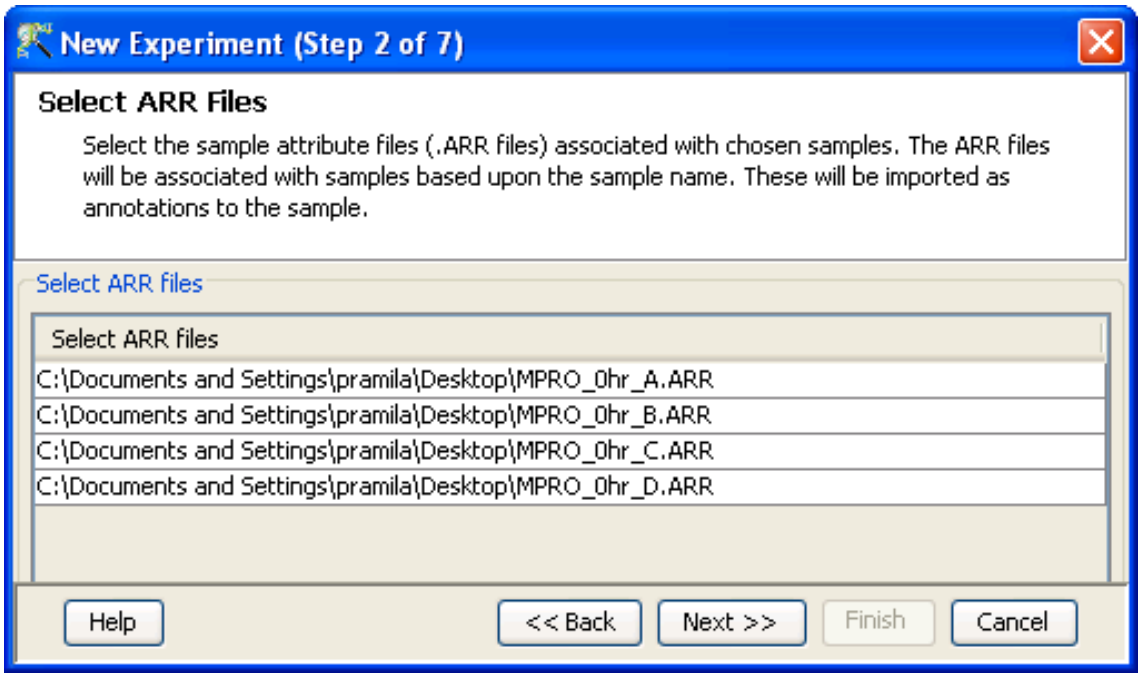


Figure 17.7: Select ARR files

New Experiment (Step 3 of 7): Pairing of transcript and probeset level files This step of the wizard is used in the case of Affymetrix Exon Splicing experiment type.

New Experiment (Step 4 of 7): Preprocess Baseline Options This step is specific for CEL files. Any one of the Summarization algorithms provided from the drop down menu can be chosen to summarize the data. The available summarization algorithms are:

- The RMA Irazarry *et al.* [Ir1, Ir2, Bo].
- The PLIER16 Hubbell [Hu2].
- The IterativePLIER16

Subsequent to probeset summarization, baseline transformation of the data can be performed. The baseline options include:

- ***Do not perform baseline***
- ***Baseline to median of all samples:*** For each probe the median of the log summarized values from all the samples is calculated and subtracted from each of the samples.
- ***Baseline to median of control samples:*** For each sample, an individual control or a set of controls can be assigned. Alternatively, a set of samples designated as controls can be used for all samples. For specifying the control for a sample, select the sample and click on ***Assign value***. This opens up the ***Choose Control Samples*** window. The samples designated as Controls should be moved from the *Available Items* box to the *Selected Items* box. Click on ***Ok***. This will show the control samples for each of the samples.
  In *Baseline to median of control samples*, for each probe the median of the log summarized values from the control samples is first computed and then this is subtracted from the sample. If a single sample is chosen as the control sample, then the probe values of the control sample are subtracted from its corresponding sample.

This step also enables the user to select the Confidence level, as Core, Extended or Full, using which the summarization is done.

1. Core comprising transcript clusters groups unique exon level probe sets with strong annotation support e.g., from RefSeq and other GenBank alignments of 'complete CDS' transcripts, into transcript clusters.
2. Extended comprising transcript clusters groups unique exon level probe sets with empirical annotation support into transcript clusters. This includes cDNA transcripts, mapped syntenic mRNA from rat and mouse, and probe sets which are part of the Ensembl or Vega transcript annotation set.
3. Full transcript clusters including ab-initio predictions from Geneid, Genscan, GENSCAN Suboptimal, Exoniphy, RNAgene, SgpGene and TWINSCAN.

Clicking ***Finish*** creates an experiment, which is displayed as a Box Whisker plot in the active view. Alternative views can be chosen for display by navigating to ***View*** in Toolbar. Figure 17.8 shows the Step 4 of 7 of Experiment Creation.

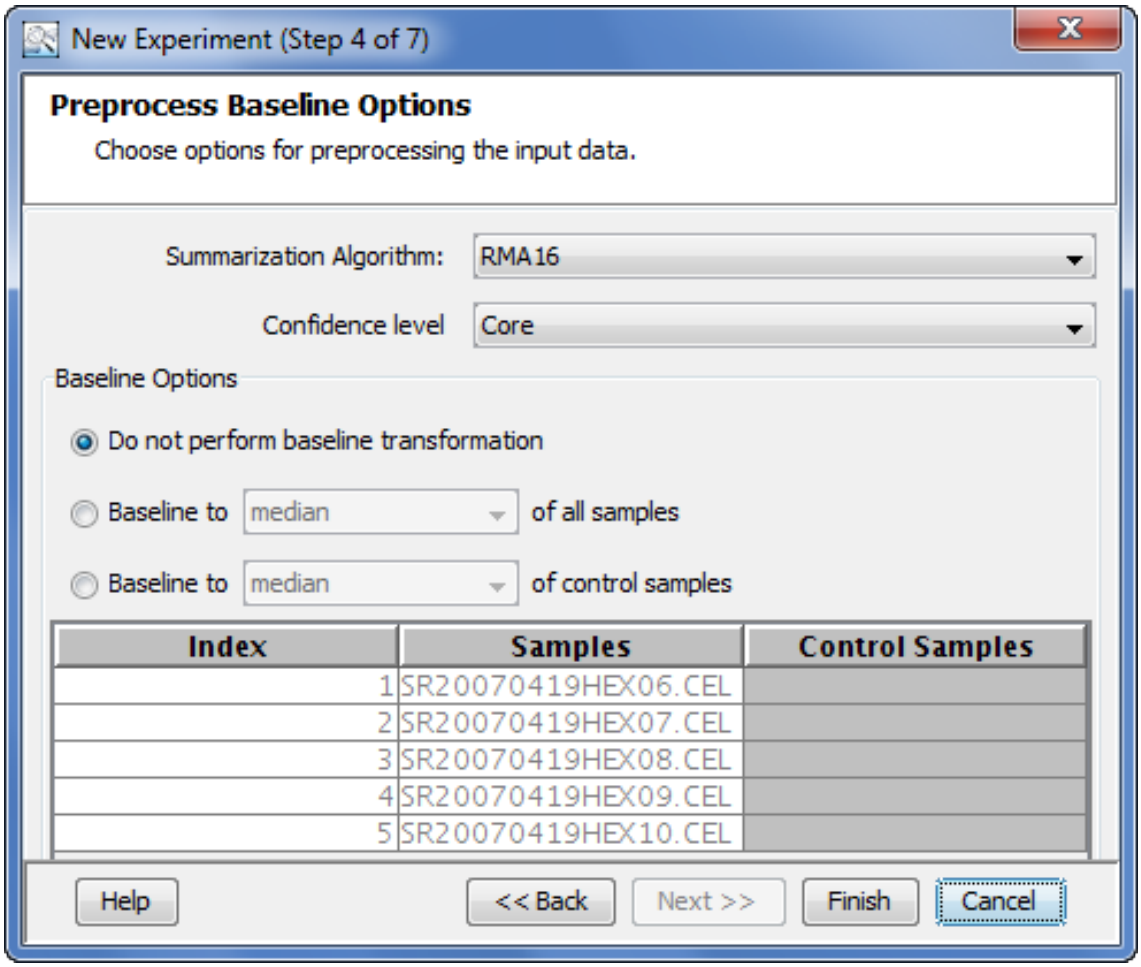


Figure 17.8: Summarization Algorithm

New Experiment (Step 5 of 7): This step is specific for CHP files only. See Figure 17.9. It gives the user the following normalization options:

- **Percentile Shift:** On selecting this normalization method, the **Percentile Target** box gets enabled allowing the user to enter a specific percentile value. The default value is set at 75^th percentile.
- **Scale:** On selecting this normalization method, the user is presented with an option to scale it to the median of percentile target/mean of means of 'all' or 'control' samples. On choosing the latter, the user has to select the control samples from the available samples in the **Choose Samples** box. The **Percentile Target** box is set at a default value of 75.

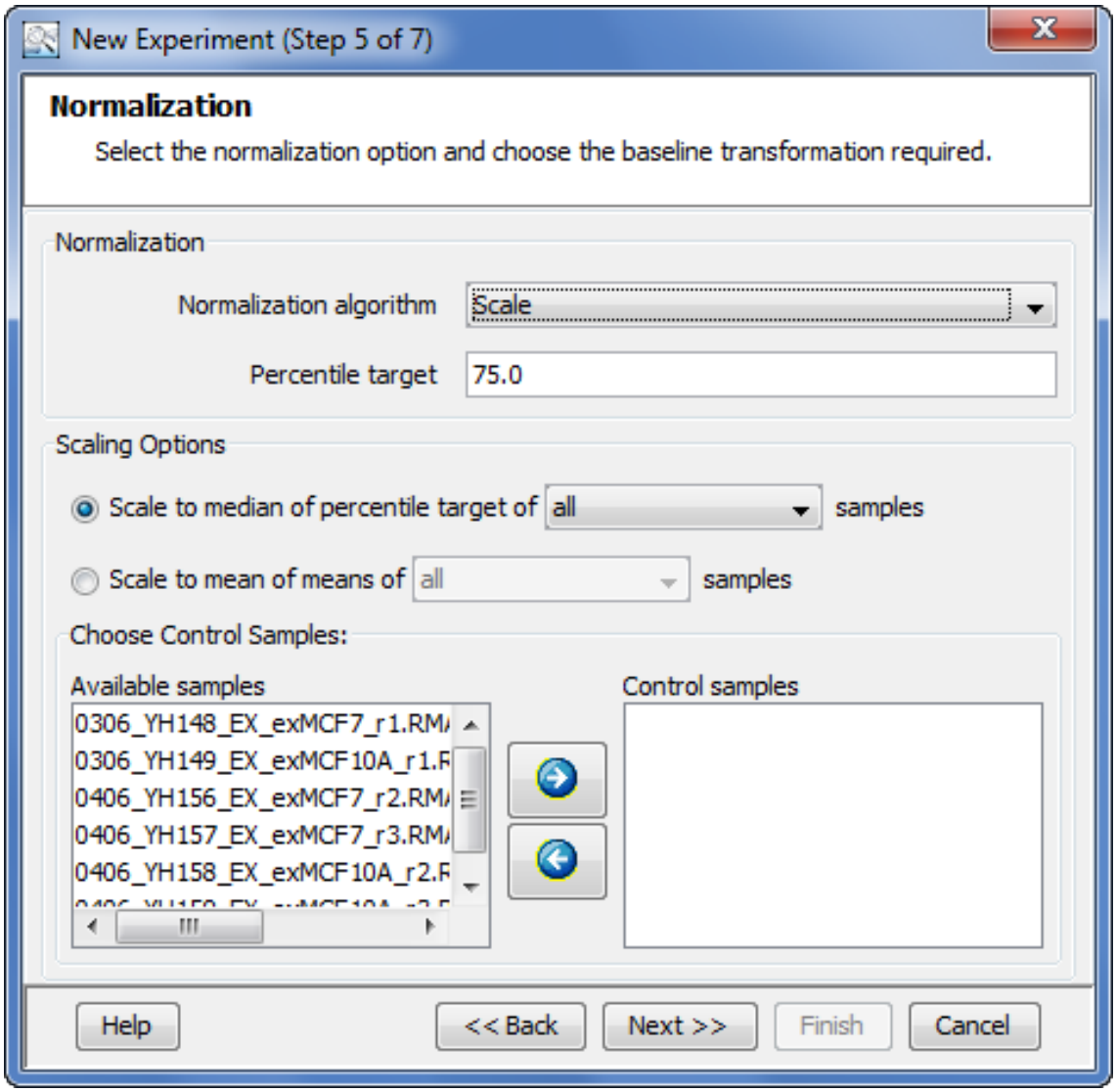


Figure 17.9: Normalization

- **Normalize to control genes:** After selecting this option, the user has to specify the control genes in the next wizard. The **Percentile Target** box is disabled.
- **Normalize to External Value:** This option will bring up a table listing all samples and a default scaling factor of '1.0' against each of them. The user can use the *'Assign Value'* button at the bottom to assign a different scaling factor to each of the sample;

multiple samples can be chosen simultaneously and assigned a value. The **Percentile Target** box is disabled.

For details on the above normalization methods, refer to section on Normalization Algorithms.

New Experiment (Step 6 of 7): If the **Normalize to control genes** option is chosen, then the list of control entities can be specified in the following ways in this wizard:

- By choosing a file(s) (.txt, .csv, or .tsv) which contains the control entities of choice denoted by their probe id. Any other annotation will not be suitable.
- By searching for a particular entity by using the ***Choose Entities*** option. This leads to a search wizard in which the entities can be selected. All the annotation columns present in the technology are provided and the user can search using terms from any of the columns. The user has to select the entities that he/she wants to use as controls when they appear in the **Output Views** page and then click ***Finish***. This will result in the entities getting selected as control entities and will appear in the wizard. See Figures 17.10, 17.11 and 17.12.

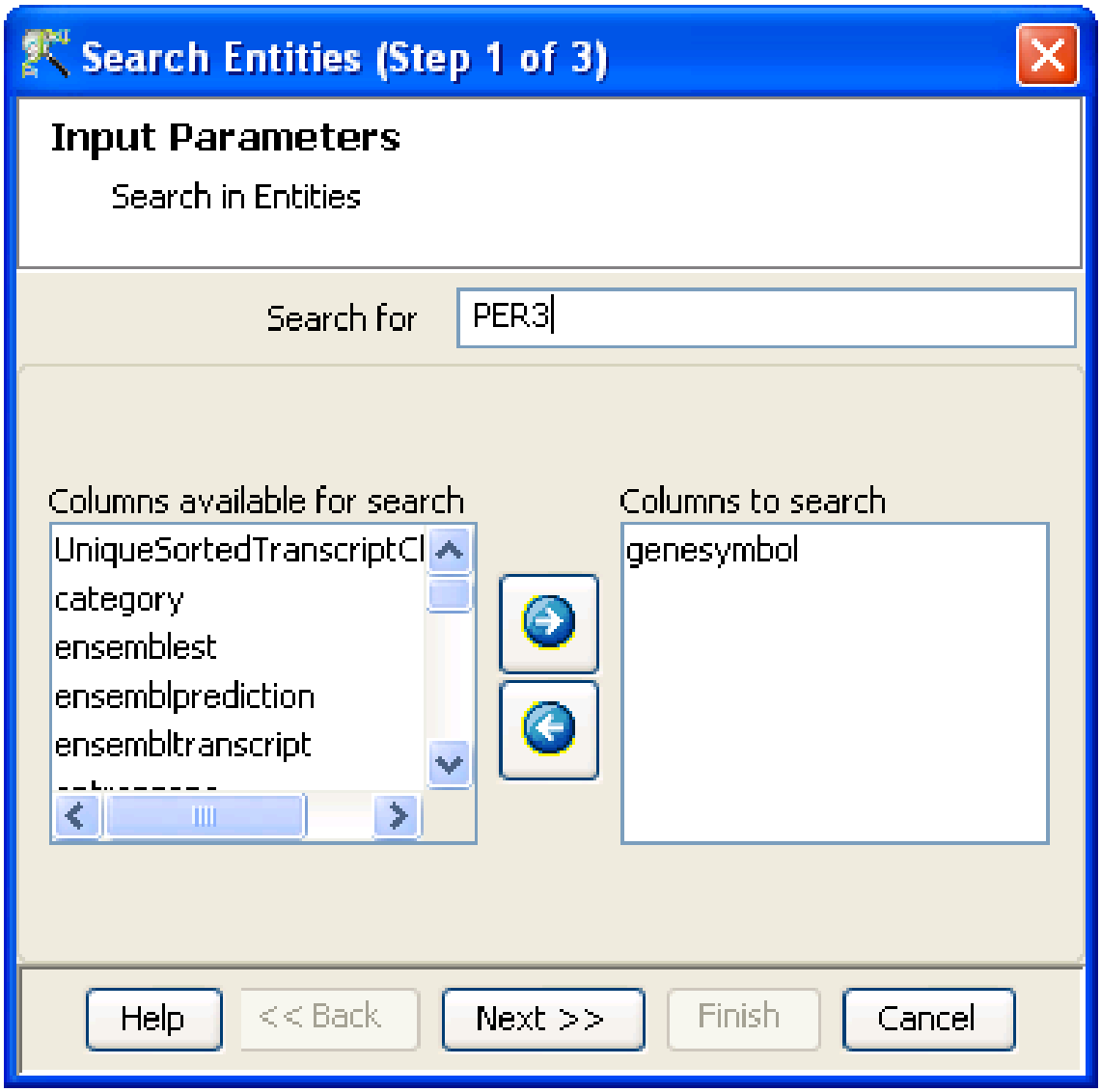


Figure 17.10: Search Entities

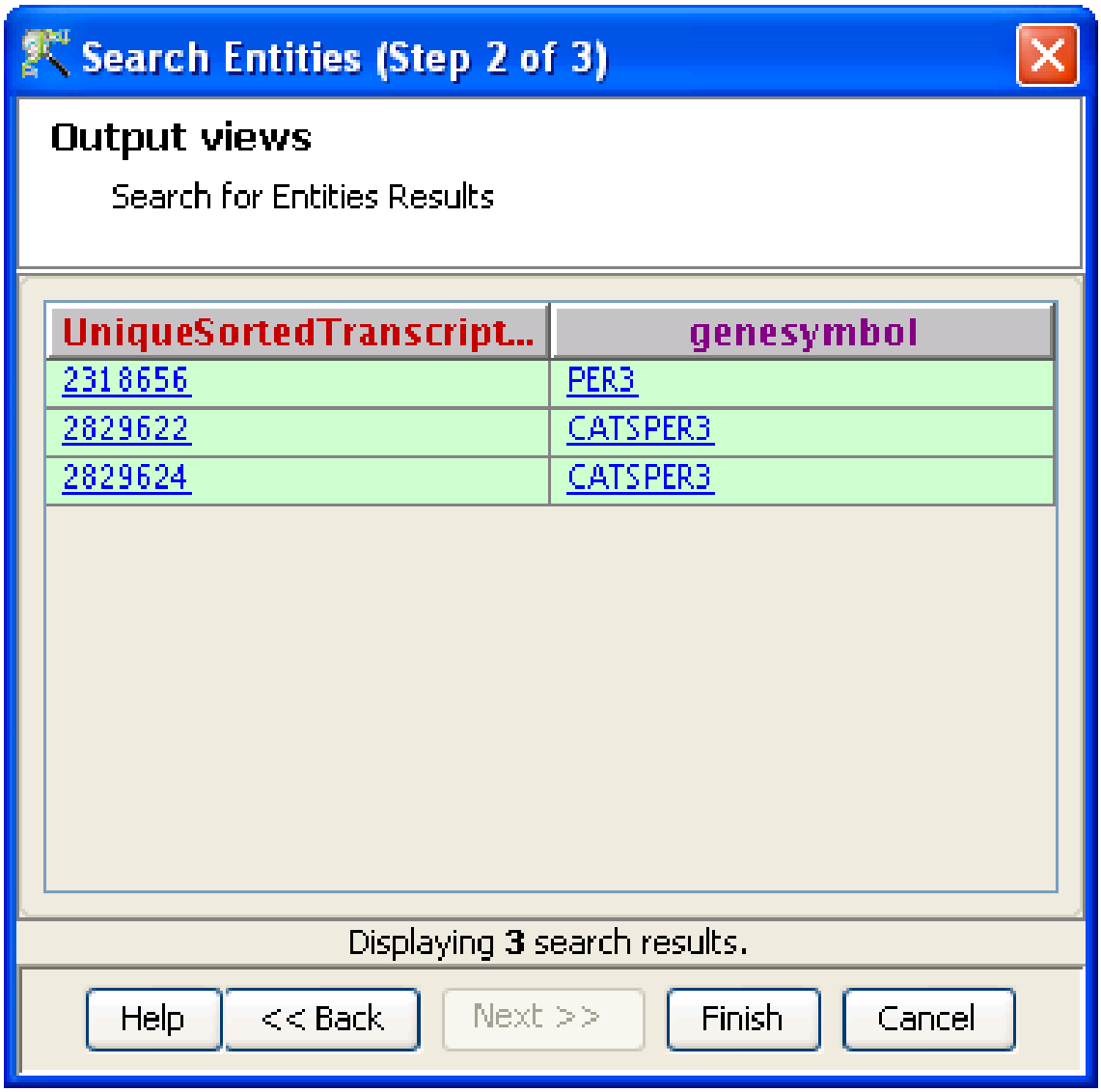


Figure 17.11: Output Views

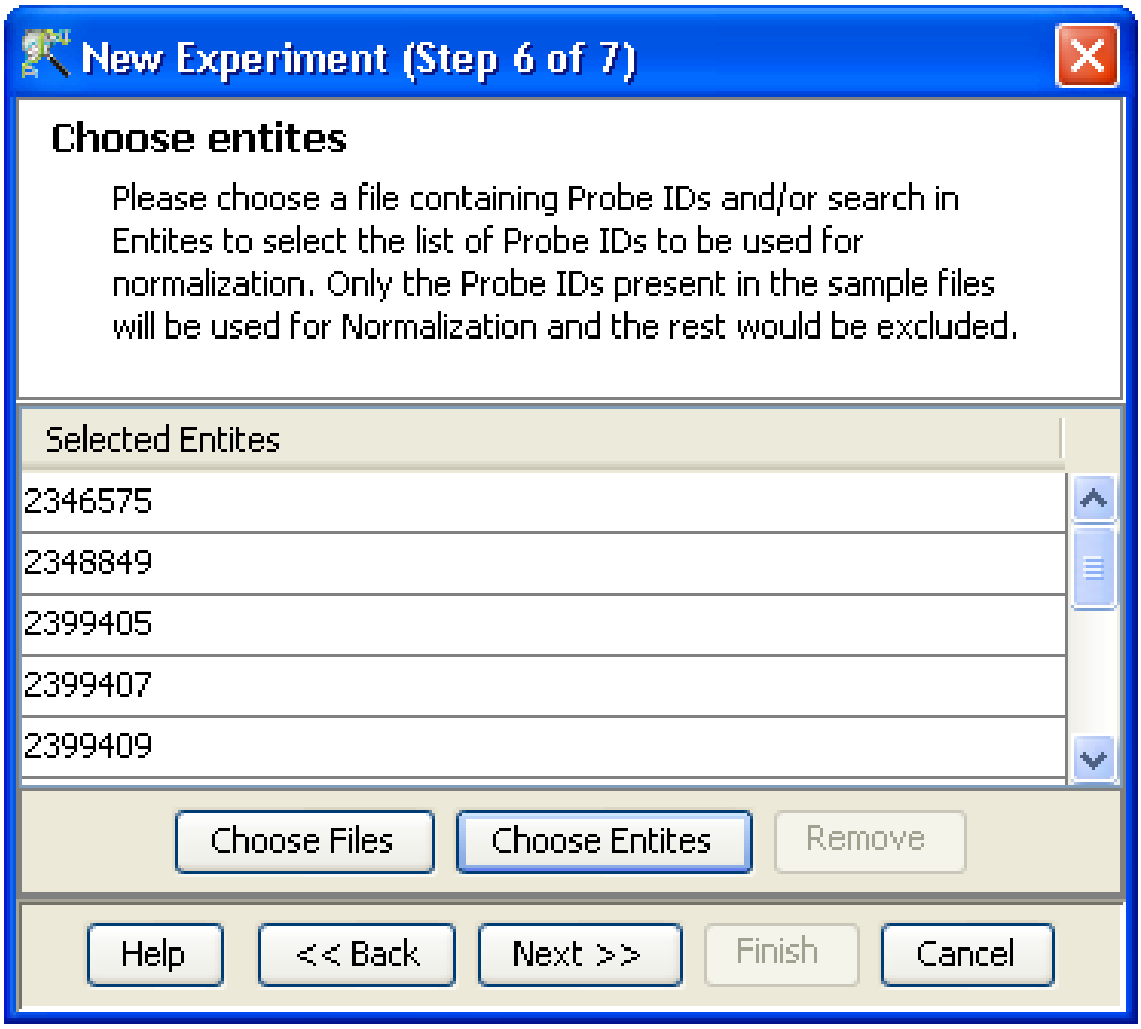


Figure 17.12: Choose Entities

The user can choose either one or both the options to select his/her control genes. The chosen genes can also be removed after selecting the same.

In case the entities chosen are not present in the technology or sample, they will not be taken into account during experiment creation. The entities which are present in the process of experiment creation will appear under matched probe ids whereas the entities not present will appear under unmatched probe ids in the experiment notes in the experiment inspector.

New Experiment (Step 7 of 7): This step allows the user to perform baseline transformation. The methods available are the same as those used for CEL files in Step 4 of 7.

Clicking ***Finish*** creates an experiment, which is displayed as a Box Whisker plot in the active view. Alternative views can be chosen for display by navigating to ***View*** in Toolbar. The final step of Experiment Creation (CHP file specific) is shown in Figure 17.13.

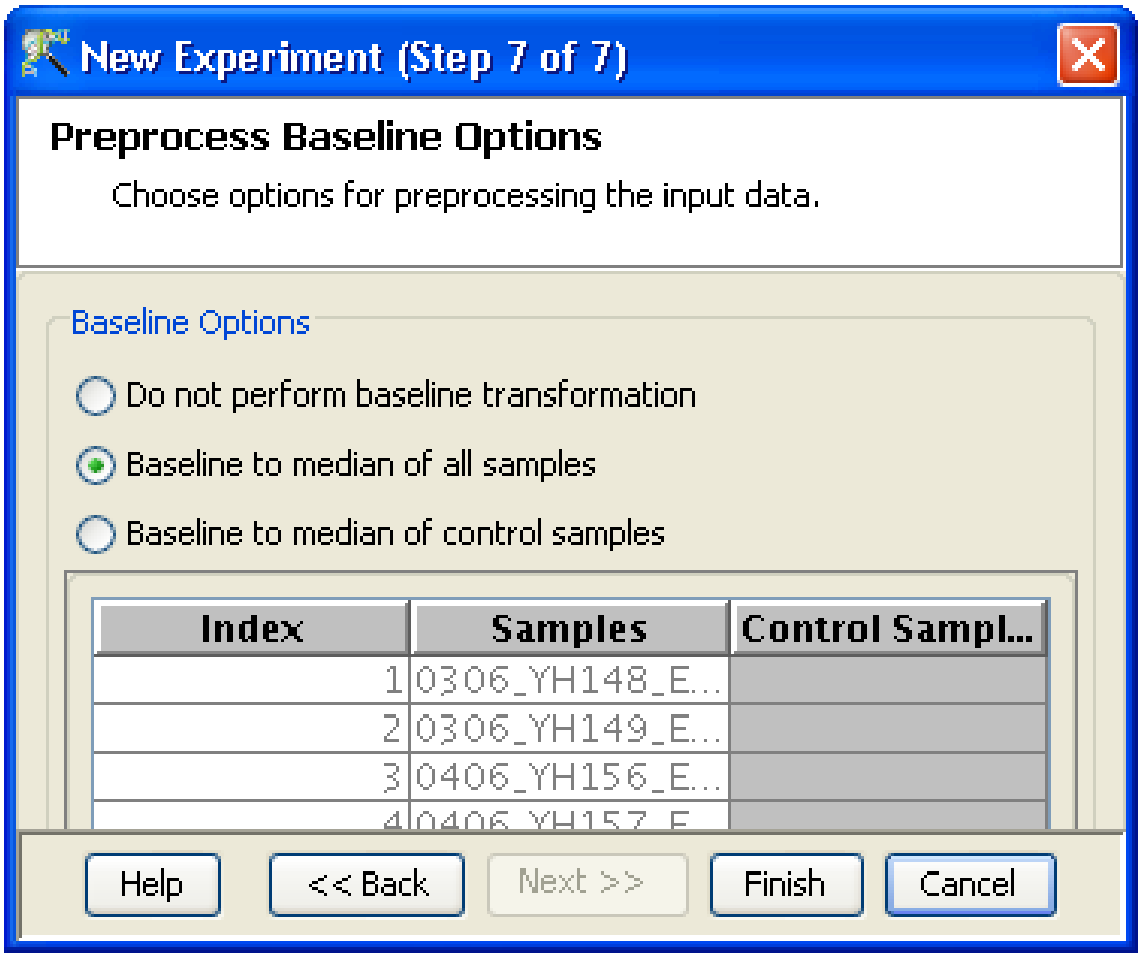


Figure 17.13: Normalization and Baseline Transformation

17.4.2 Experiment setup

Once an experiment is created, the *Data Import Wizard* steps appear on the right hand side. This gives options to view quick start guide, create experiment grouping and interpretation and the utility to create gene level experiment from probe level experiment. See section Experiment Setup for complete details.

17.4.3 Quality Control

- **Quality Control on Samples**

  Quality Control (QC) allows you decide which samples are ambiguous and which are passing the quality criteria. Based on the QC results, unreliable you can remove samples from further downstream analysis. The QC dialog displays three tiled windows:

– 3D PCA scores, Correlation Analysis - Samples, and a Correlation Coefficients tab.
– Experiment Grouping, Hybridization Controls (displayed only for .CEL files).
– Legend.

Figure 17.14 has tiled windows which reflect the QC on samples.

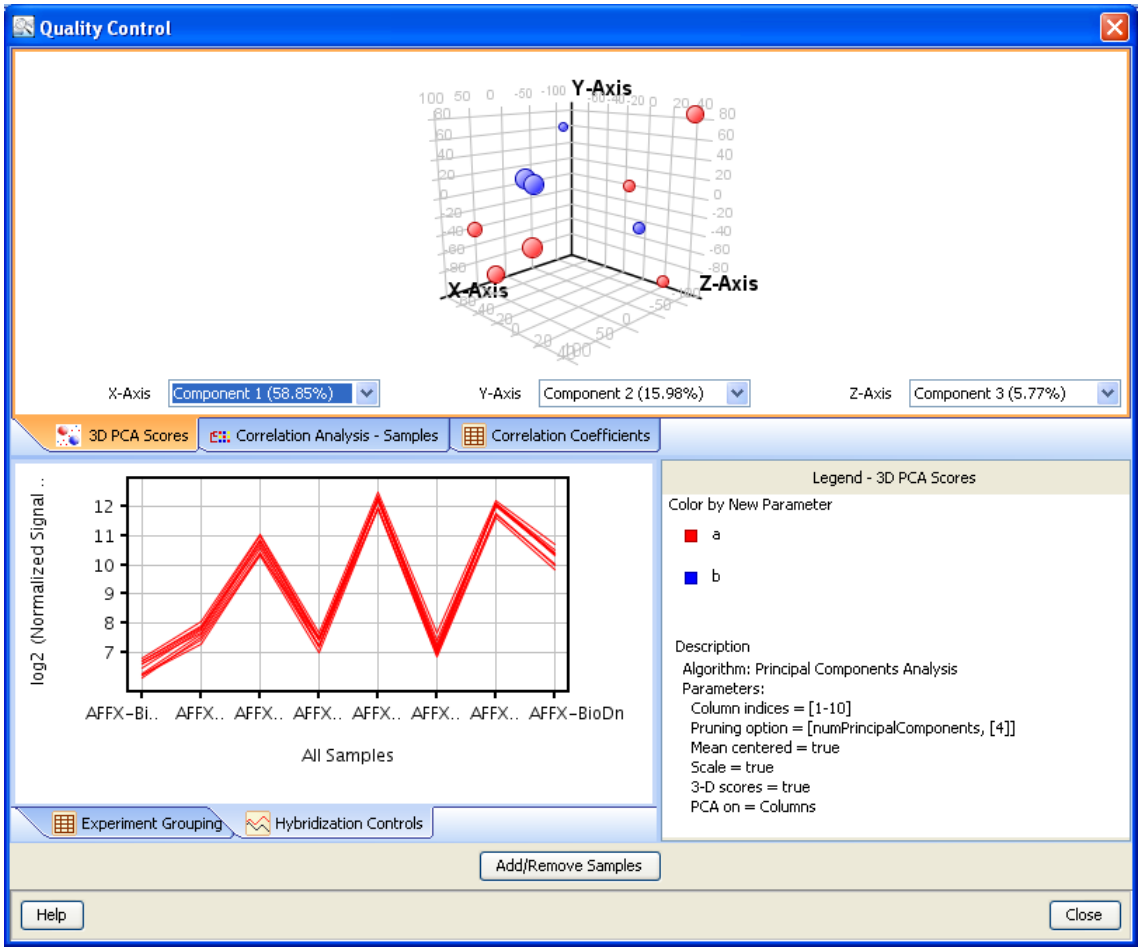


Figure 17.14: Quality Control Dialog

In the **Principal Component Analysis (PCA)** tab, PCA scores are visually represented in a 3D scatter plot. These scores are used to check data quality. It shows one point per array and is colored by the *Experiment Factors* provided earlier in the *Experiment Groupings* view. This allows viewing of separations between groups of replicates. Ideally, replicates within a group should cluster together and separately from arrays in other groups. The PCA components, represented in the X, Y, and Z axes are numbered 1, 2, 3... according to their decreasing significance. The 3D PCA scores plot can be customized via ***Right-Click*⟶*Properties***. To zoom into a 3D Scatter plot, press the Shift key and simultaneously hold down the left mouse button and move the mouse upwards. To zoom out, move the mouse downwards instead. To rotate, press the Ctrl key, simultaneously hold down the left mouse button and move the mouse around the plot.

The **Correlation Analysis - Samples** tab shows the correlation analysis heatmap across arrays (samples) (See How to Perform Correlation Analysis on Samples on page 750). The

correlation coefficient for each pair of arrays are displayed in the **Correlation Coefficients** tab. Correlation coefficients are calculated using Pearson Similarity Metric (see Pearson on page 799).

Pearson Similarity Metric: Calculates the mean of all elements in vector **a**. Then it subtracts that value from each element in **a** and calls the resulting vector **A**. It does the same for **b** to make a vector **B**. Correlation Coefficient is then computed as **A.B**/(—**A**——**B**—).

In a Correlation Analysis heatmap:

- Sample Order- In a correlation heatmap, the initial order of samples is the same as the input sample order at the time of experiment creation.
- Clustering - Perform clustering analysis on the correlation heatmap using the **Run clustering** icon from the correlation analysis heatmap toolbar (See To run clustering on a correlation heatmap on page 760). Samples are clustered based on correlation coefficients using hierarchical clustering algorithm and plotted in the order they cluster together. The linkage rule used is average and the distance metrics used is Euclidean.

For more information on the available toolbar icons in a correlation analysis heatmap, see Correlation Heatmap View Toolbar on page 776. See Properties on 792 to change the properties of the correlation analysis heatmap.

NOTE Correlation coefficients are computed on raw, non-normalized data and in linear scale.

Experiment Grouping shows the parameters and parameter values for each sample.

The *Hybridization Controls* view depicts the hybridization quality. Hybridization controls are composed of a mixture of biotin-labelled cRNA transcripts of bioB, bioC, bioD, and cre prepared in staggered concentrations (1.5, 5, 25, and 100pm respectively). This mixture is spiked-in into the hybridization cocktail. bioB is at the level of assay sensitivity and should be called Present at least 50% of the time. bioC, bioD and cre must be Present all of the time and must appear in increasing concentrations. The X-axis in this graph represents the controls and the Y-axis, the log of the Normalized Signal Values.

The third window shows the legend of the active QC tab.

Unsatisfactory samples or those that have not passed the QC criteria can be removed from further analysis, at this stage, using *Add/Remove Samples* button. Once a few samples are removed, re-summarization of the remaining samples is carried out again. The samples removed earlier can also be added back. Click on **OK** to proceed.

- **Filter Probe Set by Expression** Entities are filtered based on their signal intensity values. For details refer to the section on Filter Probesets by Expression
- **Filter Probe Set by Flags** No flags are generated during creation of exon expression experiment.
- **Filter Probesets by Error:** Entities can be filtered based on the standard deviation or coefficient of variation using this option. For details refer to the section on Filter Probesets by Error

17.4.4 Analysis

- **Statistical Analysis**
  For details refer to section Statistical Analysis in the Data Import Wizard section.
- **Filter on Volcano Plot**
  For details refer to section Filter on Volcano Plot.
- **Fold Change**
  For details refer to section Fold change.
- **Clustering**
  For details refer to chapter Clustering Analysis.
- **Correlation**
  For details refer to chapter Correlation Analysis.
- **Find Similar Entities**
  For details refer to section Find similar entities.
- **Filter on Parameters**
  For details refer to section Filter on Parameters.
- **Principal Component Analysis**
  For details refer to section PCA.

17.4.5 Class Prediction

- **Build Prediction Model** For details refer to section Build Prediction Model.
- **Run Prediction** For details refer to section Run Prediction.

17.4.6 Results

- **Gene Ontology (GO) analysis**
  For more details see GO Analysis.
- **Gene Set Enrichment Analysis (GSEA)**
  For more details see Performing GSEA in **GeneSpring**.
- **Gene Set Analysis (GSA)**
  For more details see Gene Set Analysis.
- **Find Similar Entity Lists**
  This feature is discussed in a separate section called Find Similar Entity Lists.
- **Single Experiment Analysis or Multi-Omic Analysis**
  This feature is discussed in detail in the section Single Experiment and Multi-Omic Analysis.

- **Export to IPA**

  This feature is discussed in detail in the section Ingenuity Pathways Analysis (IPA) Connector.

- **Import IPA Entity List**

  This feature is discussed in detail in the section Import IPA Entity List.

- **NLP Network Discovery**

  This feature is discussed in detail in the section NLP Network Discovery.

- **MeSH Network Builder**

  This feature is discussed in detail in the section MeSH Network Builder.

- **Extract Relations via NLP**

  This feature is discussed in detail in the section Extract Relations via NLP.

17.4.7 Utilities

- **Import Entity list from File**

  For details refer to section Import Entity list from File.

- **Analysis: Biological Significance**

  For details refer to section Analysis: Biological Significance.

- **Filter On Entity List**

  For further details refer to section Filter On Entity List.

- **Remove Entities with missing signal values**

  For details refer to section Remove Entities with missing signal values.

17.4.8 Algorithm Technical Details

Here are some technical details of the Exon RMA16, Exon PLIER16, and Exon IterPLIER16 algorithms.

Exon RMA 16. Exon RMA does a GC based background correction (described below and performed only with the PM-GCBG option) followed by Quantile normalization followed by a Median Polish probe summarization, followed by a Variance Stabilization of 16. The computation takes roughly 30 seconds per CEL file with the *Full* option.

GCBG background correction bins background probes into 25 categories based on their GC value and corrects each PM by the median background value in its GC bin.

Exon PLIER 16. Exon PLIER does Quantile normalization followed by the PLIER summarization using the PM or the PM-GCBG options, followed by a Variance Stabilization of 16. The PLIER implementation and default parameters are those used in the Affymetrix Exact 1.2 package.

PLIER parameters can be configured from *Tools* $\longrightarrow$ *Options* $\longrightarrow$ *Affymetrix Exon Summarization Algorithms* $\longrightarrow$ *Exon PLIER/IterPLIER*.

Exon IterPLIER 16. Exon IterPLIER does Quantile normalization followed by the IterPLIER summarization using the PM or the PM-GCBG options, followed by a Variance Stabilization of 16. IterPLIER runs PLIER multiple times, each time with a smaller subset of the probes obtained by removing outliers from the previous PLIER run. IterPLIER parameters can be configured from *Tools* $\longrightarrow$ *Options* $\longrightarrow$ *Affymetrix Exon Summarization Algorithms* $\longrightarrow$ *Exon PLIER/IterPLIER*.

Chapter 18

Analyzing Affymetrix Exon Splicing Data

Alternative splicing is defined as variations in RNA splicing mechanisms resulting in multiple splice variants, each specific to a stage or condition of the cell. Affymetrix Exon chips are used for studying the alternative splicing of genes. A large population of human mRNAs undergo alternative splicing which generates splice variants that produce proteins with distinct and sometimes even antagonistic functions. Also changes in splicing signals or in sequences regulating splicing have been implicated as the cause for certain genetic mutations which result in human diseases. Thus measuring changes in splicing patterns is integral to understanding the disease mechanism or biological process under study. **GeneSpring** supports Exon Splicing analysis using the Affymetrix Exon Arrays.

18.1 Running the Affymetrix Exon Splicing Workflow

Upon launching **GeneSpring**, the startup is displayed with 3 options.

1. Create new project
2. Open existing project
3. Open recent project

Either a new project can be created or else a previously generated project can be opened and re-analyzed. On selecting ***Create new project***, a window appears in which details (Name of the project and Notes) can be recorded. Press ***OK*** to proceed. An Experiment Selection Dialog window then appears with two options

1. Create new experiment

2. Open existing experiment

Selecting ***Create new experiment*** allows the user to create a new experiment (steps described below). ***Open existing experiment*** allows the user to use existing experiments from any previous projects in the current project. Choosing ***Create new experiment*** opens up a **New Experiment dialog** in which experiment name can be assigned. The *Analysis Type* (exon) and the *Experiment Type* (Affymetrix Exon Splicing) should then be specified. The **Data Import Wizard** is the only option for the Affymetrix Exon Splicing experiment.

Upon clicking ***OK***, the Affymetrix Exon Splicing experiment creation wizard appears.

18.1.1 Creating an Affymetrix Exon Splicing Experiment

Data Import Wizard analysis can be done using either CEL or CHP files. However, a combination of both file types cannot be used. If CHP files are being used for analysis, then both transcript (gene) summarized and probeset (exon) summarized files need to be present for a sample.

New Experiment (Step 1 of 7): Load data An experiment can be created either using data files or using samples. **GeneSpring** differentiates between a data file and a sample. A data file refers to the hybridization data obtained from a scanner. A sample, on the other hand is created within the tool, when it associates the data file with its appropriate technology. For more details, refer to the section on Technology. Thus a sample created within a technology cannot be used in an experiment of another technology. These samples are stored in the system and can be used to create another experiment of the same technology.

- For loading new CEL/CHP files, use ***Choose Files***.
- If the CEL/CHP files have been previously used in experiments ***Choose Samples*** can be used.

18.1.1.1 Selecting technology for Affymetrix Exon Splicing Experiments

Before creating the experiment, user should download the technology or technologies from the menu **Annotations** → **Create Technology** → **From Agilent Server or From Agilent eArray**. Users can also update the technology annotations as described in section Update Technology Annotations. After loading data, **GeneSpring** will prompt the user to select a technology for creating the Affymetrix Exon Splicing experiment. See Figure 18.1.

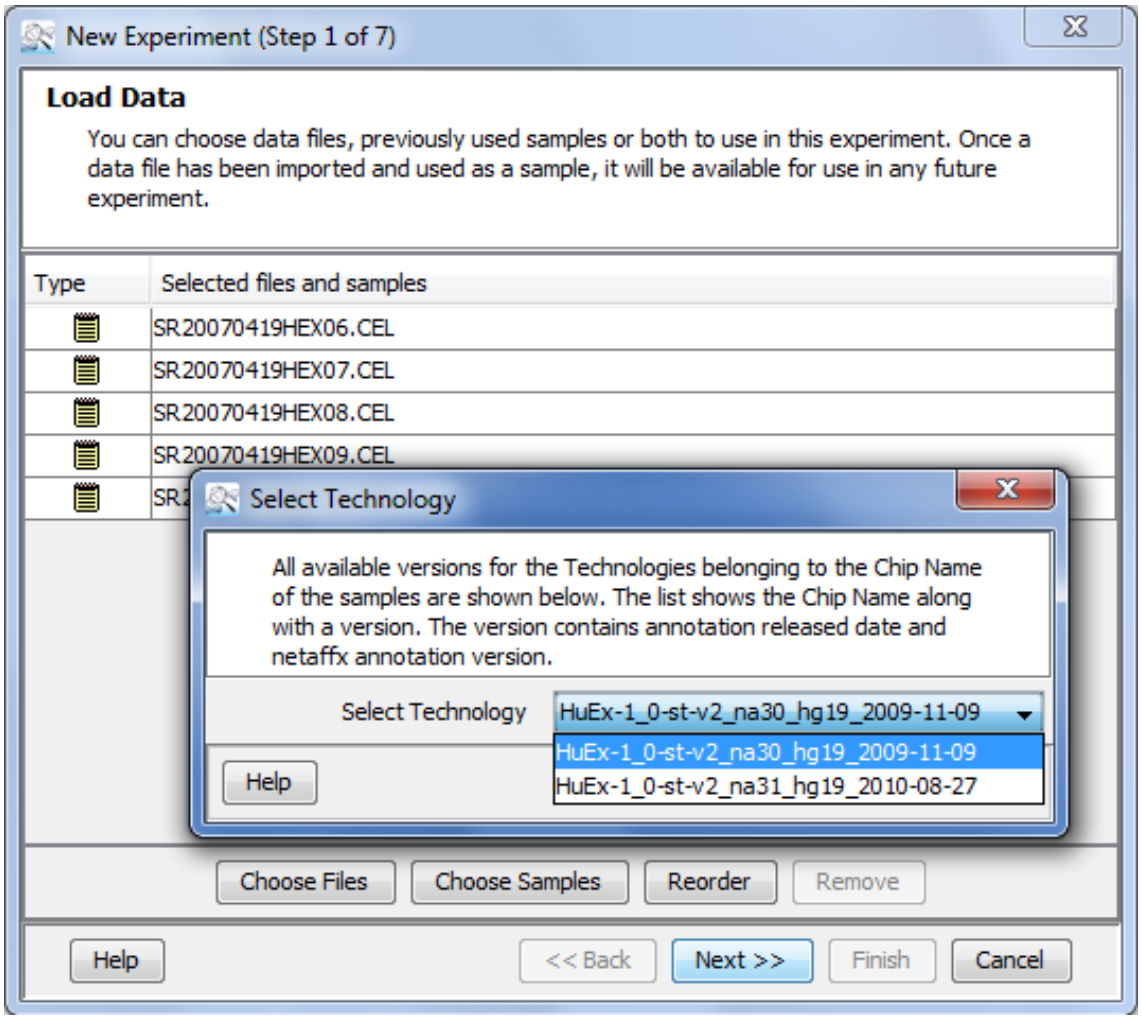


Figure 18.1: Select Technology

The drop down menu in 'Select Technology' step shows a list of all available technologies which have been downloaded. Upon selecting one and clicking OK, the Affymetrix Exon Splicing workflow appears.

Note: In **GeneSpring Exon Splicing Workflow**, experiment creation using CHP files requires 2 types of CHP files per array i.e., the transcript level CHP file and the probeset level CHP file. This is necessary as **GeneSpring** requires the probeset level data for splicing analysis. If the user has not provided the required files for each array, **GeneSpring** prompts the user to provide the necessary files. Refer Figure 18.3. Additionally if the same experiment is created again from ***Project Navigator*** $\longrightarrow$ ***Experiment Name*** $\longrightarrow$ ***Right click*** $\longrightarrow$ ***Create New Experiment***, the files taken into account are only the transcript level files and the user needs to provide the probeset level files. These files can be loaded from the ***Choose Samples*** option.

Step 1 of 7 of Experiment Creation, the **Load Data** window, is illustrated in Figure 18.2.

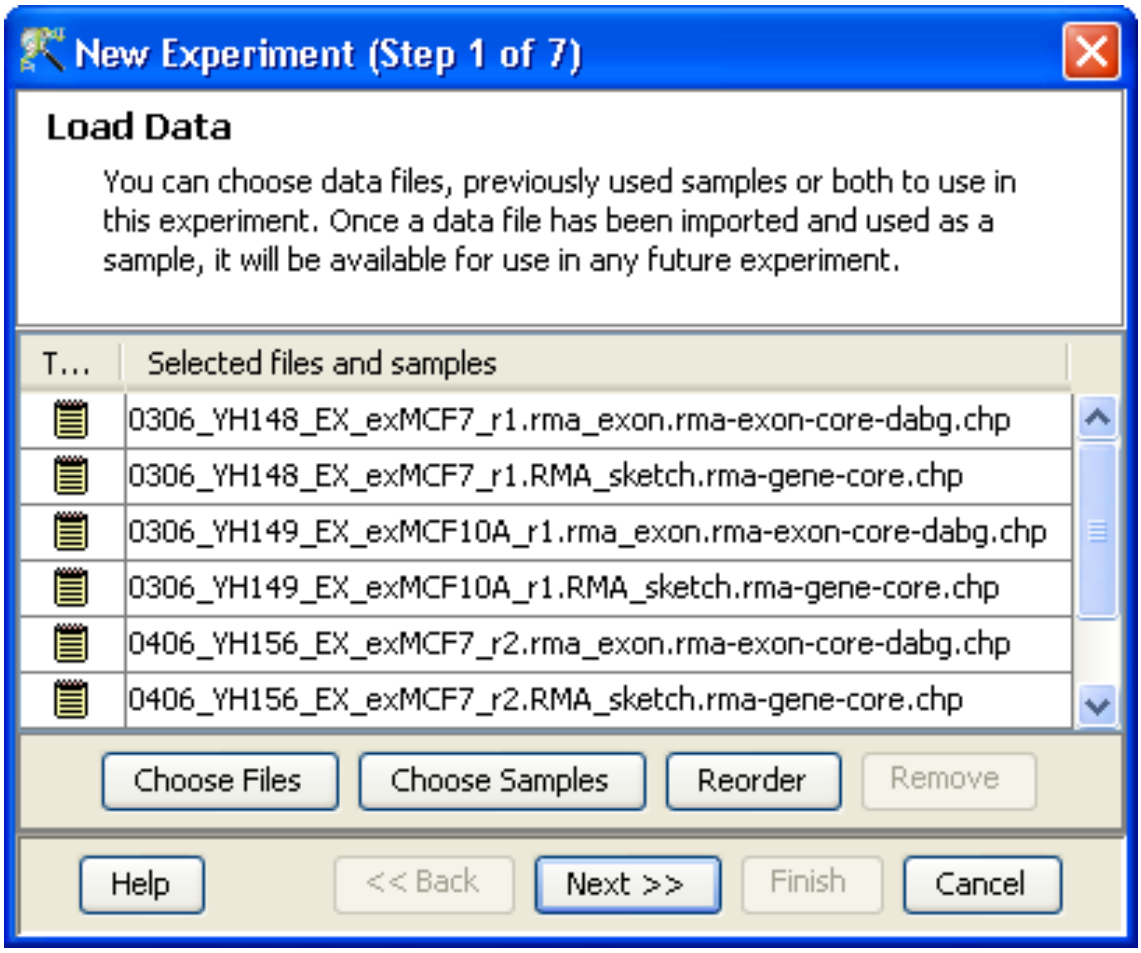


Figure 18.2: Load Data

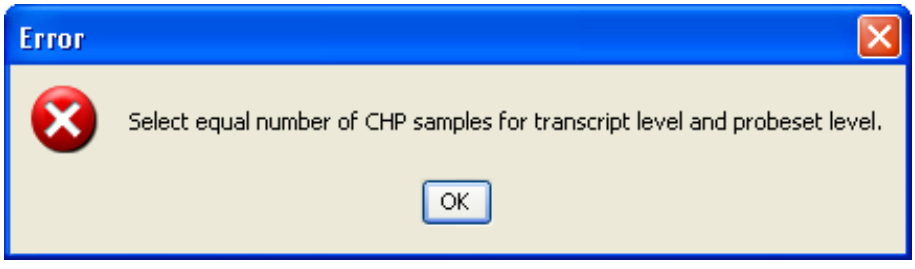


Figure 18.3: Error Message

New Experiment (Step 2 of 7): Selecting ARR files ARR files are Affymetrix files that hold annotation information for each sample's CEL and CHP file and are associated with the sample based on the sample name. These are imported as annotations to the sample. Click on ***Next*** to proceed to the next step.

Step 2 of 7 of Experiment Creation, the **Select ARR files** window, is depicted in the Figure 18.4.

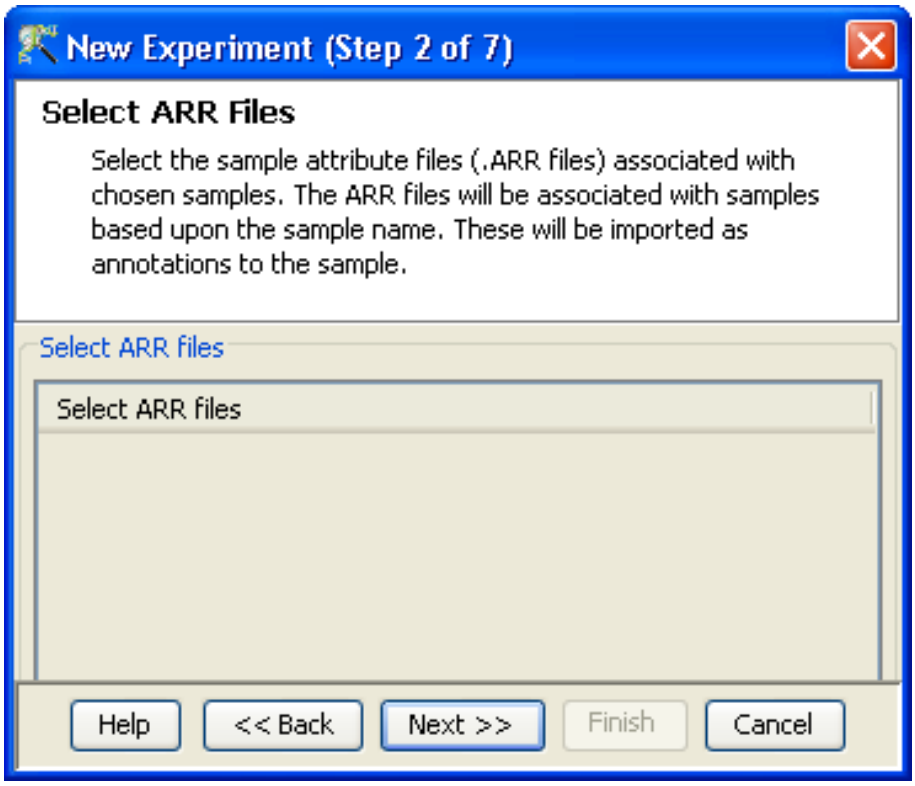


Figure 18.4: Select ARR files

New Experiment (Step 3 of 7): Pairing of transcript and probeset level files This step is specific to CHP files. The tool pairs both the CHP files of a sample automatically, based on the file names. But in case the naming is different and the pairing done is incorrect, the user can change the pairing by selecting the file and moving it with the help of the buttons provided on the right side of the wizard. See Figure 18.5.

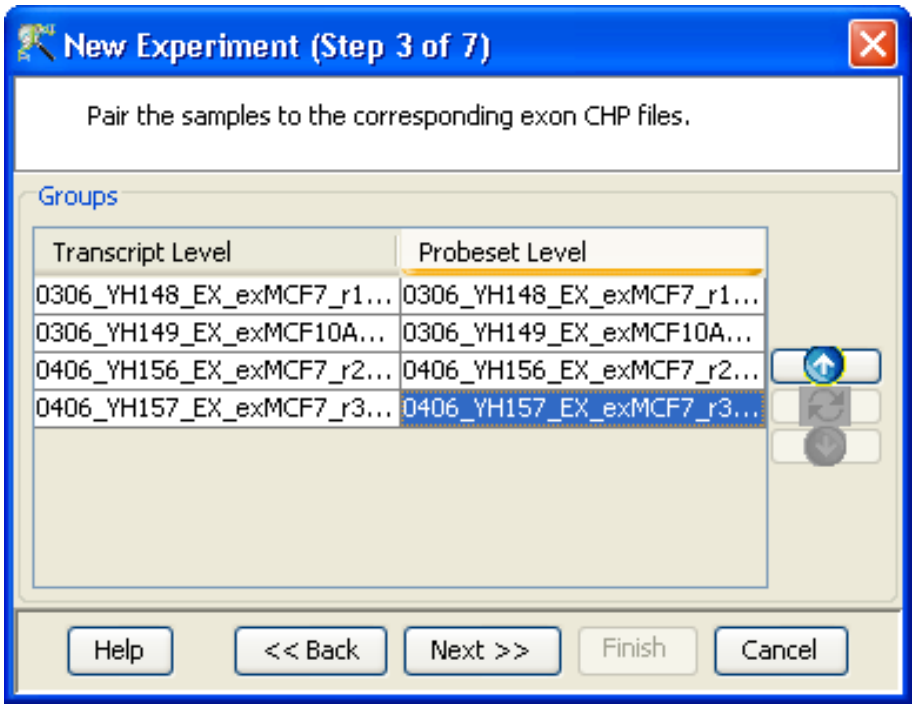


Figure 18.5: Pairing of CHP files

New Experiment (Step 4 of 7): Preprocess Baseline Options Specific to the CEL files, step 4 provides three summarization algorithms. The suffix 16 in these algorithms denotes a variance stabilization addition of 16 to the result of each algorithm.

- RMA16 Irazarry *et al.* [Ir1, Ir2, Bo].
- PLIER16 Hubbell [Hu2].
- IterativePLIER16

This step also enables the user to select the Confidence level, as Core, Extended or Full, using which the summarization is done.

1. Core comprising transcript clusters groups unique exon level probe sets with strong annotation support e.g., from RefSeq and other GenBank alignments of 'complete CDS' transcripts, into transcript clusters.
2. Extended comprising transcript clusters groups unique exon level probe sets with empirical annotation support into transcript clusters. This includes cDNA transcripts, mapped syntenic mRNA from rat and mouse, and probe sets which are part of the Ensembl or Vega transcript annotation set.
3. Full transcript clusters including ab-initio predictions from Geneid, Genscan, GENSCAN Suboptimal, Exoniphy, RNAgene, SgpGene and TWINSCAN.

The full list includes both the core and extended lists while the extended list contains the core probeset list.

NOTE: Exon Splicing support for Affymetrix Gene ST Arrays have been added in **GeneSpring** . However, **GeneSpring** will consider 'All' transcript clusters and probesets irrespective of the confidence levels.

Subsequent to probeset summarization, Baseline Transformation of the data can be performed. Baseline Transformation is carried out row-wise across all samples. This data processing step

is particularly useful when visualizing the results in a profile plot or heat map. The baseline transformation options, available in **GeneSpring** are:

- ***Do not perform baseline***
- ***Baseline to median of all samples:*** For each row (probe), the median of the log summarized values across all the samples is calculated. This value is then subtracted from the probe value for all samples.
- ***Baseline to median of control samples:*** Here control samples are used to calculate the median value for each probe. This value is then subtracted from the probe value for all samples. The controls could be an individual control for each sample or it could be a set of controls. Alternatively, a set of samples can be used as controls for all samples. For specifying the control for a sample, select the sample and click on ***Assign value***. This opens up the ***Choose Control Samples*** window from where the samples designated as *Controls* should be moved from the *Available Items* box to the *Selected Items* box. Click on ***Ok***. This will show the control samples for each of the samples. In *Baseline to median of control samples*, for each probe the median of the log summarized values from the control samples is first computed and then this is subtracted from the sample. If a single sample is chosen as the control sample, then the probe values of the control sample are subtracted from its corresponding sample.

Clicking ***Finish*** creates an experiment, which is displayed as a Box Whisker plot in the active view. Alternative views can be chosen for display by navigating to ***View*** in Toolbar. Figure 18.6 shows the Step 4 of 7 of Experiment Creation.

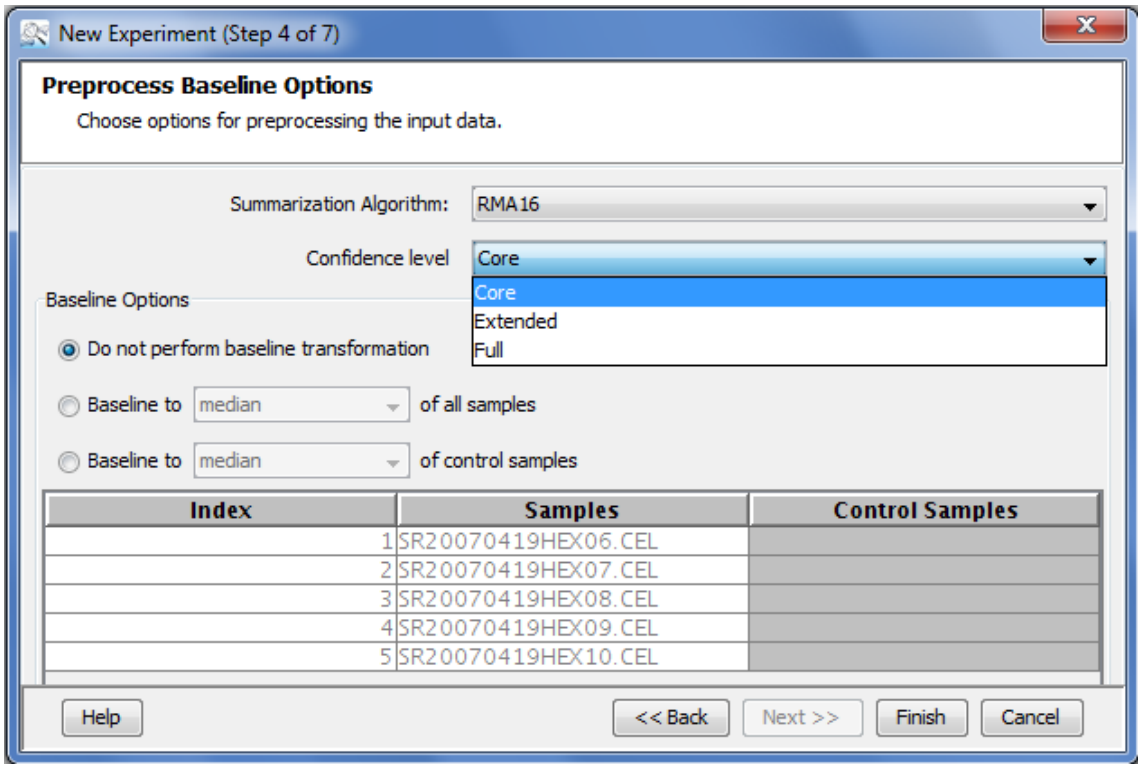


Figure 18.6: Summarization Algorithm

New Experiment (Step 5 of 7): This step is specific for CHP files only. It gives the user the following normalization options. See Figure 18.7.

- **Percentile Shift:** On selecting this normalization method, the **Percentile Target** box gets enabled allowing the user to enter a specific percentile value. The default value is set at 75^th percentile.
- **Scale:** On selecting this normalization method, the user is presented with an option to scale it to the median of percentile target/mean of means of 'all' or 'control' samples. On choosing the latter, the user has to select the control samples from the available samples in the **Choose Samples** box. The **Percentile Target** box is set at a default value of 75.

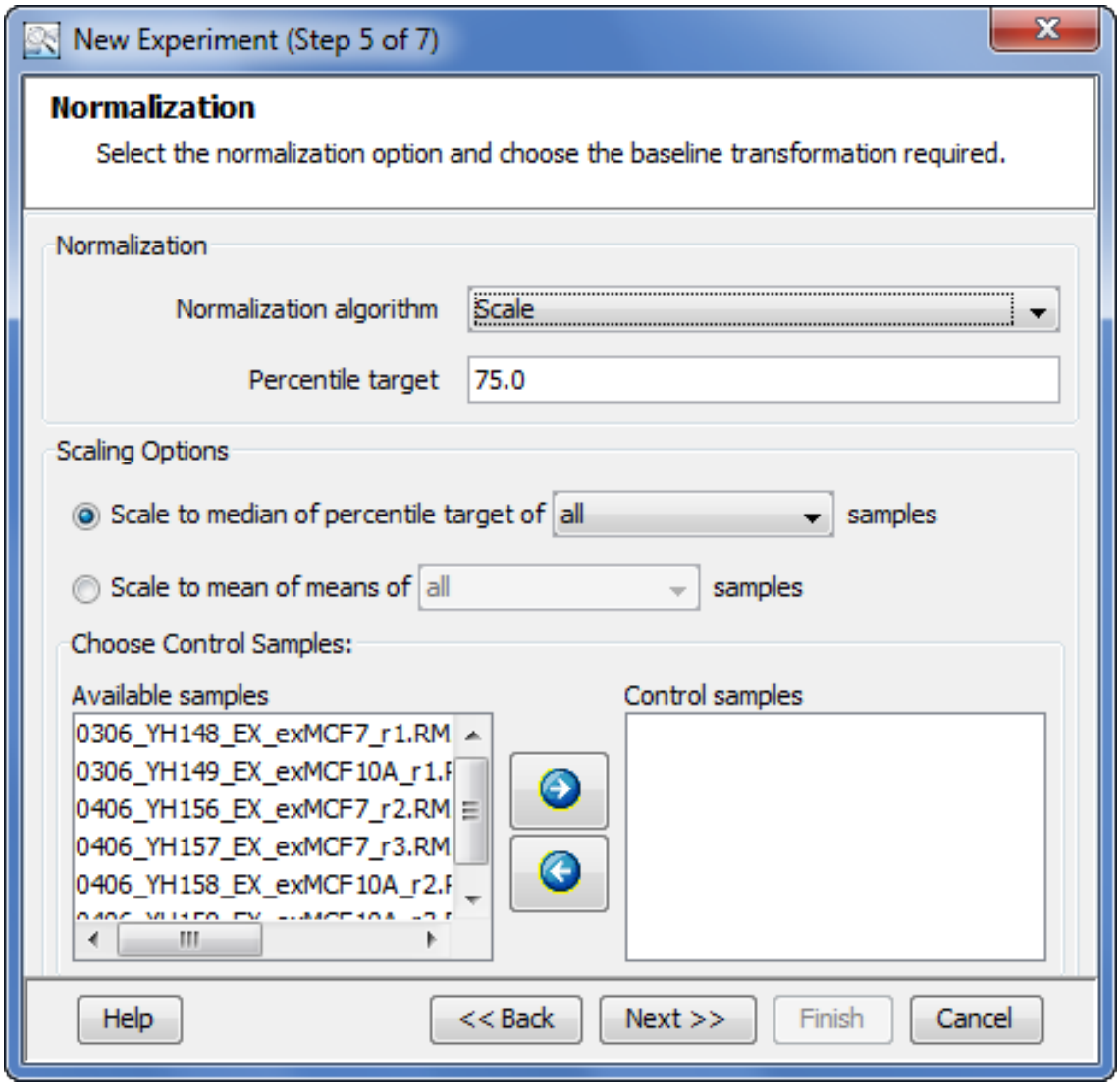


Figure 18.7: Normalization

- **Normalize to control genes:** After selecting this option, the user has to specify the control genes in the next wizard. The **Percentile Target** box is disabled.
- **Normalize to External Value:** This option will bring up a table listing all samples and a default scaling factor of '1.0' against each of them. The user can use the *'Assign Value'* button at the bottom to assign a different scaling factor to each of the sample;

multiple samples can be chosen simultaneously and assigned a value. The **Percentile Target** box is disabled.

For details on the above normalization methods, refer to section Normalization.

New Experiment (Step 6 of 7): If the **Normalize to control genes** option is chosen, then the list of control entities can be specified in the following ways in this wizard:

- By choosing a file(s) (txt, csv or tsv) which contains the control entities of choice denoted by their probe id. Any other annotation will not be suitable.
- By searching for a particular entity by using the ***Choose Entities*** option. This leads to a search wizard in which the entities can be selected. All the annotation columns present in the technology are provided and the user can search using terms from any of the columns. The user has to select the entities that he/she wants to use as controls when they appear in the **Output Views** page and then click ***Finish***. This will result in the entities getting selected as control entities and will appear in the wizard.

The user can choose either one or both the options to select his/her control genes. The chosen genes can also be removed after selecting the same.

In case the entities chosen are not present in the technology or sample, they will not be taken into account during experiment creation. The entities which are present in the process of experiment creation will appear under matched probe ids whereas the entities not present will appear under unmatched probe ids in the experiment notes in the experiment inspector. See Figure 18.8.

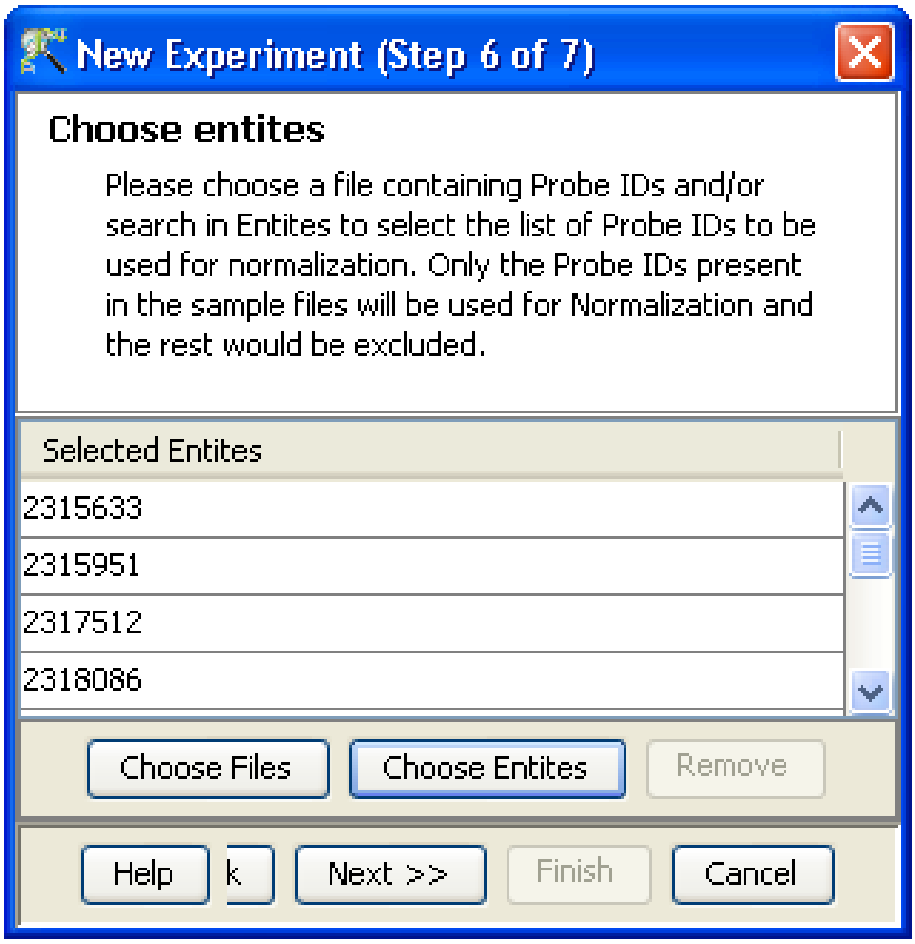


Figure 18.8: Normalize to control genes

New Experiment (Step 7 of 7): This step allows the user to perform baseline transformation. The methods available are the same as those used for CEL files in Step 4 of 7.

Clicking ***Finish*** creates an experiment, which is displayed as a Box Whisker plot in the active view. Alternative views can be chosen for display by navigating to ***View*** in Toolbar. The final step of Experiment Creation (CHP file specific) is shown in

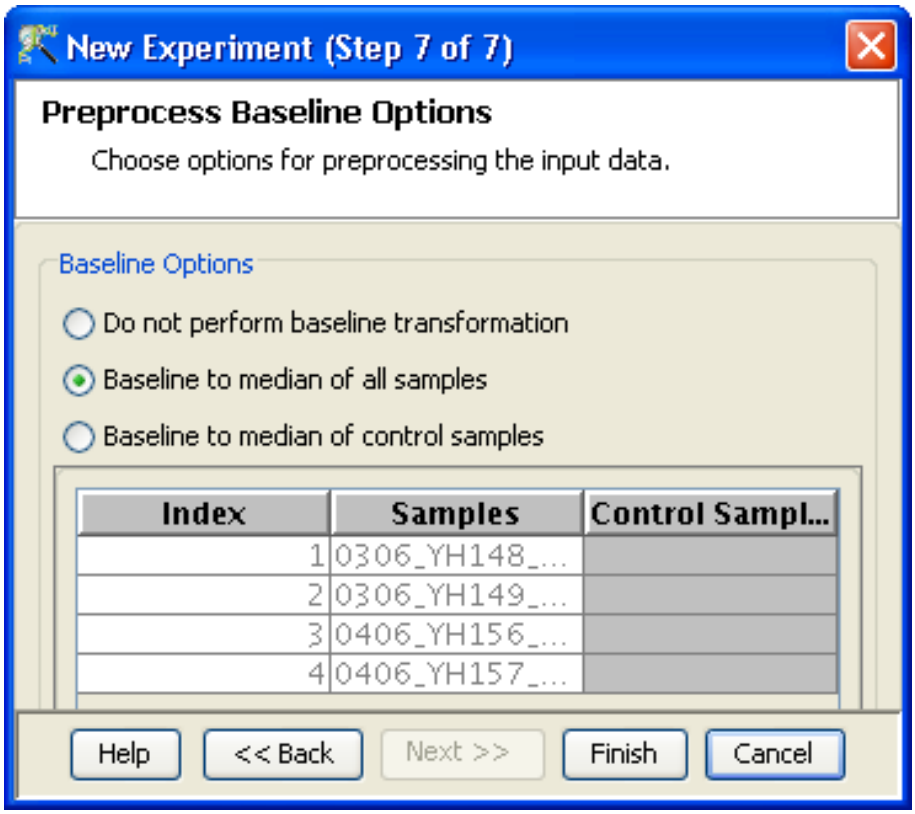


Figure 18.9: Normalization and Baseline Transformation

18.1.2 Data Processing for Exon arrays

This section describes the data processing which occurs during the experiment creation step. During the experiment creation steps, probeset level and transcript level data are processed simultaneously. Even though only the transcript level processing is user specified, the probe set level processing uses the exact same set of parameters. Thus, while for exon summarization the probes in each individual probe set are summarized, for generating transcript level data, all the probes within the transcript are summarized. DABG is then performed. DABG stands for "detection above background". It is calculated by comparing Perfect Match(PM) probes to a distribution of background probes. This comparison is used to generate a p-value. This is then combined into a probe set level p-value using the Fischer equation. This option allows the user to filter the transcripts(genes) having low expression values by correlating it with the probeset(exon) level data.

1. **File formats**: The data file should be present either as a CEL file or a CHP file. However while creating an experiment; only one type of file (CEL/CHP) can be used.
2. **Raw signal values (CEL files)**: In an Affymetrix Exon Expression experiment, the term ''raw'' signal values refers to the linear data which has been summarized using a summarization algorithm (RMA16, PLIER 16 and Iterative PLIER 16). All summarization algorithms also do variance stabilization by adding 16. This is applicable to both the transcript and the probeset level data. Raw values for both are shown in the entity inspector.

3. **Raw signal values (CHP files)**: In an Affymetrix Exon Expression experiment, the term "raw" files refers to the linear data obtained from the CHP files. This is applicable to both the transcript and the probeset level data. Raw values for both are shown in the entity inspector.

4. **Normalized signal values (CEL files)**: "Normalized'' values are generated after the log transformation and baseline transformation step. This is applicable to both the transcript and the probeset level data and the same transforms are performed on both. The normalized signal value of the probeset can be viewed under the signal value tab of the splicing visualization link.

5. **Normalized signal values (CHP files)**: The term "Normalized" refers to values generated after log transformation, normalization (Percentile Shift, Scale or Normalize to control genes) and baseline transformation. This is applicable to both the transcript and the probeset level data. The normalized signal value of the probe set can be viewed under the signal value tab of the splicing visualization link.

6. **Gene-level Normalized intensity**: It is the difference of normalized exon-level signal and its normalized gene-level signal

7. **Treatment of on-chip replicates**: Not Applicable.

8. **Flag values**: Not Applicable.

9. **Treatment of Control probes**: Not Applicable.

10. **Empty Cells**: Not Applicable.

11. **Sequence of events (CEL files)**: The sequence of events involved in the processing of a CEL file is: Summarization$\longrightarrow$ Log Transformation$\longrightarrow$ Baseline Transformation. This is applicable to both the transcript and the probeset level data.

12. **Sequence of events (CHP files)**: If the data in the CHP file is already log transformed, then **GeneSpring** detects it and proceeds with the normalization step. This is applicable to both the transcript and the probeset level data.

18.1.3 Experiment setup

Once an experiment is created, the *Data Import Wizard* steps appear on the right hand side. This gives options to view quick start guide, create experiment grouping and interpretation. It is not possible to create gene level experiment for splicing experiments. See section Experiment Setup for complete details.

Note: All links in the Workflow Browser work on transcript level data; the only exception are links in the section on Splicing Analysis. All entity lists store only transcript level data as well.

18.1.4 Quality Control

- **Quality Control on Samples**

Quality Control (QC) allows you decide which samples are ambiguous and which are passing the quality criteria. Based on the QC results, unreliable you can remove samples from further downstream analysis. The QC dialog displays three tiled windows:

- 3D PCA scores, Correlation Analysis - Samples, and a Correlation Coefficients tab.
- Experiment Grouping, Hybridization Controls (displayed only for .CEL files).
- Legend.

Figure 18.10 has tiled windows which reflect the QC on samples.

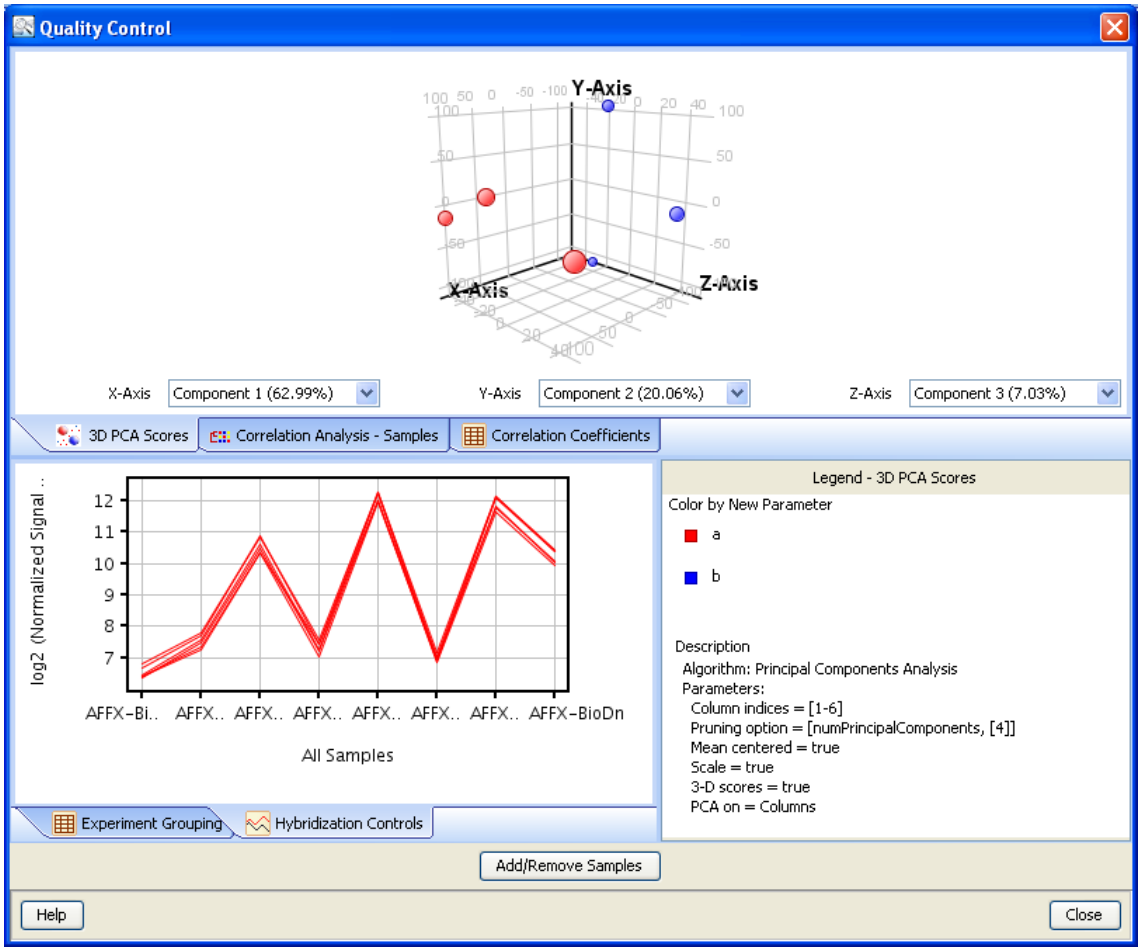


Figure 18.10: Quality Control Dialog

In the **Principal Component Analysis (PCA)** tab, PCA scores are visually represented in a 3D scatter plot. These scores are used to check data quality. It shows one point per array and is colored by the *Experiment Factors* provided earlier in the *Experiment Groupings* view. This allows viewing of separations between groups of replicates. Ideally, replicates within a group should cluster together and separately from arrays in other groups. The PCA components, represented in the X, Y, and Z axes are numbered 1, 2, 3... according to their decreasing significance. The 3D PCA scores plot can be customized via ***Right-Click***⟶ ***Properties***. To zoom into a 3D Scatter plot, press the Shift key and simultaneously hold down the left mouse button and move the mouse upwards. To zoom out, move the mouse

downwards instead. To rotate, press the Ctrl key, simultaneously hold down the left mouse button and move the mouse around the plot.

The **Correlation Analysis - Samples** tab shows the correlation analysis heatmap across arrays (samples) (See How to Perform Correlation Analysis on Samples on page 750). The correlation coefficient for each pair of arrays are displayed in the **Correlation Coefficients** tab. Correlation coefficients are calculated using Pearson Similarity Metric (see Pearson on page 799).

Pearson Similarity Metric: Calculates the mean of all elements in vector **a**. Then it subtracts that value from each element in **a** and calls the resulting vector **A**. It does the same for **b** to make a vector **B**. Correlation Coefficient is then computed as **A.B**/(—**A**——**B**—).

In a Correlation Analysis heatmap:

- Sample Order- In a correlation heatmap, the initial order of samples is the same as the input sample order at the time of experiment creation.
- Clustering - Perform clustering analysis on the correlation heatmap using the **Run clustering** icon from the correlation analysis heatmap toolbar (See To run clustering on a correlation heatmap on page 760). Samples are clustered based on correlation coefficients using hierarchical clustering algorithm and plotted in the order they cluster together. The linkage rule used is average and the distance metrics used is Euclidean.

For more information on the available toolbar icons in a correlation analysis heatmap, see Correlation Heatmap View Toolbar on page 776. See Properties on 792 to change the properties of the correlation analysis heatmap.

NOTE Correlation coefficients are computed on raw, non-normalized data and in linear scale.

Experiment Grouping displays the parameters and parameter values for each sample.

The *Hybridization Controls* view depicts the hybridization quality. Hybridization controls are composed of a mixture of biotin-labelled cRNA transcripts of bioB, bioC, bioD, and cre prepared in staggered concentrations (1.5, 5, 25, and 100pm respectively). This mixture is spiked-in into the hybridization cocktail. bioB is at the level of assay sensitivity and should be called Present at least 50% of the time. bioC, bioD and cre must be Present all of the time and must appear in increasing concentrations. The X-axis in this graph represents the controls and the Y-axis, the log of the Normalized Signal Values.

The third window shows the legend of the active QC tab.

Unsatisfactory samples or those that have not passed the QC criteria can be removed from further analysis, at this stage, using *Add/Remove Samples* button. Once a few samples are removed, re-summarization of the remaining samples is carried out again. The samples removed earlier can also be added back. Click on **OK** to proceed.

- **Filter Probe Set by Expression** Entities are filtered based on their signal intensity values. For details refer to the section on Filter Probesets by Expression
- **Filter Probe Set by Flags** No flags are generated during creation of exon splicing experiment.

- **Filter Probesets by Error:** Entities can be filtered based on the standard deviation or coefficient of variation using this option. For details refer to the section on Filter Probesets by Error

18.1.5 Analysis

- **Statistical Analysis**

  For details refer to section Statistical Analysis in the Data Import Wizard section.

- **Filter on Volcano Plot**

  For details refer to section Filter on Volcano Plot.

- **Fold Change**

  For details refer to section Fold change.

- **Clustering**

  For details refer to chapter Clustering Analysis.

- **Correlation**

  For details refer to chapter Correlation Analysis.

- **Find Similar Entities**

  For details refer to section Find similar entities.

- **Filter on Parameters**

  For details refer to section Filter on Parameters.

- **Principal Component Analysis**

  For details refer to section PCA.

18.1.6 Exon Splicing Analysis

This analysis section is specific to the Affymetrix Exon Splicing Workflow. The following options are provided:

- **Filter transcripts on DABG:**

  DABG is performed at the time of experiment creation. For Exon Splicing analyis, the transcripts can be filtered on DABG results. This occurs through a four step wizard and the filtering considers only core probesets (for the purpose of calling a transcript as Present) even though DABG values were generated initially for all the probe sets. The DABG values are stored only for probesets that are a part of exon summarization or if they are marked core. This change does not affect the downstream analysis and only reduces the memory usage.

  1. The first step allows the user to choose the entity list and interpretation. See Figure 18.11.

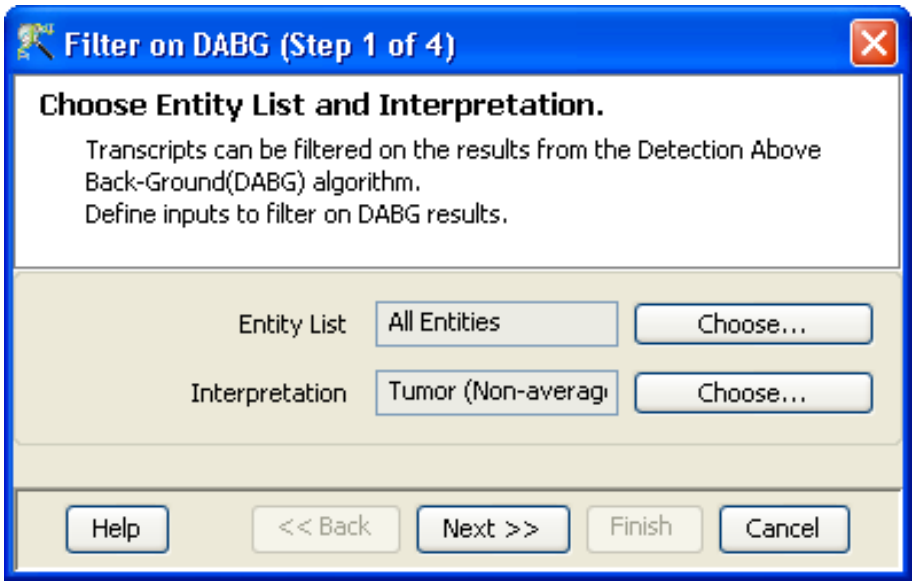


Figure 18.11: Input Data

2. In the second step, the filtering options can be specified. Probe sets are defined as Present based upon a p-value cut-off which was generated during executing of DABG algorithm. Secondly, the minimum percentage of **core** exons that should be present in a gene in a sample to mark it as Present should be given. For a transcript to be called as Present, a substantial number of core probe sets should be "Present" (as designated by the DABG generated p-value). The default value specifies 50% of core probe sets to be 'Present'. The percentage of samples (within a condition) in which a gene must be present for it to be retained is set at 50% and can be increased for more stringency. See Figure 18.12.

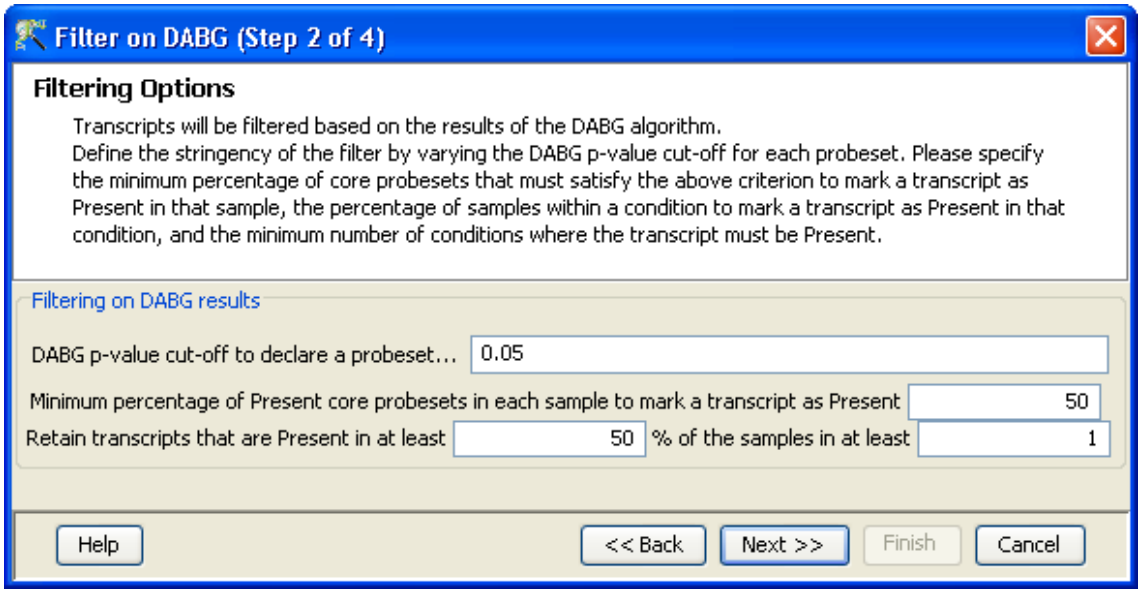


Figure 18.12: Filtering Options

3. This step shows the entities which have passed the filter, in the form of a spreadsheet(along with their normalized values) and a profile plot. The number of entities passing the filter is mentioned at the top of the panel. See Figure 18.13

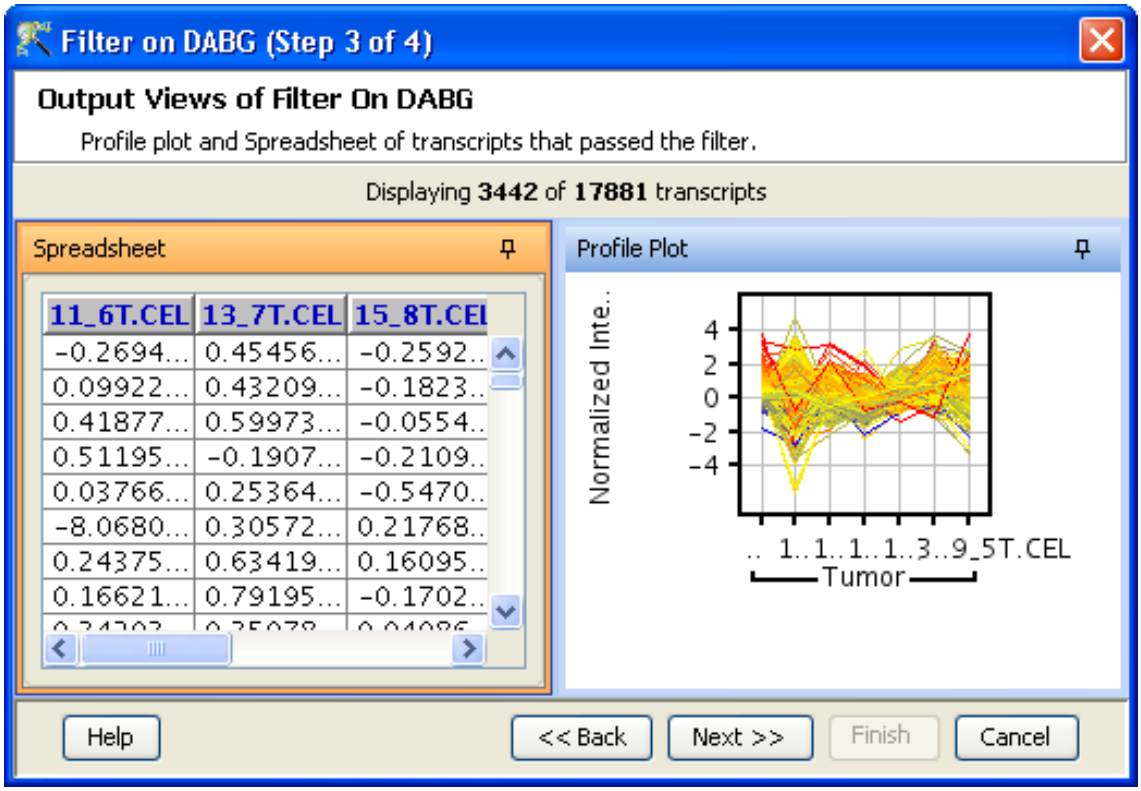


Figure 18.13: Output Views

4. The last step shows all the entities passing the filter along with their annotations. It also shows the details (regarding creation date, modification date, owner, number of entities, notes etc.) of the entity list. Click *Finish* and an entity list will be created corresponding to entities which satisfied the cut-off. Double clicking on an entity in the *Entities* table opens up an *Entity Inspector* giving the annotations corresponding to the selected profile. Additional tabs in the *Entity Inspector* give the raw and the normalized values for that entity. The name of the entity list will be displayed in the experiment navigator. Annotations being displayed here can be configured using *Configure Columns* button. See Figure 18.14.

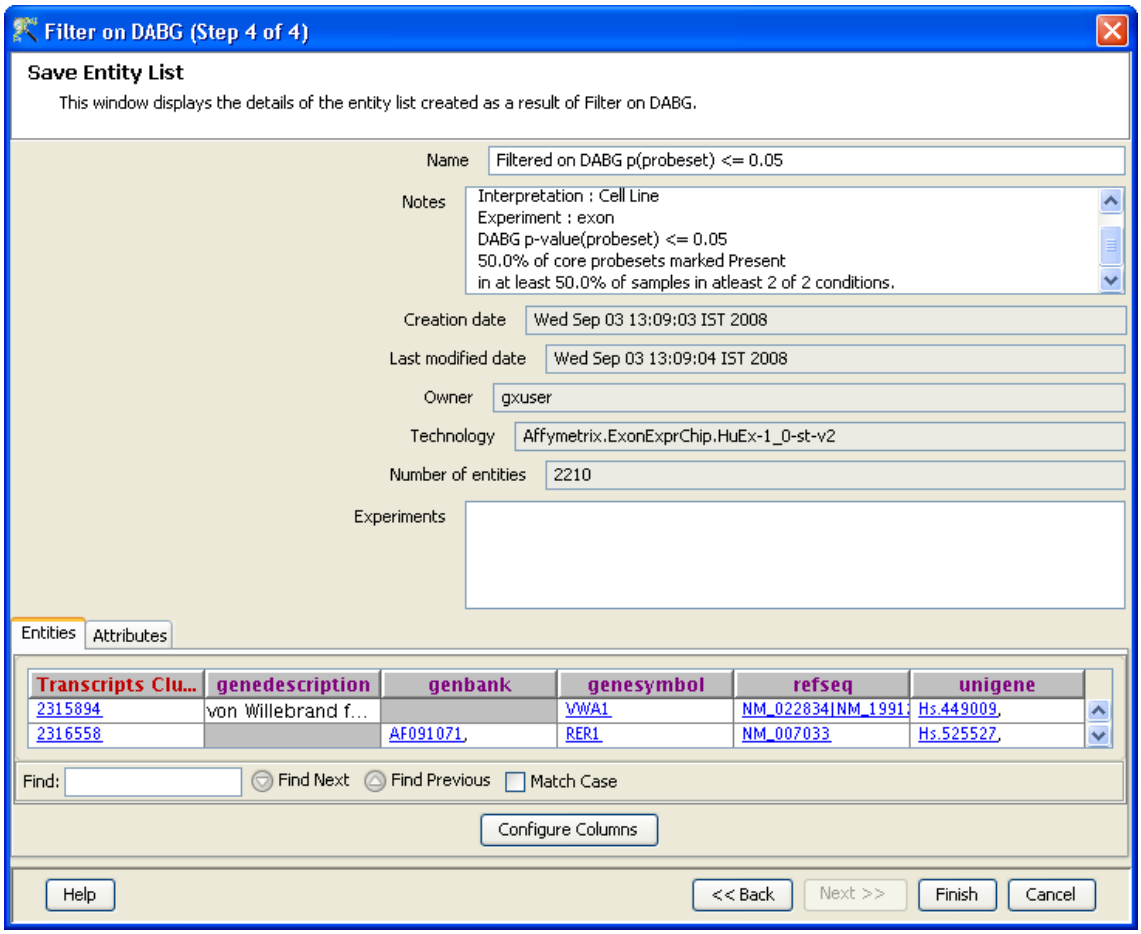


Figure 18.14: Save Entity List

For more details on DABG and on the defaults used in the filtering option, refer to Affymetrix white paper [2].

- **Splicing ANOVA:**

  Splicing ANOVA initially calculates the gene-level normalized intensities for each of the probesets (i.e., the difference between probeset level signal and transcript level signal). Then it runs a $(n+1)$-way ANOVA where n denotes the number of parameters in the chosen interpretation and the plus 1 is on account of the added probeset parameter. Currently, **GeneSpring** supports values of only 1 or 2 for n. The alternative splicing p-value is given by the p-value for the probeset*parameter term when there is only one parameter. In the event that there two parameters, individual p-values are output for each of probeset*parameter1 and probeset*parameter2 as well as probeset*parameter1*parameter2. Additionally, in situations where parameter1 and parameter2 are such that each replicate group has only one sample, only the probeset*parameter1 and probeset*parameter2 p-values are output. This could happen in paired experiments, i.e., if you have paired tumor and normal samples and 2 parameters, one indicating disease-state and other indicating the individual from whom the sample is derived; in such cases, the p-value of interest is probeset*disease-state. Note that both balanced and unbalanced designs are supported but balanced designs will run faster. Unbalanced designs will progress slowly for transcripts with many probesets and cancelling will cause display of results on all transcripts which have completed so far.

Splicing ANOVA can be executed in 5 steps:

– This step requires the user to provide the entity list and the interpretation. See Figure 18.15.

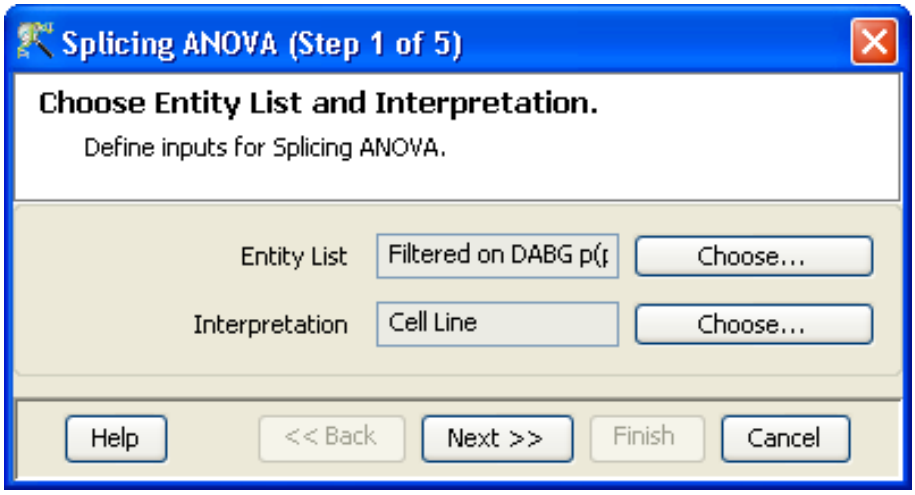


Figure 18.15: Input Data

– Filtering criteria specified are provided here. See Figure 18.16.

1. This specifies the probe set list on which Splicing ANOVA should be calculated. This option is dependant on the probe set list initially used for summarization. For example, if the full list was selected then both the core and extended lists are available and if extended was selected, then core would also appear in the list of options.
2. Filtering of probe sets is based upon the results of DABG algorithm. For more details refer to the section on Filter transcripts on DABG. Note that unlike the filter transcripts steps, the goal here is to identify which probesets for a transcript should be carried into splicing ANOVA.
3. Gene-level Normalized Intensity: Probesets with large gene-level normalized intensities are excluded from ANOVA. The default is set at 5.0, which means that if the gene level normalized intensity of any probeset is greater than log (base-2.0) of 5.0 in a minimum of one sample, then that probe set will be excluded from splicing ANOVA. This filter is implemented to weed out probes with high background and cross hybridization potential.
4. Cross-hybridization: Probe sets with high cross-hybridization potential are removed from the analysis. Only probesets with value of 1 have been recommended to be included in analysis. Refer to [2] for more details.

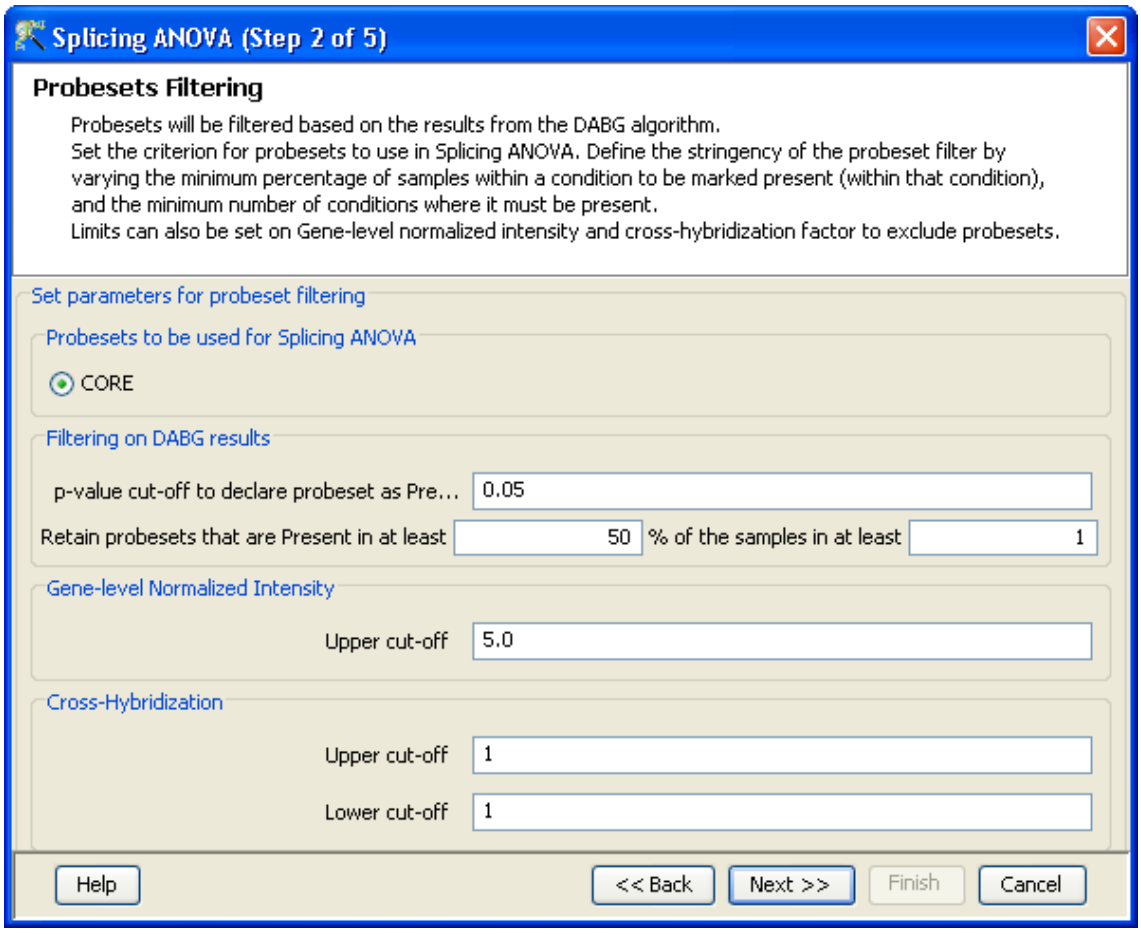


Figure 18.16: Filtering of Probesets

– The multiple testing correction to be implemented for p-value computation is chosen here. See Figure 18.17.

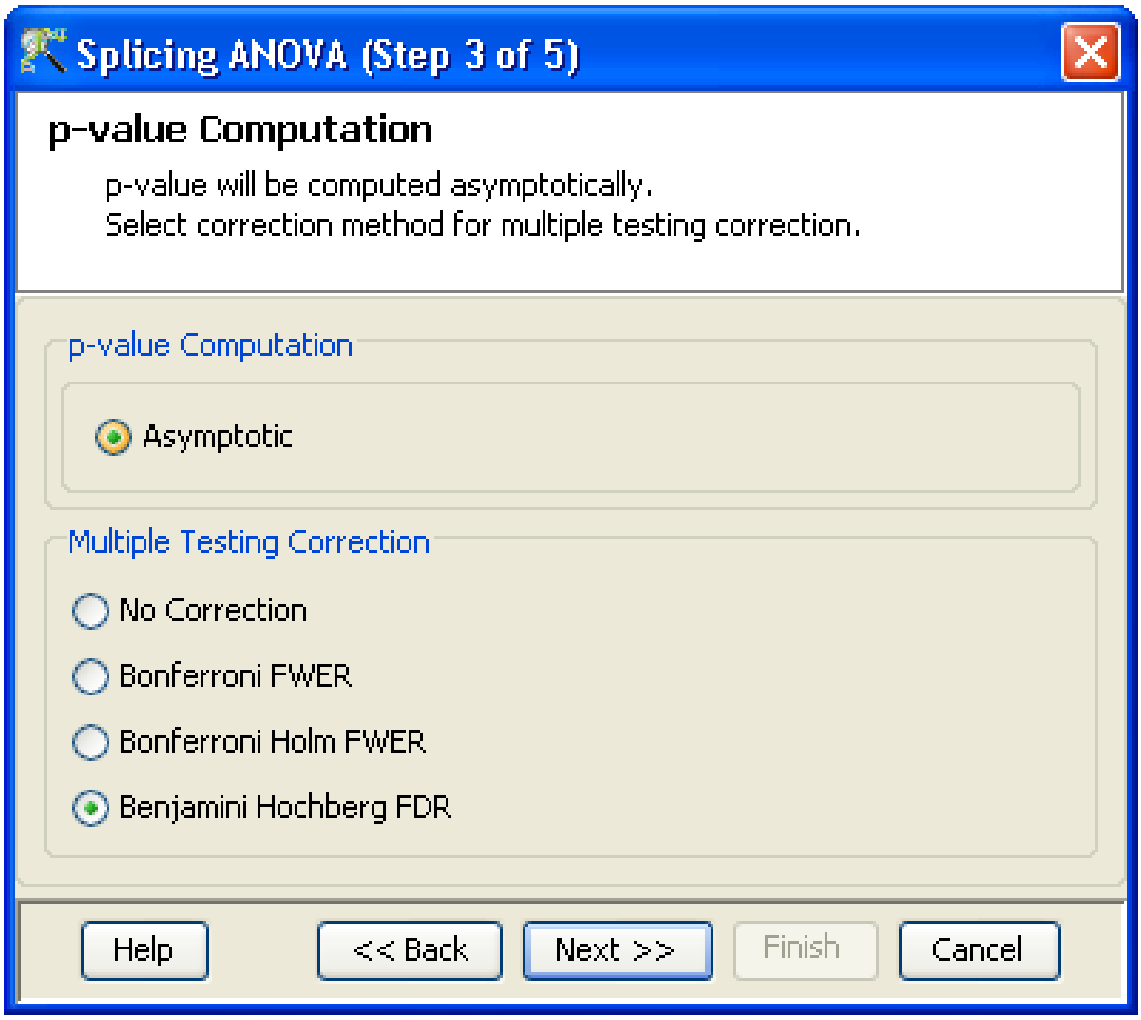


Figure 18.17: Multiple Testing Correction

– This step shows the results of the Splicing ANOVA in the form of a spreadsheet. For each transcript, the p-value, corrected p-value and the number of probesets that were for performing Splicing ANOVA are shown. If multiple p-value are computed, then the list of transcripts shown are exactly those for which any one of the p-values is within the specified threshold. The default p-value cut-off used is 0.05 but it can be reconfigured using the ***Change p-value cut-off*** button. Transcripts in which only one probeset has passed the previously applied filters are automatically excluded. The term PROBESET in the p-value names indicates that the p-value is for an interaction term with the probeset parameter. See Figure 18.18.

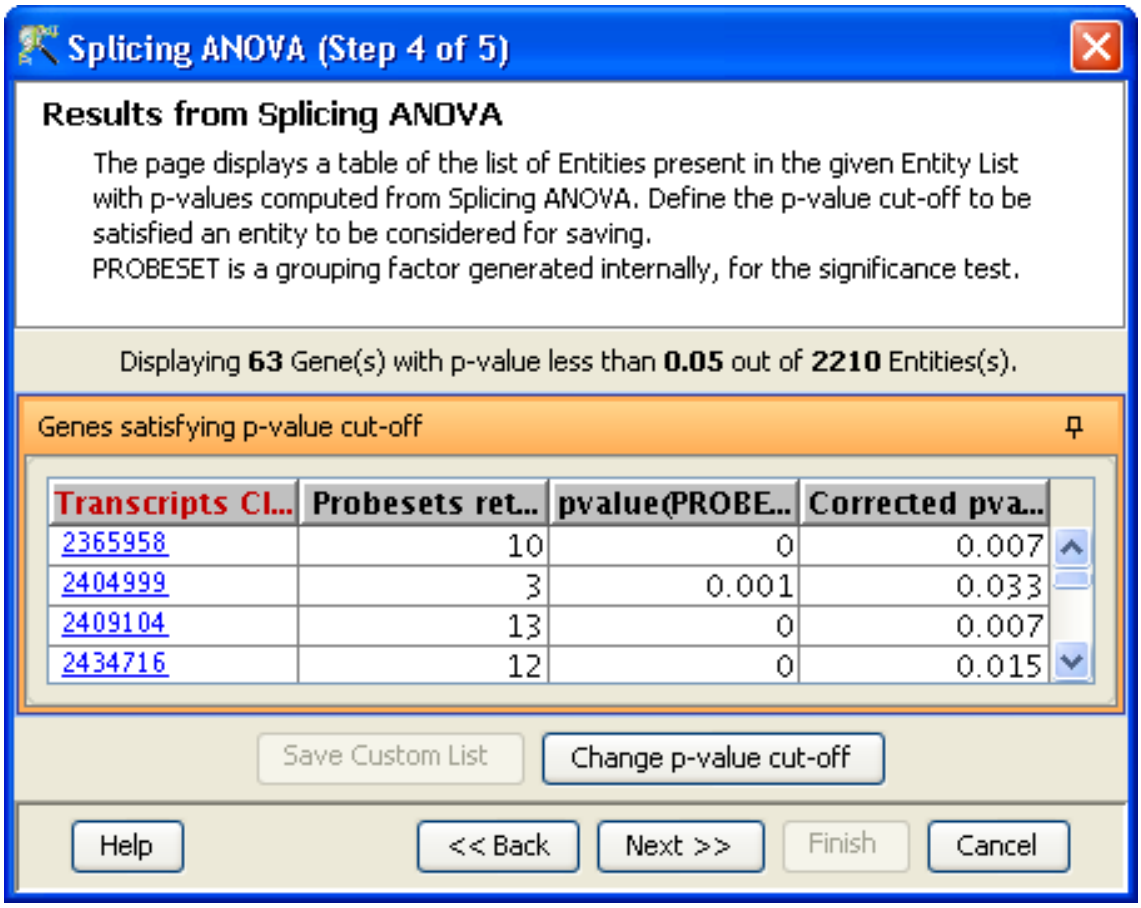


Figure 18.18: Results

– The **S**ave Entity List window shows the details of the entity list that is created as a result of the above analysis. It also shows information regarding creation date, modification date, owner, number of entities, notes etc. of the entity list. Annotations can be configured using Configure Columns button. Selecting Finish results in an entity list being created containing entities which satisfied the cut off. The name of the entity list will be displayed in the experiment navigator. The Entity List generated as a result of Splicing ANOVA has an attachment associated with it. The attachment remembers which probesets were used to perform splicing ANOVA for each transcript. The 'Splicing Visualization' step will use this attachment to show relevance probesets for a transcript in the variance plots. This attachment is also carried over when a custom list is created while performing 'Filter on Splicing Index' or during the 'Splicing Visualization' steps. See Figure 18.19.

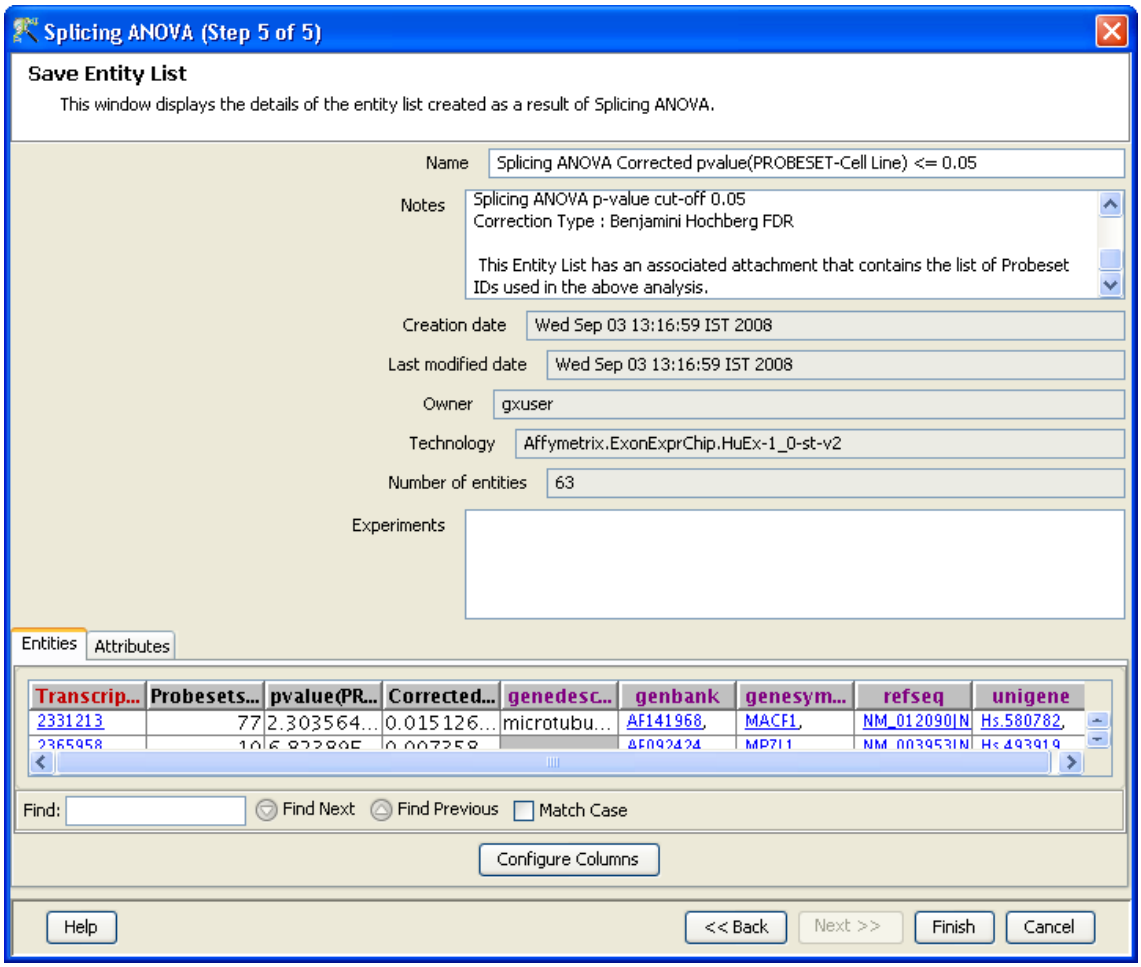


Figure 18.19: Save Entity List

For more details on Splicing ANOVA and the defaults specified in this option, refer to Affymetrix white papers [4, 2]

- **Filter on Splicing Index:**

  Splicing Index is essentially a fold change analysis step wherein difference between the gene normalized signal intensities for 2 conditions are computed as follows:

  – For a given transcript, this difference is computed for each probeset; if any of the probesets has an absolute value difference greater than the specified threshold (0.5 by default) then the transcript will pass this filter.

  – In situations where the interpretation has only 1 condition, the Splicing Index is computed against zero.

  This analysis is executed in four steps:

  1. This step requires the user to provide the entity list and the interpretation. See Figure 18.20.

Figure 18.20: Input Data

2. The second step in the wizard asks the user to select pairing options based on parameters and conditions in the selected interpretation. In case of two or more groups, the user can evaluate either pairwise or with respect to a control. In the latter situation, the condition to be used as control needs to be specified. The order of conditions can also be flipped (in case of pairwise conditions) using an icon. See Figure 18.21.

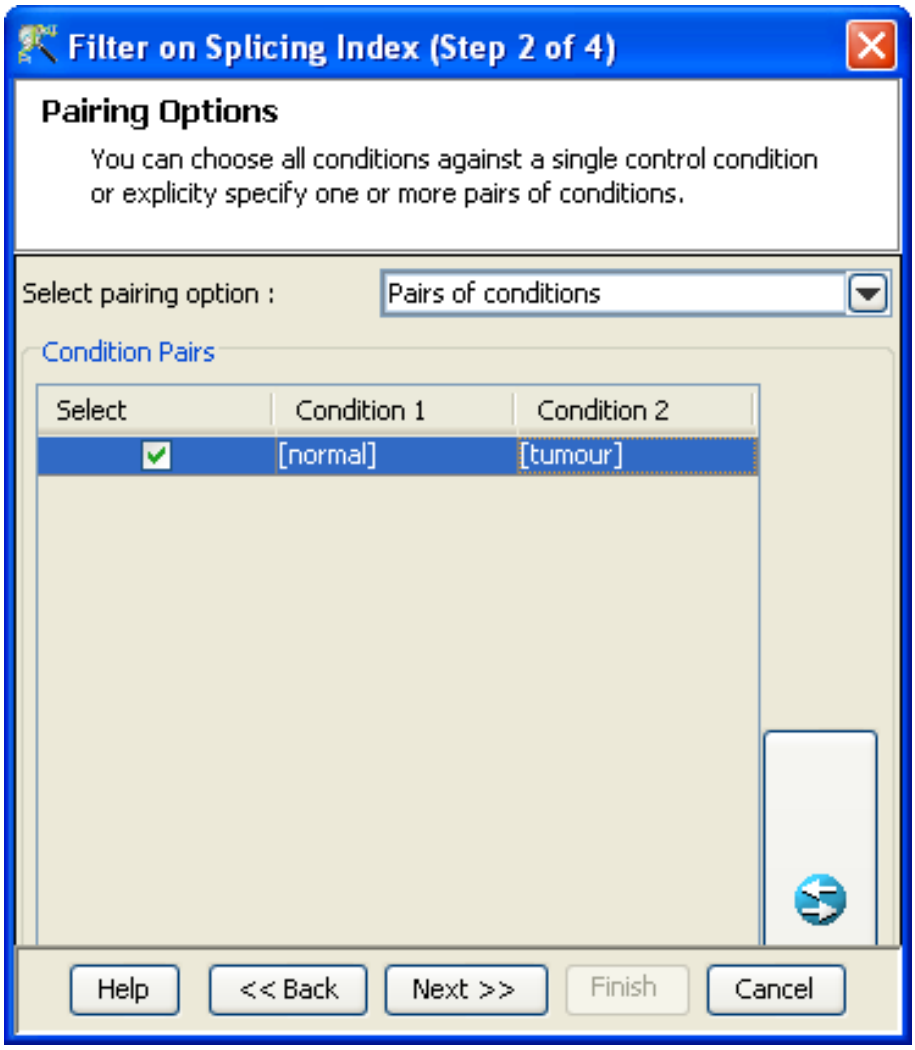


Figure 18.21: Pairing Options

3. This step shows the results of the analysis in the form of a spreadsheet. The transcripts that have passed the cut-off are shown along with the Splicing Index. It also displays the probesets considered for each transcript, for calculating the Splicing Index. The cut-off can be changed using the ***Change Splicing Index cut-off button***. See Figure 18.22. For information on how Cross-hybridization values are derived from the probeset annotation files, please refer to section on Affymetrix Exon Splicing Technology

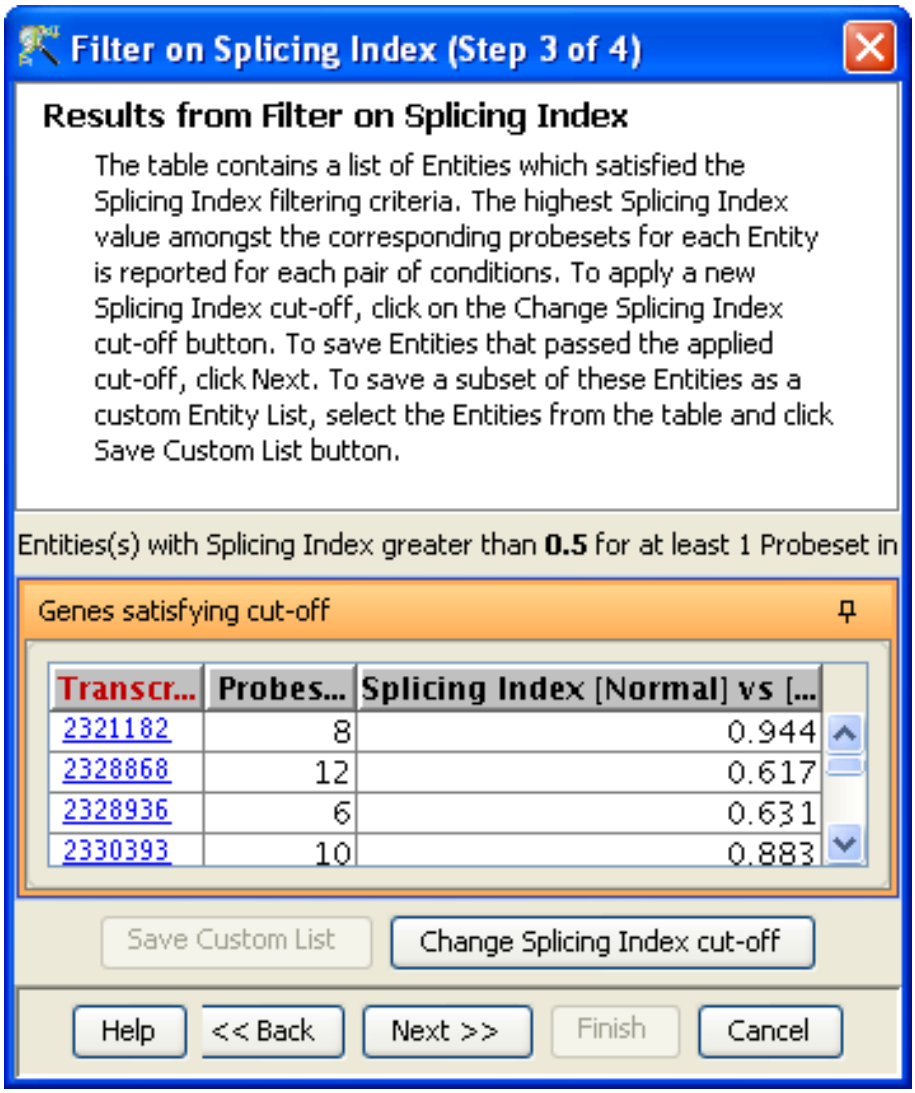


Figure 18.22: Results

4. The last step shows all the entities passing the filter along with their annotations. It also shows the details (regarding creation date, modification date, owner, number of entities, notes etc.) of the entity list. Click ***Finish*** and an entity list will be created corresponding to entities which satisfied the cutoff. Double clicking on an entity in the **Entities table** opens up an **Entity Inspector** giving the annotations corresponding to the selected profile. Additional tabs in the **Entity Inspector** give the raw and the normalized values for that entity. The name of the entity list will be displayed in the experiment navigator. Annotations being displayed here can be configured using ***Configure Columns*** button. See Figure 18.23.

   Please note that when a entity list is generated as a result of Splicing ANOVA is used for Splicing Index analysis, the Splicing Index is computed for those probesets that were retained after filtering. However, if the entities are not put through Splicing ANOVA before running 'Filter on Splicing Index', the Splicing Index is computed for all probesets mapping to the entities in the technology. Therefore, the Splicing Indices may differ.

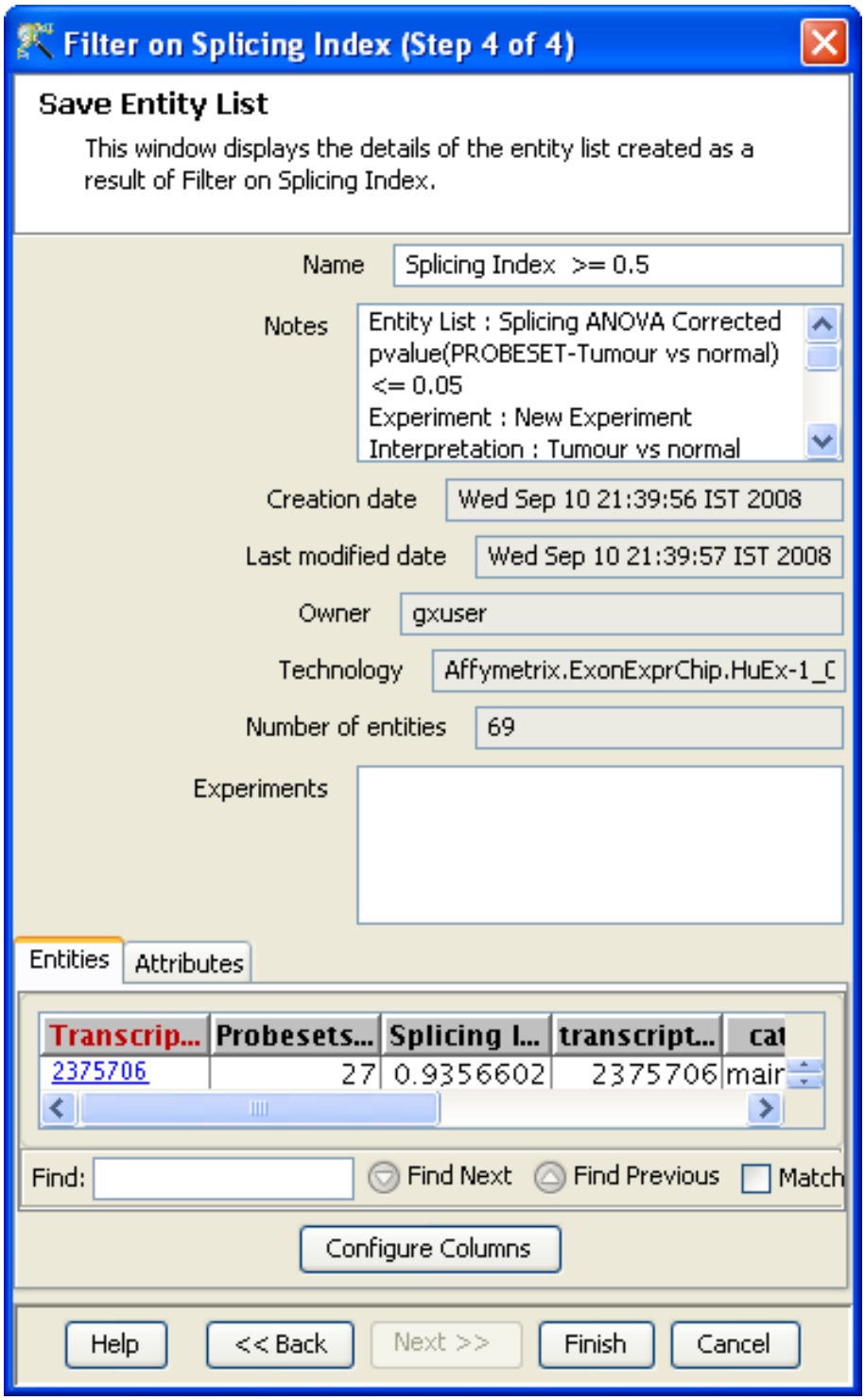


Figure 18.23: Save Entity List

- **Splicing Visualizations:**

  The results of splicing analysis can be viewed as 6 tabs under the 'Splicing Visualization' link in the workflow.

– This step requires the user to provide the entity list and the interpretation. See Figure 18.24.

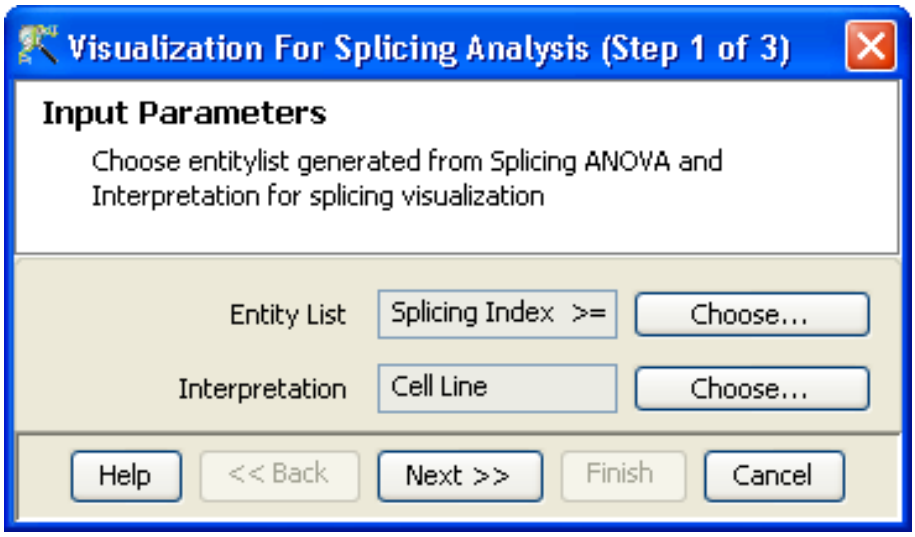


Figure 18.24: Input Data

– The next step presents the data in 6 views. The first three show normalized probeset signals while the last three show gene normalized probeset signals. The legend for the current view is present on the left and a message on the top shows the gene being displayed. The tabs for the views are present at the bottom along with the option of selecting the transcript(gene) to be viewed. Only one transcript can be viewed at a time. Clicking on ***Save Transcript*** adds the transcript in view to a cache, which is then saved as a new entity list when the wizard concludes. See Figure 18.25.

  * **Signal Values**: This displays the normalized intensity values of probesets in the selected transcript. The exons corresponding to the probesets as well as associated annotation information on the probesets such as chromosomal location and level are also given. In case the entity list used is obtained after *Splicing ANOVA*, a column containing information on whether the probeset was filtered out or used for splicing ANOVA is also given.
  * **Probeset Profile Plot**: This shows a profile plot of the probesets in the selected transcript. When run on an output list from Splicing ANOVA or Splicing Index, the grayed out profiles, if any, indicate probesets filtered out in *Splicing ANOVA*. The data used in the plot is the probeset level normalized data.
  * **Probeset Variance Plot**: This shows the variance in the probesets across different conditions. The X-axis contains the probeset ID ordered by chromosomal location and the Y-axis is the mean of the probeset normalized intensity values across samples in a condition. The plot shows each point on the profile as a shape (where the shape determines the exon as described in the legend). The plot also shows error bars for each of the points in the profile, where the error bars indicate the standard error of mean within the corresponding condition (the standard error of mean is defined as the standard deviation divided by the square root of number of samples in the condition minus 1)

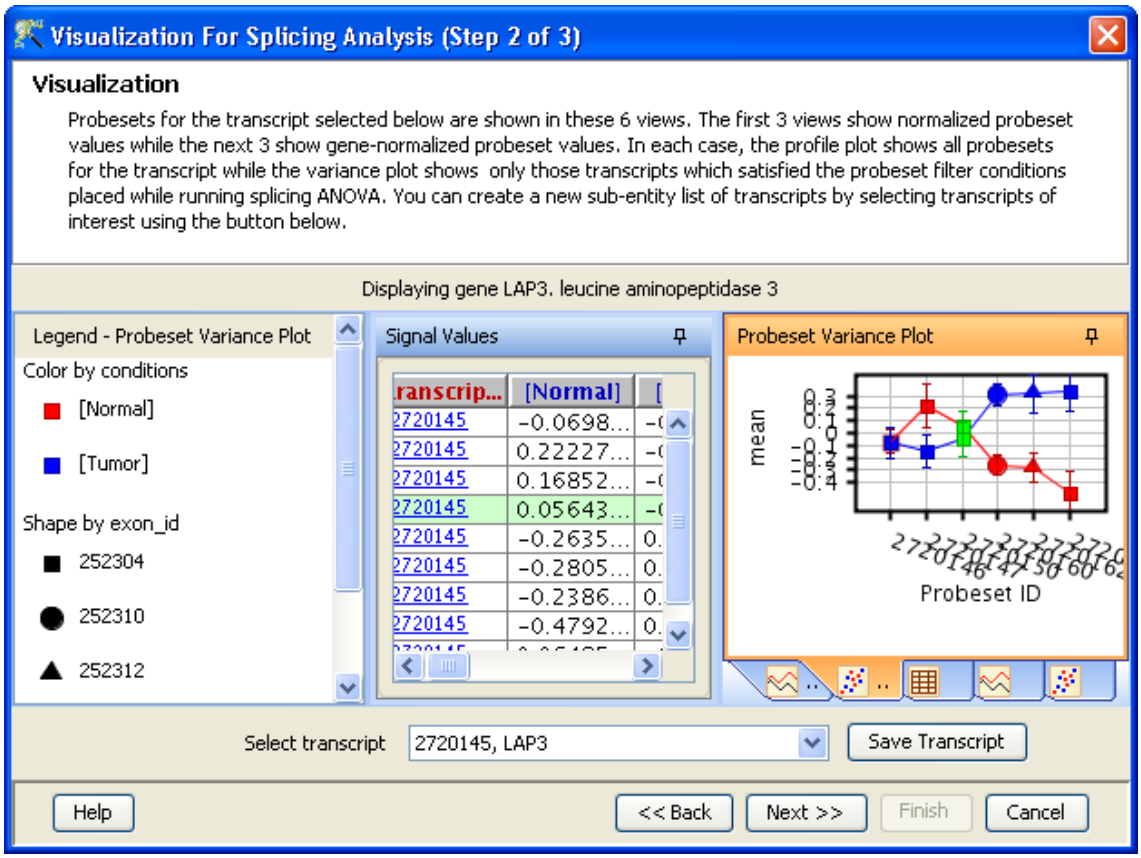


Figure 18.25: Visualization

* **Gene Normalized Signal Values**: This is similar to the *Signal Values* view except that the intensities shown are 'Gene Normalized Signals'. The gene normalized signal refers to the difference between the probeset level normalized signal and the transcript level normalized signal.
* **Gene Normalized Profile Plot**: This shows a profile plot of the probesets in the selected transcript. The greyed out profiles belong to the exons filtered out in *Splicing ANOVA*. The data used in the plot is the gene normalized data.
* **Gene Normalized Variance Plot**: This is similar to the *Probeset Variance Plot* except that the intensity values used are the gene-level normalized intensities. This is often the most useful plot for viewing splicing and therefore also the default view.

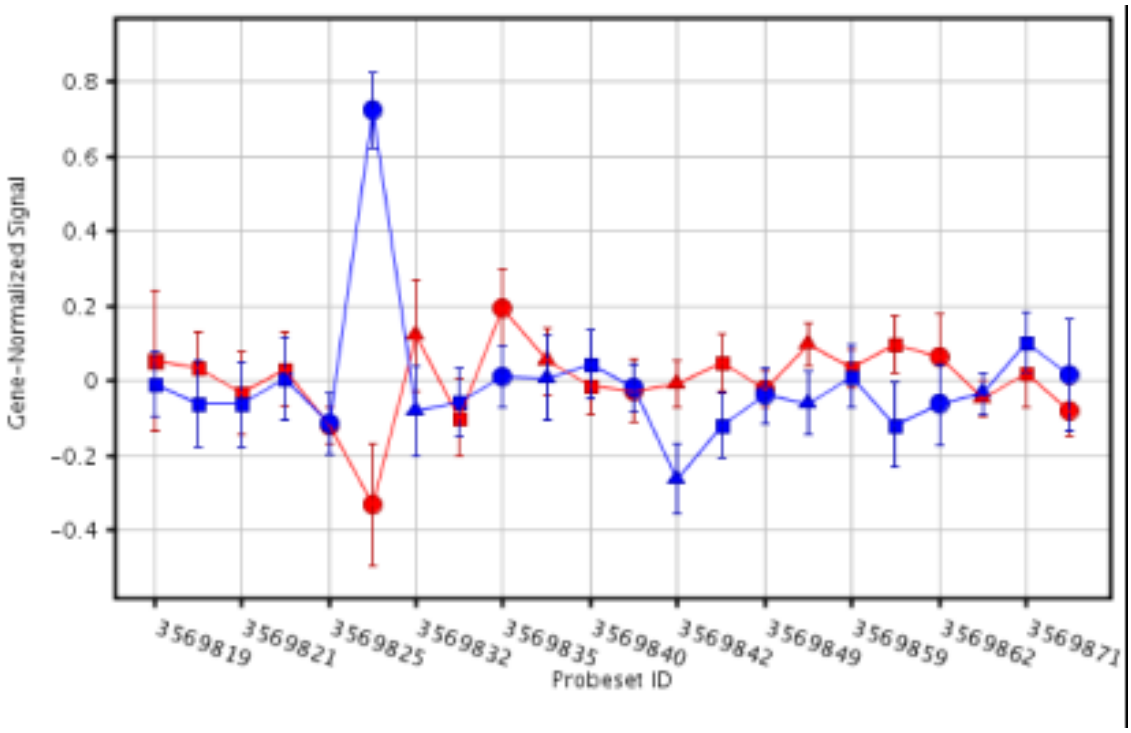


Figure 18.26: Visualization

In case of various probeset IDs corresponding to the same exon in a transcript, they usually have similar values across a given condition. In case there is a significant difference in the expression levels for a particular probeset when compared to others (all of them having the same exon ID), then it could either mean:

1. The expression level of that particular probeset might be erroneous owing to noise or other experimental factors.
2. The particular probeset might actually correspond to a sub-exon within the exon.

– The third step appears only when a transcript is saved. The **S**ave Entity List window shows the details of the entity list that is created as a result of the above analysis. It also shows information regarding creation date, modification date, owner, number of entities, notes etc. of the entity list. Annotations can be configured using ***Configure Columns*** button. Selecting ***Finish*** results in an entity list being created containing selected entities. The name of the entity list will be displayed in the experiment navigator. See Figure 18.27.

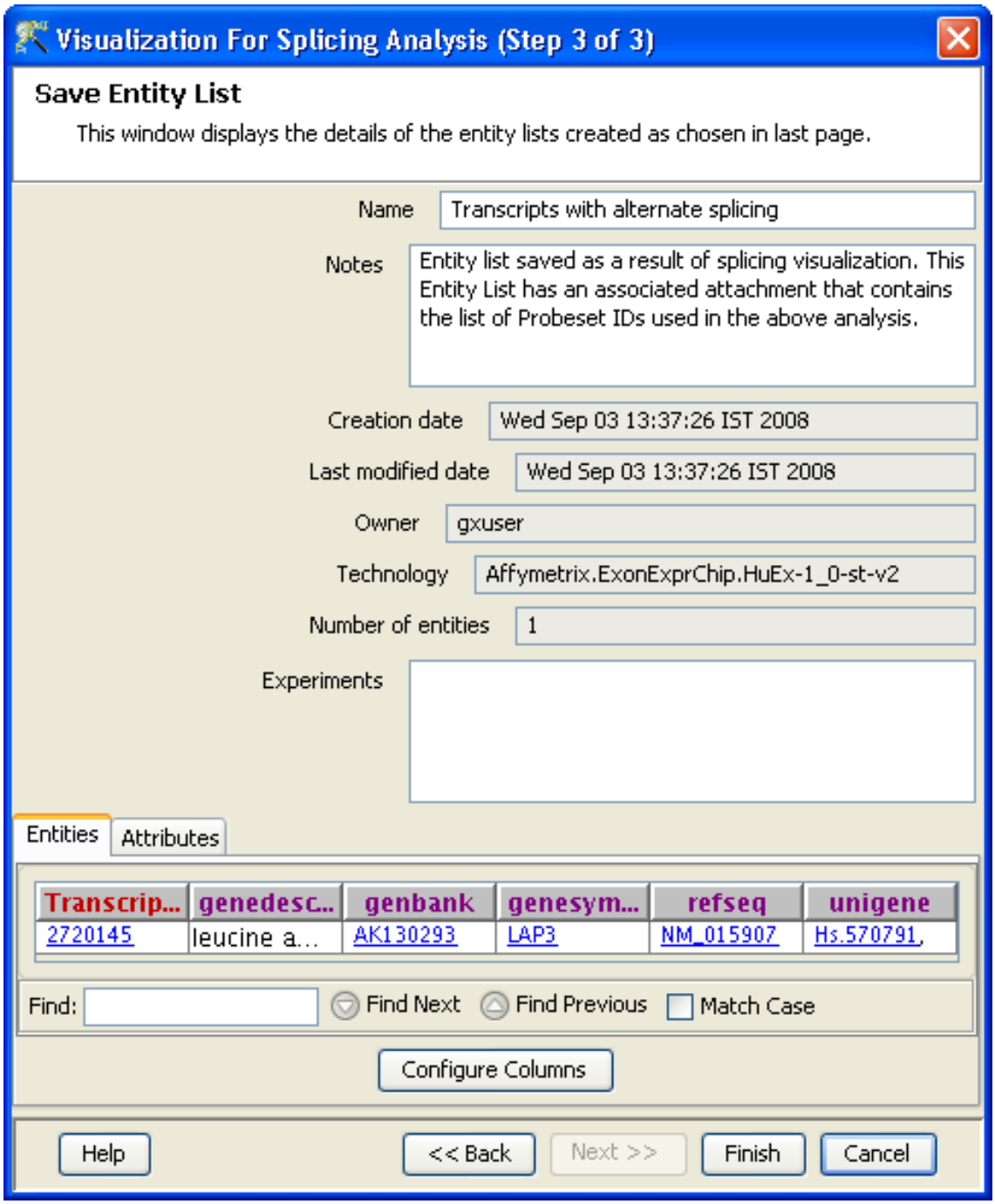


Figure 18.27: Save Entity List

18.1.7 Class Prediction

- **Build Prediction Model** For details refer to section Build Prediction Model.
- **Run Prediction** For details refer to section Run Prediction.

18.1.8 Results

- **Gene Ontology (GO) analysis**
  For more details see GO Analysis.
- **Gene Set Enrichment Analysis (GSEA)**
  For more details see Performing GSEA in **GeneSpring**.
- **Gene Set Analysis (GSA)**
  For more details see Gene Set Analysis.
- **Find Similar Entity Lists**
  This feature is discussed in a separate section called Find Similar Entity Lists.
- **Single Experiment Analysis or Multi-Omic Analysis**
  This feature is discussed in detail in the section Single Experiment and Multi-Omic Analysis.
- **Export to IPA**
  This feature is discussed in detail in the section Ingenuity Pathways Analysis (IPA) Connector.
- **Import IPA Entity List**
  This feature is discussed in detail in the section Import IPA Entity List.
- **NLP Network Discovery**
  This feature is discussed in detail in the section NLP Network Discovery.
- **MeSH Network Builder**
  This feature is discussed in detail in the section MeSH Network Builder.
- **Extract Relations via NLP**
  This feature is discussed in detail in the section Extract Relations via NLP.

18.1.9 Utilities

- **Import Entity list from File**
  For details refer to section Import Entity list from File.
- **Analysis: Biological Significance**
  For details refer to section Analysis: Biological Significance.
- **Filter On Entity List**
  For further details refer to section Filter On Entity List.
- **Remove Entities with missing signal values**
  For details refer to section Remove Entities with missing signal values.

18.1.10 Algorithm Technical Details

Here are some technical details of the Exon RMA16, Exon PLIER16, and Exon IterPLIER16 algorithms.

Exon RMA 16. Exon RMA does RMA background correction followed by Quantile normalization followed by a Median Polish probe summarization, followed by a Variance Stabilization of 16. An option for GCBG background correction is available from *Tools* $\longrightarrow$ *Options* $\longrightarrow$ *Affymetrix Exon Summarization Algorithms* $\longrightarrow$ *Exon RMA*. GCBG background correction bins background probes into 25 categories based on their GC value and corrects each PM by the median background value in its GC bin. Only antigenomic probes are used by default for GCBG calculation.

Exon PLIER 16. Exon PLIER does Quantile normalization followed by the PLIER summarization using the PM or the PM-GCBG options (the latter is default), followed by a Variance Stabilization of 16. The PLIER implementation and default parameters are those used in the Affymetrix Exact 1.2 package. PLIER parameters can be configured from *Tools* $\longrightarrow$ *Options* $\longrightarrow$ *Affymetrix Exon Summarization Algorithms* $\longrightarrow$ *Exon PLIER/IterPLIER.*

Exon IterPLIER 16. Exon IterPLIER does Quantile normalization followed by the IterPLIER summarization using the PM or the PM-GCBG options (the latter is default), followed by a Variance Stabilization of 16. IterPLIER runs PLIER multiple times, each time with a smaller subset of the probes obtained by removing outliers from the previous PLIER run. IterPLIER parameters can be configured from *Tools* $\longrightarrow$ *Options* $\longrightarrow$ *Affymetrix Exon Summarization Algorithms* $\longrightarrow$ *Exon PLIER/IterPLIER.*

Note:

- By default, only anti-genomic probes are used for background correction for RMA 16, PLIER 16 and IterPLIER 16. This can be changed by the user by going to Tools$\longrightarrow$ Options$\longrightarrow$ Affymetrix Exon Summarization Algorithms. The choice made for background probes here is applicable for the DABG p-value calculation as well.
- When RMA 16 is chosen as the transcript level summarization algorithm, the same algorithm will also be used for exon-level summarization. If PLIER 16 or IterPLIER 16 is chosen for transcript level summarization, then PLIER 16 is used for exon-level summarization.

18.2 Affymetrix Exon Splicing Technology

For each probeset, NetAffx lists the probe sequences and the consensus sequence interrogated by the probes.

18.2.1 Inputs for Technology Creation

Mainly, Affymetrix technology creation involves 2 steps:

- Download files from NetAffx: Affymetrix Exon Technologies are created with updated annotation files downloaded from NetAffx.
- Create technology from downloaded files.

Files of extention '.ps' and '.mps' are not used from Affymetrix, but are created internally by extracting the information in these columns from the 'probeset_annot' file. Extracted columns are: probeset_id, probe_count, transcript_cluster_id, crosshyb_type, level, probeset_type. Technology Id format is Affymetrix.ExonExprChip.ChipName_net-affx build_GenomeBuild_Probeset Annotation file release date. Eg: Affymetrix.ExonExprChip.HuGene-1_0-st-v1_na30_1_hg19_2010-01-22

Cross-hybridization potential of the probeset is determined at the time of array design. This information is obtained from 'crosshyb_type' column that is present in '_probeset_annot.csv.zip' file. This column is extracted and modified. This column contains values in range 1, 2, 3, 0, "---". The value of 1, indicates all unique probes in the probeset that perfectly match only one sequence in the putatively transcribed array design content. The value of 2, indicates 'similar' probes that match more than one sequence and 3 indicates partial match to more than one sequence. If any missing value or "---" or "0" is present, then the value is modified to 0; typically, this happens for control probesets, which don't have transcript cluster ids.

18.2.2 Technology creation

Briefly, the technology creation involves the following steps :

1. Technology name creation
2. Library creation
3. Annotation Library creation
4. Data Library creation
5. Upload to update server

Please note that **GeneSpring** has removed Experiment Creation support for older Affymetrix Exon technologies, i.e, those that were created prior to GX11.5 release. Add/Remove samples will not work for experiments created with these legacy technologies.

18.3 Tutorial for Exon Splicing Analysis

GeneSpring provides a unique analysis tool for analyzing Affymetrix exon chip to study exon splicing. The following tutorial describes the steps in Exon Splicing Analysis using the tool. The dataset used in the tutorial can be downloaded from http://www.affymetrix.com/support/technical/sample_data/exon_array_data.affx. Using the tutorial, the splicing events described in the paper "Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array" by Turpaz *et al.* , 2006 can be observed.

The tutorial will not go into the details of the experiment creation as it has been described in detail above.

1. **Experiment Creation:**

   - Create a new experiment with experiment type 'Affymetrix Exon Splicing'
   - Choose 'PLIER 16' as the Summarization algorithm in Step 4.
   - Choose 'core' for both the transcript and probeset level and select the 'Baseline to median of all samples' option.
   - Click ***Finish***.

   Experiment creation will now commence. Experiment creation includes background correction of all probe sets using the DABG algorithm.

2. **Experiment Grouping:**

   - Group your data into 2 groups, Normal and Tumor, using 'Experiment Grouping' in the Workflow. All files whose filename ends with '_N.cel' are healthy tissue files, whereas the ones with the suffix '_T.cel' are tumor tissue files.
   - Create an Interpretation with these 2 conditions.

3. **Exon Splicing Analysis:**

   There are 4 stages in Exon Splicing Analysis.

   - **Filter Transcripts on DABG:**
     DABG algorithm is executed on all the probesets at the time of experiment creation. The p-values generated as a result of DABG is used at this step to filter out transcripts before proceeding for ANOVA.
     (a) Click on **Filter Transcripts on DABG** in the Workflow.
     (b) Choose 'All Entities' as the entity list and 'Tumor vs Normal' as the interpretation.
     (c) Use the default parameters in the Step 2.
     (d) In Step 3, the filtering results will be displayed as a Profile Plot, showing the profiles of the transcript clusters that passed the filter criteria.
     (e) Continue on to Step 4 and click ***Finish***. A new entity list named 'Filtered on DABG p(probeset) ≤ 0.05' will appear in the Analysis folder in the Project Navigator.
   - **Splicing ANOVA:**
     Among the transcripts identified, the probesets can be eliminated or retained for Splicing ANOVA based upon the DABG generated p-values. Additional filtering can also be performed at this stage to eliminate probes with high background and cross-hybridization potential.

(a) Click on **Splicing ANOVA** in the Workflow.

(b) Choose 'Filtered on DABG p(probeset) ≤ 0.05' as the entity list, and 'Tumor vs Normal' as the interpretation.

(c) Retain the defaults provided in steps 2 and 3. The transcripts that have a p-value ≤ 0.05 after running the Splicing ANOVA test will be shown.

(d) Continue through the steps to save this list as an entity list named 'Splicing ANOVA corrected pvalue(PROBESET-tumor vs normal)'.

- **Filter on Splicing Index:**

  Splicing Index is defined as the difference between the gene normalized signal intensities of the probesets for the normal and tumor samples. For each transcript, this fold change value is computed for all the probesets that have passed the splicing ANOVA.

  (a) Click on **Filter on Splicing Index** in the Workflow.

  (b) Choose 'Splicing ANOVA corrected pvalue(PROBESET-tumor vs normal)' as the entity list , and 'Tumor vs Normal' as the interpretation.

  (c) Continue through the process and save the entity list named 'Splicing Index ≥ 0.5'.

- **Splicing Visualizations:**

  (a) To visualize the results of this analysis, click on the final step in this section, **Splicing Visualizations**.

  (b) Choose 'Splicing Index ≥ 0.5' as the entity list, and 'Tumor vs Normal' as the interpretation.

  (c) A visualization results window containing 6 tabs, opens up. 3 of the tabs contain views of the gene normalized data, whereas the other 3 show the raw data.

  (d) Click on the 'Gene normalized variance plot' tab. This plot will be most useful in finding exons that vary between the 2 experiment conditions (Normal and Tumor). To view plots for different transcripts, select the particular transcript cluster ID in the drop down box at the bottom of the visualization panel.

  In case of colon cancer, splicing occurs in transcripts involved in cytoskeletal organization, ACTN1 being one of them. The Gene Normalized Variance Plot for the ACTN1 transcript shows a clear variance for the 2 conditions for Exon 3569830, whereas the other exons are fairly invariant between the 2 conditions (Figure 18.28). This indicates that this exon is spliced out in one of the conditions.

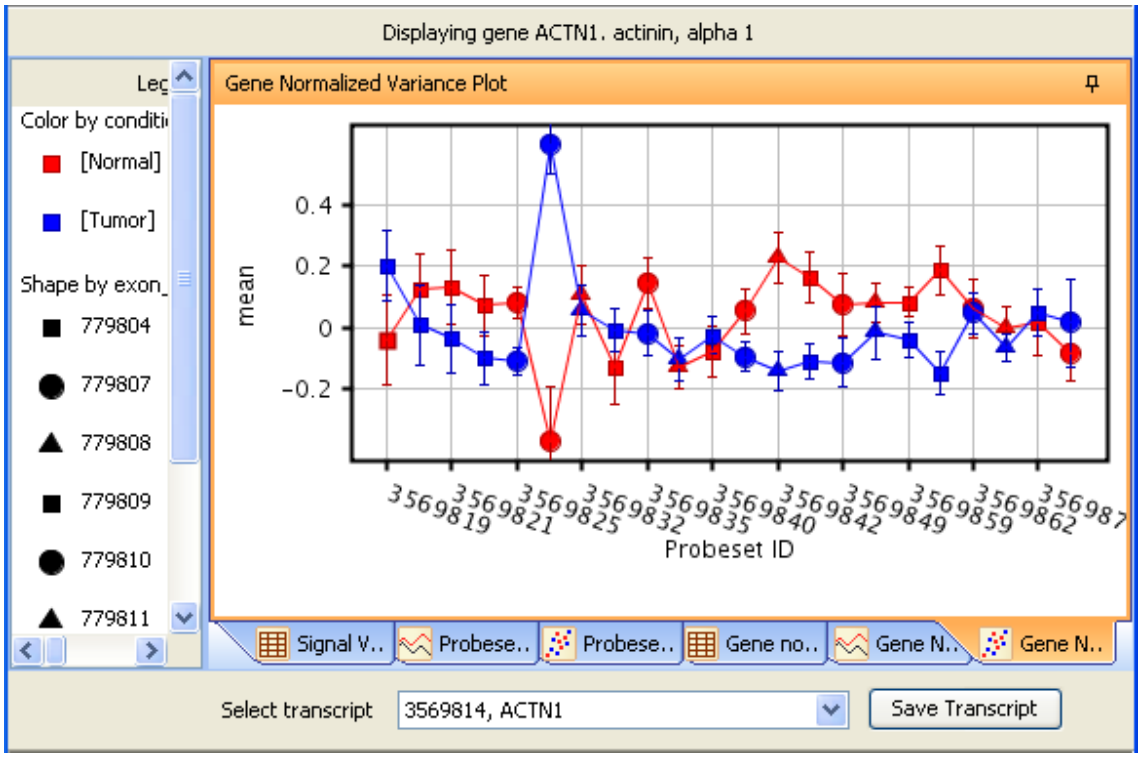


Figure 18.28: Gene Normalized Variance Plot

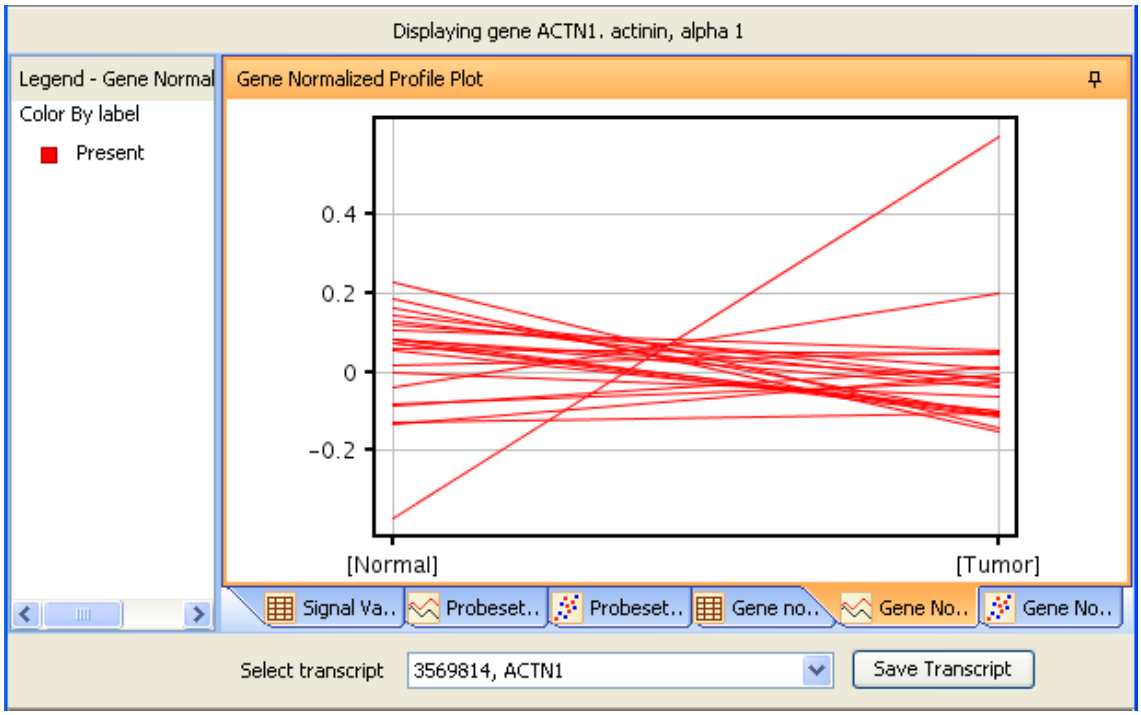


Figure 18.29: Gene Normalized Profile Plot

Click on the tab immediately preceding this, i.e., the Gene Normalized Profile Plot. This plot shows the exons (if any) which were filtered out, in gray ((Figure 18.29).

To conclude the exon splicing analysis, select each transcript you wish to save in a list, one by one in the 'Select transcript' drop down box, and click on ***Save transcript*** to save it. All transcripts thus chosen will appear in a new entity list 'transcripts with alternative splicing' in the Project Navigator.

Chapter 19

Analyzing Real Time PCR Data

Real Time PCR also called Quantitative Real-Time PCR (qPCR) is used to rapidly measure the quantity of DNA, cDNA, or RNA present in a sample. It is the most sensitive technique for mRNA detection and quantitation currently available. Compared to the two other commonly used techniques for quantifying mRNA levels, Northern blot analysis and RNase protection assay, RT-PCR can be used to quantify mRNA levels from much smaller samples. In fact, this technique is sensitive enough to enable quantitation of RNA from a single cell.

GeneSpring supports all version of the ABI's 7900HT RT-PCR system. The columns that are imported into **GeneSpring** from the original data file are the Sample, Detector, Task and the Avg Ct or Ct Avg column. In addition, the tool also creates a Gene symbol and a synonyms (of the GeneSymbol) column.

19.1 Running the Real Time PCR Workflow

Upon launching **GeneSpring**, the startup is displayed with 3 options.

1. Create new project
2. Open existing project
3. Open recent project

Either a new project can be created or else a previously generated project can be opened and re-analyzed. On selecting ***Create new project***, a window appears in which details (Name of the project and Notes) can be recorded. Press ***OK*** to proceed. An Experiment Selection Dialog window then appears with two options:

1. Create new experiment

2. Open existing experiment

Selecting ***Create new experiment*** allows the user to create a new experiment (steps described below). ***Open existing experiment*** allows the user to use existing experiments from any previous projects in the current project. Choosing ***Create new experiment*** opens up a **New Experiment dialog** in which experiment name can be assigned. The analysis type should then be specified as *RTPCR*.

Upon clicking ***OK***, the Real Time PCR experiment creation wizard appears. This wizard requires details such as name of the technology, organism under study and the sample files for experiment creation. See Figure 19.1.

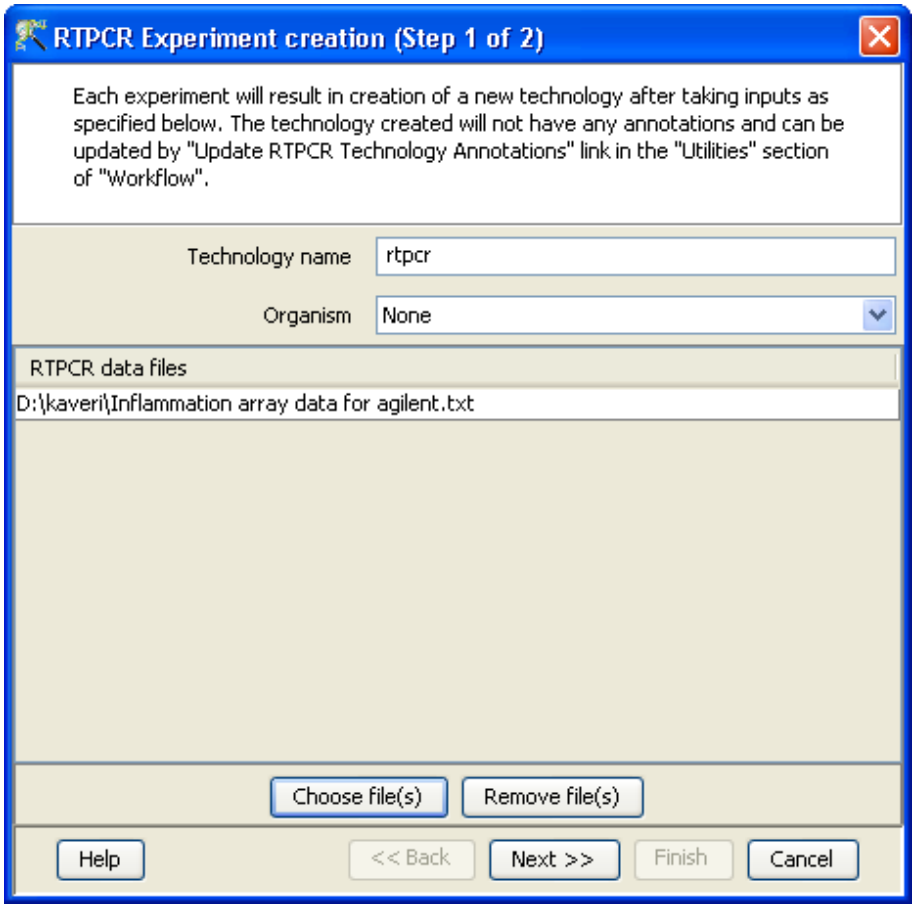


Figure 19.1: Experiment Creation

The next step allows the user to perform baseline transformation. See Figure 19.2. Baseline Transformation is carried out row-wise across all samples. This data processing step is particularly useful when visualizing the results in a profile plot or heat map. The baseline transformation options, available in **GeneSpring** are:

- ***Do not perform baseline***

- ***Baseline to median of all samples:*** For each row (probe), the median of the log summarized values across all the samples is calculated. This value is then subtracted from the probe value for all samples.

- ***Baseline to median of control samples:*** Here control samples are used to calculate the median value, for each probe. This value is then subtracted from the probe value for all samples. The controls could be an individual control for each sample or it could be a set of controls. Alternatively, a set of samples can be used as controls for all samples. For specifying the control for a sample, select the sample and click on ***Assign value***. This opens up the ***Choose Control Samples*** window from where the samples designated as *Controls* should be moved from the *Available Items* box to the *Selected Items* box. Click on ***Ok***. This will show the control samples for each of the samples.

  In *Baseline to median of control samples*, for each probe the median of the log summarized values from the control samples is first computed and then this is subtracted from the sample. If a single sample is chosen as the control sample, then the probe values of the control sample are subtracted from its corresponding sample.

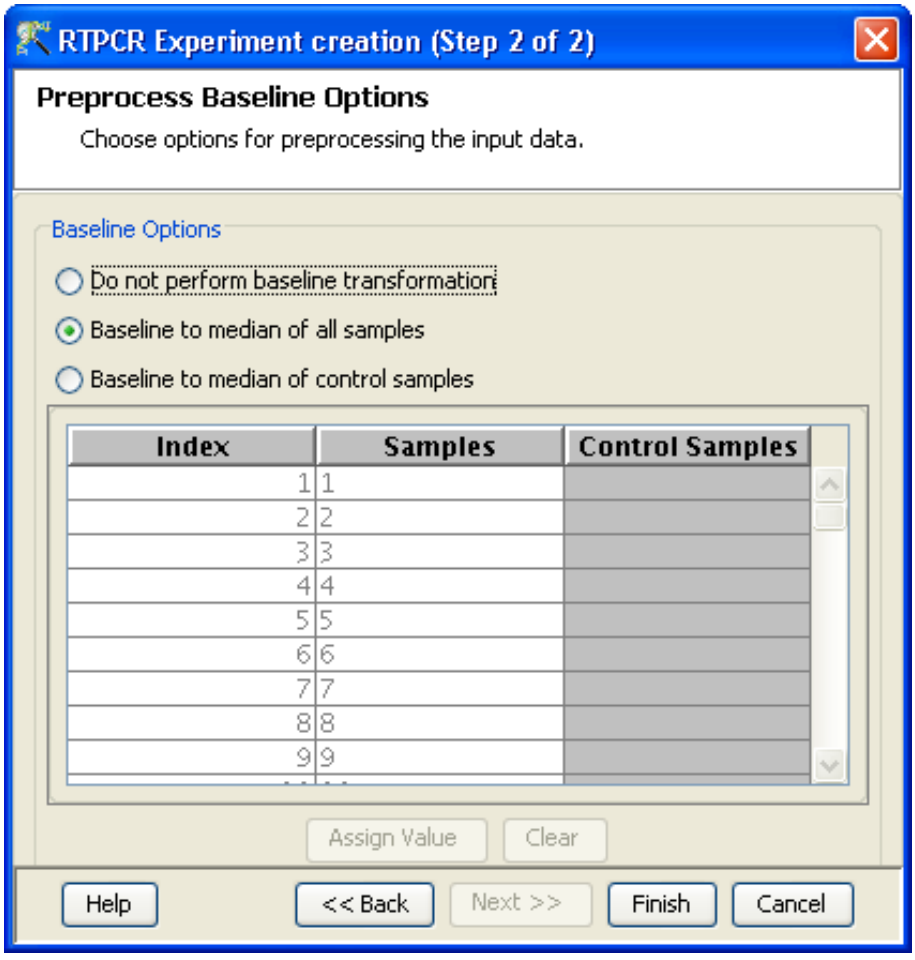


Figure 19.2: Baseline Transformation Options

Clicking ***Finish*** creates an experiment, which is displayed as a Box Whisker plot in the active view. Alternative views can be chosen for display by navigating to ***View*** in Toolbar.

19.1.1 Technology Creation in RT-PCR experiments

For each experiment, a new Technology is created within **GeneSpring**, which is unique and cannot be used in another experiment. The Technology name is in the format ABI.RTPCR.RQ(version)-(name) where version corresponds to the supported version of the RQ Manager software and name corresponds to what is provided by the user during the experiment creation process. **GeneSpring** can detect the following Annotations - Detector, Assay ID, Gene Symbol and Synonyms for the Technology out of the supported RQ data files based on the following guidelines :

All the above mentioned Annotations are derived from the Detector column and the following format is assumed

Synonym(Gene Symbol)-Assay ID: Please note that only the Assay ID is mandatory in the above format.

Here are a few examples

- If the user wants to bring in the Gene Symbol GS1 with an Assay ID Assay1 then the Detector column should be specified as: (GS1)-Assay1
- If the user wants to bring in the Synonym S1, Gene Symbol GS1 with an Assay ID Assay1 then the Detector column should be specified as: S1(GS1)-Assay1
- For example, if the value in the Detector column is ATIR(AGTRI)-HS00241341_m1, then AGTRI will be the Gene Symbol, ATIR will be the synonym and HS00241341 will be the Assay ID

19.1.2 Data Processing

1. **File formats:** The files should be in text (.txt) format.

   The columns that are imported into GeneSpring from the original data file are the

   (a) Sample,

   (b) Detector,

   (c) Task and

   (d) Avg Ct/Ct Avg column

   Note: The column names need to be exactly the same as the ones given above. If there are columns that are analogous to the above ones, but have a different name, they need to be changed in the sample file before experiment creation.

2. **Raw:** The term ''raw'' signal values refer to the data after averaging **Avg Ct** or **Ct Avg** column within a sample (summarization).

3. **Normalized:** The term **Normalized** signal value refers to a difference between the summarized Averaged counts of Endogenous controls and the target within a sample. It also reflects the baseline transformation performed.

4. **Treatment of on-chip replicates:** Replicates of a target are averaged to compute their total intensity values as described above.

5. **Flag values:** Not applicable.

6. **Treatment of Control probes:** The control probes that are taken into account are the endogenous control probes which are identified by the "Task" column in the original data file.

7. **Empty Cells:** Empty cells might be present in the intensity values column for certain genes in the data file. These genes are brought in **GeneSpring** . These can be removed from the entity lists during analysis from ***Utilities*** $\longrightarrow$ ***Remove Entities with missing signal values***.

8. **Sequence of events:** The sequence of events involved in the processing of the data files is: summarization, normalization and baseline transformation. No log transformation is performed as the data is already in log scale.

19.1.3 Experiment Setup

Once an experiment is created, the **Data Import Wizard** steps appear on the right hand side. Following is an explanation of the various workflow links

- **Quick Start Guide:** Clicking on this link will take you to the appropriate chapter in the on-line manual giving details of loading files into **GeneSpring**, the **Data Import Wizard**, the method of analysis, the details of the algorithms used and the interpretation of results
- **Experiment Grouping: Experiment Parameters** define the grouping or the replicate structure of the experiment. For details refer to the section on Experiment Grouping
- **Create Interpretation:** An interpretation specifies how the samples would be grouped into experimental conditions for display and used for analysis. For details refer to the section on Create Interpretation

19.1.4 Quality Control

- **Quality Control on Samples:**

  Quality Control (QC) allows you decide which samples are ambiguous and which are passing the quality criteria. Based on the QC results, unreliable you can remove samples from further downstream analysis. The QC dialog displays three tiled windows:

  – 3D PCA scores, Correlation Analysis - Samples, and a Correlation Coefficients tab.
  – Experiment Grouping.
  – Legend.

  Figure 19.3 illustrates the tiled windows for QC on samples.

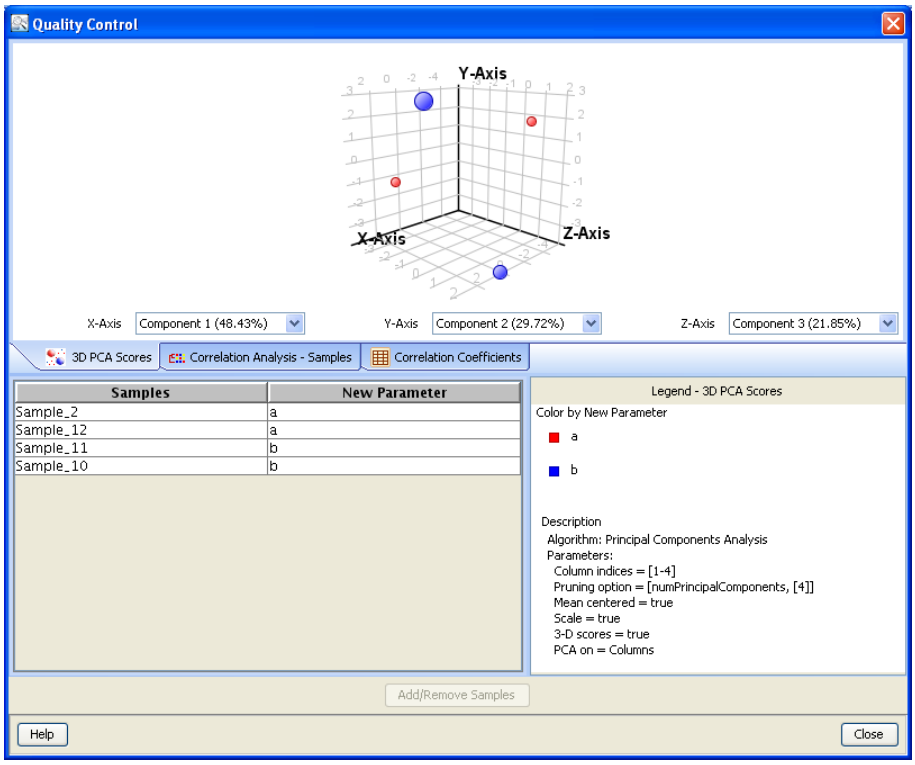


Figure 19.3: Quality Control

In the **Principal Component Analysis (PCA)** tab, PCA scores are visually represented in a 3D scatter plot. These scores are used to check data quality. It shows one point per array and is colored by the *Experiment Factors* provided earlier in the *Experiment Groupings* view. This allows viewing of separations between groups of replicates. Ideally, replicates within a group should cluster together and separately from arrays in other groups. The PCA components, represented in the X, Y, and Z axes are numbered 1, 2, 3... according to their decreasing significance. The 3D PCA scores plot can be customized via ***Right-Click*⟶*Properties***. To zoom into a 3D Scatter plot, press the Shift key and simultaneously hold down the left mouse button and move the mouse upwards. To zoom out, move the mouse downwards instead. To rotate, press the Ctrl key, simultaneously hold down the left mouse button and move the mouse around the plot.

The **Correlation Analysis - Samples** tab shows the correlation analysis heatmap across arrays (samples) (See How to Perform Correlation Analysis on Samples on page 750). The correlation coefficient for each pair of arrays are displayed in the **Correlation Coefficients** tab. Correlation coefficients are calculated using Pearson Similarity Metric (see Pearson on page 799).

Pearson Similarity Metric: Calculates the mean of all elements in vector **a**. Then it subtracts that value from each element in **a** and calls the resulting vector **A**. It does the same for **b** to make a vector **B**. Correlation Coefficient is then computed as **A.B**/(—**A**——**B**—).

In a Correlation Analysis heatmap:

- Sample Order- In a correlation heatmap, the initial order of samples is the same as the input sample order at the time of experiment creation.
- Clustering - Perform clustering analysis on the correlation heatmap using the **Run clustering** icon from the correlation analysis heatmap toolbar (See To run clustering

on a correlation heatmap on page 760). Samples are clustered based on correlation coefficients using hierarchical clustering algorithm and plotted in the order they cluster together. The linkage rule used is average and the distance metrics used is Euclidean.

For more information on the available toolbar icons in a correlation analysis heatmap, see Correlation Heatmap View Toolbar on page 776. See Properties on 792 to change the properties of the correlation analysis heatmap.

NOTE Correlation coefficients are computed on raw, non-normalized data and in linear scale.

Experiment Grouping displays the parameters and parameter values for each sample.

The third window shows the legend of the active QC tab.

NOTE Samples cannot be removed after QC. To exclude samples from analysis, a new interpretation should be created, containing only the samples of interest.

- **Filter Probe Set by Expression:** Entities are filtered based on their signal intensity values. For details refer to the section on Filter Probesets by Expression.
- **Filter Probe Set by Flags:** This is not applicable as flags are not created in this experiment type.
- **Filter Probesets on Data Files:** Entities can be filtered based on values in a specific column of the original data files. For details refer to the section on Filter Probesets on Data Files.
- **Filter Probesets by Error:** Entities can be filtered based on the standard deviation or coefficient of variation using this option. For details refer to the section on Filter Probesets by Error.

19.1.5 Analysis

- **Statistical Analysis**

  For details refer to section Statistical Analysis in the Data Import Wizard section.
- **Filter on Volcano Plot**

  For details refer to section Filter on Volcano Plot.
- **Fold Change**

  For details refer to section Fold change.

- **Clustering**
  For details refer to chapter Clustering Analysis.
- **Correlation**
  For details refer to chapter Correlation Analysis.
- **Find Similar Entities**
  For details refer to section Find similar entities.
- **Filter on Parameters**
  For details refer to section Filter on Parameters.
- **Principal Component Analysis**
  For details refer to section PCA.

19.1.6 Class Prediction

- **Build Prediction Model** For details refer to section Build Prediction Model.
- **Run Prediction** For details refer to section Run Prediction.

19.1.7 Results

- **Gene Ontology (GO) analysis**
  For more details see GO Analysis.
- **Gene Set Enrichment Analysis (GSEA)**
  For more details see Performing GSEA in **GeneSpring**.
- **Gene Set Analysis (GSA)**
  For more details see Gene Set Analysis.
- **Find Similar Entity Lists**
  This feature is discussed in a separate section called Find Similar Entity Lists.
- **Single Experiment Analysis or Multi-Omic Analysis**
  This feature is discussed in detail in the section Single Experiment and Multi-Omic Analysis.
- **Export to IPA**
  This feature is discussed in detail in the section Ingenuity Pathways Analysis (IPA) Connector.
- **Import IPA Entity List**
  This feature is discussed in detail in the section Import IPA Entity List.
- **NLP Network Discovery**
  This feature is discussed in detail in the section NLP Network Discovery.
- **MeSH Network Builder**
  This feature is discussed in detail in the section MeSH Network Builder.
- **Extract Relations via NLP**
  This feature is discussed in detail in the section Extract Relations via NLP.

19.1.8 Utilities

19.1.8.1 Update RT-PCR Technology Annotations

This option enables the user to update the annotations of the created technology. Selecting this option,opens up a wizard having two steps:

1. **Step 1 of 2** - Here the user specifies the technology as well as the source from which it has to be updated. See Figure 19.4. The technology can be updated either from a file or from the **Biological Genome** of that organism. If the **Biological Genome** of that organism does not exist, then the user can create a genome from ***Annotations⟶ Create Biological Genome***. For more details on the creation of a genome, refer to Biological Genome. Please note that to be able to utilize this feature, the technology should have a column in common with the **Biological Genome**. If the user chooses to update from a file, then it should be chosen accordingly via the ***Choose file*** option. The file from which the update is to be performed has to be in a tabular format.

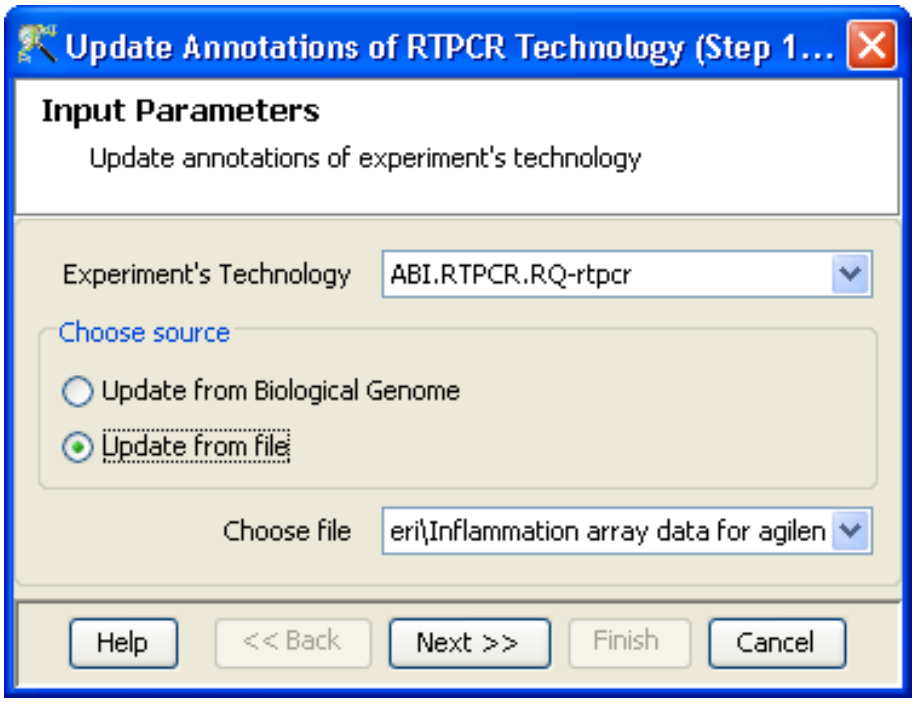


Figure 19.4: Input Parameters

2. **Step 2 of 2** - The annotation columns are merged with the existing technology using a technology identifier. This step asks the user to specify the identifier and to choose the column to be updated from the annotation file/genome. While specifying the columns, column marks should be assigned. See Figure 19.5. It is recommended that the user chooses a column with unique values (for e.g : Entrez-ID) as the identifier. Three kinds of updates are possible:
   - Append to the existing information,
   - Overwrite

- Fill in the missing values.

Appending the values will retain the original value as well as add the new value. Overwrite will replace the original value with the newer one, whereas fill in missing values will add values at places where previously there were none.

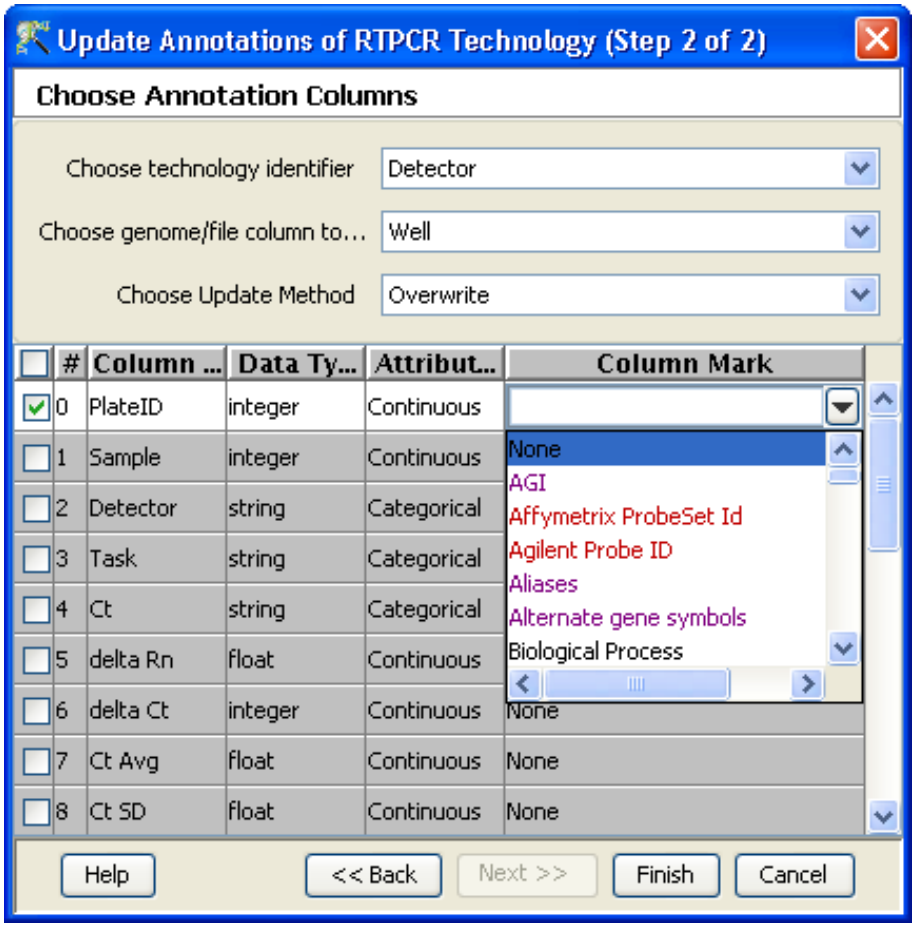


Figure 19.5: Choose Annotation Columns

Alternatively, the technology can also be updated from ***Annotations⟶ Update Technology Annotations⟶ From file or Biological Genome***. More details on this feature can be obtained here 3.2.2.

See Chapter Utilities for more options.

Chapter 20

Multi-Omic Analysis in GeneSpring

Multi-Omic Analysis was introduced to **GeneSpring** in the form of multi-omic pathway analysis, allowing you to select entity lists from two different experiment types and query curated/canonical pathways as well as NLP networks for any overlap (see What is Multi-Omic Pathway Analysis? on page 896). This was enabled by the implementation of extensive mapping of identifiers across technologies as well as organisms using BridgeDb, Homologene and custom ID Mapper databases (see BridgeDb - ID Mapping on page 1012).

As the importance of integrating results from the different 'omic' disciplines has grown in the field, we have also sought to enhance **GeneSpring**'s capabilities to match the demand for cross-technology analysis. We have therefore expanded the concept of the Multi-Omic Analysis (MOA) experiment introduced with multi-omic pathway analysis and opened it up to also perform the latest type of analysis added to **GeneSpring** - Correlation Analysis - in a multi-omic environment in addition to the single technology analysis type (see Correlation Analysis on page 747).

Looking to the future we are aiming to expand this feature further by adding more analysis types and increasing the number of experiments beyond the currently enabled two experiments. If you are working in a multi-omic research environment, let us know how we can enhance this particular feature of **GeneSpring** further to support your work and put in your feature request at `informatics_support@agilent.com` or contact your Agilent Sales Representative.

20.1 Introduction to Multi-Omic Analysis in GeneSpring

This section describes the prerequisites for multi-omic analysis in **GeneSpring** as well as the supported analysis and experiment types.

20.1.1 Prerequisites for Multi-Omic Analysis

The only prerequisite for performing multi-omic analysis in **GeneSpring** is that the two experiments that you want to use for multi-omic analysis have to be in the same project.

20.1.2 Supported experiment types

The creation of an MOA experiment is supported for the following experiment and technology types:

- Agilent Expression Single Color
- Agilent Expression Two Color
- Agilent Exon Single Color
- Agilent Exon Two Color
- Agilent miRNA
- Affymetrix Expression
- Affymetrix Exon Expression
- Affymetrix Exon Splicing
- Generic Single Color
- Generic Two Color
- Illumina Single Color
- All types of Mass Profiler Professional (MPP) experiments
- RT-PCR experiment
- Pathway experiment
- Gene-Level experiment
- Imported Strand NGS experiments

NOTE The availability of either Pathway Analysis or Correlation Analysis in an MOA experiment is subject to the following exceptions:

- **Agilent miRNA experiments**: Multi-omic Pathway Analysis is only available for entity lists resulting from a TargetScan (see "To choose an experiment" on page 900 for more details).
- **Pathway experiments created from entity lists without normalized signal values**: Only multi-omic Pathway Analysis is available. However, due to the lack of normalized signal values in such experiments no data is overlaid as heatstrips or heatmaps in the resulting pathways. Correlation Analysis is not available as it always requires normalized signal values to run.
- **ChIP-Seq experiments imported from Strand NGS** are not supported for either of the Multi-Omic Analysis types.
- **DNA-Seq and Methyl-Seq experiments imported from Strand NGS**: Only multi-omic Pathway Analysis is available. However, due to the lack of normalized signal values in such experiments no data is overlaid as heatstrips or heatmpas in the resulting pathways. (see "To choose an experiment" on page 900 for more details). Correlation Analysis is not available as it always requires normalized signal values to run.

20.1.3 Supported analysis workflows

At present, multi-omic analysis is supported for:

- Pathway Analysis (see What is Multi-Omic Pathway Analysis? on page 896)
- Correlation Analysis (see How to Perform Correlation Analysis in a MOA experiment on page 752).

20.2 Creating a Multi-omic Analysis Experiment

At present there are three ways to create a Multi-omic Analysis (MOA) Experiment. The fastest way to create an MOA experiment is to use the dedicated toolbar icon or the **New MOA Experiment** option from the **Project** menu. However, if the first multi-omic analysis you want to perform is multi-omic pathway analysis you can also choose the **Multi-Omic Analysis** option from the **Workflow Browser** in any supported experiment type.

This section explains each of these three methods.

NOTE Once created, an MOA experiment remains independent of the two parent experiments. Thus any further analysis done in either of the parent experiments (new entity lists created as a result of further analysis, deleting entity lists, etc.) will not be reflected in the MOA experiment.

20.2.1 How to create an MOA experiment using the toolbar icon

Use this procedure to create an MOA experiment.

1. Open the project that contains the two experiments with the data on which you want to perform multi-omic analysis.
2. Click the **New MOA experiment** icon from the main toolbar.
   The **New Multi-Omic Analysis Experiment** window opens.
3. Provide a **Name** and **Notes** for the new MOA experiment in the respective fields.
4. Select the two desired experiments from the list of available experiments. **GeneSpring** lists only those experiments that have a valid **GeneSpring** module license and are of a supported type and technology.
5. Click **Create MOA**.

GeneSpring creates an MOA experiment that contains all the interpretations and entity lists from the two source experiments and saves it as part of the open project. The results of both Pathway Analysis and Correlation Analysis will be stored in respective dedicated folders that are created the first time you complete either analysis. Results derived from any of the available right-click options will be saved in a hierarchical fashion below the input entity list.

20.2.2 How to create an MOA experiment using the Project menu option

Use this procedure to create an MOA experiment.

1. Open the project that contains the two experiments with the data on which you want to perform multi-omic analysis.
2. Select **Project** > **New MOA Experiment** from the main menu.
   The **New Multi-Omic Analysis Experiment** window opens.
3. Provide a **Name** and **Notes** for the new MOA experiment in the respective fields.
4. Select the two desired experiments from the list of available experiments. **GeneSpring** lists only those experiments that have a valid **GeneSpring** module license and are of a supported type and technology.
5. Click **Create MOA**.

GeneSpring creates an MOA experiment that contains all the interpretations and entity lists from the two source experiments and saves it as part of the open project. The results of both Pathway Analysis and Correlation Analysis will be stored in respective dedicated folders that are created the first time you complete either analysis. Results derived from any of the available right-click options will be saved in a hierarchical fashion below the input entity list.

20.2.3 How to create an MOA experiment by performing Multi-Omic Pathway Analysis

GeneSpring also creates an MOA experiment when you perform **Multi-Omic Analysis** in the experiment of your choice (see Single Experiment and Multi-Omic Analysis on page 900).

1. Select **Multi-Omic Analysis** from the Pathway Analysis section of the Workflow Browser in the experiment of your choice (see Single Experiment and Multi-Omic Analysis on page 900)
2. Select the two desired experiments and provide a **Name** and **Notes** in the **New Multi-Omic Analysis Experiment** window. Click **OK**. **GeneSpring** lists only those experiments that have a valid **GeneSpring** module license and are of a supported type and technology.

GeneSpring creates an MOA experiment that contains all the interpretations and entity lists from the two selected experiments and saves the MOA experiment as part of the open project.

You can continue the MOA workflow and complete multi-omic pathway analysis or click **Cancel** to close the Multi-Omic-Analysis window to review the newly created

MOA experiment and perform multi-omic Correlation or Pathway Analysis at a later time. The results of the MOA workflow and any future MOA workflows you are going to perform in this MOA experiment are stored in a dedicated Pathway Analysis folder. Results derived from any of the available right-click options will be saved in the traditional hierarchical fashion below the input entity list.

NOTE When creating a new MOA experiment, you might encounter an error message, "**Error in reading ShellExperiment**". Check if there is a leading space in the name of the nodes in the experiment tree of either of the experiments. Remove the space and try creating a new MOA experiment.

20.3 Working in a Multi-omic Analysis Experiment

This section describes the available operations in an MOA experiment. Due to the fact that the MOA experiment contains interpretations and entity lists from different experiment types or technologies as well as pathway lists if Pathway Analysis was performed, only a limited number of operations are available:

- From the **Workflow Browser**, the only available options are **Pathway Analysis** and **Correlation Analysis**, the two currently supported analysis types for multi-omic analysis.
- The main **Toolbar** options are the creation of new experiments or adding existing experiments to the project as well as drawing a Venn diagram on pathway lists in the project (see The Venn Diagram on page 173).

NOTE The functionality to create a Venn diagram on entity lists is not enabled in a MOA experiment. However, if you launch a Venn Diagram on Entity Lists from another single open experiment in the same project as the MOA experiment, you can drag-and-drop entity lists from the open MOA experiment to add to the open Venn diagram (see Drag and drop operations on Venn diagram).

- From the right-click menu for entity lists (see Entity Lists on page 42), pathway lists (see Pathway List on page 50), and Correlation Analysis results (see Correlation Node on page 51) all the usual operations are available.

20.4 Multi-Omic Analysis Screen Reference

This section describes the user interface elements for creating an MOA experiment and the MOA experiment.

20.4.1 New Multi-Omic Analysis Experiment Window

Enter details and select experiments required for the creation of a Multi-Omic Analysis (MOA) Experiment.

Name

Field to provide a name for the new MOA experiment.

Description

Field to provide a description of the new MOA experiment.

Select Experiments for Multi-Omic Analysis

A table that contains all the experiments in the open project that are available for the creation of an MOA experiment. Select two experiments from this table.

Help button

To find out more about MOA experiment creation, click Help. The user manual opens in the default browser of your computer.

Create MOA button

Prompts **GeneSpring** to create the new MOA experiment from the selected two experiments. The experiment is saved in the **Project Navigator** below the other existing experiments.

Cancel button

Closes the **New Multi-Omic Analysis Experiment** window and returns to the **GeneSpring** desktop view.

20.4.2 MOA Experiment Navigator

The MOA Experiment Navigator contains folders for each selected experiment. Each of the folders for the input experiments contains a respective folder for the Interpretations of each experiment and an Analysis folder holding the entity lists of each experiment. When either Pathway Analysis or Correlation Analysis are performed, **GeneSpring** creates a new folder for each analysis type respectively to store the results.

20.4.3 MOA Experiment Toolbar

The MOA Experiment Toolbar contains the icons for the Project and Experiment sections, the Venn Diagram on pathway lists icon, and the Hide Sidebars icon.

Chapter 21

Importing Strand NGS experiments into GeneSpring

GeneSpring allows the import of Next Generation Sequencing (NGS) experiments from **Strand NGS** to facilitate Integrated Biology (IB) analysis. You can import **Strand NGS** experiments into **GeneSpring** and perform Multi-Omic Analysis (MOA) on experiments including expression experiments of **GeneSpring** and NGS experiments of **Strand NGS**.

In a Multi-Omics experiment (see Multi-Omic Analysis in **GeneSpring** on page 603), you can perform **Pathway Analysis** (see Pathway Analysis on page 859) and **Correlation Analysis** (see Correlation Analysis on page 747). The ability to analyse NGS experiments along with expression experiments is useful for:

- Identifying common pathways or pathway entities that are of significance in either or both expression and sequencing experiments.
- Correlation Analysis between NGS read density and micro-array expression data or metabolite abundance measured using mass spectrometry can be performed by first identifying the genes in regions with high sequencing reads density and comparing them with the differentially abundant genes or metabolites.

21.1 Introduction to Importing Strand NGS Experiments for Multi-Omic Analysis in GeneSpring

This section describes the prerequisites for importing **Strand NGS** experiments into **GeneSpring** for multi-omic analysis as well as the supported analysis and experiment types.

21.1.1 Prerequisites for importing Strand NGS experiments

The prerequisites for importing **Strand NGS** experiments into **GeneSpring** are:

- The NGS experiment is exported from **Strand NGS** in a **GeneSpring** recognized format (see How to Export a NGS Experiment from **Strand NGS**).
- The build, reference, and annotations used for creating the NGS experiment in **Strand NGS** are downloaded in **GeneSpring** via the **Annotations Manager** before importing the NGS experiment. If the build (as a minimum) is not present in **GeneSpring** when the NGS experiment is imported, you will not be able to open that experiment in **GeneSpring**.

See Prerequisites for Multi-Omic Analysis on page 604 for more information on performing multi-omic analysis in **GeneSpring**.

21.1.2 Supported Strand NGS experiment types for importing into GeneSpring

GeneSpring supports the import of following **Strand NGS** experiment types:

- Targeted Sequencing DNA Variant Analysis
- Targeted Sequencing Methyl-Seq Analysis
- Targeted Sequencing RNA-Seq Analysis
- Whole Genome ChIP-Seq Analysis (for Genome Browser viewing only, MOA is not supported)
- Whole Genome DNA Variant Analysis
- Whole Genome Methyl-Seq Analysis
- Whole Genome RNA-Seq Analysis
- Whole Genome Small RNA Analysis

NOTE The Alignment experiment type in **Strand NGS** is not supported for import into **GeneSpring**.

The imported **Strand NGS** experiment is added to the Project Navigator and **GeneSpring** opens it by default. All available read tracks of the imported **Strand NGS** experiment are displayed in the **Genome Browser**.

All the exported objects for the experiment are present in the same folder structure as in **Strand NGS**. The **Samples** folder contains all the samples, the **Interpretations** folder contains all the interpretations, and the **Analysis** folder contains all the entity lists of the original experiment. In addition, all the selected read lists and region lists are also present in the Analysis folder, however the original hierarchy will only have been maintained if all the lists were exported. If you only selected some of the child lists without their parent lists, **GeneSpring** tries to recreate a hierarchy as close as possible to the original by looking for the nearest parent and placing it underneath that object.

Importantly, if you did not select the 'All Aligned Reads' read list for export, **GeneSpring** recreates an 'All Aligned Reads' read list that only contains the reads present in the imported read lists. Correspondingly, the number of reads present in each of the samples (which you can look up by inspecting each sample using the right-click option **Inspect Sample**) will reflect the coverage for each sample by the imported read lists only. The original number of reads in each sample in the **Strand NGS** experiment in **Strand NGS** is recorded in the **Notes** for each sample.

21.1.3 Supported MOA analysis workflows for imported Strand NGS experiments

At present, multi-omic analysis including **Strand NGS** imported experiments is supported for:

- Pathway Analysis (see What is Multi-Omic Pathway Analysis? on page 896)
- Correlation Analysis (see How to Perform Correlation Analysis in a MOA experiment on page 752).

NOTE

The availability of either Pathway Analysis or Correlation Analysis in an MOA experiment including **Strand NGS** imported experiments is subject to the following exceptions:

- **ChIP-Seq** experiments imported from **Strand NGS** are not supported for either of the Multi-Omic Analysis. Imported Chip-Seq experiments can be viewed in **GeneSpring** using all the available Genome Browser options (see Genome Browser on page 195 for more information).
- **DNA-Seq** and **Methyl-Seq** experiments imported from **Strand NGS**: Only multi-omic Pathway Analysis is available. However, due to the lack of normalized signal values in such experiments no data is overlaid as heatstrips or heatmaps in the resulting pathways (see "To choose an experiment" on page 900 for more details). The lack of normalized signal values in these experiment types also means that you cannot perform **Correlation Analysis**.

21.2 Importing a Strand NGS Experiment and creating an MOA experiment

You can import **Strand NGS** experiments into **GeneSpring** from the **Project** menu. After the completion of import, supported NGS experiment types will be available for creating a multi-omic analysis experiment (see Creating a Multi-omic Analysis Experiment on page 606).

21.2.1 How to Export a NGS Experiment from Strand NGS

If you are a **Strand NGS** customer, you can use the advanced **Pathway Analysis** and **Correlation Analysis** features of **GeneSpring** for your NGS experiments. Use the following procedure to export a NGS experiment from **Strand NGS**:

1. In **Strand NGS**, right-click on the experiment name in the **Project Navigator** of the experiment that you want to export for MOA in **GeneSpring** and select **Export Experiment for GeneSpring**. You can only export one experiment at a time. Repeat these steps for all the experiments that you would like to export.

   The **Export Experiment for GeneSpring** dialog opens. By default, all **Samples** and **Interpretations** are selected for **Export to GeneSpring**. In the **Analysis** folder, all **Region Lists**, **Read Lists**, and **Entity Lists** are selected by default for **Export to GeneSpring**.

   a. Deselect any **Region Lists** or **Read Lists** that you do not want to export for use in **GeneSpring**.

   b. Specify where you want to save the exported experiment file on your system using the **Browse** button next to the **Location** field. Please note that if you enter a name other than the experiment's name in **Strand NGS** for the exported experiment file during this process, **GeneSpring** will retrieve and use the original experiment name when the experiment is imported.

NOTE **Gene Quantification Nodes** of **Strand NGS** experiments are not available for **Export to GeneSpring**.

---

2. Click **Export**.

   A **Space Information** dialog opens with information on the **Required space for import** and the **Available system space**. Please note that the required space exceeds the size of the exported experiment file as the export process requires a large amount of temporary space that is released again upon completion of the experiment export.

NOTE If you do not have sufficient **Available system space**, you cannot proceed with the export of a **Strand NGS** experiment.

---

3 Click **Yes** to proceed with the export.
On successful completion of export, an information message is displayed confirming the location of the ∗.*expt* file.

4 You can **Import** this file into **GeneSpring** for multi-omic analysis.

NOTE It is recommended to complete all the data analysis required for multi-omic analysis in **GeneSpring** on your NGS experiments in **Strand NGS**, so that the corresponding read lists, entity lists, and region lists are exported to **GeneSpring**. **GeneSpring** does not provide any data analysis workflows for **Strand NGS** experiments beyond the enabled multi-omic analyses, i.e. Pathway Analysis and Correlation Analysis.

CAUTION

- For a given **Read List**, if you do not select to export all the parent read lists until the **All Aligned Reads** list, **Mate Status** of few reads might differ from **Strand NGS** to **GeneSpring**.
- Before importing a **Strand NGS** experiment into **GeneSpring**, you have to download the same **Build**, **Reference**, and all **Annotations** used for the **Strand NGS** experiment creation and analysis from **GeneSpring Annotations Manager**. This will ensure that the **Read Metrics** are same in both **Strand NGS** and **GeneSpring**.

21.2.2 How to Import a Strand NGS Experiment into GeneSpring

Use the following procedure to import a **Strand NGS** experiment:

1 Open the project to which you want to add the **Strand NGS** experiment.

2 Select **Project** > **Import Strand NGS Experiment** from the main menu.
An **Open** file window opens.

3 Select an ∗.*expt* file exported from **Strand NGS** to import into **GeneSpring**. Please note that you can only import one **Strand NGS** experiment at a time.
Click **Open**.

4 A **Space Information** dialog opens with information on **Required space for import** and **Available system space**. Please note that the required space exceeds the size of the exported experiment file as the import process requires a large amount of temporary space that is released again upon completion of the experiment import.

   Click **Yes** to proceed with the import.

5 If the required build annotations for the **Strand NGS** experiment to be imported are not present in **GeneSpring**, a message is displayed at this point with instructions on how to download the required build annotations from the Annotations Manager. Without these annotations you will not be able to open your **Strand NGS** experiment. We advise you to also download the same Gene and Transcript annotations (Ensembl, RefSeq, or UCSC) that you had used to create the **Strand NGS** experiment for best access to all the Genome Browser visualization functionalities in **GeneSpring**.

6 On successful completion of the experiment import, an information message is displayed. Click **OK**.

NOTE If you do not have sufficient **Available system space**, you cannot proceed with the import of a **Strand NGS** experiment.

21.2.3 How to create an MOA experiment with the imported Strand NGS Experiment

From an open **Strand NGS** experiment, you can use the dedicated toolbar **New MOA experiment** icon or the **New MOA Experiment** option from the **Project** menu to create an MOA experiment.

See How to create an MOA experiment using the toolbar icon on page 606 and How to create an MOA experiment using the Project menu option on page 607 for more information.

21.3 Working in an imported Strand NGS Experiment

This section describes the available operations in an imported **Strand NGS** experiment. As a different kind of downstream analysis is required for NGS experiments, **GeneSpring** allows only a limited number of operations on imported **Strand NGS** experiments.

NOTE **GeneSpring** supports **Elastic Genome Browser** functionalities. See Elastic Genome Browser on page 214 for more information.

21.3.1 Workflow Browser

The following analysis options are enabled from the Workflow Browser in imported **Strand NGS** experiments. Availability of each option depends on the NGS experiment type and is indicated in Table 21.1.

- **Experiment Setup:** Use the **Experiment Setup** functions **Experiment Grouping** and **Create Interpretation** to create interpretations for **Strand NGS** experiments which can be used for the analyses available in a MOA experiment (see Supported analysis workflows on page 605).
- **Expression Analysis:**
  - Use the **Clustering** (see Clustering Analysis on page 669) option to identify entities that cluster together based on their expression values. The **Metadata Framework** to display sample parameters alongside clustered entities is also enabled (see How to Visualize Sample Metadata in the Dendrogram View on page 680).
  - Use the **Correlation Analysis** to find correlating entities (see Correlation Analysis on page 747).
- **Utilities:**
  - Use the **Import Region List** function to bring region lists exported from Strand NGS in the *.bed, *.txt, *.tsv, or *.csv formats into **GeneSpring** for visualization in the Genome Browser alongside other results brought into and generated in **GeneSpring**.
  - Use the **Cluster Regions** function to cluster regions in any region list that was imported into **GeneSpring** as part of the experiment, or imported using the **Import Region List** function. The **Metadata Framework** (see How to Visualize Sample Metadata in the Dendrogram View on page 680) to display sample parameters alongside the clustered heatmap view is also enabled for the resulting region cluster trees. For details on region clustering, please refer to the **Strand NGS** user manual.

Table 21.1: Availability of analysis options available from the Workflow Browser by experiment type (applicable to both Targeted Sequencing and Whole Genome experiments

NGS experiment type	Experiment Setup	Expression Analysis	Utilities
ChIP-Seq Analysis	YES	NO	YES
DNA Variant Analysis	YES	NO	YES
Methyl-Seq Analysis	YES	NO	YES
RNA-Seq Analysis	YES	YES	YES
Small RNA Analysis	YES	YES	YES

21.3.2 Toolbar

The following functionalities are available from the main Toolbar:

- Closing the active project
- Creation of new **GeneSpring** experiments
- Creation of new MOA experiments
- Adding existing experiments to the project
- Creation of new entity list from selection
- Inspecting selected entities
- Import entity list from file or Dropbox or Google Drive
- Drawing different views for visualizing plots, charts, heatmap, summary statistics on the active interpretation and entity list in RNA-Seq and Small RNA experiments ((see Data Visualization on page 115).
- Launching the Genome Browser
- Drawing a Venn diagram on entity lists in the project (see The Venn Diagram on page 173).

21.3.3 Right-click options on objects in an imported NGS experiment

The various right-click options available on the experiment object listed in the Project Navigator, region lists, entity lists, and read lists are explained under Imported **Strand NGS** Experiment Screen Reference on page 621.

21.4 Imported Strand NGS Experiment Screen Reference

This section describes the available workflow, toolbar, and right-click options for the objects in an imported **Strand NGS** experiment.

21.4.1 Imported Strand NGS Experiment Navigator

The imported **Strand NGS** Experiment Navigator contains Samples, Interpretations, and an Analysis folder holding all the objects that were selected for export from **Strand NGS** (e.g. read lists, region lists, entity lists).

21.4.2 Imported Strand NGS Experiment Toolbar

The available Experiment Toolbar icons for an imported **Strand NGS** are summarized in Table 21.2. The toolbar icons related to launching various views on the active entity list and interpretation are only available for RNA-Seq and Small RNA experiments as these are the only experiment types that contain the required signal values.

Table 21.2: Experiment Toolbar icons available for an imported **Strand NGS** experiment

Icon	Function
	Closes the active **Project**.
	Opens a **Create New Experiment** dialog.
	Opens a **Create New MOA Experiment** dialog.
	Allows you to add experiments to the active project from all the available experiments in **GeneSpring**.
	Allows you to create a new entity list of any selected entities.
	*Opens the **Entity Inspector** to inspect any selected entities.
	Allows you to import a new entity list into the active experiment.
	*Opens the Scatter Plot view for the active interpretation.
	*Opens the Matrix Plot view for the active interpretation.
	*Opens the MvA Plot view for the active interpretation.
	*Opens the Profile Plot view for the active interpretation.
	*Opens the Spreadsheet View for the active interpretation.
	*Opens the Histogram view for the active interpretation.
	*Opens the Heatmap view for the active interpretation.
	*Opens the Box Whisker Plot view for the active interpretation.
	*Opens the Violin Plot view for the active interpretation.
	*Opens the Summary Statistics metrics as a spreadsheet for the active interpretation.
	Opens all the available read tracks in the Genome Browser view.
	Allows you to create a Venn Diagram on all the entity lists in the active project.

* only available in RNA-Seq and Small RNA **Strand NGS** experiments

21.4.3 Right-click options on an imported Strand NGS experiment

The available right-click options on an imported **Strand NGS** experiment from the **Project Navigator** are summarized in Table 21.3.

Table 21.3: Menu options available on right-click of an imported **Strand NGS** experiment

Option	Description
Open Experiment	Opens the experiment in the main window of **GeneSpring**.
Inspect Experiment	Opens the **NGS Experiment Inspector** with all the experiment samples.
Remove Experiment	Removes the experiment from the active project.
Delete Experiment	Deletes the experiment from **GeneSpring**.

21.4.4 Right-click options on a Region List

The available right-click options on a region list of an imported **Strand NGS** experiment are summarized in Table 21.4.

Table 21.4: Menu options available on right-click over a region list of an imported **Strand NGS** experiment

Option	Description
Inspect Region List	Opens the Region List Node Inspector.
Show in Genome Browser	Opens the region List in the Genome Browser.
Navigate in Genome Browser	Displays the regions of a region list as a spreadsheet in the Navigate Spreadsheet tab in the Genome Browser.
Export Region List	Allows you to export a region list as a *.tsv, *.bed, or *.fa files.
Export List as VCF	Exports the region list in the Variant Call Format (only available for region lists resulting from SNP-related workflows.
Remove Region List	Removes the region list from the active **Strand NGS** experiment.
Delete Region List	Deletes the region list from **GeneSpring**.
Export to eArray	Allows the export of GenBank Accession data or Genomic co-ordinates data from the region list to Agilent's eArray portal

21.4.5 Right-click options on an Entity List

From the right-click menu for entity lists (see Entity Lists on page 42) all the usual operations are available.

21.4.6 Right-click options on a Read Lists

From the right-click menu for read lists you can:

- Inspect the Read List
- Launch Genome Browser
- Delete the Read List (except the 'All Aligned Reads' read list).

Chapter 22

Loading Experiment from NCBI GEO

22.1 Introduction

The Gene Expression Omnibus hosted at the NCBI (http://www.ncbi.nlm.nih.gov/geo/) is a public repository of functional genomics data sets submitted by the scientific community. **GeneSpring** can import the expression data sets, directly from the main interface, by providing it with a GSE or GEO Series identifier. The data sets will be downloaded directly from the NCBI and a new experiment will be created from the data. The experimental parameters will be extracted and used to annotate the experiment.

To load the data into **GeneSpring**, first find the GSE identifier for the experiment. If the data is described in a publication, the author will probably list the GSE identifier in the material and methods or other section of the paper. If the GSE identifier is not provided, you can search for the identifier at the Gene Expression Omnibus webpage, http://www.ncbi.nlm.nih.gov/geo/.

22.1.1 Load a GSE dataset

To load a GEO dataset, select the menu option **Tools** → **Import NCBI GEO Experiment**. A dialog appears allowing you to enter a GSE identifier, such as GSE3541. See Figure 22.1.

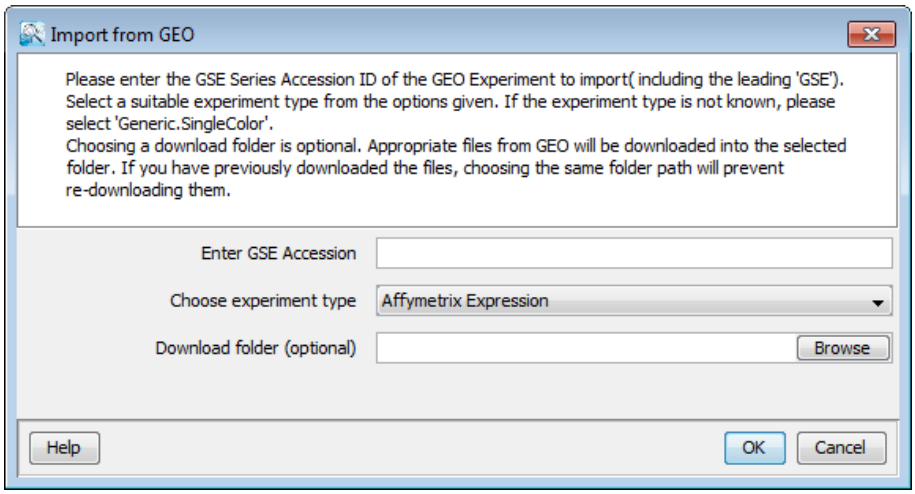


Figure 22.1: GEO Identifier Entry Dialog

The experiment type should be chosen from the drop down menu. Currently, only the following experiment formats are supported in their native forms:

1. Affymetrix expression
2. Agilent Single Color
3. Agilent Two Color
4. Generic Single Color

For Affymetrix expression data sets, only *.CEL files are supported. Pivot files are not supported in the GEO import. For Agilent expression data sets, files created with Feature Extraction version 8.5 and later are supported.

Experiments in a different format can still be loaded, but need to be loaded as 'Generic SingleColor'. This will create a technology on-the-fly (if it does not already exist) and create an experiment. If you are unsure of the experiment type, choose 'Generic SingleColor' since that will work most of the time. **GeneSpring** will also switch to importing as a 'Generic SingleColor' if the wrong experiment type is chosen (for instance, when the experiment is an Agilent experiment, but the user chooses 'Affymetrix', the experiment will be loaded as a 'Generic SingleColor' experiment.

Choosing a download folder is optional. Appropriate files from GEO will be downloaded into the selected folder. If you have previously downloaded the files, choosing the same folder path will prevent re-downloading them.

Press *OK* to continue. The data sets will be downloaded directly from the NCBI FTP site and a progress bar will be displayed.

After the data set has been successfully downloaded, a New Experiment creation window will be presented. The Experiment Name and Experiment Notes sections will be pre-populated with the

information from the data set. The Experiment Name is rather long since it is based on the title of the experiment and it is suggested to choose a shorter name (although the long names will be OK in most cases).

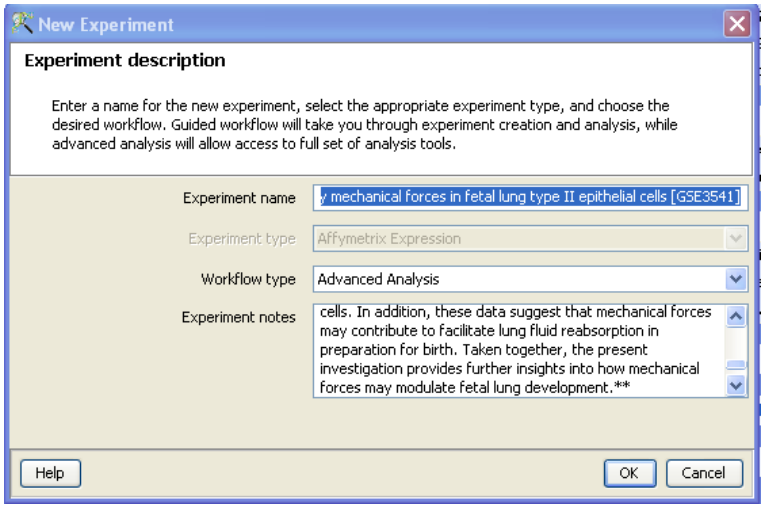


Figure 22.2: Create New Experiment Dialog

Press *OK* to start the experiment creation. After successful creation of the experiment, the Information window will show which technology was used in the creation of the experiment and how many probes matched the probes in the technology.

22.1.2 Experiment Parameters

After an experiment is created and the data set had a corresponding GDS data set (Not guaranteed for every GSE set), the experimental grouping data is automatically copied to the experiment. Open the Experiment Grouping window from the Workflow section to see the experiment grouping information.

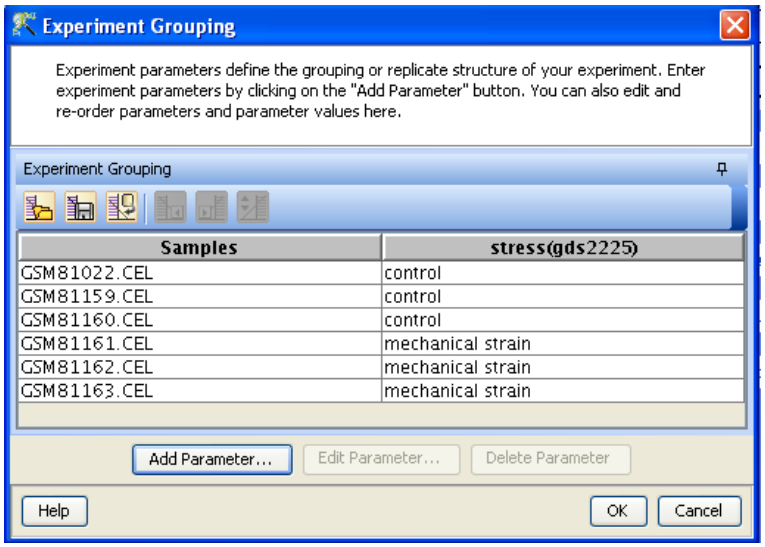


Figure 22.3: Experiment Grouping Information is automatically copied over

In this example the parameter 'stress' was copied from the GDS set 'GDS2225' and the values for each sample, such as 'control' and 'mechanical strain' are provided.

22.1.2.1 Duplicate Experiment Parameters

Sometimes, GEO creates two (or more) GDS sets from one submission GSE data set. In this case, two (or more) experiment parameters could be copied as experiment grouping parameters. An example of this is shown in Figure 22.4.

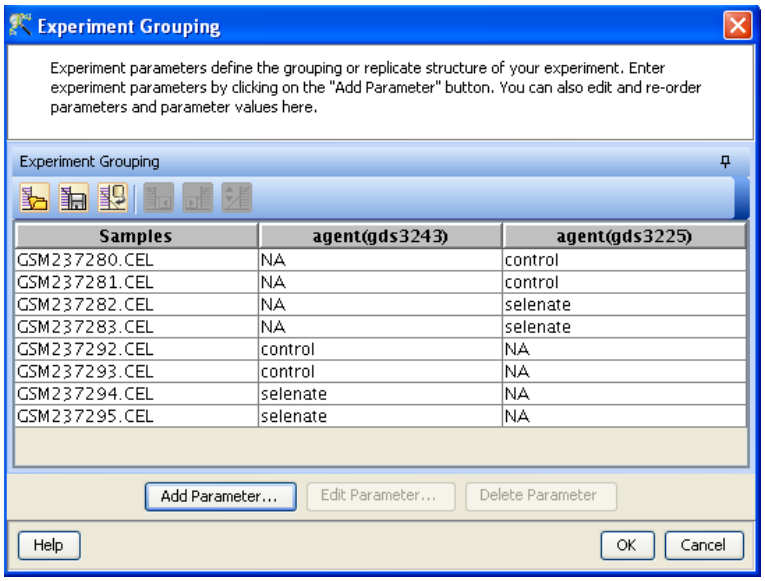


Figure 22.4: Duplicate Experiment Parameters

The 'agent' parameter was used in both GDS sets, but somehow the GEO curators felt that it should not be the same parameter, since it was separated into two separate Geo Data Sets (GDS). In these situations it is often useful to examine the sample attributes. The sample attributes are saved with each sample and are the attributes originally submitted by the author and they may have some more information on the origin of the samples. To review the sample attributes, select the 'Import Parameters from Samples' icon in the Experiment Grouping window.

In this example it turns out that the experimental design included both shoots and roots of the *Arabidopsis* plant, something that was not recorded in the GDS sets.

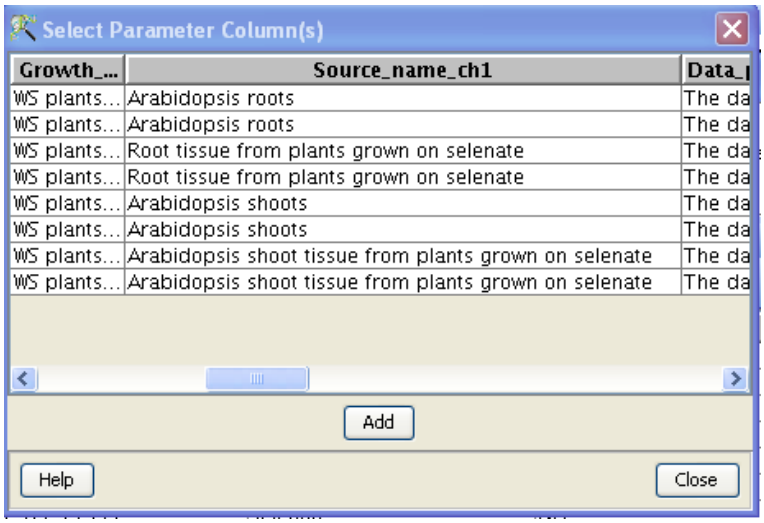


Figure 22.5: Duplicate Parameters

The sample attribute 'Source' (Source_name_ch1) seems a good candidate for an experimental parameter and should be added to the experiment grouping for this experiment, by selecting the column and pressing <Add>. Some editing of the values is required, to make these useful parameters. The two 'agent' parameters can then be combined into one parameter 'agent' and the experiment is now ready for further analysis.

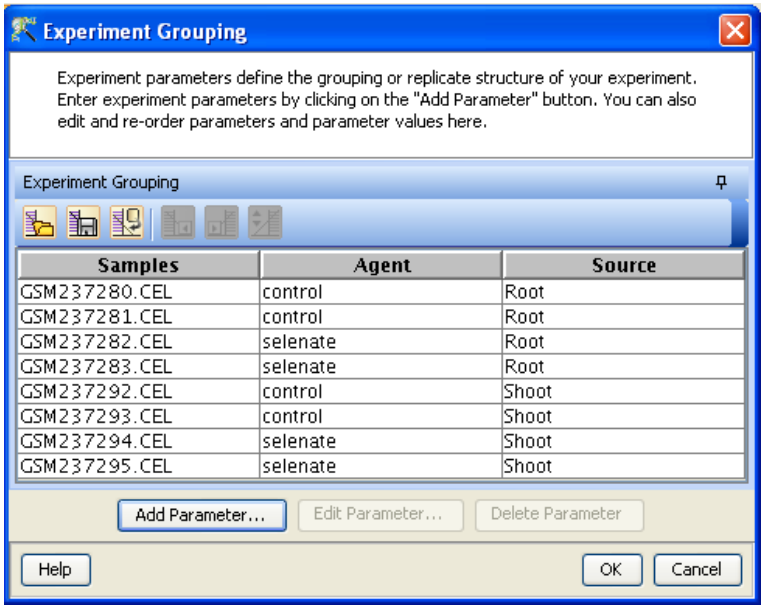


Figure 22.6: Final Experiment Grouping

22.2 Possible Error Messages

Invalid GSEid provided: **GeneSpring** currently only accepts the GSE identifiers. The GSE identifiers represent the original dataset submitted by the author. If an invalid identifier is used, the error message *Do not enter GSD or other GEO accession numbers* is displayed.

Technology type does not match: When an incorrect experiment type was selected in the first dialog box, a warning dialog may be presented. This will happen most likely if the GSE data set is NOT an Affymetrix data set, but something else. **GeneSpring** will still be able to import the data set as a Generic Single Color experiment. Press *YES* to import the experiment as a generic single color experiment or press NO to cancel the import.

If the GSE data set is an Affymetrix data set and the user chooses 'Agilent Single (or Two) color' in the experiment type, this window will not appear, but the Sample Chooser 'New Experiment (Step 1 of 5)' will be empty. This is an indication that the experiment type is not an Agilent experiment. Cancel the experiment creation and execute the GEO importer again, choosing 'Affymetrix' or 'Generic SingleColor' as the experiment type. The experiment type should also be obvious from the information on the GSE data set on the GEO website. Consult the GEO website for more information on the chosen GSE data set if unsure about the origin of the data set.

Supplementary files are unavailable: Some GSE data sets do not have any of the original data files for the experiment. When the user chooses 'Affymetrix' or 'Agilent' as the experiment type and the GSE data set does not contain any original data files, a warning dialog is shown.

GeneSpring will still be able to import the data set as a Generic Single Color experiment. Press *YES* to import the experiment as a generic single color experiment or press NO to cancel the import.

Experiment creation failed: This message may appear when no connection can be made to the GEO FTP site or other network problems. Check the network connection. **GeneSpring** attempts to connect to the FTP site at the ftp.ncbi.nlm.nih.gov. Check to see if the FTP site can be reached with other tools.

Unable to validate Experiment information: If this window appears, the most likely reason is the fact that the GSE data set contains data for two different chips or technologies. At this point only data sets from a single technology (chip) can be loaded directly from GEO.

22.3 Experiment Parameters and Sample Attributes

Experiment parameters, such as treatment, source etc. are usually not part of the GSE data sets as submitted by the research community. The experiment parameters are created by the curation staff of GEO and are part of the GEO Data Set or GDS. These GDS are not available for every data set that is submitted to GEO. The curation is a complicated and time-consuming process and the GEO staff is currently experiencing a considerable backlog. Also, not every data set submission to GEO is suitable for curation. Therefore, it is not guaranteed that **GeneSpring** will be able to extract the Experiment Parameters for every GSE data set that is available on GEO.

22.3.1 Create Experiment Parameters from Sample Attributes

During the download of the GSE data sets, each sample from the data set is annotated with a number of attributes. These attributes are usually provided directly by the submitter and sometimes (but not always!) contains information on the experimental design or experimental parameters.
These sample attributes can potentially be used as experiment parameters. The Experiment Grouping window will allow one or more columns to be used as samples attributes. Open the Experiment Grouping window from the Workflow section **Experiment Setup** and click on **Import Parameters from Samples** icon.

This window will show all the sample attributes. Many of the sample attributes are the same for all samples, such as 'Platform' or 'Submission Date' and would not be useful experiment parameters, but 'Title' usually contains some indication of the experimental conditions that are important for the analysis (such as 'Myoblast (1) G1' and 'Myotube (D1)' in the example below).

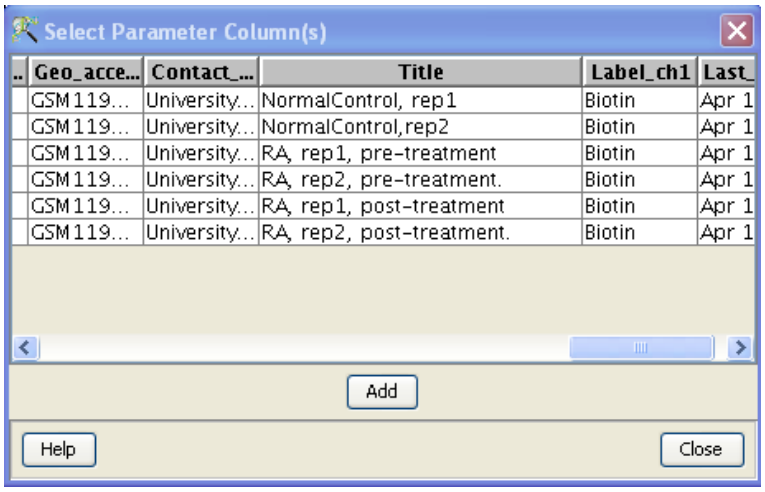


Figure 22.7: Sample attributes that can be chosen as Experiment Parameters

The actual value of the 'Title' column would not be suitable as an Experiment Grouping parameter, since none of the values for an experimental condition are the same, but the column can be loaded as a parameter and later edited in the Experimental Grouping window.

Choosing the 'Title' as the Experiment Parameter for this experiment makes it easier to edit the values to their proper value and avoids errors. Select *Add* to add the selected column 'Title' as an experiment parameter and edit the parameter values to contain correct experiment parameter values that can be used in the creation of Interpretations and perform statistical analysis.

Chapter 23

Exporting Expression Experiments in MAGE-TAB Format

23.1 Introduction

MAGE-TAB format [43] uses a number of different files to capture information about a microarray experiment. The common components of MAGE-TAB are:

- IDF: Investigation Description Format
- SDRF: Sample and Data Relationship Format
- ADF: Array Design Format
- Raw and processed data files, including "data matrix" files

GeneSpring allows exporting the contents of micro array expression experiments in MAGE-TAB format. **GeneSpring** writes out IDF and SDRF files by default from the files listed above and gives additional option to write out the Raw and processed data files. The MAGE-TAB exporter in **GeneSpring** does not create an Array Design Format (ADF) file. The SDRF file includes a pointer to the appropriate array design accession that is there in Array Express.

The following sections describe the details of each of the MAGE-TAB format files that **GeneSpring** writes out and some options to customize them.

23.2 IDF file

The Investigation Design Format (IDF) file is the primary output of the MAGE-TAB export. The IDF in turn consists of some components which are described below.

23.2.1 IDF Model file

The starting point for the IDF generation is a textual representation of the template for an IDF file. The model file reflects the IDF format is described at http://tab2mage.sourceforge.net/docs/idf.html. Each line in the model file describes one of the keys of the IDF file. The IDF file is divided into groups of keys, for example the ‘‘Person” group includes the keys ‘‘Person Name”, ‘‘Person Email”, etc.

Creating an IDF file involves specifying which of the keys from the model to output in the IDF file, and how these fields map to data stored in **GeneSpring** . This is done using the preferences file and the mappings file described below. In the absence of the preferences file and the mappings file, the program will output an IDF file with all the keys and with some basic default values from the information contained in **GeneSpring** .

23.2.2 IDF Preferences File

The preferences file can be used to customize what fields should be present in the IDF output generated by **GeneSpring** . It contains a simple list of all the IDF 'Internal Key Identifiers' from the IDF model file that should be output. For example, if the output IDF file needs to contain only the Person?s last name, first name and email information from within the Person group, this can be easily customized using the preferences file.

The default preferences file that is used by the tool can be found in the folder /Installation Directory/GeneSpring/app/MageTab/CustomFiles/. To use a different preferences file, one needs to simply copy this file, edit the contents of the file, and put the new file in the the path `INSTALLATION DIRECTORY/bin/packages/marray/genespring/<Version No>/plugins`.

Given a preferences file that contains only a subset of the IDF keys, and given no mappings file, the program will output an IDF file with (a subset of) keys as specified in the preferences file, and with some basic default values.

23.2.3 IDF Mappings

The mappings file is used to map the data stored in GeneSpring to the corresponding keys of the IDF file. The data from GeneSpring that can be mapped to the corresponding IDF keys include experiment attributes like the name of the experiment, the notes of the experiment, etc., and sample attributes like the sample preparation protocol. The experiment attributes are extracted

from the experiment and output 'as is' in the IDF file, whereas for sample attributes, the unique set of values are extracted from all the samples of the experiment and output in the IDF file as multiple values for the corresponding IDF key.

The sample mappings file can be found in the folder /Installation Directory/GeneSpring/app/MageTab/CustomFiles/. To use a different mappings file, one needs to simply copy this file, edit the contents of the file, and put this new file in the path `GENESPRING_INSTALL_DIR/bin/packages/marray/genespring/<Version No>/plugin`

Given a preferences file that contains only a subset of the IDF keys, and given a mappings file that contains mappings for a subset of these keys, the program will output an IDF file with all the keys as specified in the preferences file, and with values wherever a mapping exists in the mappings file, in addition to default values.

23.2.4 IDF Defaults

GeneSpring outputs values for some of the fields in the IDF file by default, based on the information it has for the experiment. The list of these fields and what they contain is described below:

Table 23.1: IDF Defaults

Investigation Title	Name of the project in GX
Experiment Description	Name of the experiment in GX
Date of Experiment	Date on which the experiment was created in GX
Comment [**Notes**]	Notes of the experiment in GX
Experimental Factor Name	The names of the parameters in the experiment grouping in GX
Experimental Factor Type	Defaults to the ontology keyword "Test" (`MO_197`) from MGEDOntology
Person First Name	Name of the (Operating System) user Login

Normalization Type	One of the following from MGEDOntology:
	`median_log_centering (MO_811)`
	`mean_log_centering (MO_398)`
	`lowess_normalization (MO_720)`
	`quantile_normalization (MO_951)`
	If the experiment has some other kind of normalization, for example percentile shift, then this field is not populated.

If any of these fields are specified in the mappings file, then that mapping is used instead of the

default described above. Similarly, if the mappings file is missing or empty, the IDF file generated will have exactly the fields described above populated with values. All other fields will be left blank.

23.2.5 Ontology mappings

Some of the fields in the IDF file are expected to come from ontologies along with the specification of the ontology and the accession number in that ontology. Examples of such fields include the 'Experimental Design', 'Experimental Factor Type', 'Person Roles', etc. The MAGE-TAB exporter detects such fields that have a value by virtue of GeneSpring defaults or by virtue of a user defined mapping, and automatically queries ontologies to populate the subsidiary information. The ontologies that are integrated include the 'Experimental Factor Ontology' or EFO from EBI (http://www.ebi.ac.uk/efo/), and the MGEDOntology from the MGED Network (http://mged.sourceforge.net/ontologies/MGEDontology.php).

To give an example, suppose the 'Experimental Design' IDF field is mapped to the `'DESIGN_OF_EXPERIMENT'` sample attribute, and suppose the exporter determines the value of this field to be 'Compound Treatment Design', then the 'Experimental Design Term Source REF', and the 'Experimental Design Term Accession Number' fields are automatically populated as `'EFO'` and `'EFO_0001755'` respectively by the exporter in **GeneSpring** .

The exporter integrates EFO using the web-services provided by the Ontology Lookup Service (OLS) from EBI (http://www.ebi.ac.uk/ontology-lookup/browse.do?ontName=EFO). For every term that needs an ontology mapping, the exporter first checks if EFO contains the term, and if not, it then checks the MGEDOntology. The MGEDOntology is packaged within GeneSpring, with the contents derived from the xls file at http://mged.sourceforge.net/ontologies/MO-class-individual.xls.

23.3 SDRF file

The Sample and Data Relationship Format (SDRF) is the other key output of the MAGE-TAB export. The general philosophy of the design is the same as that of the IDF file - namely the model file that describes the template, the preferences file that customizes the keys in the output, and the mappings file that customizes the mapping from GeneSpring.

23.3.1 SDRF Model File

The starting point for the SDRF generation is a textual representation of the template for an SDRF file. Each line in the model file describes one of the keys of the SDRF file. Compared to the IDF model however, the SDRF model is a little more complex, since it can contain the same SDRF key in multiple places, **each signifying a different meaning depending on the position of occurrence**. For example, a 'Protocol REF' key occurring after the source related columns and before the 'Sample Name' column in the output is expected to contain the

details of the sample preparation protocol. The same 'Protocol REF' key occurring after the sample related columns and before the 'Extract Name' column is expected to contain the details of the extraction protocol. To avoid ambiguity, the **'Internal Key Identifiers'** in the SDRF file are specified in a fully qualified manner. Thus, for the example above, the two 'Protocol REF' fields are distinguished using the internal keys `'sample_preparation_protocol'`, and `'extraction_protocol'`, and the 'Unit' keys under each of these are distinguished using the internal keys `'sample_preparation_protocol.parameter.unit'`, and `'extraction_protocol.parameter.unit'` respectively.

Each line in the model file has multiple tab-separated fields, which correspond to 'Internal Group Identifier', 'Internal Key Identifier', 'Display Key Name', and 'Multiplicity'. The model file reflects the SDRF format described at http://tab2mage.sourceforge.net/docs/sdrf.html.

Creating an SDRF file involves specifying which of the keys from the model to output in the SDRF file, and how these fields map from data stored in **GeneSpring** . This is done using the preferences file and the mappings file described below. In the absence of the preference file and the mappings file, the program will output an SDRF file with all the keys and with some basic default values.

23.3.2 SDRF Preferences

The preferences file can be used to customize what fields should occur in the SDRF output generated by **GeneSpring** . The file contains a simple list of all the SDRF 'Internal Key Identifier' tokens from the SDRF model file that should be output.

The sample preferences file can be found in the folder /Installation Directory/GeneSpring/app/MageTab/CustomFiles/. To use a different preferences file, one needs to simply copy this file, edit the contents of the file, and put the new file in the path `GENESPRING INSTALLATION DIRECTORY /bin/packages/marray/genespring/<Version No>/plugins`.

Given a preferences file that contains only a subset of the SDRF keys, and given no mappings file, the program will output an SDRF file with (a subset of) keys as specified in the preferences file and with some basic default values.

23.3.3 SDRF Mappings

The mappings file is used to map the data stored in **GeneSpring** to the corresponding keys of the SDRF file. Any sample attribute data from **GeneSpring** can be mapped to the corresponding SDRF key.

The sample mappings file can be found in the folder /Installation Directory/GeneSpring/app/MageTab/CustomFiles/. To use a different preferences file, one needs to simply copy this file, edit the contents of the file, and put the new file in the path `GENESPRING INSTALLATION DIRECTORY/bin/packages/marray/genespring/<Version No>/plugins` folder.

Given a preferences file that contains only a subset of the SDRF keys, and given a mappings file that contains mappings for a subset of these keys, the program will output an SDRF file with all the keys as specified in the preferences file, and with values wherever a mapping exists in the mappings file in addition to the basic default values.

23.3.4 SDRF Rows

In the case of single color experiments, the output SDRF file contains one row of data for each sample in the experiment, and as many columns as specified in the preferences file. The name of the source, the sample, the extract, the labeled extract, the hybridization and the scan are all derived from the name of the raw file corresponding to the GeneSpring sample by trimming the extension, unless explicitly specified in the mappings file.

In the case of two color experiments, the output SDRF file contains two rows of data for each sample of the experiment - one each for each of the channels. In this case, the name of the source, the sample, the extract, the labeled extract, the hybridization and the scan are derived from the name of the raw file and are appended with an indicator of the channel they correspond to. The control rows are appended with the keyword `'_control'`, and the treated rows are left as is. In case an array is marked as a dye-swapped array, the Label field indicator is reversed.

If some samples have been removed from the experiment using a QC - Add/remove samples step, these samples are omitted from the SDRF output.

23.3.5 Array Design REF

One of the columns in the SDRF file is the Array Design REF that contains the reference to the array design used for the hybridization. This column contains the ArrayExpress accession number corresponding to the GeneSpring technology of the sample. The exporter automatically populates this information using a plugin file, unless explicitly specified in the mappings file using a different sample attribute. Further, if this information is missing from both the SDRF mappings file and from the plugin file, it outputs the name of the technology for that sample in GeneSpring.

23.3.6 SDRF Defaults

Similar to the IDF file, **GeneSpring** outputs values for some of the fields in the SDRF file by default, based on the information it has for the samples of the experiment. If any of these fields are specified in the mappings file, then that mapping is used instead of the default described below. Similarly, if the mappings file is missing or empty, the SDRF file generated will have exactly the fields described below populated with values. All other fields will be left blank.

The list of these fields and what they contain is described below:

Table 23.2: SDRF Defaults

SDRF Key	Default Content
Source Name	Name of the raw file without the extension: Affymetrix CEL: S1 for a sample S1.CEL Affymetrix CHP: S1 for a sample S1.CHP Agilent 1-color: S1 for a sample S1.txt Agilent 2-color: two rows with S1 and `S1_control` for a sample S1.txt Illumina: S1 for a sample with a lane S1 Generic 1-color: S1 for a sample S1.extn Generic 2-color: S1 and `S1_control` for a sample S1.extn
Characteristics[Organism]	The organism mentioned in the technology for the sample in GX
Sample Name **Extract Name** **Labeled Extract Name** **Hybridization Name** **Scan Name**	Same as the rule for Source Name listed above
Array Design REF	Mapping to ArrayExpress Array Design from the technology of the sample in GX
Label	Name of the labeling dye: Affymetrix: biotin Agilent 1-color, Generic 1-color, Illumina: Cy3 Agilent 2-color, Generic 2-color: Cy5 (for the treated channel) and Cy3 (for the control channel). Swapped for the case of dye-swapped arrays.
Normalization Name	Derived from the name of the raw file without the extension
Comment[Parameters]	A description of the normalization parameters, extracted from the notes of the experiment in GX
Derived Array Data Matrix File	The name of the file that stores the data matrix for the normalized data from all the samples of the experiment
Factor[F]	One column for each of the parameters in the experiment grouping in GX, and the groups within the parameter

23.3.7 Ontology mappings

Similar to the IDF file, some of the fields in the SDRF file are expected to come from ontologies along with the specification of the ontology and the accession number in that ontology. Examples of such fields include the 'Material Type', 'Protocol REF', 'Characteristics', etc.

The MAGE-TAB exporter automatically detects such fields that need ontology annotations, and includes the appropriate 'Term Source REF', and 'Term Accession Number' in the SDRF file by integrating EFO and MGEDOntology. It also ensures that the appropriate Term Source occurs in the IDF file.

23.4 Array Design Format

The MAGE-TAB Exporter in **GeneSpring** does not write out an explicit 'Array Design Format' (ADF) file. Instead, it populates a field 'Array Design REF' in the SDRF file with the ArrayExpress Accession Number corresponding to the GeneSpring technology of the sample. This is extracted using a plugin file, unless explicitly specified in the SDRF mappings file. If the information cannot be extracted using the plugin and is not specified in the mappings file, then the exporter outputs the name of the technology for that sample in **GeneSpring** .

For the standard arrays, this should not create any problem in exporting MAGE-TAB and in reusing it in ArrayExpress; but it may not be sufficient to handle custom arrays.

23.5 Data files

In addition to the IDF and SDRF files described above, the MAGE-TAB exporter in **GeneSpring** gives options to write out the raw and the processed data files from the experiment.

23.5.1 Raw Files

The MAGE-TAB exporter dumps all the raw files for the samples of the experiment in their native vendor format. In addition, it populates the 'Array Data File' column in the SDRF file with the names of these files. An exception to this is for the Illumina experiments, in which the native raw file is not maintained in **GeneSpring** . The MAGE-TAB exporter exports a single data matrix file for this case, and populates the 'Array Data Matrix File' column in the SDRF file with the name of this file.

23.5.2 Raw Array Data Matrix File

This option would export out data that is created and available just after it was brought into **GeneSpring** . It would imply the following for various input files:

- Affymetrix CEL files - linear RMA/MAS5/PLIER/etc data.
- Affymetrix CHIP - the linear data contained in the CHP files.
- Agilent Single Colour - the linear data summarized over probes.
- Agilent Two Colour - the linear data summarized over probes for both channels.
- Illumina - the linear data summarized over probes.

23.5.3 Derived Array Data Matrix File

The MAGE-TAB exporter extracts the final normalized data for the experiment from **GeneSpring** and exports it in the appropriate data matrix format. This data is what the views and analysis in GeneSpring are performed with - namely the data after all preprocessing steps including summarization, normalization, thresholding, log transformation, ratio of channels, and baseline transformation as performed by the user. In addition, it populates the 'Derived Array Data Matrix file' column in the SDRF file with the name of this file.

This file includes references to the 'Normalization Name' column in the SDRF and has a new 'quantitation type' called `'log_normalized_signal'`.

23.6 Validations

23.7 Points to note

- The SDRF export captures the most common experiment designs for microarray experiments with single or two color arrays. However, more complex experiment designs like pooling of samples into an extract is not handled.
- There is no handle in the UI to define mappings for the IDF file; in the SDRF file, the existing default mappings are primitive.
- **GeneSpring** does not write out Array Design Format (ADF) file.
- MAGE-TAB exported from **GeneSpring** should ideally work in ArrayExpress in case of standard arrays. However, handling custom arrays is not in the purview of this functionality.
- Since **GeneSpring** does not contain the raw files for the Illumina experiments, the MAGE-TAB exporter exports a single data matrix file for this case, and populates the 'Array Data Matrix File' column in the SDRF file with the name of this file.

Chapter 24

Normalization, Statistical Hypothesis Testing, and Differential Expression Analysis

Various normalization methods and statistical tests available in **GeneSpring** are described in this chapter. For a simple simple introduction to the statistical tests, refer [23].

24.1 Threshold

Thresholding is a substitution step where all expression values below a certain user specified value are made constant, the constant being the specified value. Thresholding is done to remove very small expression values or negative values in the data before considering the log values of the data. This step is carried out to ensure that there would not be large negative values or missing values in the log transformed data.

The default in **GeneSpring** is to threshold the expression values to 1.

NOTE If you suppose that the bulk of the spots have low expression values, reduce the threshold value to < 1.

.

24.2 Normalization Algorithms

Normalization algorithms to normalize microarray data consisting of many hybridization samples in an experiment available in **GeneSpring** are detailed in this section. These include Percentile Shift Normalization, Scale, Quantile, Normalize to Control Genes, Normalize to External Value, and Lowess.

24.2.1 Percentile Shift

Percentile Shift normalization is a global normalization method where the location of all the spot intensities in an array are adjusted. This normalization method takes each column in an experiment independently, and computes the n^{th} percentile of the expression values for this array across all spots (where n has a range from 0-100 and n=50 is the median). Then this value is subtracted from the expression value of each entity.

In **GeneSpring**, log transformation is done on the dataset before the normalization and hence the percentile is subtracted from the expression value.

NOTE For data on linear scale, the expression value would be divided by the percentile.

24.2.2 Scale

This option is useful ton overcome the inter array differences. **GeneSpring** provides scaling option to median of percentile target or mean of means of all samples/control samples. The percentile should be specified in the 'Percentile target' box present above the Scaling Options.

If scaling to mean is chosen, the normalization method calculates the mean of samples (either control samples or all the samples depending on the specification) individually and then calculates the mean of those means. For example, if S1, S2 , S3, and S4 are the samples and the option chosen is to scale the values to the mean of means of all samples, then the mean of S1, S2 , S3, and S4 are calculated as M1, M2, M3, and M4. The next step is to calculate the mean (M) of M1, M2, M3, and M4.

A scaling factor is then calculated by subtracting the individual means (M1, M2, M3 and M4) from the Mean of means (M).

Scaling Factor = M - M1; M - M2; M - M3; M-M4 and so on

If the user wants to scale to median of nth percentile, the procedure still remains the same, but with the nth percentile being calculated for the samples and then the median being calculated for these percentiles. This scaling factor is then added to every intensity value on the array.

NOTE For data in linear scale, the intensity value would be multiplied by the scaling factor instead of being added.

24.2.3 Quantile

Quantile normalization is a method of normalization which make the distribution of expression values of all samples in an experiment the same. Thus after normalization, all statistical parameters of the sample, i.e., mean, median, and percentiles of all samples will be identical. Quantile normalization works quite well at reducing variance between arrays.

Quantile normalization is performed as follows:

- The expression values of each sample is sorted in an ascending order and placed next to each other.
- Each column is sorted in an ascending order. The mean of the sorted order across all samples is taken. Thus each row in this sorted matrix has value equal to the previous mean.
- The modified matrix as obtained in the previous step is rearranged to have the same ordering as the input matrix.

Quantile normalization takes care of missing values in the dataset.

24.2.4 Normalize to Control Genes

This option allows to normalize using control genes which can be any of the genes in the array (Rank invariant genes are usually recommended). This option is usually exercised in the case of arrays populated with only specific genes of interest or arrays having less than 1000 spots. It is not advisable to use this normalization if the control genes vary across the samples.

This normalization takes each sample in an experiment independently. It calculates the median of the control genes in each sample and this value is subtracted from all the genes in the sample.

24.2.5 Normalize to External Value

This option is to enable to scale the intensity value for each of the sample. Provided a scaling factor for each of the sample, the algorithm subtracts the scaling factor from the signal intensity value in case of data in log scale. For data in linear scale, the signal intensity value is divided by the scaling factor.

If *Normalization to External Value* is chosen, **GeneSpring** will bring up a table listing all samples and a default scaling factor of '1.0'. You can change this value by using the 'Assign Value' option at the bottom after highlighting the sample in the table. Multiple samples can be chosen simultaneously to assign a value.

24.2.6 Lowess Normalization

In two-color experiments, two fluorescent dyes (red and green) are used and intensity-dependent variation in dye bias may introduce spurious variations in the collected data. Lowess normalization merges two-color data, applying a smoothing adjustment that removes such variations.

Characteristics of Lowess normalization are:

- This normalization can be applied to a two-color array expression dataset.
- All samples in the dataset are corrected independently.
- This normalization can be applied to complete or partial datasets. It can be performed independently on each block or portion of the array, or on the whole array.

Lowess regression, or locally weighted least squares regression, is a technique for fitting a smoothing curve to a dataset. The degree of smoothing is determined by the window width parameter. A larger window width results in a smoother curve, a smaller window results in more local variation.

This method involves the following steps:

- Determine the smoothing widows as the percentage of the total number of points or expression values to be considered.
- For the central point in the smoothing window, compute a locally weighted least square regression. Thus points closer to the central point will be given a higher weight and points away from the central point will be given lower weight in the regression. Use this as the value for the central point.
- Move the smoothing window by one point and compute the locally weighted least square regression value for the next central point.
- Repeat this and compute a Lowess normalized expression value for each point of entity in the sample.

The default smoothing parameter for Lowess normalization is 0.2. A sliding window of length 20% of the total number of spots in a grid is used to perform weighted linear regression. Twenty percent of the expression values of all the entities are used to run the locally weighted least square regression. In case the number of spots in a grid are too few (< 250), then a sliding window of length 50 is used to calculate the mean instead of regression.

GeneSpring supports Lowess normalization for the whole array or block by block (sub-grid) in all two-color experiments.

24.3 Details of Statistical Tests in GeneSpring

Table 24.1 details the statistical tests available in **GeneSpring** for analysis to compare data organized into conditions (groups) and parameters (See Experiment Grouping, Parameters and Parameter Values). Determine which conditions and parameters you want to use by creating a corresponding Interpretation (see Conditions and Interpretations) and select this Interpretation when performing statistical analysis.

NOTE You can only perform Statistical Analysis if all the groups (and combinations of groups for multiple parameters) in the chosen Interpretation have replicates. If a group or combination of groups has only one sample (or none), the standard error cannot be calculated.

Table 24.1: Available statistical tests by experimental conditions and grouping

No. of Groups/ Conditions	**Statistical Tests**	**Comparison**
1	*t*-Test against Zero, Mann-Whitney Test against Zero	Condition 1 *vs* None
2	Moderated *t*-Test, Unpaired *t*-Test, Paired *t*-Test, Unpaired Unequal Variance *t*-Test (Welch *t*-Test), Unpaired Mann-Whitney Test, Paired Mann-Whitney Test	Condition 1 *vs* Condition 2 or Condition 2 *vs* Condition 1
≥ 3	One-way ANOVA, Unequal Variance ANOVA, Kruskal-Wallis Test, Repeated Measures ANOVA, Repeated Measures Friedman	Between all conditions
$\geq$ 2 Experiment parameters	N-way ANOVA	Between all conditions

24.3.1 The t-Test against Zero for a Single Group

This test can be applied to compare one group/condition against zero to compute the t-statistic using the formula:

t-statistic for a gene g in Condition 1;

$$t_g = \frac{m_1}{\sqrt{\sigma_1^2/n_1}}$$

Here,

t_g - t-statistic for each gene g

m_1 - Mean expression values for gene g within Condition 1

σ_1 - Standard deviation

n_1 - Sample size

- **GeneSpring** translates this t-statistic to obtain a p-**value** (from a look up table).
- The p-value cut-off filter in Step 8 of 9 of the Statistical Analysis workflow is applied on this p-value to pass a gene g.

24.3.2 The Unpaired t-Test for Two Groups

Qualitatively, this t-statistic has a high absolute value for a gene if the means within the two sets of replicates are very different and if each set of replicates has small standard deviation. Thus, the higher the t-statistic is in absolute value, the greater the confidence with which this gene can be declared as being differentially expressed.

This test can be applied on two groups to compute the t-statistic using the formula:

$t_g = \frac{m_1 - m_2}{s_{m_1 - m_2}}$
where, $s_{m_1-m_2}$ (the unbiased pooled variance estimate) $= \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}(\frac{1}{n_1}+\frac{1}{n_2})}$

Here for groups 1 and 2,

- m_1, m_2 - mean expression values for gene g within the respective groups
- s_1, s_2 - standard deviations for the respective groups
- n_1, n_2 - sample sizes of the respective groups

NOTE This test is a more sophisticated measure than the commonly used fold-change measure (which would just be $m_1 - m_2$ on the log-scale). It considers large fold-change in conjunction with small variances in each group. The effectiveness of this statistic lies in differentiating between true differential expression and differential expression due to random effects as the sample sizes n_1 and n_2 increase.

24.3.3 The Paired t-Test for Two Groups

The paired t-test is done in two steps. Let $a_1 \dots a_n$ be the values for gene g in the first group and $b_1 \dots b_n$ be the values for gene g in the second group.

- First, the paired items in the two groups are subtracted, i.e., $a_i - b_i$ is computed for all i.
- A t-test against Zero is performed on this single group of $a_i - b_i$ values.

Please note that if even one sample has a missing value in a pair, the entity/gene is dropped from the paired statistical analysis.

24.3.4 The Moderated t-Test for Two Groups

The moderated t-test is a modification of the Unpaired t-Test. It is optimized to handle the following corner cases in a better way:

- Two conditions with a small difference between their means and very low variance within each condition results in the unpaired t-test calling the two conditions differentially expressed, which may not be correct.
- Two conditions with a large difference between their means and very high variance within each condition, which might result in the conditions not passing the unpaired t-test.

The moderated t-test attempts to estimate the variance, v_{global} and a degree of freedom, d_{global} of the assumed normal distribution across genes. The details of the process can be obtained from [25].

Following this, the moderated variance for each gene is computed by combining the variance for that gene across the given condition and the 'global variance' calculated.

The variance across the condition is calculated as:

$$s_{m_1-m_2} = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{d_{f(m_1-m_2)}}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$d_{f(m_1-m_2)} = n_1 + n_2 - 2$

Here for groups 1 and 2,

m_1, m_2 - mean expression values for gene g within the respective groups

s_1, s_2 - standard deviations for the respective groups

n_1, n_2 - sample sizes of the respective groups

The moderated variance (s_{mod}) is then calculated as follows:

$s_{mod} = \frac{s_{m_1-m_2} * d_{f(m_1-m_2)} + s_{global} d_{global}}{d_{f(m_1-m_2)} + d_{global}}$

Once the moderated variance has been calculated, it is used for calculating the corresponding t-statistic values as:

$t_{gmod} = \frac{m_1-m_2}{s_{mod}}$

NOTE This test works with the premise of unpaired samples and equal variances for both the groups involved. This test has not been made available against zero because the zero condition test essentially is like a paired t-test and the availability of literature for using moderated t-test for the paired case is sparse.

24.3.5 The Unpaired Unequal Variance t-Test (Welch t-Test) for Two Groups

The standard t-test assumes that the variance of the two groups under comparison is equal. The Welch t-test is applicable when the variances are significantly different.

Welch's t-test defines the statistic t by the following formula:

$$t_g = \frac{m_1 - m_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

Here for groups 1 and 2,

- m_1, m_2 - mean expression values for gene g within the respective groups

- s_1, s_2 - standard deviations for the respective groups
- n_1, n_2 - sample sizes of the respective groups

The degrees of freedom associated with this variance estimate is approximated using the Welch-Satterthwaite equation:

$$df = \frac{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}{\frac{s_1^4}{n_1^2 - df_1} + \frac{s_2^4}{n_2^2 - df_2}}$$

24.3.6 The Mann-Whitney Test against Zero for a Single Group

This is a non-parametric test that can be applied to compare one group/condition against zero to compute the U-statistic using the formula:

U-statistic for Condition 1;

$$U_1 = R_1 - \frac{(n_1 * (n_1 + 1)}{2}$$

Here,

U_1 - U-statistic

R_1 - Rank sum of the observations arranged in an ascending order.

n_1 = Total number of observations

For large samples, U_n is assumed to be approximately normally distributed and this U-statistic is used to compute the standardized *z-value*.

z-statistic for a gene g in Condition 1 from U_1;

$$z_g = \frac{U_1 - m_{U_1}}{\sigma_{U_1}}$$

Here,

z_g - z-statistic for each gene g

U_1 - U-statistic

m_1 - Mean expression values for gene g within Condition 1

σ_1 - Standard deviation

- **GeneSpring** translates this *z-value* for a both-sided test (+/- z) and applies a normal approximation to obtain a ***p*-value** (from a look up table).
- The p-value cut-off filter in Step 8 of 9 of the Statistical Analysis workflow is applied to this p-value to pass a gene g.

NOTE Apply this non-parametric statistical hypothesis test to assess a condition that contains a relatively larger number of samples (at least twenty or more).

24.3.7 The Unpaired Mann-Whitney Test for Two Groups

The t-Test assumes that the gene expression values within groups 1 and 2 are independently and randomly drawn from the source population **and** follow a normal distribution. If the latter assumption may not be reasonably supposed, the preferred test is the non-parametric Mann-Whitney test, also referred to as the Wilcoxon Rank-Sum test. It only assumes that the data within a sample are obtained from the same distribution but requires no knowledge of that distribution. The test combines the raw data from the two samples of size n_1 and n_2 respectively into a single sample of size $n = n_1 + n_2$. It then sorts the data and provides ranks based on the sorted values. Ties are resolved by giving averaged values for ranks. The data thus ranked is returned to the original sample group 1 or 2. All further manipulations of data are now performed on the rank values rather than the raw data values. The probability of erroneously concluding differential expression is dictated by the distribution of T_i, the sum of ranks for group i, $i = 1, 2$. This distribution can be shown to be normal mean $m_i = n_i(\frac{n+1}{2})$ and standard deviation $\sigma_1 = \sigma_2 = \sigma$, where σ is the standard deviation of the combined sample set.

24.3.8 The Paired Mann-Whitney Test for Two Groups

The samples being paired, the test requires that the sample size of groups 1 and 2 be equal, i.e., $n_1 = n_2$. The absolute value of the difference between the paired samples is computed and then ranked in increasing order, apportioning tied ranks when necessary. The statistic U, representing the sum of the ranks of the absolute differences taking non-zero values obeys a normal distribution with mean $m = \frac{1}{2}(n_1\frac{(n_1+1)}{2}) - S_0)$, where S_0 is the sum of the ranks of the differences taking value 0, and variance given by one-fourth the sum of the squares of the ranks.

NOTE Even if one sample has a missing value in a pair, the entity/gene is dropped from the paired statistical analysis.

The Mann-Whitney Test and t-Test described previously address the analysis of two groups of data; in case of three or more groups, the following tests are applicable.

24.3.9 The One-Way ANOVA

When comparing data across three or more groups, the obvious option of considering data one pair at a time presents itself. The problem with this approach is that it does not allow one to draw any conclusions about the dataset as a whole. While the probability that each individual pair yields significant results by mere chance is small, the probability that any one pair of the entire dataset does so is substantially larger. The One-Way ANOVA takes a comprehensive approach in analyzing data and attempts to extend the logic of t-tests to handle three or more groups concurrently. It uses the mean of the sum of squared deviates (SSD) as an aggregate measure of variability between and within groups. NOTE: For a sample of n observations $X_1, X_2, ...X_n$, the sum of squared deviates is given by

$$SSD = \sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}$$

The numerator in the t-statistic is representative of the difference in the mean **between** the two groups under scrutiny, while the denominator is a measure of the random variance **within** each group. For a dataset with k groups of size $n_1, n_2, ...n_k$, and mean values $M_1, M_2, ..., M_k$ respectively, One-Way ANOVA employs the SSD between groups, SSD_{bg}, as a measure of variability in group mean values, and the SSD within groups, SSD_{wg} as representative of the randomness of values within groups. Here,

$$SSD_{bg} \equiv \sum_{i=1}^{k} n_i (M_i - M)^2$$

and

$$SSD_{wg} \equiv \sum_{i=1}^{k} SSD_i$$

with M being the average value over the entire dataset and SSD_i the SSD within group i. (Of course it follows that sum $SSD_{bg} + SSD_{wg}$ is exactly the total variability of the entire data).

Again drawing a parallel to the t-test, computation of the variance is associated with the number of degrees of freedom (df) within the sample, which as seen earlier is $n-1$ in the case of an n-sized sample. One might then reasonably suppose that SSD_{bg} has $df_{bg} = k - 1$ degrees of freedom and SSD_{wg}, $df_{wg} = \sum_{i=1}^{k} n_i - 1$. The mean of the squared deviates (MSD) in each case provides a measure of the variance between and within groups respectively and is given by $MSD_{bg} = \frac{SSD_{bg}}{df_{bg}}$ and $MSD_{wg} = \frac{SSD_{wg}}{df_{wg}}$.

If the null hypothesis is false, then one would expect the variability between groups to be substantial in comparison to that within groups. Thus MSD_{bg} may be thought of in some sense as

$MSD_{hypothesis}$ and MSD_{wg} as MSD_{random}. This evaluation is formalized through computation of the

$$F - ratio = \frac{MSD_{bg}/df_{bg}}{MSD_{wg}/df_{wg}}$$

It can be shown that the F-ratio obeys the F-distribution with degrees of freedom df_{bg}, df_{wg}; thus p-values may be easily assigned.

The One-Way ANOVA assumes independent and random samples drawn from a normally distributed source. Additionally, it also assumes that the groups have approximately equal variances, which can be practically enforced by requiring the ratio of the largest to the smallest group variance to fall below a factor of 1.5. These assumptions are especially important in case of unequal group-sizes. When group-sizes **are** equal, the test is amazingly robust, and holds well even when the underlying source distribution is not normal, as long as the samples are independent and random. If the assumptions stated above do not hold and the group sizes are extremely unequal, the Unequal Variance (Welch) ANOVA is more appropriate. Similarly, when the normality assumption does not hold, the Kruskal-Wallis Test should be chosen.

24.3.10 Post Hoc Testing of ANOVA Results

A significant ANOVA result only suggests rejecting the null hypothesis H0 = ''means are the same''. It does not reveal which groups' means are significantly different. For a given gene, if any of the group pairs are significantly different, then in an ANOVA test the null hypothesis is rejected. Post hoc tests are multiple comparison procedures commonly used on only those genes that are significantly different on the basis of the ANOVA F-test. If the F-value for a factor is not significant, **GeneSpring** does not perform post hoc testing. This 'protects' the post hoc tests from being (ab)used too liberally. Post hoc tests are designed to keep the experiment wise error rate to acceptable levels.

The most common post hoc test is **Tukey's** Honestly Significant Difference or HSD test . Tukey's test calculates a new critical value that can be used to evaluate whether differences between any two pairs of means are significant. One simply calculates one critical value and then the difference between all possible pairs of means. Each difference is then compared to the Tukey critical value. If the difference is larger than the Tukey value, the comparison is significant. The formula for the critical value is:

$HSD = q\sqrt{\frac{MS_{error}}{n}}$, where q is the studentized range statistic (similar to the t-critical values, but different). MS_{error} is the mean square error from the overall F-test, and n is the sample size for each group. Error df is the df used in the ANOVA test.

SNK (Student-Newman-Keuls) test is a less stringent test compared to the Tukey HSD.

$SNK = q_r\sqrt{\frac{MS_{error}}{n}}$

Different cells have different critical values. The r value is obtained by taking the difference in the number of steps between cells and q_r is obtained from a standard table. In Tukey HSD the q value is identical to the lowest q from the SNK post hoc test.

24.3.11 The Unequal Variance (Welch) ANOVA

ANOVA assumes that the populations from which the data came all have the same variance, regardless of whether or not their means are equal. Heterogeneity in variance among different groups can be tested using Levine's test (not available in **GeneSpring**). If you suspect that the variance may not be equal and the number of samples in each group is not the same, then the Welch ANOVA should be used for statistical analysis.

In the Welch ANOVA, each group is weighted by the ratio of the number of samples and the variance of that group. If the variance of a group equals zero, the weight of that group is replaced by a large number. When all groups have zero variance and equal mean, the null hypothesis is accepted, otherwise for unequal means the null hypothesis is rejected.

24.3.12 The Kruskal-Wallis Test

The Kruskal-Wallis (KW) test is the non-parametric alternative to the One-Way independent samples ANOVA, and is in fact often referred to as ''ANOVA by rank''. The preliminaries for the KW test follow the Mann-Whitney procedure almost verbatim. Data from the k groups to be analyzed are combined into a single set, sorted, ranked and then returned to the original group. All further analysis is performed on the returned ranks rather than the raw data. Now, departing from the Mann-Whitney algorithm, the KW test computes the **mean** (instead of simply the sum) of the ranks for each group, as well as over the entire dataset. As in the One-Way ANOVA, the sum of squared deviates between groups, SSD_{bg}, is used as a metric for the degree to which group means differ. As before, the understanding is that the group means will not differ substantially in case of the null hypothesis. For a dataset with k groups of sizes $n_1, n_2, ..., n_k$ each, $n = \sum_{i=1}^{k} n_i$ ranks will be accorded. Generally speaking, apportioning these n ranks among the k groups is simply a problem in combinatorics. Of course SSD_{bg} will assume a different value for each permutation/assignment of ranks. It can be shown that the mean value for SSD_{bg} over all permutations is $(k-1)\frac{n(n-1)}{12}$. Normalizing the observed SSD_{bg} with this mean value gives us the H-ratio, and a rigorous method for assessment of associated p-values: The distribution of the

$$H - ratio = \frac{SSD_{bg}}{\frac{n(n+1)}{12}}$$

may be neatly approximated by the chi-squared distribution with $k-1$ degrees of freedom.

24.3.13 The Repeated Measures ANOVA

Two groups of data with inherent correlations may be analyzed via the paired t-Test and Mann-Whitney. For three or more groups, the Repeated Measures ANOVA (RMA) test is used. The RMA test is a close cousin of the basic, simple One-Way independent samples ANOVA, in that it treads the same path, using the sum of squared deviates as a measure of variability between and within groups. However, it also takes additional steps to effectively remove extraneous sources of variability that originate in pre-existing individual differences. This manifests itself in a third sum of squared deviates that is computed for each individual set or row of observations. In a dataset with k groups, each of size n,

$$SSD_{ind} = \sum_{i=1}^{n} k(A_i - M)^2$$

where M is the sample mean, averaged over the entire dataset and A_i is the mean of the $k$values taken by individual/row i. The computation of SSD_{ind} is similar to that of SSD_{bg}, except that values are averaged over individuals or rows rather than groups. The SSD_{ind} thus reflects the difference in mean per individual from the collective mean, and has $df_{ind} = n - 1$ degrees of freedom. This component is removed from the variability seen within groups, leaving behind fluctuations due to "true" random variance. The F-ratio, is still defined as $\frac{MSD_{hypothesis}}{MSD_{random}}$, but while $MSD_{hypothesis} = MSD_{bg} = \frac{SSD_{bg}}{df_{bg}}$ as in the garden-variety ANOVA.

$$MSD_{random} = \frac{SSD_{wg} - SSD_{ind}}{df_{wg} - df_{ind}}$$

Computation of p-values follows as before, from the F-distribution, with degrees of freedom $df_{bg}, df_{wg} - df_{ind}$.

24.3.14 The Repeated Measures Friedman Test

As has been mentioned before, ANOVA is a robust technique and may be used under fairly general conditions, provided that the groups being assessed are of the same size. The non-parametric Kruskal-Wallis test is used to analyze independent data when group-sizes are unequal. In case of correlated data however, group-sizes are necessarily equal. What then is the relevance of the Friedman test and when is it applicable? The Friedman test may be employed when the data is a collection of ranks or ratings, or when it is measured on a non-linear scale.

To begin with, data is sorted and ranked **for each individual or row** unlike in the Mann-Whitney and Kruskal-Wallis tests, where the entire dataset is bundled, sorted and then ranked. The remaining steps for the most part, mirror those in the Kruskal-Wallis procedure. The sum of squared deviates between groups is calculated and converted into a measure quite like the H measure; the difference however, lies in the details of this operation. The numerator continues to be SSD_{bg}, but the denominator changes to $\frac{k(k+1)}{12}$, reflecting ranks accorded to each individual or row.

24.3.15 The N-way ANOVA

The N-Way ANOVA is used to determine the effect due to N parameters concurrently. It assesses the individual influence of each parameter, as well as their net interactive effect.

GeneSpring uses the Type III sums of squares (SS) method when performing N-way ANOVA [53, 51]. This is equivalent to the method of weighted squares of means or complete least square method of Overall and Spiegel. The Type III SS is defined as follows:
Let A and B be the factors, each having several levels. The complete effects model for these two factors is
$y_{ijk} = \mu + a_i + b_j + t_{ij} + e_{ijk}$,
where y_{ijk} is the k-th observation in ij-th treatment group, μ is the grand mean, $a_i(b_j)$ is additive combination and t_{ij} is the interaction term and e_{ijk} is the error term, which takes into account the variation in y that cannot be accounted for by the other four terms on the right hand side of the equation. The difference in residual sums of squares (RSS) of the models
$y_{ijk} = \mu + a_i + t_{ij} + e_{ijk}$,
and
$y_{ijk} = \mu + a_i + b_j + t_{ij} + e_{ijk}$, is the SS corresponding to factor B. Similarly, for other factors we take the difference of RSS of the model excluding that factor and the full model.

The **GeneSpring** n-way ANOVA can handle both balanced and unbalanced designs, although only full factorial design is allowed. For more than three factors, terms are only calculated up to 3-way interactions due to computational complexity. Moreover, **GeneSpring** calculates a maximum of 1000 levels, i.e., if the total number of levels for a 3-way interactions model is more than 1000 (main + doublet + triplet), then **GeneSpring** calculates only up to 2-way interactions. If the number of levels is still more than 1000, **GeneSpring** calculates only the main effects.

For full factorial designs with no replicates **GeneSpring** excludes the highest level interaction (with previous constraints) to avoid over-fitting.

Missing values are handled in the **GeneSpring** n-way ANOVA. If at least one sample in a condition has a value, then the entity is included for ANOVA. But, if all the samples in a condition have missing values, then those values (entities) are excluded for p-value computation and a separate list titled 'Excluded Entities' is output at the end of the analysis. See Figure 24.1

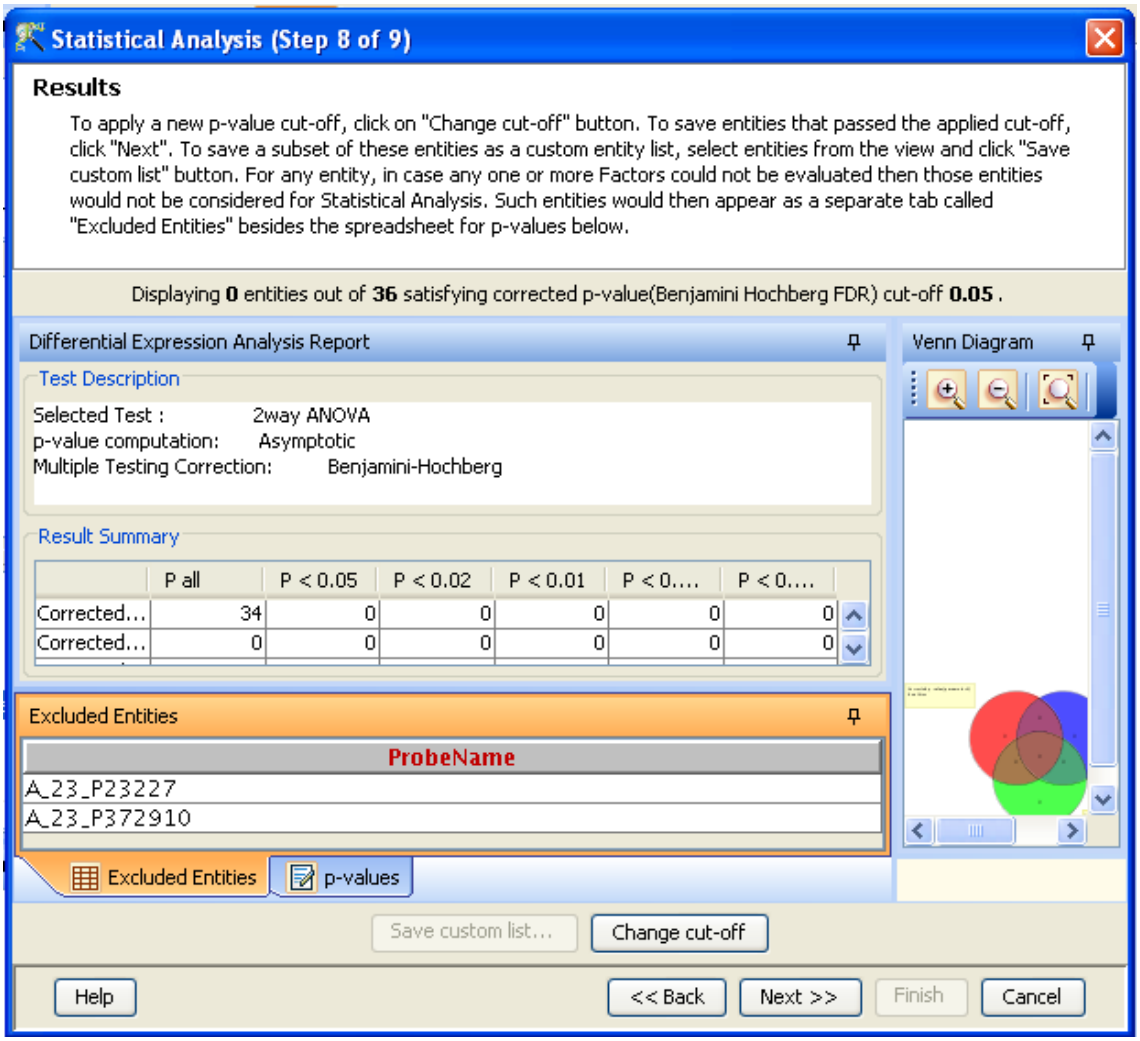


Figure 24.1: Anova result showing 'Excluded Entities' because of missing values

24.4 Obtaining p-Values

Each of the statistical tests detailed in Details of Statistical Tests in **GeneSpring** will generate a test value or statistic called the **test metric** for each gene. In general, the larger the test-metric, the more significant is the differential expression for the gene in question. To identify all differentially expressed genes, you could just sort the genes by their respective test-metrics and then apply a cut-off. However, determining that cut-off value would be easier if the test-metric could be converted to a more intuitive p-value which gives the probability that the gene g appears as differentially expressed purely by chance. So a p-value of .01 would mean that there is a 1% chance that the gene is not really differentially expressed but random effects have conspired to make it look so. Clearly, the actual p-value for a particular gene will depend on how expression values within each set of replicates are distributed. These distributions may not always be known.

Under the assumption that the expression values for a gene within each group are normally distributed and that the variances of the normal distributions associated with the two groups are the same, the computed test-metrics for each gene can be converted into p-values, in most cases using closed form expressions. This way of deriving p-values is called **Asymptotic** analysis. However, if you do not want to make the normality assumptions, a Permutative analysis method can be used.

24.4.1 p-values via Permutative Tests

As described in Dudoit et al. [20], this method does not assume that the test-metrics computed follows a certain fixed distribution.

For example:

- A spreadsheet consists of genes along the rows and arrays along columns, with the first n_1 columns belonging to the first group of replicates and the remaining n_2 columns belonging to the second group of replicates.
- The left to right order of the columns is now shuffled several times.
- In each trial, the first n_1 columns are treated as if they comprise the first group and the remaining n_2 columns are treated as if they comprise the second group; the t-statistic is now computed for each gene with this new grouping.
- This procedure is ideally repeated $\binom{n_1+n_2}{n_1}$ times, once for each way of grouping the columns into two groups of size n_1 and n_2, respectively.
- However, if this is computationally too expensive, a large enough number of random permutations are generated alternatively.

p-values for genes are now computed as follows:

- Each gene has a computed test metric and several permutation test metrics.
- For a particular gene, its p-value is the fraction of permutations in which the test metric computed is larger in absolute value than the actual test metric for that gene.

24.5 Adjusting for Multiple Comparisons - Corrected p-values

Microarray experiments typically measure several thousand genes concurrently across multiple conditions. For example, this leads to the following problem where:

- p-value for each gene has been computed and all genes with a p-value of less than .01 are considered.
- Let k be the number of such genes and each of these genes has a less than 1 in 100 chance of appearing to be differentially expressed by random chance.
- However, the chance that **at least** one of these k genes appears differentially expressed by chance is much higher than 1 in 100 (As an analogy, consider fair coin tosses, where, each toss produces heads with a 1/2 chance, but the chance of getting at least one heads in a hundred tosses is much higher). In fact, this probability could be as high as $k * .01$ (or in fact $1 - (1 - .01)^k$ if the p-values for these genes are assumed to be independently distributed).

Thus, a p-value of .01 for k genes does not translate to a 99 in 100 chance of all these genes being truly differentially expressed. However, assuming so could lead to a large number of false positives. To be able to apply a p-value cut-off of .01 and claim that all the genes which pass this cut-off are indeed truly differentially expressed with a .99 probability, an adjustment needs to be made to these p-values.

Refer Dudoit et al. [20] and the book by Glantz [23] for detailed descriptions of various algorithms for adjusting the p-values.

The simplest methods called the Holm step-down (Bonferroni Holm FWER) method and the Benjamini Hochberg step-up (Benjamini Hochberg FDR) methods are motivated by the example description in the previous paragraph. **GeneSpring** offers different types of Multiple Testing Correction to correct for family-wise error rate (FWER) (Bonferroni FWER, Bonferroni Holm FWER, Westfall Young Permutative) and correct false discovery rate (FDR) (Benjamini Hochberg FDR, Storey with Bootstrapping, Storey with Curve Fitting).

NOTE The Benjamini-Yekutieli method is only available for **GO** analysis.

The Multiple Testing Correction methods available in **GeneSpring** are:

1. No Correction
2. Bonferroni FWER
3. Bonferroni Holm FWER
4. Westfall Young Permutative

5. Benjamini-Yekutieli method (Only for **GO** Analysis)
6. Benjamini Hochberg FDR
7. Storey with Bootstrapping
8. Storey with Curve Fitting

NOTE The corrected p-value is the value obtained after a Multiple Testing Correction (MTC) has been applied on the p-value.
q-value is the name given to the adjusted p-values found using the FDR approach. So, the q-value is defined to be the FDR analogue of the p-value.

The methods are listed in order of their stringency, with the Bonferroni method being the most stringent, and the Benjamini and Hochberg FDR method being the least stringent. The more stringent a multiple testing correction, the less false positive genes are allowed. The trade-off of a stringent multiple testing correction is that the rate of false negatives (genes that are called non-significant when they are) is very high.

In the examples provided for each method, an error rate of 0.05 and a gene list of 1000 genes are assumed.

24.5.1 Bonferroni FWER

The Bonferroni method is a single step procedure, where each p-value is corrected independently of others. The p-value of each gene is multiplied by the number of genes in the gene list. If the corrected p-value is still below the error rate, the gene is considered significant.
Corrected p-value= p-value $\times n$ (number of genes in test)
As a consequence, if testing 1000 genes at a time at the same p-value, the highest accepted individual p-value for declaring significance is 0.00005, making the correction very stringent. With a FWER of 0.05 (i.e., the probability of at least one error in the family), the expected number of false positives will be 0.05.

24.5.2 Bonferroni Holm FWER (Holm step-down)

Holm's test is a stepwise method, also called a sequential rejection method, because it examines each hypothesis in an ordered sequence, and the decision to accept or reject a null hypothesis depends on the results of the previous hypothesis tests.

Entities are sorted in increasing order of p-values. The p-value of the j^{th} gene in this order is now multiplied by $(m - j + 1)$ to get the adjusted p-value. Because it is a little less corrective as the p-value increases, this correction is less conservative.

Example:

Gene Name	p-value before correction	Rank	Correction	Is gene significant after correction
A	0.00002	1	$0.00002 \times 1000 = 0.02$	$0.02 < 0.05 \rightarrow Yes$
B	0.00004	2	$0.00004 \times 999 = 0.039$	$0.039 < 0.05 \rightarrow Yes$
C	0.00009	3	$0.00009 \times 998 = 0.0898$	$0.0898 > 0.05 \rightarrow No$

24.5.3 Westfall Young Permutative

The Westfall and Young [58] permutation method takes advantage of the dependence structure between genes, by permuting data in genes between groups (conditions) in all the genes at the same time. Suppose the tests are two-sample t-tests with n_1, n_2 samples in groups (conditions) 1 and 2 respectively. Then the $n_1 + n_2$ sample units are permuted by allotting at random n_1, n_2 of the units to the two conditions. The same permutation of units will apply to all genes. Thus in all the genes, the data in the two groups will also be permuted likewise. The same test is conducted as before in each of the genes. Thus each permutation will give rise to a different set of p-values for the genes. A similar approach is used for any other test.

The Westfall and Young procedure is a permutation procedure in which genes are first sorted by increasing t-statistic obtained on unpermuted data. Then, for each permutation, the test metrics obtained for the various genes in this permutation are artificially adjusted so that the following property holds: if gene i has a higher original test-metric than gene j, then gene i has a higher adjusted test metric for this permutation than gene j. The overall corrected p-value for a gene is now defined as the fraction of permutations in which the adjusted test metric for that permutation exceeds the test metric computed on the unpermuted data. Finally, an artificial adjustment is performed on the p-values so a gene with a higher unpermuted test metric has a lower p-value than a gene with a lower unpermuted test metric; this adjustment simply increases the p-value of the latter gene, if necessary, to make it equal to the former. Alhough not explicitly stated, a similar adjustment is usually performed with all other algorithms described here as well.

NOTE This method is very time consuming because of the number of permutations to be considered.

24.5.4 Benjamini-Yekutieli method

For more general cases, in which positive dependency conditions do not apply (see Benjamini Hochberg FDR), Benjamini and Yekuteili [7] showed that replacing q with $q/\sum_{i=1}^{m}(\frac{1}{i})$ will provide control of the FDR. This control is typically applied in GO analysis, since the GO terms have both positive and negative regression dependency.

24.5.5 Benjamini Hochberg FDR

This method [6] assumes independence of p-values across genes. However, Benjamini and Yekuteili showed that the technical condition under which the test holds is that of positive regression dependency on each test statistics corresponding to the true null hypothesis. In particular, the condition is satisfied by positively correlated normally distributed one sided test statistics and their studentized t-tests. Furthermore, since up-regulation and down-regulation are about equally likely to occur, the property of FDR control can be extended to two sided tests.

This procedure makes use of the ordered p-values $P_{(1)} \leq \ldots \leq P_{(m)}$. Denote the corresponding null hypotheses $H_{(1)}, \ldots, H_{(m)}$. For a desired FDR level q, the ordered p-value $P_{(i)}$ is compared to the critical value $q.\frac{i}{m}$. Let $k = maxi : P_{(i)} \leq q.\frac{i}{m}$. Then reject $H_{(1)}, \ldots, H_{(k)}$, if such k exists.

24.5.6 The Storey method

We begin with a practical description of how Storey's method for multiple testing correction refines the Benjamini Hochberg method. Given m p-values, the Benjamini-Hochberg method essentially multiplies the ith largest p-value by m/i (If we are taking i from the highest value, then the multiplicand is $m/(m-i)$) to obtain a corrected p-value; the multiplication by m reflects the fact that the number m_0 of entities that do not behave differentially is at most m. On some datasets, m_0 could be much smaller than m, in which case, a less conservative correction could be done by multiplying by m_0 instead of m. The challenge of course is that m_0 is unknown. The mechanics of Storey's method centers largely on estimating m_0 (or alternatively, the fraction π_0 where $m_0 = \pi_0 * m$) from the given p-values. Here is a conceptual framework used for this estimation.

Consider the histogram of p-values (say, bin these p-values in bins of width .1). If none of the entities are differentially expressed, then this distribution (under certain randomness assumptions) will be a uniform distribution, i.e., the histogram will be flat. In practice however, the histogram will be typically skewed towards the left because the entities at hand are a mix of ones that are differentially expressed (so low p-values) and ones that are not (so uniformly distributed p-values). A rough way to determine π_0 is then to find the breakpoint b where the leftward skew on the histogram subsides into a uniform distribution on the right as shown in the Figure 24.2; then multiplying the fraction of entities to the right of the breakpoint by $1/(1-b)$ yields an estimate of π_0.

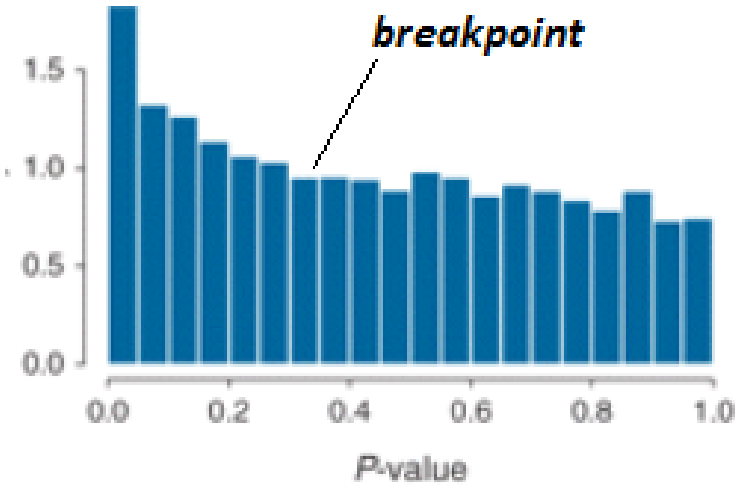


Figure 24.2: Histogram of p-values

An alternative way is to consider a different but equivalent view where one plots λ against $\frac{\#p-values>\lambda}{(1-\lambda)m}$ where λ goes from 0 to 1. If there are no differentially expressed entities then this plot will be a flat horizontal line. However, since the data at hand is a mix, with differentially expressed genes having small p-values, this plot will typically decrease as one goes to the right, often hitting a horizontal asymptote at π_0, as shown in Figure 24.3 (from [36]).

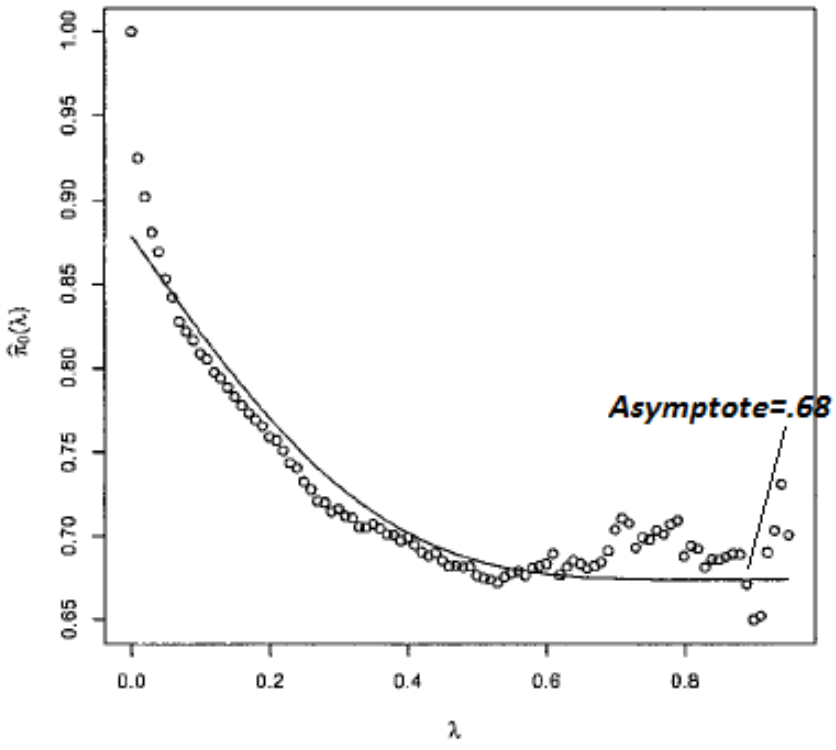


Figure 24.3: Asymptote Plot

The original work by Storey [35] used a bootstrapping method to estimate π_0. Subsequently, Storey and Tibshirani [36] obtained another method (called the smoothing method) which found the asymptote of the plot in Figure 24.3. The bootstrapping method is reliable but sometimes produces a conservative π_0 estimate of 1. The smoothing method produces better estimates but is a little more vulnerable if the data does not quite have the expected asymptotic pattern in Figure 24.3.

In **GeneSpring**, both the **bootstrapping** and **curve fitting** options are available. However, the π_0 values computed are not shown in the user interface, they are provided in the log file (`<INSTALLATION_DIRECTORY>/logs`) associated with the analysis. The GENESPRING_GX.log in the latest log folder contains this information.

A script is available (see Script to Validate Storey Expectation Maximization) to allow you to obtain the plot described in Figure 24.3 (the p-values are contained in the entity list obtained from the bootstrapping approach and this script will obtain the p-values from there) and verify if an asymptote exists and if the π_0 value computed matches this asymptote.

The eventual correction algorithm simply considers the p-values in:

- Original - decreasing
- Changed - increasing order and the corrected value q_i for the ith largest p-value p_i is obtained as $\min(\pi_0 * m/i * p_i, q_{i+1})$. (Note: Formula above is correct for increasing order.)

Unlike most other multiple testing methods, where for a given entity, the p-value is always smaller than the corrected p-value, in Storey's methods, the q-value corresponding to the p-value may be either greater or smaller depending on the position of the entity in the list sorted by ascending order of p-values. This results in less significant entities being penalized less and hence, ensures that the number of false-negatives remains minimal.

Also, please note that Storey's MTC methods are most appropriate for large entity lists which are not selectively enriched or filtered based on differential expression or splicing.

24.5.6.1 Script to Validate Storey Expectation Maximization

The script should be pasted into the Script Editor accessible from **Tools**⟶ **Script Editor**. Then click on the **Run** icon to execute the script.

```
import sys
import __builtin__
ERR_MSG = '''p-Value Column not found.
Please run Statistical Analysis and
set the p-value cut-off to 1 before finishing
'''
P_VALUE_MARK = "Stat.pvalue"
```

```
def getPValueColumn():#{{{

        activeEntityList = script.marray.project.getActiveProject().getActiveEntityList()
        dataset = activeEntityList.getDataset()

        pValueColumn=None
        colCount=dataset.getColumnCount()

        for i in range(colCount):
                column=dataset.getColumn(i)
                if column.getMetaData().getMark() == P_VALUE_MARK :
                        pValueColumn = column
                        break

        if not pValueColumn:
                script.omega.error(ERR_MSG)
                return None
        else:
                return pValueColumn

#}}}

def computeAxes(pValueColumn):#{{{

        pValueArray = []
        for i in range (0, pValueColumn.getSize()):
                pValueArray.append(pValueColumn.getFloat(i))
        pValueArray.sort()
        #pValueArray = [i*1.0/len(pValueArray1) for i in range(0, len(pValueArray1))]

        lambd = [x/100.0 for x in range(0,99)]
        yAxis = []
        for i in (lambd):
                values = [n for n in range(0, len(pValueArray)) if pValueArray[n] > i]
                yAxis.append(len(values)/(len(pValueArray)*(1-i)))

        return lambd, yAxis
#}}}

def plotGraph(lambd, yAxis):#{{{

        col1 = script.dataset.createFloatColumn("lambda", lambd)
        col2 = script.dataset.createFloatColumn("pi0", yAxis)
        dataset = script.dataset.createDataset("data", [col1,  col2])

        script.view.ScatterPlot(dataset=dataset, title="Asymptote Plot").show()
#}}}

pValueColumn = getPValueColumn()
if pValueColumn:
        lambd, yAxis = computeAxes(pValueColumn)
        plotGraph(lambd, yAxis)
```

24.5.7 Recommendations

1. The default multiple testing correction method is the Benjamini and Hochberg FDR method. It is the least stringent of all corrections and provides a good balance between discovery of statistically significant genes and limitation of false positive occurrences.
2. The Bonferroni FWER method is the most stringent test of all, but offers the most conservative approach to control for false positives.
3. The Westfall and Young Permutation method and Benjamini and Yekuteili method are the only correction methods accounting for genes' coregulation. However, the Westfall and Young Permutation method is slow and also very conservative.

24.5.8 FAQ

1. **Q.** Why do I get more genes with a smaller gene list than with all genes list when I perform a one-way ANOVA using a Multiple Testing Correction?

   A. As multiple testing corrections depend on how many genes are tested, the larger the gene list, the more stringent the correction will be. For instance, the Bonferroni correction will multiply the p-values of each gene by the number of tests performed. The more tests (or the more genes, since there is one test per gene), the smaller the p-value must be to pass the restriction.

2. **Q.** Why should I use a Multiple Testing Correction? If I select one, no genes pass the restriction.

   A. Even though no genes pass the statistical restriction, it is important to keep in mind that genes that pass a restriction without multiple testing correction might all be false positives, thus not significant at all. If you have 10,000 genes in your genome, and perform a statistical analysis, a p-value cut-off of 0.05 allows a 5% chance of error. That means that 500 genes out of 10,000 could be found to be significant by chance alone.

3. **Q.** What should I do if no genes pass the statistical test when I apply the multiple testing correction?

   A. To improve your statistical results, try one or more of the following suggestions:

   - Increase the p-value cut-off or error rate.
   - Increase the number of replicates in your experiment.
   - Select a smaller list of genes to use with your analysis. The smaller the list, the less stringent the multiple testing correction will be.
   - Select a less stringent or no multiple testing correction. If you choose to apply no multiple testing correction, rank the genes by their p-values to inspect them manually. Genes with the smallest p-values will be the most reliable.

4. **Q.** When I increase the p-value cut-off, suddenly a lot of genes pass a critical value. What is the reason for this?

   A. Typically this case arises when a permutative test is performed. If the number of permutations is small then the minimum uncorrected p-value is large, say only 0.03. Hence a

large number of genes can artificially have a p-value of 0.03. When you increase the p-value cut-off from 0.01 to 0.03 then this large number of genes will pass the cut-off.

Chapter 25

Clustering Analysis

This chapter describes in detail the different clustering analysis options available in **GeneSpring**. Clustering analysis is an efficient way to organize genes/entities and conditions in the dataset into clusters. Clusters can be based on the similarity of their expression profiles using different clustering and data characterization techniques. The results of clustering analysis are represented visually in the form of cluster trees, cluster sets, or U-matrices.

NOTE

For imported **Strand NGS** experiments (see Importing Strand NGS experiments into GeneSpring on page 613, region clustering on any imported region lists is also enabled. Please refer to the **Strand NGS** user manual for a description of this functionality.

25.1 Getting Started

To begin Clustering Analysis in **GeneSpring**, you can change the default settings for:

1. Dendrogram visualization
2. Clustering Algorithms

This section describes the options available to modify the default parameters for dendrogram visualization and the **GeneSpring** clustering algorithms.

25.1.1 Dendrogram visualization

Use **Tools** ⟶ **Options** ⟶ **Views** ⟶ **Dendrogram** and modify the default settings as required. You can modify these settings at any time.

NOTE Once you modify the default settings, the view on re-launching a previously created dendrogram will reflect the altered settings.

25.1.2 Default options for Clustering Algorithms

Use **Tools** ⟶ **Options** ⟶ **Clustering Algorithms** and modify the default settings of any of the three available algorithms as required. You can modify these settings at any time. See Hierarchical Clustering, K-Means Clustering, or Self-Organizing Map Clustering for details on how to perform Clustering Analysis using these algorithms.

NOTE Once you modify the default settings, only the subsequent results of Clustering Analysis will reflect the altered settings.

25.2 Clustering Analysis in GeneSpring

This section describes in detail the required steps for each of the available clustering analyses. This section also describes the available options for analyzing the saved results in the form of a cluster tree or cluster set once they have been saved as part of the experiment.

CAUTION The results of clustering analysis on samples or conditions are sensitive to baseline transformation performed during experiment creation. Create a new experiment with a modified baseline transformation parameter, or during experiment creation choose **Do not perform baseline** to change the input data for clustering analysis.

25.2.1 How to Perform Hierarchical Clustering Analysis

Hierarchical clustering is used as a method of showing relationships among the expression levels of entities over a series of conditions, or the expression levels of samples or conditions. In **GeneSpring** you can also perform hierarchical clustering on both entities and conditions combined. In the resulting cluster trees, entities or conditions with similar expression patterns are placed closer to each other.

Use the following procedure to set up and run hierarchical clustering analysis.

1. Select **Clustering** from the **Analysis** section of the **Workflow** browser. The **Clustering** window opens. See "Clustering Window Common Features" on page 699.
2. Click **Choose...** next to the **Entity List** field and select the entity list that contains the set of entities you want to analyze in the **Choose Entity List** dialog box. Click **OK**. See "Choose Entity List" on page 712.

NOTE Clustering is performed only on entities in the selected entity list. It is recommended (but not mandatory) to perform clustering on a set of entities filtered for reliability and/or statistically significant changes in expression by t-test or ANOVA. The selected entity list in the **Experiment Navigator** (the one that appears in bold) is selected by default.

---

3. Click **Choose...** next to the **Interpretation** field and select the interpretation that you want to analyze in the **Choose Interpretation** dialog box. Click **OK**. See "Choose Interpretation" on page 713.

NOTE Interpretations, other than the **All Samples** interpretation, have to be created before starting the **Clustering** workflow (see "Create Interpretation") to be available for selection in this workflow step. The selected interpretation in the Experiment Navigator (the one that appears in bold) is selected by default.

---

4 Select **Hierarchical** from the **Clustering Algorithm** drop-down list. See "Clustering Window - Hierarchical Clustering Parameters" on page 701.

5 Select by which type of data you want to perform clustering: **Normalized intensity values** or **Associated values**.

For **Normalized intensity values**:

a Select one of the following parameters from the **Cluster on** drop-down list: **Conditions**, **Entities**, **Both entities and conditions**.

b Select one of the following parameters from the **Distance metric** drop-down list: **Eucledian**, **Squared Eucledian**, **Manhattan (City-block)**, **Maximum (Chebychev)**, **Minimum**, **Differential**, **Canberra**, **Harmonic**, **Pearson's Centered**, **Pearson's Uncentered (Cosine)**, **Pearson's Centered - Absolute**, or **Pearson's Uncentered - Absolute**. See "Distance Metric" on page 742.

c Select one of the following parameters from the **Linkage rule** drop-down list: **Ward's**, **Centroid**, **Median**, **Single**, **Complete**, or **Average**. See "Linkage Rule" on page 744.

d Select whether you want to **Cluster within conditions** or not. This option is only available if you selected a non-averaged interpretation for clustering.

For **Associated values**:

e Select the desired **Associated values**.

NOTE At least two items (associated value types) are required to perform clustering on conditions or clustering on both entities and conditions.

---

f Select the remaining parameters as described in Steps 5a - c of this section.

6 Click **Preview**.

GeneSpring performs clustering analysis according to your chosen parameters and displays the result as a cluster tree in the results pane on the right side of the **Clustering** window.

7 Review the cluster tree.

If you selected **Conditions** or **Entities** from the **Cluster on** drop-down list in Step 4a of this section, a cluster tree is displayed with the following two panes:

- Summary view (to the left for Conditions and on the top for Entities)
- Detailed view (to the right for Conditions and at the bottom for Entities)

If you selected **Both entities and conditions** a combined tree is displayed with two panes:

- Summary view (left pane)
- Detailed view (right pane)

To navigate the combined cluster tree, use the right-click options to switch between the **Condition Tree Navigation Mode** and the **Entity Tree Navigation Mode**.

See "Clustering result visualizations in **GeneSpring**" on page 696.

8 You can change any of the parameter settings in the pane on the left and rerun the analysis by clicking **Preview**.

9 When you are ready to save your clustering result in the experiment, click **OK**.

10 Review the details that will be saved in our experiment along with the cluster tree in the **Inspector** that opens. You can edit the default **Name** and **Notes**. See "Inspector" on page 725.

11 Click **OK** to save the hierarchical clustering tree in your experiment.

The cluster tree is saved as an object in your experiment hierarchy below the selected input entity list.

25.2.2 How to Perform K-Means Clustering Analysis

K-means clustering divides entities into groups based on their expression patterns. The goal is to produce groups of entities with high degree of similarity within each group and low degree of similarity between groups. K-means clusters are constructed so that the average behavior in each group is distinct from any of the other groups. For example, in a time series experiment you could use K-means clustering to identify unique classes of entities that are upregulated or downregulated in a time-dependent manner.

Use the following procedure to set up and run a K-means clustering analysis.

1 Select **Clustering** from the **Analysis** section of the **Workflow** browser. The **Clustering** window opens. See "Clustering Window Common Features" on page 699.

2 Click **Choose...** next to the **Entity List** field and select the entity list that contains the set of entities you want to analyze in the **Choose EntityList** dialog box. Click **OK**. See "Choose Entity List" on page 712.

NOTE Clustering is performed only on entities in the selected entity list. It is recommended (but not mandatory) to perform clustering on a set of entities filtered for reliability and/or statistically significant changes in expression by t-test or ANOVA. The selected entity list in the **Experiment Navigator** (the one that appears in bold) is selected by default.

3 Click **Choose...** next to the **Interpretation** field and select the interpretation that you want to analyze in the **Choose Interpretation** dialog box. Click **OK**. See "Choose Interpretation" on page 713.

NOTE Interpretations, other than the **All Samples** interpretation, have to be created before starting the **Clustering** workflow (see "Create Interpretation") to be available for selection in this workflow step. The selected interpretation in the Experiment Navigator (the one that appears in bold) is selected by default.

4 Select **K-Means** from the **Clustering Algorithm** drop-down list.

The options under Parameters change to the relevant parameters for K-Means clustering. See "Clustering Window - K-Means Clustering Parameters" on page 703.

5 Select one of the following parameters from the **Cluster on** drop-down list: **Conditions**, **Entities**, **Both entities and conditions**.

6 Select one of the following parameters from the **Distance metric** drop-down list: **Eucledian**, **Squared Eucledian**, **Manhattan (Cityblock)**, **Maximum (Chebychev)**, **Minimum**, **Differential**, **Canberra**, **Harmonic**, **Pearson's Centered**, **Pearson's Uncentered (Cosine)**, **Pearson's Centered - Absolute**, or **Pearson's Uncentered - Absolute**. See "Distance Metric" on page 742.

7 Enter values for the following options for the clustering analysis:

- **Number of Clusters** - The number of clusters to make. The default setting is 3.
- **Number of Iterations** - The maximum number of times that each centroid is recalculated after entities are reassigned to groups with the most similar centroids. The default setting is 50.

NOTE For clustering on Conditions and Both entities and conditions (see Step 5), the number of clusters has to be lower than the number of experimental conditions in the selected interpretation. For clustering on Entities, the number of clusters has to be lower than the number of entities in the selected entity list.

8 Review the clustering result in the results pane on the right:

- If you selected **Conditions** in Step 4, one or several cluster set views, depending on the **Number of clusters** specified in Step 6, are displayed in the results panel. All the entities (from the input entity list) appear on the x-axis and are grouped into clusters based on their proximity of expression profiles.

- If you selected **Entities**, one or several cluster set views, depending on the Number of clusters specified in Step 6, are displayed in the results panel. All the conditions (from the selected interpretation) appear on the x-axis.
- If you selected **Both entities and conditions**, the output view contains two tabs: one containing the cluster set view(s) for **Entities**, the other containing the cluster set view(s) for **Conditions**.

See "Clustering result visualizations in **GeneSpring** on page 696.

9 You can change any of the parameter settings in the pane on the left and rerun the analysis by clicking **Preview**.

10 When you are ready to save your clustering result in the experiment click **OK**.

11 Review the details that will be saved in your experiment along with the cluster tree or cluster set representing the K-Means clustering result in the **Inspector** that opens. You can edit the default **Name** and **Notes**. See "Inspector" on page 725.

12 Click **OK** to save the cluster sets in your experiment.

The cluster tree, cluster sets, or both are saved as objects in your experiment hierarchy below the selected input entity list.

25.2.3 How to Perform Self-Organizing Map Clustering Analysis

The self-organizing map (SOM) is a clustering technique that illustrates the relationship between groups by arranging them in a two-dimensional map and by dividing entities into groups based on expression patterns. SOMs are useful for visualizing the number of distinct expression patterns in your data and determining which of these patterns are variants of one another.

Use the following procedure to set up and run self-organizing map clustering analysis.

1 Select **Clustering** from the **Analysis** section of the **Workflow** browser. The **Clustering** window opens. See "Clustering Window Common Features" on page 699.

2 Click **Choose...** next to the **Entity List** field and select the entity list that contains the set of entities you want to analyze in the **Choose EntityList** dialog box. Click **OK**. See "Choose Entity List" on page 712.

NOTE Clustering is performed only on entities in the selected entity list. It is recommended (but not mandatory) to perform clustering on a set of entities filtered for reliability and/or statistically significant changes in expression by t-test or ANOVA. The selected entity list in the **Experiment Navigator** (the one that appears in bold) is selected by default.

3 Click **Choose...** next to the **Interpretation** field and select the interpretation that you want to analyze in the **Choose Interpretation** dialog box. Click **OK**. See "Choose Interpretation" on page 713.

NOTE Interpretations, other than the **All Samples** interpretation, have to be created before starting the **Clustering** workflow (see "Create Interpretation") to be available for selection in this workflow step. The selected interpretation in the Experiment Navigator (the one that appears in bold) is selected by default.

---

4 Select **Self-Organizing Map** from the **Clustering Algorithm** drop-down list.

The options under **Parameters** change to the relevant parameters for SOM clustering.

5 Select one of the following parameters from the **Cluster on** drop-down list: **Conditions**, **Entities**, **Both entities and conditions**.

6 Select one of the following parameters from the **Distance metric** drop-down list: **Angular**, **Eucledian**, **Squared Eucledian**, **Manhattan (City-block)**, **Maximum (Chebychev)**, **Differential**, **Pearson's Centered - Absolute**, or **Pearson's Centered**, **Pearson's Uncentered (Cosine)**. See "Distance Metric" on page 742.

7 Enter values for the following clustering options:

- **Maximum number of iterations** - the number of times to move the nodes. This parameter also indirectly defines how many times each entity is examined. For example, if there are 10,000 entities and 60,000 iterations are specified, each entity is examined six times.
- **Number of grid rows** - the number of rows in your map.
- **Number of grid columns** - the number of columns in your map.
- **Initial learning rate** - the learning rate at the start of iterations. This parameter determines the extent of adjustment of the reference vectors and decreases monotonically to zero with each iteration.
- **Initial neighborhood radius** - the number of nodes moving towards a data point at the beginning of the iterations. This parameter is an indicator of how similar the profiles are for each node.

NOTE A good way to estimate the optimum number of rows and columns is to try to predict how many distinct classes of entities are affected by the conditions in your experiment. With small data sets, the algorithm may generate a number of empty nodes. To avoid this, try using a smaller map.

---

8 Select one of the following parameters from the **Grid topology** drop-down list: **Hexagonal** or **Rectangular**.

9 Select one of the following parameters from the **Neighborhood type** drop-down list: **Bubble** or **Gaussian**. A **Bubble** neighborhood defines a fixed circular area, whereas a **Gaussian** neighborhood defines an infinite area. Only nodes lying in the neighborhood are updated when a gene is assigned to a winning node. However, the update adjustment decreases exponentially as a function of distance from the winning node.

10 Select whether to **Run in batch mode** or not. **GeneSpring** performs a faster, simple version of SOM when batch mode is enabled, resulting in a quick overview.

11 Click **Preview**.

GeneSpring performs clustering analysis according to your chosen parameters and displays the result in the results pane on the right side of the **Clustering** window.

12 Review the clustering result in the results pane on the right:

- If you selected **Conditions** in Step 4 above, the result consists of two tabs: Dendrogram and U-Matrix.
- If you selected **Entities**, the result consists of three tabs: Dendrogram, Cluster Set, and U-Matrix.
- If you selected **Both entities and conditions**, the result consists of four tabs: Dendrogram, Cluster Set, U-Matrix (for entities), and U-Matrix (for conditions).

See "Clustering result visualizations in **GeneSpring**" on page 696.

13 You can change any of the parameter settings in the pane on the left and rerun the analysis by clicking **Preview**.

14 When you are ready to save your clustering result in the experiment click **OK**.

15 Review the details that will be saved in your experiment along with the dendrogram and cluster set representing the SOM clustering result in the **Inspector** that opens. You can edit the default **Name** and **Notes**. See "Inspector" on page 725.

16 Click **OK** to save the dendrogram and cluster set in your experiment.

The cluster tree, cluster set, or both are saved as objects in your experiment hierarchy below the selected input entity list.

NOTE Interpretations, other than the **All Samples** interpretation, have to be created before starting the **Clustering** workflow (see "Create Interpretation") to be available for selection in this workflow step. The selected interpretation in the Experiment Navigator (the one that appears in bold) is selected by default.

25.2.4 How to Perform Clustering Analysis on Region Lists

In **GeneSpring**, you can perform clustering analysis on region lists of imported NGS and CGH experiments. This section describes in detail the different steps of the wizard- driven Clustering on region list option for CGH experiments only.

NOTE For details on region clustering in imported NGS experiments, please refer to the **Strand NGS** user manual.

Run clustering on region list is used as a method to:

- Identify the groups of samples having similar characteristics after performing analysis.
- Identify the groups of regions having similar characteristics e.g. SNPs

Use the following procedure to set up and run hierarchical clustering analysis on region lists.

1. Right-Click on a region list in the **Experiment Navigator**and select **Run clustering**.
   Step 2 of 5 of the **Clustering on region list** wizard opens.
2. Select which **Region List Columns** you want to include or exclude for the clustering analysis by moving the desired columns from the **Available items** field to the **Selected items** field or vice versa using the move right and the move left buttons.
   Click **Next** >> to proceed to Step 3 of 5.
3. Select the parameters to perform clustering:
   a. Select one of the following parameters from the **Cluster on** drop-down list: **Regions**, **Samples**, **Both Regions and Samples**.
   b. Select one of the following parameters from the **Distance metric** drop-down list: **Eucledian**, **Squared Eucledian**, **Manhattan (City-block)**, **Maximum (Chebychev)**, **Minimum**, **Differential**, **Canberra**, **Harmonic**, **Pearson's Centered**, **Pearson's Uncentered (Cosine)**, **Pearson's Centered - Absolute**, or **Pearson's Uncentered - Absolute**. See "Distance Metric" on page 742.
   c. Select one of the following parameters from the **Linkage rule** drop-down list: **Ward's**, **Centroid**, **Median**, **Single**, **Complete**, or **Average**. See "Linkage Rule" on page 744.
   Click **Next** >> to proceed to Step 4 of 5.
4. **GeneSpring** performs clustering analysis on region list according to your chosen parameters and displays the result as a dendrogram view.
   Review the dendrogram and use the dendrogram toolbar icons to change the orientation of region tree.
   Click **Next** >> to proceed to Step 5 of 5.

5 This step displays a summary of the result that will be saved. Click **Finish** to save the result of **Region List Clustering Analysis**.

The **Hierarchical cluster tree(s)** is saved as objects in your experiment hierarchy below the selected input region list. You can launch the dendrogram by double clicking on the node or using Right-Click options on the tree.

25.2.5 How to Explore a Cluster Tree Result in the Dendrogram View

Several results from the various clustering analyses available in **GeneSpring** are saved as part of the experiment in the form of cluster trees.

You can view such a hierarchical condition, entity, or combined trees using the following operations:

1 To view the hierarchical tree in the Dendrogram View (see "The Dendrogram View" on page 707) in the main display area:

- Double-click the tree of interest in the Experiment Navigator.
- Right-click the tree of interest in the Experiment Navigator and select **Open Tree**.

2 To view the various parts of a entity tree or a condition tree:

- Click on a node to view the subtree.
- Press the <Shift> key and use the arrow keys to view neighboring nodes in a left hand tree or in a top tree.
- Click the **Redo** icon or **Undo** icon in the Dendrogram View Toolbar.

For a combined tree:

- Right-click on the tree and select **Condition Tree Navigation Mode** or **Entity Tree Navigation Mode** depending on which tree you want to explore using the same methods described for entity trees and condition trees.

3 Hover over a node, tree, or entity of interest to see the respective values (distance metric between nodes or intensity value of an entity) in the tool tip.

25.2.6 How to Visualize Sample Metadata in the Dendrogram View

GeneSpring allows visualizing sample parameters (i.e., sample metadata) alongside the sample cluster tree obtained from Hierarchical or SOM clustering analysis. Sample metadata such as phenotype, clinical parameters, demographics, time points, drug dosage, quality assessment and many more can all be imported into **GeneSpring** and visualized.

Sample metadata can be imported into **GeneSpring** via Experiment Grouping (see Experiment Grouping on page 236) as sample parameters. **GeneSpring** supports visualizing both numeric and non-numeric sample parameter values.

Sample metadata can be visualized in **GeneSpring** as Heatmap, Scatter Plot, Profile Plot, Bar Chart, and Label Plot. Heatmap and Label Plot are available for both numeric and non-numeric parameters. Scatter Plot, Profile Plot, and Bar Chart are only supported for numeric parameters.

These sample parameter plots provide visual clues to the underlying metadata patterns if any in relation to Clustering Analysis performed. In addition, the sort functionality on a plotted sample parameter can help to discover patterns within the visualized metadata.

NOTE Sample parameter plots are not supported for a cluster tree resulting from Clustering Analysis on associated values.

NOTE If you performed Clustering Analysis with an averaged interpretation as input, you can only add numerical sample parameter plots to the dendrogram. In this case, the arithmetic mean of the parameter values within each condition is plotted against the cluster tree. Non-numerical sample parameters cannot be plotted against a cluster tree if an averaged interpretation was used as input to Clustering Analysis.

NOTE The option to **Add sample parameter plots** is available only in the default orientation (**Entity Tree Top** $\longrightarrow$**Bottom**) of a cluster tree. Changing the orientation of the dendrogram disables the option to add sample parameter plots.

You can view and explore the sample parameter plots using the following operations:

1 Click on the **Add sample parameter plots** icon in the **Dendrogram View Toolbar**.
A drop-down is displayed with the available sample parameter plot views and an option to **Remove all Plots**.

2 Select a sample parameter plot view.
The sample parameter plot is displayed in the **Detailed View** pane of the dendrogram.

- You can add multiple sample parameter Heat Maps by selecting parameters from the **Choose Parameters** dialog box.
See Choose Parameters on page 718 for more information.

NOTE Once you have added a sample parameter plot to the **Dendrogram View**, you can no longer change the orientation of a cluster tree.

---

3 On a sample parameter plot, you can perform the following actions:

- In a **Sample parameter Heat Map**, click on the **Add/Remove/Re-order sample parameters** icon to add, remove, or re-order the sample parameters to be displayed.
See Add/Remove/Re-order Parameters on page 719 for more information.
- By default **Sample parameter Bar Chart**, **Scatter Plot**, and **Profile Plot** are displayed for the first available numeric parameter type in the experiment grouping information dialog. **Sample Parameter Label Plot** is displayed for the first available parameter (numeric or non-numeric) in the experiment grouping information dialog. Click on the **Change plot parameter** icon to choose a different sample parameterto be displayed.
- Click on the **Sort on parameter** icon to sort the dendrogram heatmap and visualized sample metadata on the sample parameter plot order (ascending).
- Click on the **Change sort order** button to change the sorting order (descending).
- Click on the **Reset sort order** button to reset the sorting order and return to the default sorting order.
- Click on the **Remove plot** icon to close a sample parameter plot.

NOTE Sorting on a sample parameter in any of the sample parameter plots removes the condition cluster tree of the dendrogram heatmap. There are multiple ways to return the heatmap to its clustered state (including the cluster tree):

- Click on the **Run clustering** icon in the **Dendrogram View Toolbar**
- Click the sorting icons until you have returned to the original order and the cluster tree is restored.
- Remove all sample parameter plots using the **Remove all Plots** option in the **Add sample parameter plots** icon drop-down.

---

4 Right-click on a sample parameter plot and select **Properties** to view and modify properties for each individual plot.

   See Properties of Sample parameter Heat Map on page 735, Properties of Sample parameter Scatter Plot on page 736, Properties of Sample parameter Profile Plot on page 737, Properties of Sample parameter Bar Chart on page 738, and Properties of Sample parameter Label Plot on page 739 for more information on changing properties of sample parameter plots.

5 Left-click on the border of a sample parameter plot and drag to adjust the size.

NOTE In a sample parameter heatmap, up to six parameters can be displayed within the default parameter plot width. If you chose to display more than six parameters in the same plot, increase the width of the plot to bring the additional parameter heatmaps into view.

---

6 Click on the **Reset width of sample parameter plots** icon in the Dendrogram View Toolbar to return to the default width for all the modified sample parameter plots.

NOTE Once you modify the default view of a sample parameter plot, the view on re-launching a previously created sample parameter plot will reflect the modified view settings.

25.2.7 Working with a Cluster Tree in the Dendrogram View

This section describes the viewing and analysis options that are common to a cluster tree (condition tree or combined tree) that was created from clustering analysis. These options include all the functions available in the Dendrogram View of a cluster tree through right-click menus and the Dendrogram View Toolbar.

25.2.7.1 To open a cluster tree

- Double-click the cluster tree name in the Experiment Navigator.
- Right-click over the cluster tree name and select **Open Condition/Combined Tree**.

The cluster tree opens in the Dendrogram View in the main display area. Depending on the type of cluster tree, the displayed entities and conditions are influenced by the active interpretation and active entity list in the Experiment Navigator as follows:

- **Entity Tree**: An entity tree always displays the entities from the entity list from which it was created grouped by the conditions of the active interpretation in the Experiment Navigator. Clicking on another entity list or interpretation in the Experiment Navigator after the tree has opened does not result in any changes to the open entity tree.
- **Condition Tree**: A condition tree always displays the entities and conditions based on which it was originally created when it is opened and displayed in the main display area. Changing the active entity list in the Experiment Navigator while the tree filters the displayed entities by the active entity list. As a result, only those entities that overlap between the original entity list and the active entity list are displayed in the condition tree. Changing the active interpretation in the Experiment Navigator also results in changes in the display of the condition tree. Overlapping samples are labelled (by color bar and tool-tip according to the grouping of the active interpretation. Non-overlapping samples are labelled as N/A.
- **Combined Tree**: A combined tree behaves like a condition tree with one exception. Changing the active entity list in the Experiment Navigator does not affect the display of the tree.

See the The Dendrogram View on page 707 for more information on the individual components of the view.

25.2.7.2 To inspect a cluster tree

- Right-click the cluster tree name in the Experiment Navigator and select **Inspect Entity Tree**, **Inspect Condition Tree**, or **Inspect Combined Tree**.

An Entity Tree Inspector, Condition Tree Inspector, or Combined Tree Inspector opens. The inspector holds all the available information associated with the

creation of the cluster tree, including its Name, Creation date, Last modified date, and Owner. The Notes contain all the parameter settings that were used to create the cluster tree.

You can edit the **Name** and **Notes**.

See "Entity Tree Inspector" on page 717, "Condition Tree Inspector" on page 716, and "Combined Tree Inspector" on page 715.

25.2.7.3 To copy a cluster tree to the My Favorites folder

1. Right-click over the cluster tree name in the Experiment Navigator and select **Copy**.
2. Right-click the **My Favorites** folder in the Experiment Navigator and select **Paste**.

A copy of the cluster tree is saved in the **My Favorites** folder.

25.2.7.4 To delete a cluster tree

- Right-click the cluster tree name in the Experiment Navigator and select **Delete Condition Tree**, **Delete Entity Tree**, or **Delete Combined Tree** depending on the cluster tree type.

The cluster tree is permanently deleted from **GeneSpring**.

25.2.7.5 To hide or show the summary view of a cluster tree

You can hide or show the summary view of a cluster tree. Summary view is useful to view the area displayed by the detailed view when it is in a zoomed in state.

- Click the **Hide summary view** icon in the **Dendrogram View Toolbar** to close the summary view panel.
- Click the **Show summary view** icon in the **Dendrogram View Toolbar** to open the summary view panel.

25.2.7.6 To toggle full screen mode of a cluster tree

You can view a cluster tree in a fullscreen mode. Fullscreen view is useful to maximize the view of a cluster tree.

- Click the **Toggle fullscreen mode** icon in the **Dendrogram View Toolbar** to open the cluster tree in the main display area and re-click the icon to return back to the normal mode.

25.2.7.7 To restore the condition cluster tree in the Detailed View after sorting on a sample parameter

Condition cluster trees are removed from the Detailed View of the heatmap when you sort on the parameter in a sample parameter in a sample parameter plot To add sample parameter plots to a dendrogram on page 694 and perform sorting on them. You can re-run clustering to get back the condition cluster tree of the heatmap in the Detailed View.

- Click the **Run clustering** icon in the **Dendrogram View Toolbar** to bring the condition cluster tree of the heatmap back into view.
- Restore the sorting order to the original cluster order using the sorting icons available in a parameter plot.
- Remove all sample parameter plots using the corresponding option in the **Add sample parameter plots** icon drop-down menu.

25.2.7.8 To change the orientation of a cluster tree

You can change the direction of the cluster tree from the default for better visualization of the cluster result. Changing the direction of a cluster tree can be useful, for example, if your screen size is more suited to display a large number of entities horizontally, rather than vertically.

- Click the **Orient entity tree left -> right** icon in the **Dendrogram View Toolbar**.
- Click the **Orient entity tree right > left** icon in the **Dendrogram View Toolbar**.
- Click the **Orient entity tree top > bottom** icon in the **Dendrogram View Toolbar**.
- Click the **Orient entity tree bottom > top** icon in the **Dendrogram View Toolbar**.

25.2.7.9 To change the display settings of a cluster tree

- Click the **Properties** icon in the **Dendrogram View Toolbar**.
  The Properties dialog box opens. Change the color scheme of the cluster tree, transparency, display or remove entity labels and rulers using the options provided in the Colors, Columns, Labels, Rulers, and Description tabs. See "Properties" on page 726.

25.2.7.10 To change the display settings of a sample parameter Heatmap

- Right-click on the sample parameter heatmap and select **Properties**.
  The **Properties** dialog box opens. Change the heatmap width for all parameters and color scheme for each of the selected numeric parameters. See "Properties of sample parameter Heat Map" on page 735.

NOTE You can only customize heatmap color settings for numerical parameters.

25.2.7.11 To change the display settings of a sample parameter Scatter Plot

- Right-click on the sample parameter scatter plot and select **Properties**.
  The **Properties** dialog box opens. Change the color settings for the scatter plot, choose to show or hide X- and Y-Axis labels, and change the color settings for the X- and Y-Axis of the scatter plot. See "Properties of sample parameter Scatter Plot" on page 736.

25.2.7.12 To change the display settings of a sample parameter Profile Plot

- Right-click on the sample parameter Profile Plot and select **Properties**.
  The **Properties** dialog box opens. Change the color settings for the profile plot, change the profile line width, choose to show or hide X- and Y-Axis labels, and change the color settings for the X- and Y-Axis of the profile plot. See "Properties of sample parameter Profile Plot" on page 737.

25.2.7.13 To change the display settings of a sample parameter Bar Chart

- Right-click on the sample parameter Bar Chart and select **Properties**.
  The **Properties** dialog box opens. Change the color settings for both positive and negative bar charts, change the bar width, choose to show or hide X- and Y-Axis labels, and change the color settings for the X- and Y-Axis of the bar chart. See "Properties of sample parameter Bar Chart" on page 738.

25.2.7.14 To change the display settings of a sample parameter Label Plot

- Right-click on the sample parameter Label Plot and select **Properties**.
  The **Properties** dialog box opens. Change the **Label plot color** and **Label alignment** settings. See "Properties of sample parameter Label Plot" on page 739.

25.2.7.15 To save a subtree of a cluster tree

To save a subsection of a cluster tree as a new cluster tree, use the following procedure:

1. Select the portion of the cluster tree that you want to save by clicking on the corresponding node.
2. Click the **Save subtree** icon in the **Dendrogram View Toolbar**.

The subtree is saved as a new cluster tree object below the parent tree in the Experiment Navigator.

25.2.7.16 To search and inspect selected entities in a cluster tree

1. In the Dendrogram View click Ctrl + F. A **Search** dialog opens.
2. Search for entities of interest from the **Find** field or Ctrl-select the entities of interest.
3. Click **Close** to close the Search dialog and return to the Dendrogram View. A green bar is associated with the selected entities.
4. Click the **Inspect selected entity** icon in the main toolbar.

You can inspect all available details on the selected entities in the Entity Inspector that opens.

25.2.7.17 To inspect selected entities from a cluster tree

1. Left-click and drag over the heatmap area to select the desired clusters. The zoomed view area displays only the selected clusters. A green bar is associated with the entities indicating the selection.
2. Click the **Select all in zoomed view** icon in the **Dendrogram View Toolbar**.
3. Click the **Inspect selected entity** icon in the main toolbar.

You can inspect all available details on the selected entities in the Entity Inspector that opens.

25.2.7.18 To create a new entity list from selected entities in a cluster tree

1. Left-click and drag over the heatmap area to select the desired clusters. The zoomed view area displays only the selected clusters.
2. Click the **Select all in zoomed view** icon in the Dendrogram View Toolbar.
3. Click the **Create entity list from selection** icon in the main toolbar. An **Entitylist Inspector** opens.
4. Edit the new entity list's default **Name** and **Notes** if desired.
5. Click **OK** to save the new entity list in your experiment.

The new entity list is saved in the Experiment Navigator as a child list of the input entity list that you selected to create the cluster tree.

25.2.7.19 To view entities of another entity list in a cluster tree

1. From the Experiment Navigator, select an entity list different to the entity list used to run Clustering Analysis.
2. Click the **Properties** icon in the **Dendrogram View Toolbar** or right-click on the Dendrogram View and select **Properties**. Select the **Highlight matched entities** option.

   In an hierarchical entity or combined tree:

   - Only those entities present in the selected active entity list from the experiment navigator will be displayed in the Dendrogram View. A red bar is associated with the entities indicating their presence in the active entity list.
   - Entities absent in the active entity list are hidden. You can change the transparency display settings from the Properties dialog box to visualize the hidden entities with a different transparency (See "Properties" on page 726).

   In a hierarchical condition tree:

   - The Condition Tree Heatmap View displays only entities from the selected active entity list and filters out all unmatched entities.
   - The Cluster Tree View is unchanged.

25.2.7.20 To remove condition color bars of a cluster tree

1. Click the **Properties** icon in the **Dendrogram View Toolbar** or right-click on the Dendrogram View and select **Properties**. Select the **Highlight matched entities** option.
2. Deselect **Show condition bar**.

   In an hierarchical entity, condition, or a combined tree:

   - The conditions color bars are removed from the cluster tree.

25.2.7.21 To export a cluster tree as a report

You can save the visible cluster tree as a report or download as a **.pdf* file.

Choose an option to export the cluster tree:

- **Summary View**: The current Summary View pane is exported.
- **Detailed View**: The current Detailed View pane is exported.
- **Combined View**: The current Combined View of the cluster tree is exported.

NOTE If a report contains > 100 pages (default setting), it cannot be viewed in **GeneSpring**. You can download the report as a **.pdf* or modify report settings from **Tools** ⟶**Options** ⟶**Miscellaneous** ⟶**Report**.

To export a cluster tree with entity labels:

1. Select a node in the **Summary View**.
2. Click the **Show all entity labels** icon.
3. Right-click on either of the panes, select **Export as Report** and choose one of the available options from, **Summary View**, **Detailed View**, or **Combined View**.
4. Modify the other parameters as required and click **Preview**.
5. Click **Save Report** to save the report in the experiment navigator. This report will include the **Summary View** pane, or **Detailed View** pane, or the **Combined View** along with the **Simple** or **Advanced** report parameters as selected.
6. If the Report has > 100 pages, download the report as a **.pdf* file using the **Download** icon.

NOTE To export a cluster tree without entity labels, do not click on the **Show all entity labels** icon.

NOTE Exporting a **Detailed View** or **Combined View** cluster tree as a report includes all open sample parameter plots in the report.

25.2.7.22 To export a cluster tree as an image

You can export a complete cluster tree as an image (*.tiff, *.png, *.jpg, *.jpeg, *.bmp, *.svg, *.eps).

Choose an option to export the cluster tree:

- **Summary View**: The current Summary View pane is exported.
- **Detailed View**: To export the complete Detailed View, uncheck the **Export only the visible region** and proceed. To export only the current Detailed View Pane, use the default options and proceed.
- **Combined View**: The current Combined View of the cluster tree is exported.

NOTE Exporting a complete **Detailed View** in the default *.tiff, *.svg, *or .eps format will result in a single image of the detailed cluster tree. If you select a big region or the complete **Detailed View** to export, selecting any other image format will result in a folder with a collection of images representing the Detailed View of the cluster tree.

To export a cluster tree with entity labels:

1. Select the top most node in the **Summary View**.
2. Click the **Show all entity labels** icon.
3. Right-click on either of the panes, select **Export as Image** and choose one of the available options (see above).
4. Specify a location to save the image and choose an image format.
5. Clear the option **Export only the visible region** to export the complete **Detailed View** with entity labels to the specified location.
6. Use the default setting for **Export only the visible region** to export only the visible region. This image will include the **Summary View** pane, or **Detailed View** pane, or the **Combined View** as selected.
7. Modify the other parameters as required and click **OK**.

NOTE To export a cluster tree without entity labels, do not click on the **Show all entity labels** icon.

25.2.7.23 To export a cluster tree with sample parameter plots as an image

You can export a complete cluster tree with the sample parameter plots as an image (*.tiff, *.png, *.jpg, *.jpeg, *.bmp, *.svg, *.eps).

Choose an option to export the cluster tree:

- **Detailed View**: To export the complete Detailed View, uncheck the **Export only the visible region** and proceed. To export only the current Detailed View Pane with all the sample parameter plots, use the default options and proceed.
- **Combined View**: The current Combined View of the cluster tree is exported with the sample parameter plots.

NOTE

Exporting a complete **Detailed View** in the default *.tiff, *.svg, *or .eps format will result in a single image of the detailed cluster tree with the sample parameter plots. If you select a big region or the complete Detailed View to export, selecting any other image format will result in a folder with a collection of images representing the Detailed View of the cluster tree.

To export a cluster tree with sample parameter plots:

1. Specify a location to save the image and choose an image format.
2. Clear the option **Export only the visible region** to export the complete **Detailed View** with sample parameter plots to the specified location.
3. Use the default setting for **Export only the visible region** to export only the visible region with the sample parameter plots. This image will include the **Detailed View** pane, or the **Combined View** as selected along with the sample parameter plots.
4. Modify the other parameters as required.
5. All sample parameter plots that are open in the **Dendrogram View** are selected by default in **Select plots to export**.
   Click **OK** and the dendrogram is exported with all the sample parameter plots.
6. Select which **Sample parameter plots** you want to include or exclude for the export by moving the desired plots from the **Selected plots** field to the **Available plots** field or vice versa using the move left and the move right buttons.
7. Reorder the selected plots, or refresh the order using the up , down , and refresh order buttons.
8. Click **OK**.

25.2.7.24 To export entities in cluster order

You can export the entities in a cluster tree (entity trees and combined trees only) and selected associated values in the clustered order in a tab-delimited file format (*.txt, *.tsv).

1. Right-click over the cluster tree and select **Export Entities in Cluster Order**. The **Export Options** dialog box opens. See "Export options" on page 720.
2. Click **Choose...** to select an interpretation for exporting the entities in the **Choose Interpretation** dialog box. See "Choose Interpretation" on page 713.
3. Select all the data columns you want to export, from these options: **Raw signal values**, **Normalized signal values**, **Flags**, **Entitylist data**.
4. Select which **Annotation** columns you want to export by moving the desired annotations from the **Available items** field to the **Selected items** field or vice versa using the move right and move left buttons.
5. Reorder the selected items, or refresh the order using the up , down , and refresh order buttons.
6. Click OK.
   A **Save** dialog box opens.
7. Provide a file name and determine where you want to save the list. Click **Save**.

25.2.7.25 To create a classification from a subset of clusters in an entity tree

The summary view of an entity tree includes a ruler that indicates the distance metric values. Dragging the square along the ruler highlights all the clusters with distance metric values below the value indicated on the ruler. The value selected on this ruler is also used as a threshold for creating a classification of the highlighted clusters as described in this section.

Creating a classification of entities in a cluster results in a maximum of 20 clusters. If there are more than 20 clusters with a distance metric below the selected threshold, individual cluster sets are created for the first 19 clusters (from left to right along the cluster tree) and the remaining clusters are summarized in the 20th cluster.

1. Set the desired threshold on the distance metric ruler in the summary view of the displayed entity tree.
   All the clusters with a distance metric value below the selected threshold appear highlighted in the cluster tree.
2. Click the **Create Classification** icon in the **Dendrogram View Toolbar**.
   A **Classification Inspector** opens. The **Notes** contain information about the cluster tree and a record of the selected threshold value. The inspector also indicates the **Number of Clusters** that was created. See "Classification Inspector" on page 714.

3 Click **OK**.
The classification is saved in the Experiment Navigator below the entity list that was selected to create the cluster tree.

4 Double-click the new Classification object in the Experiment Navigator to view the profile plots of each cluster in the main desktop area of **GeneSpring**.

NOTE It is possible to visualize the classification result using the following plot views: profile plot, box-and-whisker plot, violin plot, and histogram. To do so, launch the desired plot using the corresponding toolbar icon so that this plot type is the active view. Then double-click the classification result node in the Experiment Navigator and **GeneSpring** displays the corresponding view for each cluster.

25.2.7.26 To run GO Analysis on an entity tree

You can perform GO Analysis on any entity tree, or the entity tree part of a combined tree. In **GeneSpring**, there is a difference between performing GO Analysis on a cluster tree compared to performing GO Analysis on an entity list. Performing GO Analysis on an entity list identifies all the GO terms that are significantly enriched for entities in the selected entity list based on a chosen p-value cut-off. In contrast, performing GO Analysis on an entity tree identifies clusters with a minimum number of entities (as specified by you) that are significantly enriched for a selected GO term based on a chosen p-value cut-off.

Use the following procedure to perform GO Analysis on a cluster tree:

1 Click the **Run GO Analysis** icon in the **Dendrogram View Toolbar**. The **GO Analysis on entity cluster tree** dialog box opens. See "GO analysis on entity cluster tree" on page 724.

2 Select the GO term for which you want to identify enriched clusters from the GO tree. The GO tree contains all the GO terms available for the experiment technology. Use the **Search** facility below the tree to find a particular GO term for which you already know the name or the reference number.
All the entities associated with the selected GO term that are present in the open experiment are displayed under **All Genes**.

NOTE GO Analysis on a cluster tree is only performed on those entities from the entity list that was used to create the cluster tree that overlap with the entities displayed in the **All Genes** field of the **GO Analysis on entity cluster tree** dialog box.

---

3 Specify the **Minimum number of entities** using the slider or the field. If you use the field, press the Enter key to confirm the value. **GeneSpring**

only considers clusters with more entities than the specified number for GO Analysis.

4 Specify the **p-value cut-off** using the slider or the field. If you use the field, press the Enter key to confirm the value. A p-value is only calculated for clusters that satisfy the **Minimum number of entities** parameter. The default p-value cut-off is 0.1.

5 Click **OK**.

The **GO Analysis completed** information dialog indicates the number of significant nodes that was identified.

6 Click **OK**.

Significant nodes are indicated in the cluster tree by a yellow dot.

NOTE The result of GO Analysis on an entity tree is not saved when you close the cluster tree. To keep a record of the result, export the cluster tree as a report or an image using the available right-click options.

25.2.7.27 To add sample parameter plots to a dendrogram

Use the following procedure to visualize sample parameter plots on a **Dendrogram View**:

1 Click on the **Add sample parameter plots** icon in the **Dendrogram View Toolbar**.

A drop-down is displayed with the available sample parameter plot views.

2 Select a sample parameter plot view.

The sample parameter plot is displayed in the **Detailed View** pane of the dendrogram.

3 If you select the **Heat Map** option, you can select one or multiple parameter(s) to visualize as Heat Map(s) from the **Choose Parameters** dialog box as follows:

- Select which **Parameters** you want to view as **Heat Maps** by moving the desired parameters from the **Available parameters** field to the **Selected parameters** field or vice versa using the move right and move left buttons.
- Reorder the selected parameters, or refresh the order using the up , down , and refresh order buttons.
- Click **OK**.

Sample parameter Heat Maps are displayed for the selected parameters.

4 A **sample parameter Bar Chart**, **Scatter Plot**, and **Profile Plot** are displayed for the first available numeric parameter type in the experiment grouping information dialog by default. A **sample Parameter Label Plot**

is displayed for the first available parameter (numeric or non-numeric) in the experiment grouping information dialog. In a **sample parameter Bar Chart**, **Scatter Plot**, **Profile Plot**, or **Label Plot**, click on the **Change plot parameter** icon to change the sample parameters to be displayed.

5 Click **OK**.

See How to Visualize Sample Parameter Plots in the Dendrogram View on page 680 for more information.

25.2.8 Clustering result visualizations in GeneSpring

The results of clustering analysis are best represented visually and **GeneSpring** incorporates a number of intuitive visualizations for that purpose. All the result views are interactive and allow you to further explore the results. In addition you can extract selected entities from a view and create new entity lists from these entities for further analysis.

The following sections describe the purpose and functionality of each of the available result visualizations in **GeneSpring**: cluster trees, cluster sets or classifications, and U-matrices.

25.2.8.1 Cluster tree

Clustering algorithms like the hierarchical clustering algorithm produce a grouping hierarchy rather than distributing data into a fixed number of clusters. The most similar entities are merged together to form a cluster. These combined entities are treated as a unit thereafter until all the entities are grouped together. The result is a tree structure or a cluster tree, where the leaves represent individual entities and the internal branches (nodes) represent clusters of similar entities.

The leaves are the smallest units corresponding to one entity or condition each. Every entity is labelled with its identifier drawn from the Gene Symbol column in the dataset. Each node in the tree defines a cluster. The distance at which two clusters merge is a measure of dissimilarity between clusters, which is called the threshold distance. Threshold distance is determined as the height of the node from the leaf and depends on the chosen distance metric for clustering.

The Dendrogram View in **GeneSpring** shows the same tree in two panels. The view is split into two panels vertically for a condition tree and horizontally for an entity tree. The left/top panel shows the entire tree while the right/bottom panel shows a zoomed in view of the selected area in the entire tree. You can navigate along the nodes of the expanded tree to inspect certain sections of the cluster tree more closely.

Condition labels and color bars representing the parameters in each interpretation are displayed between the cluster tree and the heatmap. Condition cluster tree color can be modified to reflect any experiment grouping parameter. The type of color can be determined through **Colors** tab of the **Properties** dialog box. See Properties" on page 726.

Entity labels and color bars for each of the entities are displayed beside/below the heatmap. Entity labels can reflect any one of the associated values available for the entities in the selected entity list. The type of associated value can be determined through the **Labels** tab of the **Properties** dialog box. See Properties" on page 726.

NOTE Multiple color bands for conditions and color bars for entities are not available in the preview of a cluster tree result in the Clustering window. In the saved view, selected interpretation overlay from the experiment navigator for condition color bands is allowed only for cluster tree created on a non-averaged interpretation.

Clustering on entities creates an entity tree object. On opening an entity tree view, the tree is shown with all the entities in the selected entity list, with the columns of the active interpretation in the experiment.

Clustering on conditions creates a condition tree object. On opening a condition tree, the tree is shown with the columns being the conditions on which clustering was performed, with the rows being the active entity list in the experiment.

Clustering on both entities and conditions results in two cluster trees; a vertical cluster tree for entities, and a horizontal one for conditions. They are displayed together in the form of a combined clustering tree.

The legend shows the color range of the heat map and the conditions of the interpretation on which clustering was performed.

Hovering over the cells of the heat map shows the normalized intensity values in a tool tip. The tool tip on the row header and the column header shows the complete entity name or condition respectively. The tool tip over the tree shows the distance values corresponding to the distance measure used in the clustering algorithm. The tool tip on the condition color bar shows the conditions and the experimental parameter values for the interpretation.

See "How to Perform Hierarchical Clustering Analysis" on page 671, "How to Perform K-Means Clustering Analysis" on page 673, and "How to Perform Self-Organizing Map Clustering Analysis" on page 675.

25.2.8.2 Cluster Set or Classification

The K-Means clustering algorithm generates a fixed number of clusters k as specified by you in the parameters. The Cluster Set View resulting from performing K-Means clustering graphically displays the profile of each cluster. Clusters are labelled as Cluster 0, Cluster 1, Cluster 2, ..., Cluster k-1.

NOTE It is possible to visualize the classification result using the following plot views: profile plot, box-and-whisker plot, violin plot, and histogram. To do so, launch the desired plot using the corresponding toolbar icon so that this plot type is the active view. Then double-click the classification result node in the Experiment Navigator and **GeneSpring** displays the corresponding view for each cluster.

See "How to Perform K-Means Clustering Analysis" on page 673 and "How to Perform Self-Organizing Map Clustering Analysis" on page 675.

25.2.8.3 U-Matrix

The U-Matrix view is used to display results of the SOM clustering algorithm, which also generates a fixed number of clusters, like the K-Means algorithm. The U-Matrix view displays clusters arranged in a two-dimensional grid such that similar clusters are physically closer to each other in the grid. The grid cells can either be hexagonal or rectangular. Cells in the grid are of two types: nodes and non-nodes. Nodes and non-nodes alternate in the grid. Clusters are associated only with nodes and each node displays the reference vector or the average expression profile of all entities mapped to the node. This average expression profile is plotted in blue. The purpose of non-nodes is to indicate the similarity between neighboring nodes on a gray scale. If a non-node between two nodes is very bright, it indicates that the two nodes are very similar. If the non-node is comparatively darker, then the two nodes are very different. Furthermore, the shade of a node reflects its similarity to its neighboring nodes. Thus this view displays average cluster profiles as well as how the various clusters are related. A left-click on a node displays the Profile Plot for the associated cluster of entities.

See "How to Perform Self-Organizing Map Clustering Analysis" on page 675.

25.3 Clustering Analysis Screen Reference

This section describes the user interface elements of the Clustering window, as well as toolbar icons, dialog boxes, and right-click options associated with the workflow and its results.

25.3.1 Clustering Window Common Features

This window allows you to perform various types of clustering analyses. The purpose of clustering is to divide entities or conditions into groups that have similar expression patterns, which leads to the following advantages:

- Uncovering patterns of entity expression data and the relationships between these patterns.
- Reducing the complexity of your data.
- Discovering entities that are primarily responsible for the variation.

The various clustering options can help you identify entities that are potentially co-regulated as well as reveal coordinated responses to experimental treatments.

The Clustering window consists of two panes.

The left pane changes dynamically and provides the appropriate parameter settings for Hierarchical (see "Clustering Window - Hierarchical Clustering Parameters" on page 701), K-Means (see "Clustering Window - K-Means Clustering Parameters" on page 703), and SOM (see "Clustering Window - SOM Clustering Parameters" on page 704) clustering analyses.

The right pane provides a preview of the analysis results (see "Result Preview Pane" on page 706. If a result consists of more than one object, for example a cluster set view and a cluster tree, you can view the desired object by clicking on the respective tab. If the result is a cluster tree, the right pane also contains a toolbar with options to change the orientation of the cluster tree and move to previous or next cluster (see "Dendrogram View Toolbar" on page 708).

The following features are common to all clustering analyses.

25.3.1.1 Entity List Choose... button

The **Entity List** field displays the active entity list in the Experiment Navigator (appears in bold) by default. Click **Choose...** to change the entity list if required. Select the desired entity list in the **Choose EntityList** dialog box that opens. See "Choose Entity List" on page 712.

NOTE For best results, do not select the All Entities list for clustering analysis. Best results are obtained when the list of entities has been filtered to remove entities with measurements that are mostly in the noise range or that are mostly flagged absent.

25.3.1.2 Interpretation Choose... button

The **Interpretation** field displays the active interpretation in the Experiment Navigator by default. Click **Choose...** to change the interpretation if required. Select the desired interpretation in the **Choose Interpretation** dialog box that opens. See "Choose Interpretation" on page 713.

Some of the clustering analyses can only be performed if the chosen interpretation satisfies certain additional conditions. For example, to perform hierarchical clustering on normalized signal values with the **Cluster within conditions** option selected, the selected interpretation has to be a non-averaged interpretation with at least two conditions and at least two samples per condition.

Some of the clustering parameters provided by you are also dependent on the details of your chosen interpretation. For example, in K-Means clustering on conditions, the Number of clusters has to be less than the number of conditions in the selected interpretation when performing clustering on conditions, or both entities and conditions.

25.3.1.3 Clustering Algorithm drop-down list

Select the clustering algorithm from this drop-down list:

- Hierarchical (see "How to Perform Hierarchical Clustering Analysis" on page 671)
- K-Means (see "How to Perform K-Means Clustering Analysis" on page 673)
- Self-Organizing Map (see "How to Perform Self-Organizing Map Clustering Analysis" on page 675)

The **Parameter** section of the left pane changes according to your selection from this drop-down list to display only the relevant parameter settings for your chosen algorithm.

25.3.1.4 Preview button

Click **Preview** to start the analysis. Results appear in the preview pane on the right side of the Clustering window. After reviewing the result, you can change any of the parameters in the left panel and click Preview again to observe how the different parameters affect the result.

25.3.1.5 OK button

Click **OK** to save the result displayed in the preview pane on the right as one or several objects in your experiment. The number of saved objects depends on the selected clustering algorithm and parameters.

NOTE Only cluster trees and cluster set views are saved in the Experiment Navigator below the input entity list. If you want to keep a record of the result displayed in a U-matrix, use the right-click options to export the result as an image.

25.3.1.6 Cancel button

Closes the Clustering window. Any ongoing analysis is aborted. Your results, if any, are not saved in the experiment. **GeneSpring** does not save the last settings that you specified in the left pane. The next time you start clustering analysis, the default settings are displayed.

25.3.1.7 Help button

Click Help to find out more about clustering analysis and the available parameter settings in the Clustering window. The user manual opens in your default browser.

25.3.2 Clustering Window - Hierarchical Clustering Parameters

When you select **Hierarchical** from the **Clustering Algorithm** drop-down list in the **Clustering** window (see "Clustering Window - Common Features" on page 699), you can cluster by either normalized intensity values or associated values. Depending on which type of data you choose, the appropriate parameters are visible in the **Parameters** section of the left pane.

See "How to Perform Hierarchical Clustering Analysis" on page 671.

25.3.2.1 Cluster by normalized intensity values

Cluster on drop-down list

Select on what you want to cluster the selected entity list and interpretation:

- Conditions
- Entities
- Both entities and conditions.

Distance metric drop-down list

Select the desired method to calculate the similarity between the entities or conditions for clustering from these options: **Eucledian**, **Squared Eucledian**, **Manhattan (Cityblock)**, **Maximum (Chebychev)**, **Minimum, Differential**, **Canberra**, **Harmonic**, **Pearson's Centered**, **Pearson's Uncentered (Cosine)**, **Pearson's Centered - Absolute**, or **Pearson's Uncentered - Absolute**.

See "Distance Metric" on page 742 for the mathematical formulas for each of these options.

Linkage rule drop-down list

Select the desired method for representing the connection between clusters visually in the cluster tree from these options: **Ward's**, **Centroid**, **Median**, **Single**, **Complete**, **Average**.

See "Linkage Rule" on page 744 for more details on each of these options.

Cluster within conditions

This option is only available if you selected a non-averaged interpretation for clustering. Select this option to cluster only within the conditions of this non-averaged interpretation, rather than across all conditions, which is the default setting.

NOTE When Cluster within conditions is selected, similarity across conditions is not computed. However, in the resulting cluster tree, the clusters for each condition still appear connected. This option can be used as a form of quality control. Replicate samples within an experiment should cluster closer to each other than non-replicate samples.

25.3.2.2 Cluster by associated values

Select this option to cluster by any values other than normalized signal values that are associated with the entities in the selected entity list. In addition to the parameter options described under **Cluster by normalized intensity values** on page 701, the associated values present in the selected entity list appear in a table below the other parameter options.

25.3.3 Clustering Window - K-Means Clustering Parameters

When you select **K-Means** from the **Clustering Algorithm** drop-down list in the **Clustering** window (see "Clustering Window - Common Features" on page 699) the following parameters appear in the **Parameters** section of the left pane.

See "How to Perform K-Means Clustering Analysis" on page 673.

25.3.3.1 Cluster on drop-down list

Select on what you want to cluster the selected entity list and interpretation:

- Conditions
- Entities
- Both entities and conditions.

25.3.3.2 Distance metric drop-down list

Select the desired method to calculate the similarity between the entities or conditions for clustering from these options: **Eucledian**, **Squared Eucledian**, **Manhattan (Cityblock)**, **Maximum (Chebychev)**, **Minimum, Differential**, **Canberra**, **Harmonic**, **Pearson's Centered**, **Pearson's Uncentered (Cosine)**, **Pearson's Centered - Absolute**, or **Pearson's Uncentered - Absolute**.

See "Distance Metric" on page 742 for the mathematical formulas for each of these options.

25.3.3.3 Number of clusters

Determines the number of clusters (k) formed during K-Means clustering. This number should be less than the number of conditions in the selected interpretation for clustering on conditions and less than the number of entities in the selected entity list for clustering on entities.

25.3.3.4 Number of iterations

Determines the number of times the algorithm is applied to the entities or conditions to converge on the result.

25.3.4 Clustering Window - SOM Clustering Parameters

When you select **Self-Organizing Map** from the **Clustering Algorithm** drop-down list in the **Clustering** window (see "Clustering Window - Common Features" on page 699) the following parameters appear in the **Parameters** section of the left pane.

See "How to Perform Self-Organizing Map Clustering Analysis" on page 675.

25.3.4.1 Cluster on drop-down list

Select on what you want to cluster the selected entity list and interpretation:

- Conditions
- Entities
- Both entities and conditions.

25.3.4.2 Distance metric drop-down list

Select the desired method to calculate the similarity between the entities or conditions for clustering from these options: **Angular**, **Eucledian**, **Squared Eucledian**, **Manhattan (Cityblock)**, **Maximum (Chebychev)**, **Differential**, **Pearson's Centered - Absolute**, or **Pearson's Centered**, **Pearson's Uncentered (Cosine)**.

See "Distance Metric" on page 742 for the mathematical formulas for each of these options.

25.3.4.3 Maximum number of iterations

Determines the maximum number of times the algorithm is applied to the entities or conditions to converge on the result.

25.3.4.4 Number of grid rows

Specify a positive integer for the number of rows for the grid. The default value is 3.

25.3.4.5 Number of grid columns

Specify a positive integer for the number of columns for the grid. The default value is 4.

25.3.4.6 Initial learning rate

Specify the initial learning rate. This parameter defines the learning rate at the start of the iterations. It determines the extent of adjustment of the reference vectors and decreases monotonically to zero with each iteration. The default value is 0.03.

25.3.4.7 Initial neighborhood radius

Specify the initial neighborhood radius. This parameter defines the neighborhood extent at the start of the iterations. This radius decreases monotonically to 1 with each iteration. The default value is 5.

25.3.4.8 Grid topology

Choose either hexagonal or rectangular for visualization of the two-dimensional grid of the resulting U-Matrix. By default the grid topology is set to hexagonal.

25.3.4.9 Neighborhood type

Choose from **Bubble** or **Gaussian**. This parameter determines the extent of the neighborhood. Only nodes lying in the neighborhood are updated when a gene is assigned to a winning node. A **Bubble** neighborhood defines a fixed circular area, whereas a **Gaussian** neighborhood defines an infinite extent. However, the update adjustment decreases exponentially as a function of distance from the winning node. By default the neighborhood type is set to **Bubble**.

25.3.4.10 Run in batch mode

Select this option to run a faster, simpler version of SOM to get a quick overview of the result.

25.3.5 Result Preview Pane

The result output displayed in the Preview pane of the **Clustering Window** depends on the selected clustering algorithm and whether you chose to cluster on entities, conditions, or both. The varying output views are listed by algorithm and selected **Cluster on** option in Table 25.1. See "Clustering result visualizations in **GeneSpring**" on page 696 for a description of each of the result views.

Table 25.1: Result output views by clustering parameters

Clustering algorithm	Cluster on option	Output views
Hierarchical	Conditions	Cluster tree
	Entities	Cluster tree
	Both entities and conditions	Cluster tree
K-Means	Conditions	Cluster set view
	Entities	Cluster set view
	Both entities and conditions	Two tabs; one displaying the cluster set view of entities and the other displaying the cluster set view of conditions.
Self-Organizing Map	Conditions	Two tabs; one displaying the cluster tree and the other displaying the U-Matrix view.
	Entities	Three tabs; one displaying the cluster tree, one displaying the cluster set view, and another one displaying the U-Matrix view.
	Both entities and conditions	Four tabs; one displaying the cluster tree, one displaying the cluster set view, one displaying the U-Matrix view for entities and another one displaying the U-Matrix view for conditions.

25.3.6 The Dendrogram View

The following sections describe all the display components and functional elements of the Dendrogram View: view components, the dendrogram-specific toolbar, and right-click options.

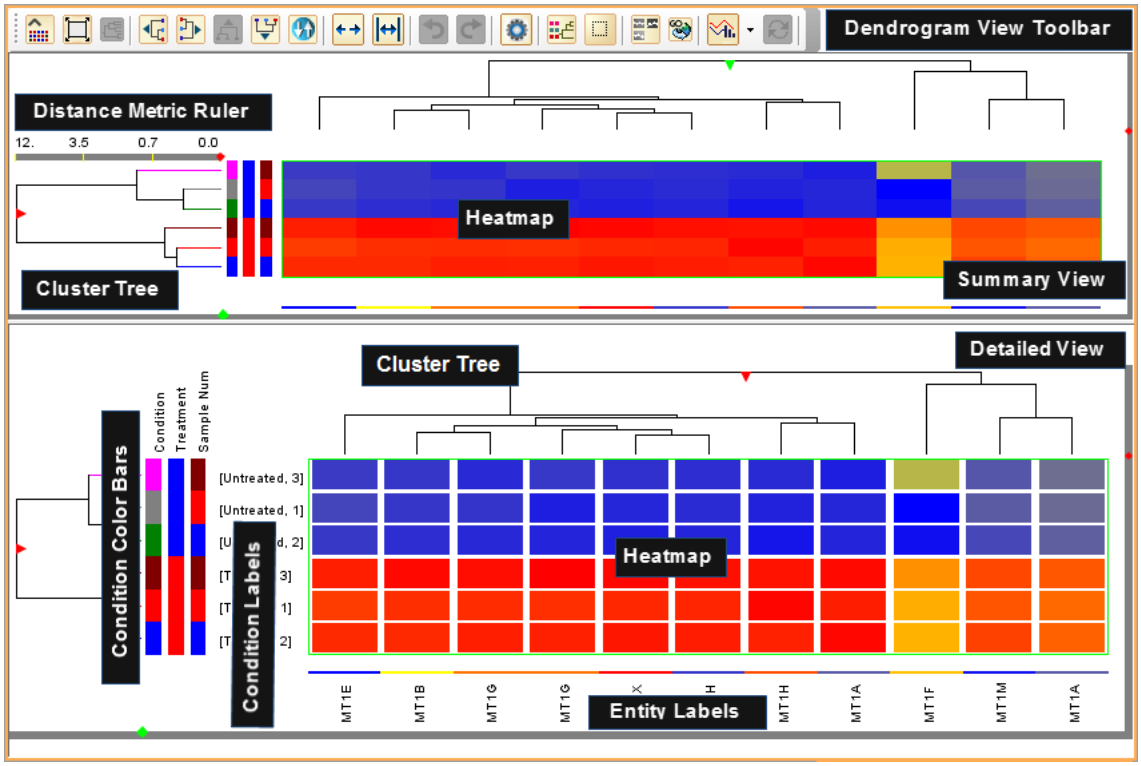


Figure 25.1: Dendrogram View of a combined tree

25.3.6.1 Dendrogram View Components

The Dendrogram View consists of a **Summary View** pane and a **Detailed View** pane.

The Summary View always displays the entire cluster tree and the distance metric ruler, if you choose to display it. To view a subtree of the cluster tree, you can use any one of the options described under "How to explore a cluster tree result in the Dendrogram View in **GeneSpring**" on page 679. The selected subtree is displayed in the Detailed View.

In the Detailed View, you can continue to zoom into the parts of the subtree that you selected in the Summary View. At any given time, the area corresponding to the part of the cluster tree that is displayed in the Detailed View is highlighted in the Summary View.

25.3.6.2 Dendrogram View Toolbar

The Dendrogram View Toolbar provides access to visualization and analysis functions, including toggle full screen mode, changing the orientation of the clustering tree, flip condition tree, display all entity labels, fit all entities to a page, undo, redo, display properties, save subtree, select all in zoomed view, create classification, and run GO analysis. The availability of each function represented by an icon in the toolbar depends on the clustering algorithm used to create the clustering tree. Table 25.2 details the icons, their utility, and availability depending on the clustering algorithm.

Table 25.2: Availability of Dendrogram View Toolbar icons by clustering algorithm

Toolbar Icon	Function	Hierarchical	K-Means	Self-Organizing Map
	Hide summary view	✓	✓	✓
	Show summary view	✓	✓	✓
	Toggle fullscreen mode	✓	✓	✓
	Run clustering	✓ +	✓ +	✓ +
	Orient tree left -> right	✓ †	✓ †	✓ †
	Orient tree right -> left	✓ †	✓ †	✓ †
	Orient tree top -> bottom	✓ †	✓ †	✓ †
	Orient tree bottom -> top	✓ †	✓ †	✓ †
	Flip condition tree	✓*	*X*	*X*
	Show all entity labels	✓	✓	✓
	Fit all entities to a page	✓	✓	✓
	Undo	✓	✓	✓
	Redo	✓	✓	✓
	Properties	✓	✓	✓
	Save subtree	✓	✓	✓
	Select all in zoomed view	✓	✓	✓
	Create Classification	✓ #	*X*	*X*
	Run GO Analysis	✓ #	*X*	*X*
	Add sample parameter plots	✓	✓	✓
	Reset sample parameter plots to default size	✓ +	✓ +	✓ +

\+ - Active only when sample parameter plots are visualized.
† - Disabled when sample parameter plots are visualized.
* - Active only for Hierarchical Combined Tree on the selected interpretation.
\# - Active only for Hierarchical Entity Tree and Hierarchical Combined Tree on the selected interpretation.

25.3.6.3 Legend

The legend at the bottom right corner of the **GeneSpring** desktop area contains the color-coding legend for the cluster tree, the legend for **Entity color by**, or the legend for **Condition color by**, or both for combined cluster trees. It also contains a summary of the parameters used to create the tree. For more information on how to change the color settings see "To change the display settings of a cluster tree" on page 685.

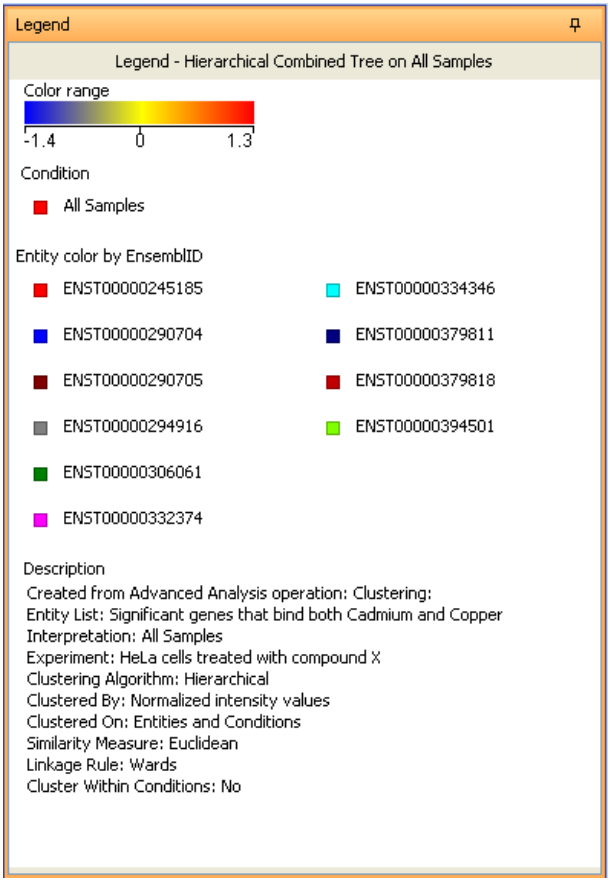


Figure 25.2: Legend of a combined cluster tree

25.3.6.4 Legend for sample parameter plots

If sample parameter plots are added to the hierarchical cluster tree view, the legend at the bottom right corner contains all the color-coding legend for the clustering tree view, summary of the parameters used to create the tree, and all the color-coding for active sample parameter plots. For more information on how to change the color settings see To change the display settings of a sample parameter Heat Map on page 686, To change the display settings of a sample parameter Scatter Plot on page 686, To change the display settings of a sample parameter Profile Plot on page 686, To change the display settings of a sample parameter

Bar Chart on page 686, and To change the display settings of a sample parameter Label Plot on page 687.

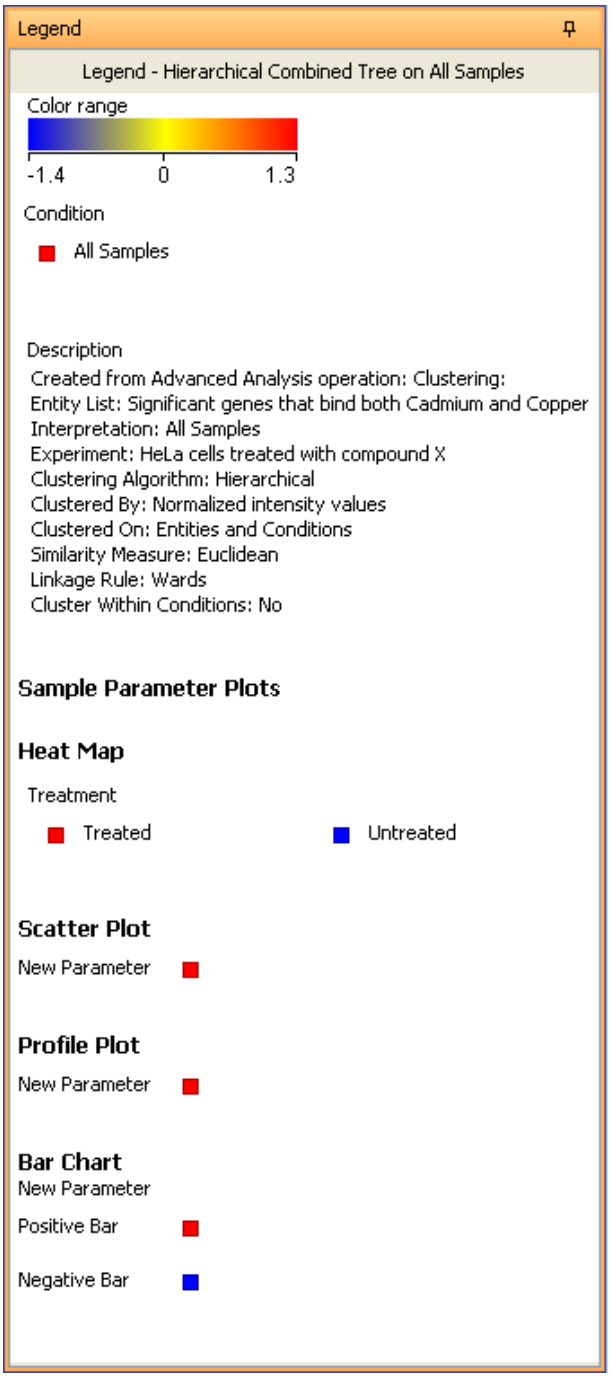


Figure 25.3: Legend of a combined cluster tree with sample parameter plots

25.3.6.5 Right-click options on a cluster tree

The available right-click options on a cluster tree are summarized in Table 25.3.

Table 25.3: Menu options available on right-click over a cluster tree view

Option	Description
Zoom Mode	Activates the zoom functions.
Selection Mode	Activates the selection of entities on the cluster tree view.
Condition Tree Navigation Mode	Activates the navigation functions for the condition tree of a combined tree.
Entity Tree Navigation Mode	Activates the navigation functions for the entity tree of a combined tree.
Select All Entities	Selects all entities in the cluster tree.
Invert Entity Selection	Inverts the entity selection.
Clear Entity Selection	Deselects any selected entities.
Select Columns	Allows you to select individual samples/columns, which consequently appear highlighted in the cluster tree. See "Select Columns" on page 740.
Invert Column Selection	Inverts the column selection.
Clear Column Selection	Deselects any selected columns.
Clear All Selections	Deselects all selected entities and columns.
Reset Zoom	Resets the zoomed level in the **Detailed View** to the selected view in the **Summary View**.
Copy View	Copies the current dendrogram view to the system clipboard.
Export As • Report • Image • HTML	Allows you to export the **Summary**, **Detailed**, or **Combined View** as a report or in any one of these image formats: *.tiff, *.png, *.jpg, *.jpeg, *.bmp, *.svg., *.eps, or the current dendrogram view in *.html format.
Export Entities in Cluster Order	Exports the entities in the open cluster tree in the cluster order to a *.txt or *.tsv file.
Properties	Allows you to change the display settings of the dendrogram view.

25.3.7 Dialog Boxes

This section describes all the dialog boxes that appear during Clustering Analysis.

25.3.7.1 Choose Entity List

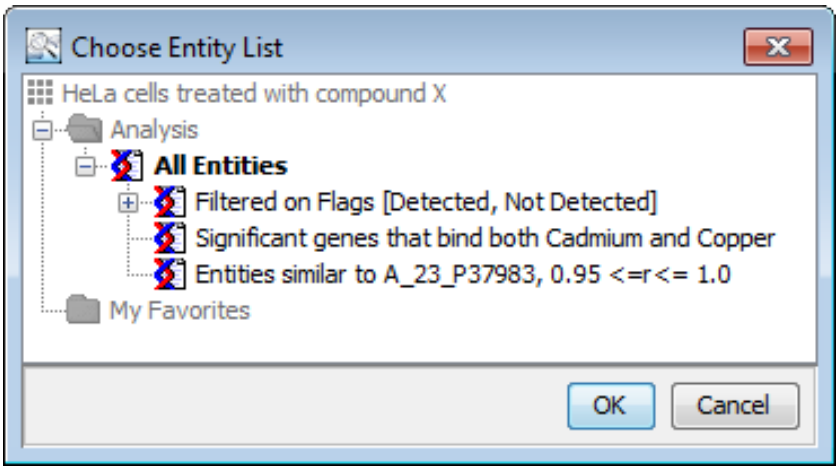


Figure 25.4: Chose Entity List dialog box

Purpose: Select the entity list that contains the entities on which you want to perform Clustering Analysis.

To open: Click **Choose...** in the Clustering Window. See "How to Perform Hierarchical Clustering Analysis" on page 671, "How to Perform K-Means Clustering Analysis" on page 673, "How to Perform Self-Organizing Map Clustering Analysis" on page 675.

OK Confirms the selected entity list, closes the dialog and returns to the Clustering Window.

Cancel Aborts the selection process and returns to the Clustering Window.

25.3.7.2 Choose Interpretation

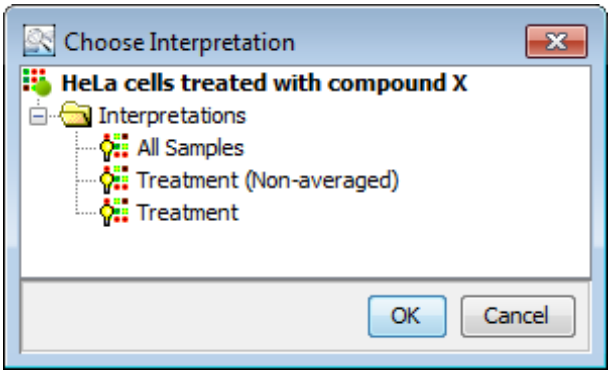


Figure 25.5: Choose Interpretation dialog box

Purpose: Select an interpretation based on which you want to perform Clustering Analysis from the available interpretations.

To open: Click **Choose...** in the Clustering Window. See "How to Perform Hierarchical Clustering Analysis" on page 671, "How to Perform K-Means Clustering Analysis" on page 673, "How to Perform Self-Organizing Map Clustering Analysis" on page 675. Click **Choose...** in the Export options dialog box. See "To export entities in cluster order" on page 692.

OK Confirms the selected interpretation, closes the dialog and returns to the Clustering Window.

Cancel Aborts the selection process and returns to the Clustering Window.

25.3.7.3 Classification Inspector

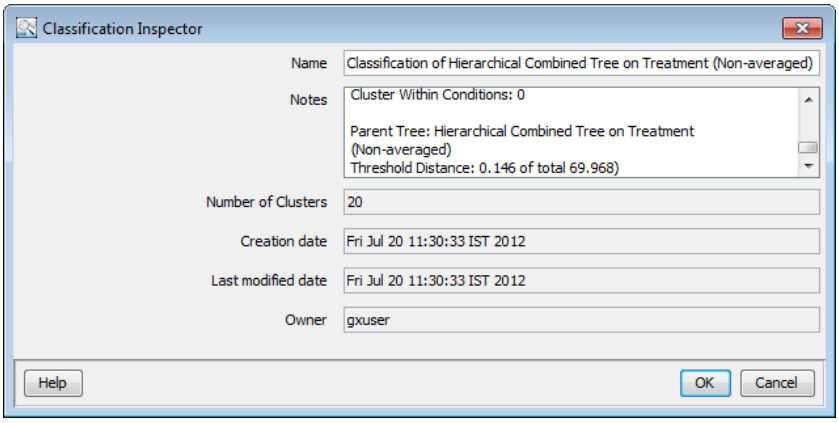


Figure 25.6: Classification Inspector dialog box

Purpose: Review the details that are recorded with the result of Create Classification on a subset of clusters in a cluster tree. You can change the default **Name** and **Notes**.

To open: Right-click the classification name in the Experiment Navigator and select **Inspect Classification**. You also see this inspector after running classification from a cluster tree by clicking the **Run Classification** icon in the **Dendrogram View Toolbar**. See "To create a classification from a subset of clusters in a cluster tree" on page 692.

Help Opens the **GeneSpring** manual in your default browser.

OK Closes the dialog box and saves the new classification in the experiment.

Cancel Closes the dialog box and returns to the cluster tree view. The classification is not saved in the experiment.

25.3.7.4 Combined Tree Inspector

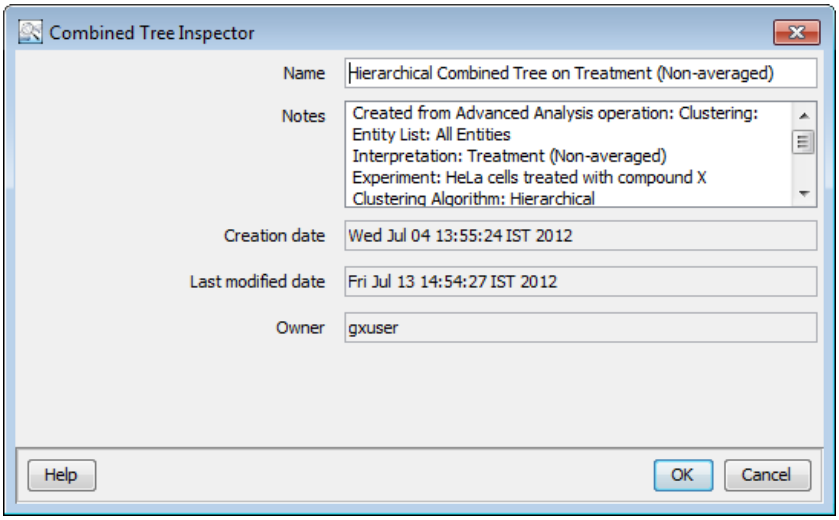


Figure 25.7: Combined Tree Inspector dialog box

Purpose: Save a subtree from selection on the active combined cluster tree on entities and conditions. You can change the default **Name** and **Notes**.

To open: Right-click the cluster tree name in the Experiment Navigator and select **Inspect Combined Tree**. See "To inspect a cluster tree" on page 683. You also see this inspector when you are saving a subtree by clicking the **Save subtree** icon in the **Dendrogram View Toolbar**. See "To save a subtree of a cluster tree" on page 687.

Help	Opens the **GeneSpring** manual in your default browser.
OK	Closes the dialog and saves the new cluster tree in the experiment.
Cancel	Closes the dialog box.

25.3.7.5 Condition Tree Inspector

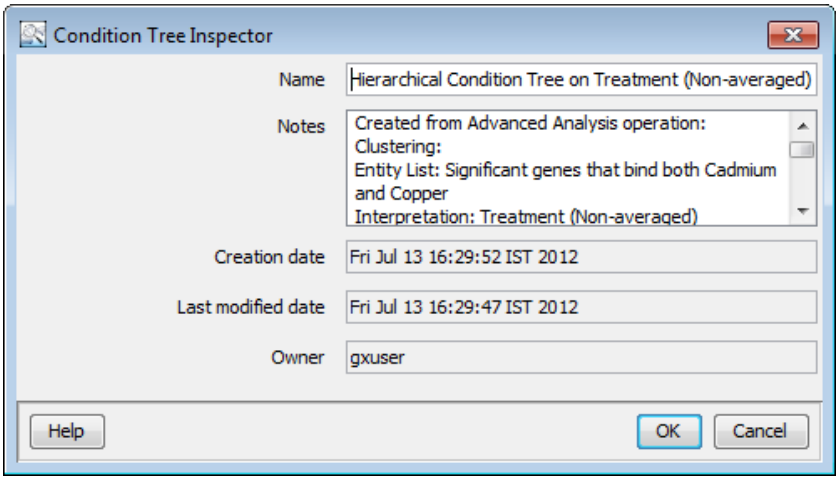


Figure 25.8: Condition Tree Inspector dialog box

Purpose: Save a subtree from selection on the active cluster tree on conditions. You can change the default **Name** and **Notes**.

To open: Right-click the cluster tree name in the Experiment Navigator and select **Inspect Condition Tree**. See "To inspect a cluster tree" on page 683. You also see this inspector when you are saving a subtree by clicking the **Save subtree** icon in the **Dendrogram View Toolbar**. See "To save a subtree of a cluster tree" on page 687.

Help	Opens the **GeneSpring** manual in your default browser.
OK	Closes the dialog and saves the new cluster tree in the experiment.
Cancel	Closes the dialog box.

25.3.7.6 Entity Tree Inspector

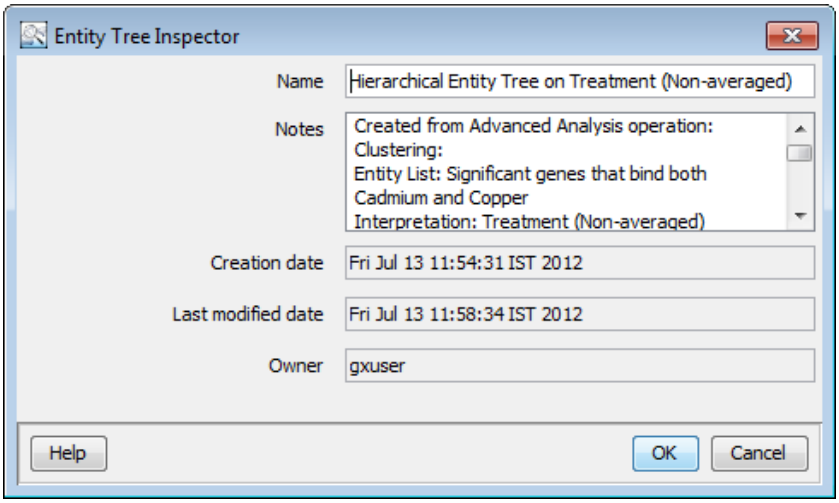


Figure 25.9: Entity Tree Inspector dialog box

Purpose: Save a subtree from selection on the active cluster tree on entities. You can change the default **Name** and **Notes**.

To open: Right-click the cluster tree name in the Experiment Navigator and select **Inspect Entity Tree**. See "To inspect a cluster tree" on page 683. You also see this inspector when you are saving a subtree by clicking the **Save subtree** icon in the **Dendrogram View Toolbar**. See "To save a subtree of a cluster tree" on page 687.

Help Opens the **GeneSpring** manual in your default browser.

OK Closes the dialog and saves the new cluster tree in the experiment.

Cancel Closes the dialog box.

25.3.7.7 Choose Parameters

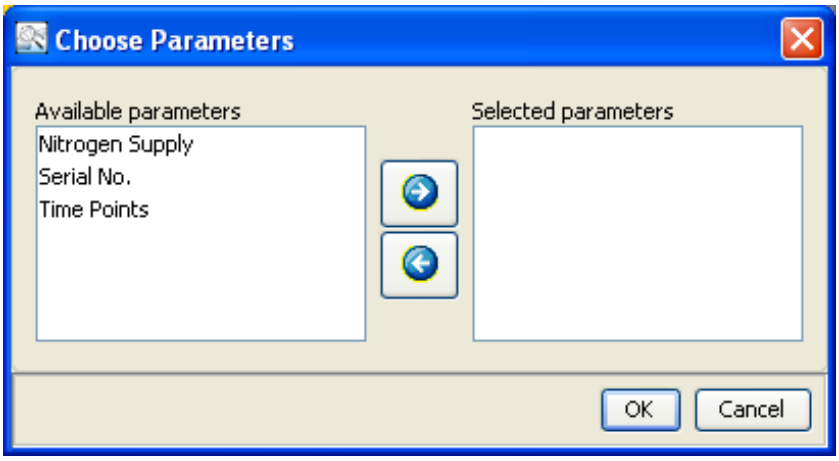


Figure 25.10: Choose Parameters dialog box

Purpose: Select from the available parameters to visualize sample parameter Heat Maps on the selected parameters.

To open: Click on the **Add sample parameter plots** icon in the **Dendrogram View Toolbar** and select **Heat Map**. See To add sample parameter plots to a Dendrogram on page 694.

OK Closes the dialog and displays the sample parameter Heat Map of the selected parameters.

Cancel Closes the dialog box.

25.3.7.8 Add/Remove/Re-order Parameters

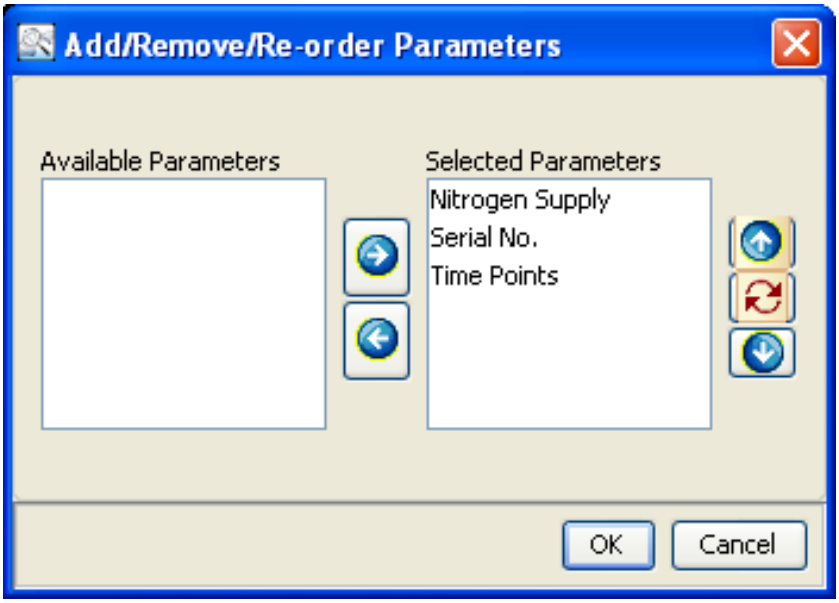


Figure 25.11: Add/Remove/Re-order Parameters dialog box

Purpose: Select from the available parameters or remove from the selected parameters to visualize sample parameter Heat Maps on the selected parameters.

To open: Click on the **Add/Remove/Re-order sample parameters** icon in a sample parameter Heat Map. See How to Visualize and Explore Sample Parameter Plots in the Dendrogram View on page 680.

OK Closes the dialog and displays the sample parameter Heat Map of the selected parameters.

Cancel Closes the dialog box.

25.3.7.9 Export options

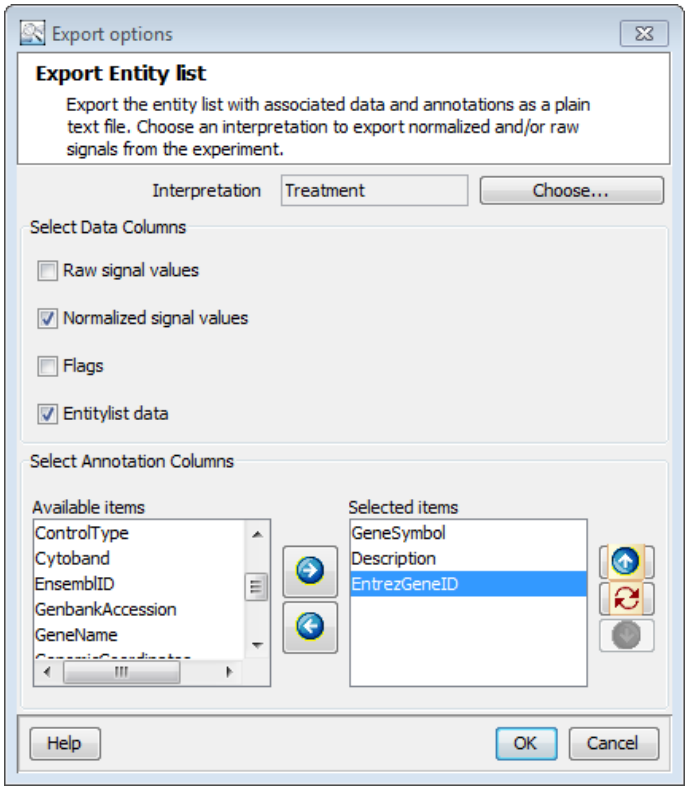


Figure 25.12: Export options dialog box

Purpose: Export the entities of a cluster tree in the cluster order. Select an interpretation as well as the data and annotation columns to be exported along with the entity identifier.

To open: Right-click the cluster tree and select **Export Entities in Cluster Order**. See "To export entities in cluster order" on page 692.

Help Opens the **GeneSpring** manual in your default browser.

OK Closes the dialog and allows you to specify the file name and location to save the entities in cluster order in a *.txt file.

Cancel Closes the dialog box and returns to the cluster tree view. The entities are not exported.

25.3.7.10 Report Exporter to Export as a Report

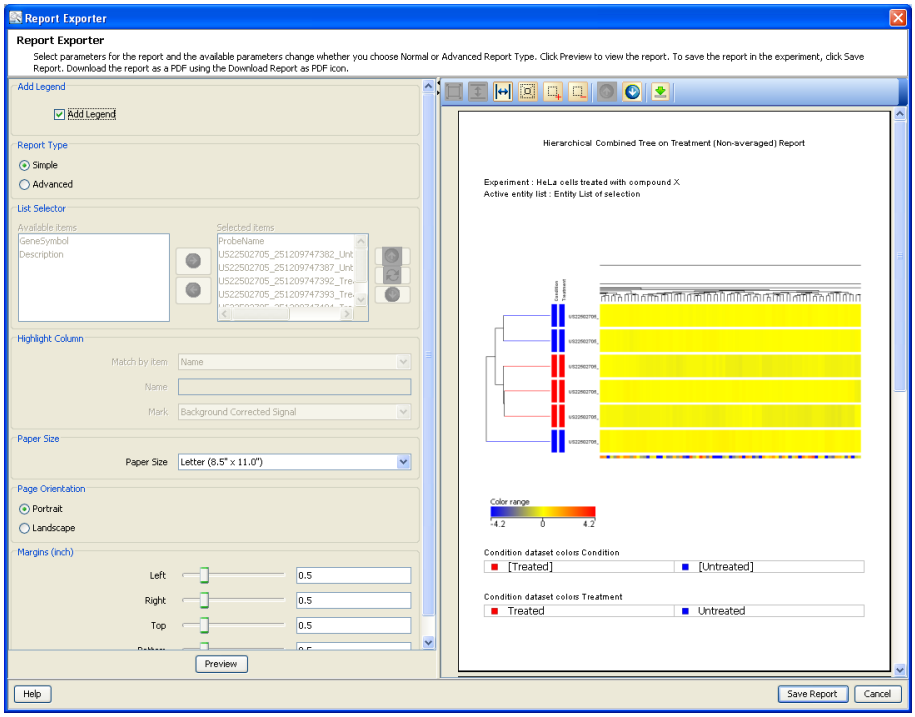


Figure 25.13: Export as Report options dialog box

Purpose: Export the cluster tree as a Report or download as a **.pdf* file.

To open: Right-click the cluster tree and select **Export as Report**. See "To export a cluster tree as a report" on page 689.

Help	Opens the **GeneSpring** manual in your default browser.
Save report	Closes the dialog and saves the cluster tree report in the experiment navigator.
Cancel	Closes the dialog box and returns to the cluster tree view. The cluster tree is not exported.

25.3.7.11 Print options to Export as an Image

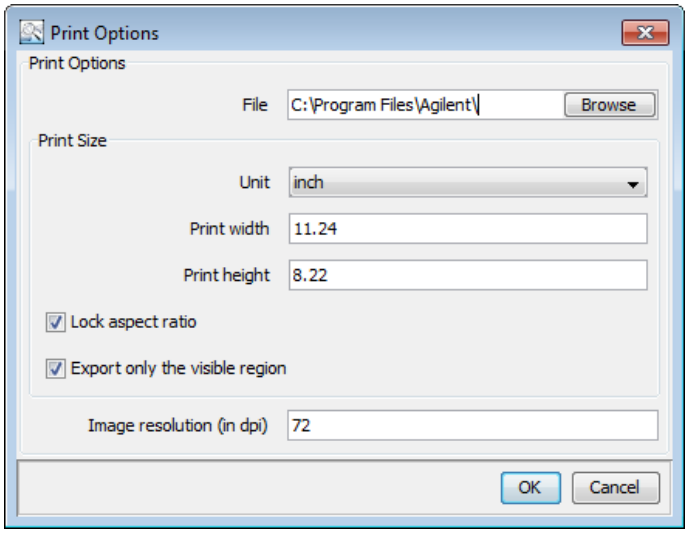


Figure 25.14: Export as Image options dialog box

Purpose: Export the cluster tree as an Image (with or without Entity labels).

To open: Right-click the cluster tree and select **Export as Image**. See "To export a cluster tree as an image" on page 690.

OK Closes the dialog and saves the cluster tree image at the specified location.

Cancel Closes the dialog box and returns to the cluster tree view. The cluster tree is not exported.

25.3.7.12 Print options to Export as an Image with sample parameter plots

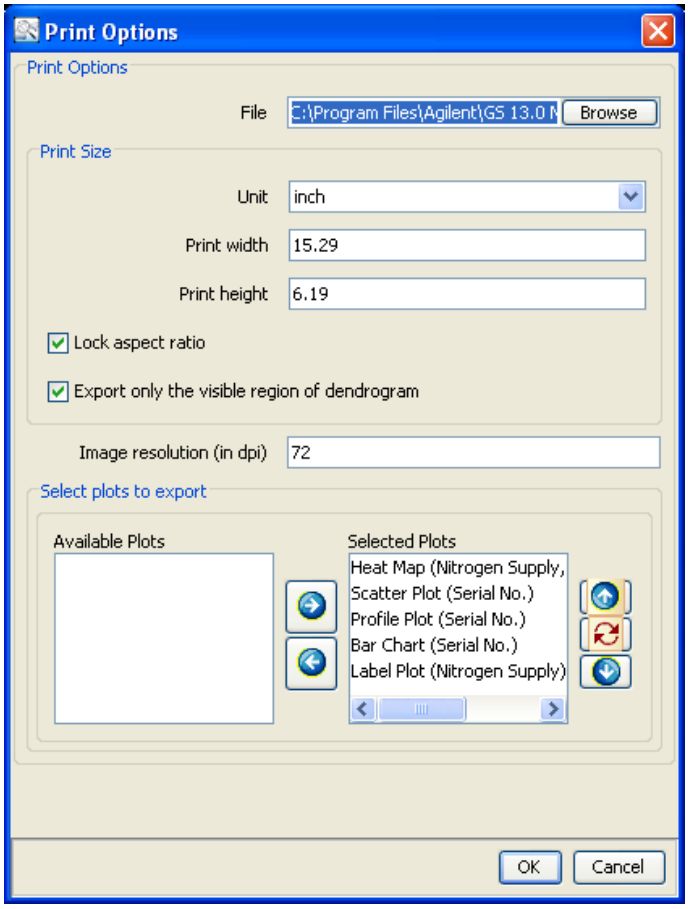


Figure 25.15: Export as Image with sample parameter plots options dialog box

Purpose: Export the **Detailed View** or **Combined View** of a cluster tree with the sample parameter plots as an Image.

To open: Right-click the cluster tree and select **Export as Image** and choose **Detailed View** or **Combined View** option. See "To export a cluster tree with sample parameter plots as an image" on page 690.

OK Closes the dialog and saves the image at the specified location.

Cancel Closes the dialog box and returns to the cluster tree view. The image is not exported.

25.3.7.13 GO analysis on entity cluster tree

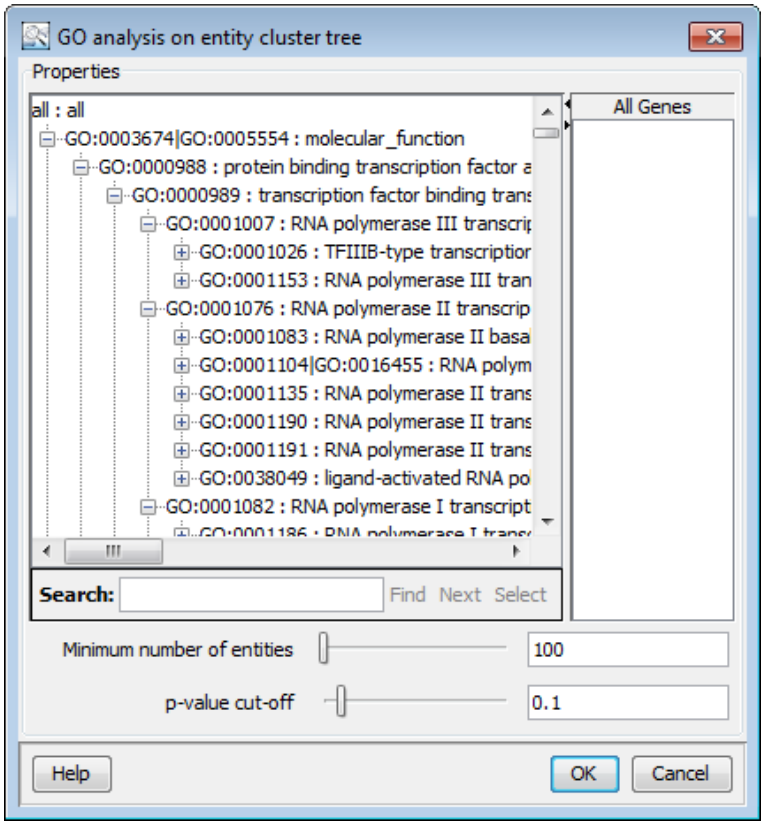


Figure 25.16: GO analysis on entity cluster tree dialog box

Purpose: Set the parameters (**Minimum number of entities** and **p-value cut-off**) for performing GO Analysis on the clusters in an entity tree or the entity tree part of a combined tree.

To open: Click the **Run GO Analysis** icon in the **Dendrogram View Toolbar**. See "To run GO Analysis on an entity tree" on page 693.

Help Opens the **GeneSpring** manual in your default browser.

OK Performs GO Analysis on the entities of an entity tree or the entity tree part of a combined tree.

Cancel Closes the dialog box. No analysis is performed.

25.3.7.14 Inspector

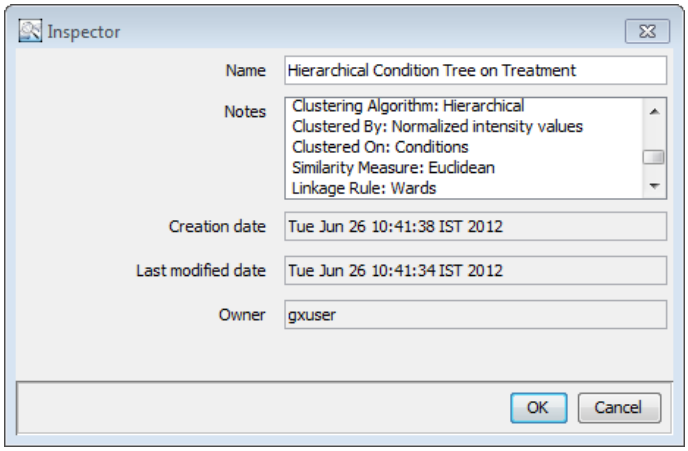


Figure 25.17: Inspector dialog box

Purpose: Review details of a clustering result object. Edit Name and Notes associated with the object.

To open: Click **OK** in the Clustering Window to save a result. See"Clustering Window - Common Features" on page 699, "How to Perform Hierarchical Clustering Analysis" on page 671,"How to Perform K-Means Clustering Analysis" on page 673, and "How to Perform Self-Organizing Map Clustering Analysis" on page 675.

OK Saves the object in the experiment as a child node of the entity list on which clustering was performed.

Cancel Aborts the saving process and returns to the Clustering Window.

25.3.7.15 Properties

Purpose: Inspect the display settings of the current clustering tree. You can change the default color settings, columns to display, labels to display, ruler settings, title and description. The Properties dialog box has five tabs: **Colors**, **Columns**, **Labels**, **Rulers**, and **Description**. Each tab is described further in the following subsections.

To open: Click the **Properties** icon in the **Dendrogram View Toolbar** or right-click on the Dendrogram View and select **Properties**. See "To change the display settings of a cluster tree" on page 685.

Help	Opens the **GeneSpring** manual in your default browser.
OK	Closes the dialog and applies any changes to the open cluster tree.
Cancel	Closes the dialog box.

Colors tab

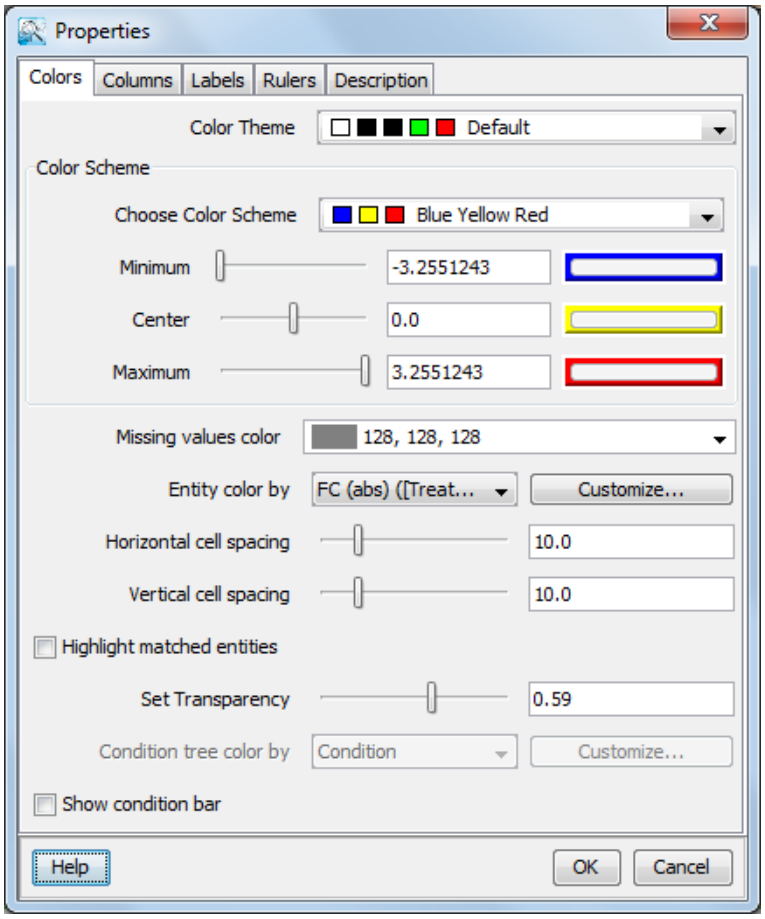


Figure 25.18: Properties dialog box - Colors tab

Customize the coloring options for the background, heatmap, entity labels, and condition labels in the Dendrogram View.

Color Theme Choose from either **Default** or create a **Custom Color Theme**(see Create New Color Theme on page 794) for the Dendrogram View .

Color Scheme Select one of the existing color schemes from the **Choose Color Scheme** drop-down list. To define a custom color scheme select the **Custom Color Scheme** option.

Specify the **Missing values color**.

Determine by what parameter the entity and condition color bars are colored from the **Entity color by** and **Condition tree color by** drop-down lists. Click **Customize** to define a custom color scheme for the entity and condition color bars.

Use the sliders to modify the value for **Horizontal cell spacing** and **Vertical cell spacing**. Cell spacing is available both in the preview and the saved image and you can view it only in the zoomed view of the dendrogram.

Select **Highlight matched entities** to match entities of the selected entity list in the experiment navigator with the entities from the Clustering Analysis input entity list with a red bar.

Use the sliders to modify the Transparency for hiding entities that are absent in a selected entity list.

Deselect **Show condition bar** to remove condition color bars.

Create New Color Theme

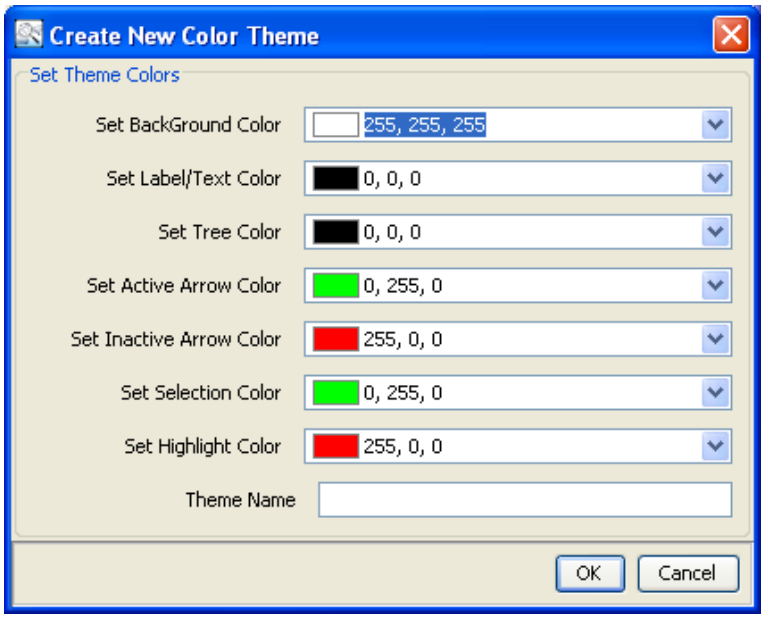


Figure 25.19: Properties dialog box - Colors tab - Create New Colors Theme

In a cluster tree, customize the colors for Background, Label or Text, Tree, Active and Inactive Arrows, Selection Bar, and Highlight Bar.

Set Color Choose a color of your choice from the available colors for each of the listed cluster tree components.

Theme Name Enter a name for the new color theme.

Columns tab

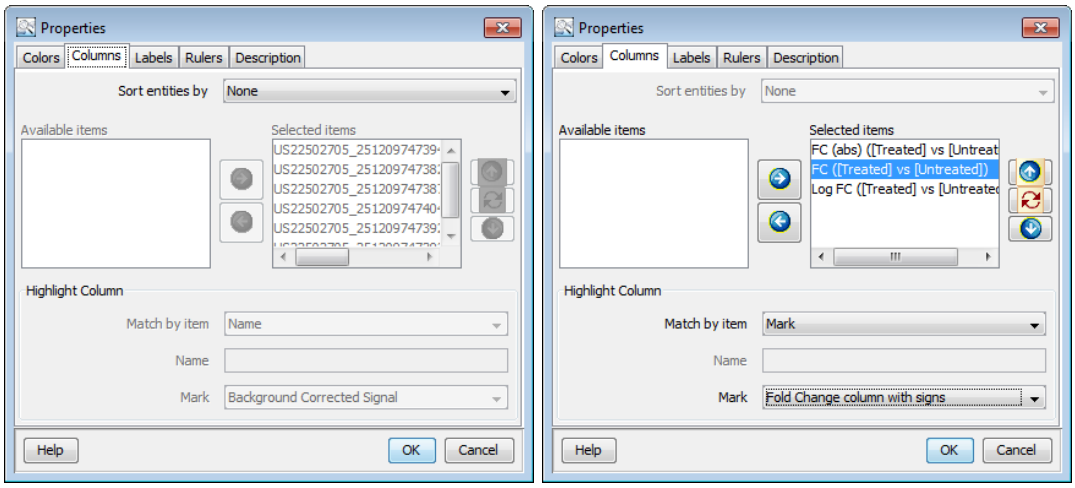


Figure 25.20: Properties dialog box - Columns tab

Sort entities in a condition tree and determine which data columns are displayed in an entity tree. The **Columns** tab options are not available for combined trees.

Sort entities by Select the data column that determines the order in which the entities appear in a condition tree. This option is not available for entity trees and combined trees.

Select and Reorder Columns Determine the data columns you want to appear in the cluster tree. Only columns listed under Selected items are displayed.

Change the order in which the selected data columns appear in the cluster tree.

Highlight Column Highlights the columns in the **Available items** or **Selected items fields** that match either a **Name** or **Mark** specified by you. This option is particularly useful when you have many data columns and you want to select all columns that have a part of their name or their mark in common.

Match by item: Select whether to match the available columns by Name or Mark.

Name: Type the exact name or a partial name and press <**Enter**> to confirm. Matching items are highlighted in the **Available items** or **Selected items** fields.

Mark: All the available marks for the data columns listed in the **Available items** or **Selected items** fields are available for selection from the **Mark** drop-down list. Columns created as a result of any analyses in **GeneSpring** are associated with a mark by default. For imported data columns, the mark has to be set at the time of importing them, to be available here.

Labels tab

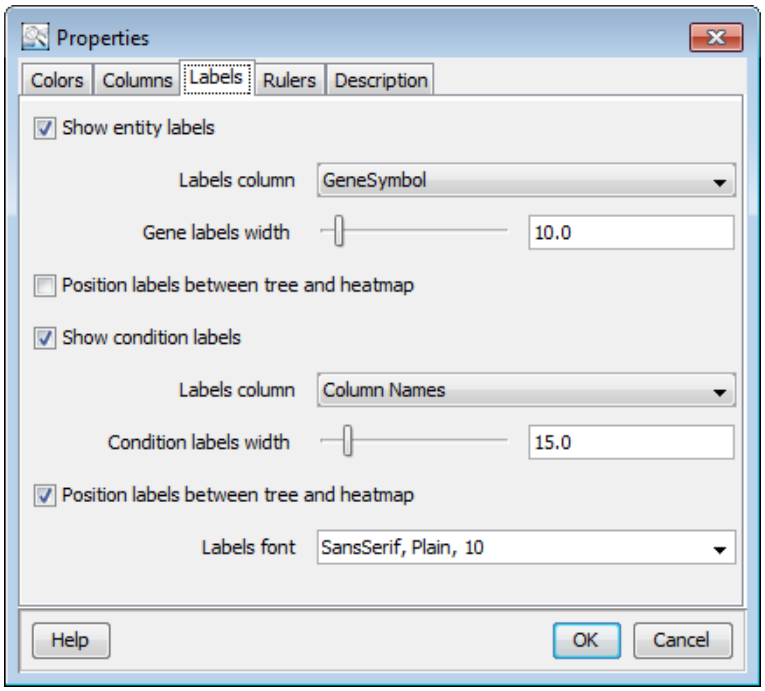


Figure 25.21: Properties dialog box - Labels tab

Select the display options for entity and condition labels.

Show entity labels	Choose to display or remove entity labels from the Dendrogram View.
	Select from which annotation column the labels should be sourced from the **Labels column** drop-down list.
	Determine the display width for the entity labels using the **Gene labels width** slider or type the value and confirm by pressing <**Enter**>.
Position labels between tree and heatmap	Select this option to display entity labels between the tree and heatmap, rather than below the heatmap.
Show condition labels	Choose to display or remove condition labels from the Dendrogram View.
	Select from which annotation column the labels should be sourced from the **Labels column** drop-down list.
	Determine the display width for the entity labels using the **Conditon labels width** slider or type the value and confirm by pressing <**Enter**>.
Position labels between tree and heatmap	Select this option to display condition labels between the tree and heatmap.
Labels font	Select the font, size, and style for both entity and condition labels. The available fonts from the drop-down list are all the fonts installed on your system.

Rulers tab

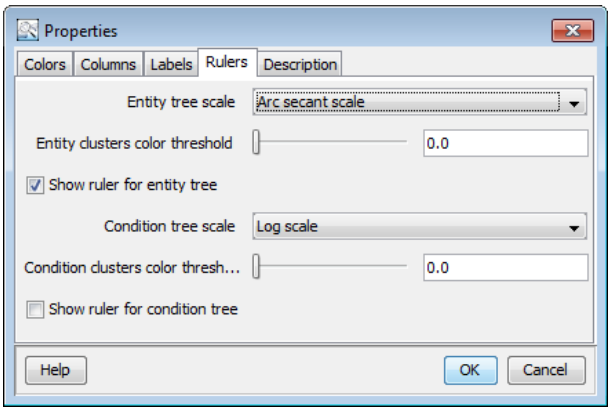


Figure 25.22: Properties dialog box - Rulers tab

Select the display settings for the entity tree and condition tree distance metric rulers.

Show ruler for entity tree Choose to display or remove the ruler for the entity tree from the Dendrogram View.

Select the scale according to which the distance metric values are displayed from the **Entity tree scale** drop-down list. Choose from these options: **Square root scale**, **Arc secant scale**, **True scale**, and **Log scale**.

Set the threshold at which a color change is applied to the cluster tree that distinguishes clusters above and below the set threshold. Use the **Entity clusters color threshold** slider or type the value and confirm by pressing <**Enter**> to set the desired threshold.

Show ruler for condition tree Choose to display or remove the ruler for the condition tree from the Dendrogram View.

Select the scale according to which the distance metric values are displayed from the **Condition tree scale** drop-down list. Choose from these options: **Square root scale**, **Arc secant scale**, **True scale**, and **Log scale**.

Set the threshold at which a color change is applied to the cluster tree that distinguishes clusters above and below the set threshold. Use the **Condition clusters color threshold** slider or type the value and confirm by pressing <**Enter**> to set the desired threshold.

Description tab

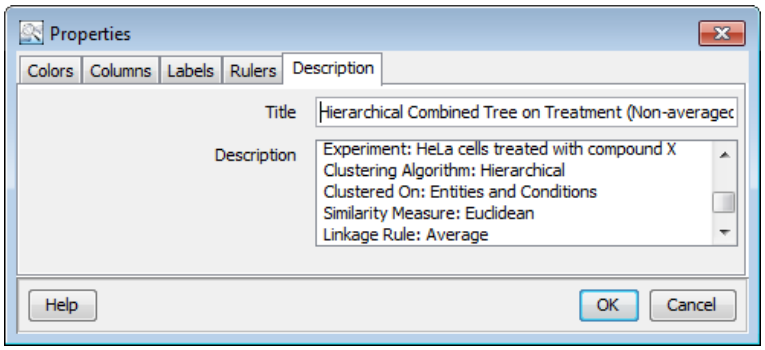


Figure 25.23: Properties dialog box - Description tab

Review the title of the cluster tree object and a description of the parameters that were used to create it.

Title Contains the name of the cluster tree object.

Description Contains information about all the parameters that were used to create the cluster tree.

NOTE You can make changes to both the **Title** and **Description** fields in this dialog. However, these changes are not saved when you close the Dendrogram View of the cluster tree.

25.3.7.16 Properties of sample parameter Heat Map

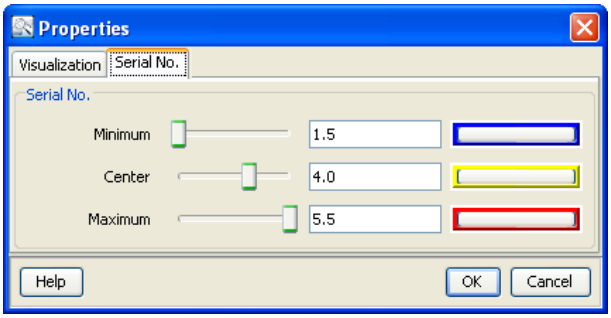


Figure 25.24: Sample parameters Heat Map Properties dialog box

Purpose: Inspect the display settings of the sample parameter Heat Map. You can change the default Heat map width for all parameters and colors for each of the selected numeric parameters. The Properties dialog box has multiple tabs: **Visualization** and a tab for each of the selected numeric parameter.

In each of the selected numeric parameters tab, define a custom color scheme for the **Heat Map**.

To open: Right-click on the sample parameter Heat Map and select **Properties**. See "To change the display settings of a sample parameter Heat Map" on page 686.

Help	Opens the **GeneSpring** manual in your default browser.
OK	Closes the dialog and applies any changes to the sample parameters Heat Map.
Cancel	Closes the dialog box. No changes are applied.

25.3.7.17 Properties of sample parameter Scatter Plot

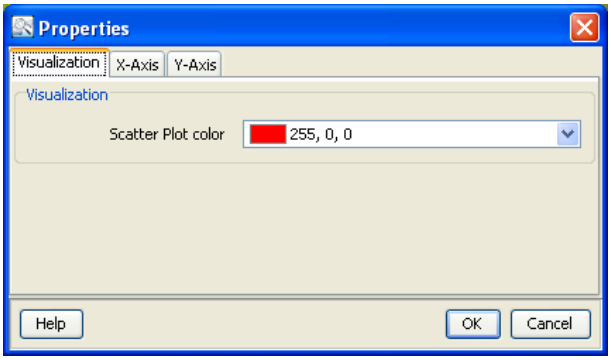


Figure 25.25: Sample parameters Scatter Plot Properties dialog box

Purpose: Inspect the display settings of the sample parameter Scatter Plot. You can change the default color settings of the scatter plot and display settings for X- and Y-axis. The Properties dialog box has multiple tabs: **Visualization**, **X-Axis**, and **Y-Axis**.

- **Visualization** - Change the color settings of the scatter plot.
- **X-Axis** - Choose to show or hide X-axis and axis labels. Change the color settings of the scatter plot X-axis and labels.
- **Y-Axis** - Choose to show or hide Y-axis and change the color settings of the scatter plot Y-axis.

To open: Right-click on the sample parameter Scatter Plot and select **Properties**. See "To change the display settings of a sample parameter Scatter Plot" on page 686.

Help	Opens the **GeneSpring** manual in your default browser.
OK	Closes the dialog and applies any changes to the sample parameters Scatter Plot.
Cancel	Closes the dialog box. No changes are applied.

25.3.7.18 Properties of sample parameter Profile Plot

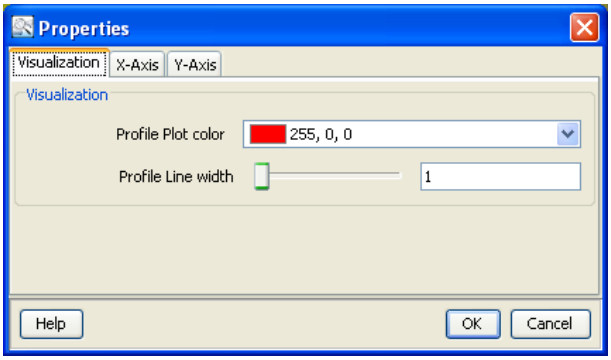


Figure 25.26: Sample parameters Profile Plot Properties dialog box

Purpose: Inspect the display settings of the sample parameter Profile Plot. You can change the default color settings of the profile plot, profile line width, and display settings for X- and Y-axis. The Properties dialog box has multiple tabs: **Visualization**, **X-Axis**, and **Y-Axis**.

- **Visualization** - Change the color settings of the profile plot and change the profile line width.
- **X-Axis** - Choose to show or hide X-axis and axis labels. Change the color settings of the profile plot X-axis and labels.
- **Y-Axis** - Choose to show or hide Y-axis and change the color settings of the profile plot Y-axis.

To open: Right-click on the sample parameter Profile Plot and select **Properties**. See "To change the display settings of a sample parameter Profile Plot" on page 686.

Help	Opens the **GeneSpring** manual in your default browser.
OK	Closes the dialog and applies any changes to the sample parameters Profile Plot.
Cancel	Closes the dialog box. No changes are applied.

25.3.7.19 Properties of sample parameter Bar Chart

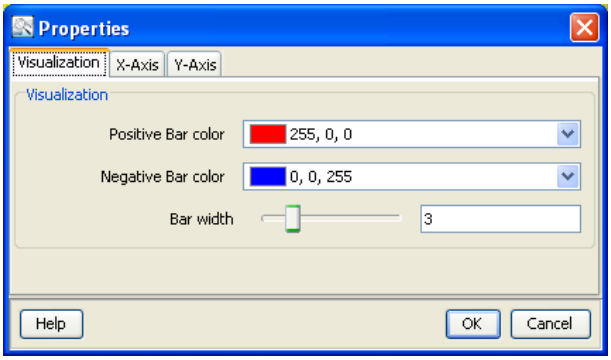


Figure 25.27: Sample parameters Bar Chart Properties dialog box

Purpose: Inspect the display settings of the sample parameter Bar Chart. You can change the default color settings of the bar chart, bar width, and display settings for X- and Y-axis. The Properties dialog box has multiple tabs: **Visualization**, **X-Axis**, and **Y-Axis**.

- **Visualization** - Change the color settings of both positive and negative bar charts and change the bar width.
- **X-Axis** - Choose to show or hide X-axis and axis labels. Change the color settings of the bar chart X-axis and labels.
- **Y-Axis** - Choose to show or hide Y-axis and change the color settings of the bar chart Y-axis.

To open: Right-click on the sample parameter Bar Chart and select **Properties**. See "To change the display settings of a sample parameter Bar Chart" on page 686.

Help	Opens the **GeneSpring** manual in your default browser.
OK	Closes the dialog and applies any changes to the sample parameters Bar Chart.
Cancel	Closes the dialog box. No changes are applied.

25.3.7.20 Properties of sample parameter Label Plot

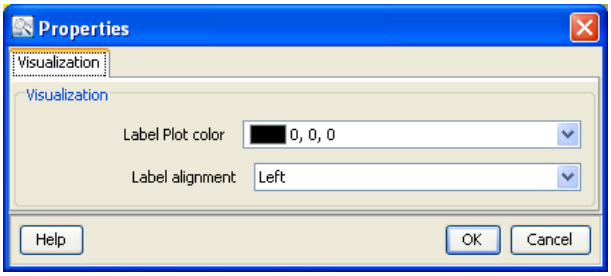


Figure 25.28: Sample parameters Label Plot Properties dialog box

Purpose: Inspect the display settings of the sample parameter Label Plot. You can change the default color settings of the labels and label alignment settings. The Properties dialog box has a **Visualization** tab only.

To open: Right-click on the sample parameter Label Plot and select **Properties**. See "To change the display settings of a sample parameter Label Plot" on page 687.

Help Opens the **GeneSpring** manual in your default browser.

OK Closes the dialog and applies any changes to the sample parameters Bar Chart.

Cancel Closes the dialog box. No changes are applied.

25.3.7.21 Select Columns

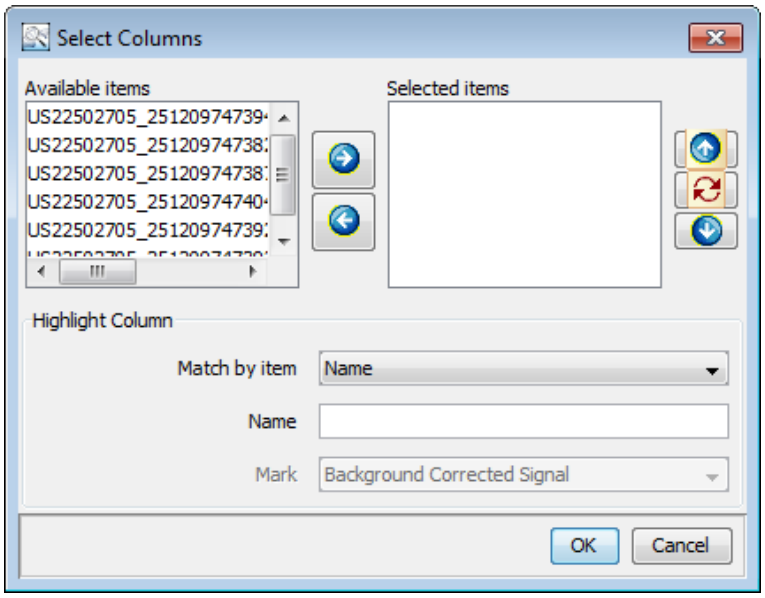


Figure 25.29: Select Columns dialog box

Purpose: Select individual samples/columns from the selected interpretation to appear highlighted in the cluster tree.

To open: Right-click the cluster tree and select **Select Columns**. See "Right-click options on a cluster tree" on page 779.

Select and Reorder Columns	Determine the samples/columns you want to appear in the cluster tree. Only columns listed under **Selected items** are displayed. Change the order in which the selected columns appear in the cluster tree.
Highlight Column	Highlights the columns in the **Available items** or **Selected items fields** that match either a **Name** or **Mark** specified by you. This option is particularly useful when you have many columns and you want to select all columns that have a part of their name or their mark in common. **Match by item**: Select whether to match the available columns by Name or Mark. **Name**: Type the exact name or a partial name and press <**Enter**> to confirm. Matching items are highlighted in the **Available items** or **Selected items** fields. **Mark**: All the available marks for the columns listed in the **Available items** or **Selected items** fields are available for selection from the **Mark** drop-down list.
OK	Closes the dialog box. The selected columns appear highlighted in the open cluster tree.
Cancel	Aborts the selection process and returns to the Dendrogram View.

25.4 Clustering Analysis - Statistical Algorithms

This section explains the statistical algorithms and options available for each algorithm involved in determining a clustering tree.

25.4.1 Distance Metric

A clustering algorithm is required to measure the similarity or difference between entities or conditions. A variety of distance metrics are available to determine the distance between intermediate clusters while computing the similarity scores. The available metric options are explained in this section along with their respective computation formulae.

See "How to Perform Hierarchical Clustering Analysis" on page 671, "How to Perform K-Means Clustering Analysis" on page 673, "How to Perform Self-Organizing Map Clustering Analysis" on page 675.

25.4.1.1 Angular

$$\cos(\theta) = \frac{x \cdot y}{\|x\|\|y\|} = \frac{\sum_{i=1}^{n} x_i \times y_i}{\sqrt{\sum_{i=1}^{n} (x_i)^2} \times \sqrt{\sum_{i=1}^{n} (y_i)^2}}$$

25.4.1.2 Euclidean

Standard sum of squared distance (L2-norm) between two entities.

$$\sqrt{\sum_{i} (x_i - y_i)^2}$$

25.4.1.3 Squared Euclidean

Square of the Euclidean distance measure. This accentuates the distance between entities. Entities that are close are brought closer, and those that are dissimilar move further apart.

$$\sum_{i} (x_i - y_i)^2$$

25.4.1.4 Manhattan (Cityblock)

This is also known as the L1-norm. The sum of the absolute value of the differences in each dimension is used to measure the distance between entities.

$$\sum_{i} |x_i - y_i|$$

25.4.1.5 Maximum (Chebychev)

This measure, also known as the L-Infinity-norm, uses the absolute value of the maximum difference in any dimension.

$$\max_i |x_i - y_i|$$

25.4.1.6 Minimum

This measure uses the absolute value of the minimum difference in any dimension.

$$\min_i |x_i - y_i|$$

25.4.1.7 Differential

The distance between two entities in estimated by calculating the difference in slopes between the expression profiles of two entities and computing the Euclidean norm of the resulting vector. This is a useful measure in time series analysis, where changes in the expression values over time are of interest, rather than absolute values at different times.

$$\sqrt{\sum_i [(x_{i+1} - x_i) - (y_{i+1} - y_i)]^2}$$

25.4.1.8 Canberra

The Canberra metric is similar to the Manhattan distance. The distinction is that the absolute difference in each dimension of the two vectors is divided by the sum of the absolute values prior to summing.

$$\sum_i \frac{|x_i - y_i|}{|x_i| + |y_i|}$$

25.4.1.9 Harmonic

It is a variation of the Euclidean distance, where the terms for the different dimensions are summed inversely (similar to the harmonic mean), and is more robust against outliers compared to the Euclidean distance.

$$\left[\frac{\sum_i \left(\frac{1}{x_i - y_i}\right)^2}{n}\right]^{-1}$$

25.4.1.10 Pearson Centered

This measure is the 1-centered variation of the Pearson Correlation Coefficient. Positively correlated entities give values of this measure close to 1; negatively correlated ones give values close to 0, and unrelated entities close to 0.5.

$$\frac{\left[\left(\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_i (x_i - \bar{x})^2)(\sum_i (y_i - \bar{y})^2)}}\right) + 1\right]}{2}$$

25.4.1.11 Pearson's Uncentered (Cosine)

This measure is similar to the Pearson Correlation coefficient except that the entities are not mean-centered. In effect, this measure treats the two entities as vectors and gives the cosine of the angle between the two vectors. Highly correlated entities give values close to 0, negatively correlated entities give values close to 2, while unrelated entities give values close to 1.

$$1 - \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2 \sum_i y_i^2}}$$

NOTE This distance metric is similar to the Pearson's Uncentered option in earlier versions of **GeneSpring**.

25.4.1.12 Pearson's Centered-Absolute

It is a variant of Pearson's centered distance metric. The highly related entities give values close to 0 while the non-correlated entities give values close to 1.

$$1 - \left| \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_i (x_i - \bar{x})^2)(\sum_i (y_i - \bar{y})^2)}} \right|$$

25.4.1.13 Pearson's Uncentered-Absolute

It is a variant of Pearson's uncentered distance metric. The highly related entities give values close to 0 while the non-correlated entities give values close to 1.

$$1 - \left| \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2 \sum_i y_i^2}} \right|$$

25.4.2 Linkage Rule

Linkage rule algorithms determine the similarity between clusters for visualization in a cluster tree as one of the parameters for the hierarchical clustering algorithm. Details on the available options in **GeneSpring** are provided in this section. See "How to Perform Hierarchical Clustering Analysis" on page 671.

25.4.2.1 Ward's method

This method is based on the ANOVA approach. It computes the sum of squared errors around the mean for each cluster. Then, two clusters are joined so as to minimize the increase in error.

25.4.2.2 Centroid Linkage

The distance between two clusters is the average distance between their respective centroids weighted by the size of the clusters.

25.4.2.3 Median linkage

The distance between two clusters is the average distance between their respective centroids. Median linkage is similar to centroid linkage, with the exception that the two clusters are given equal weight regardless of their size.

25.4.2.4 Single Linkage

The distance between two clusters is calculated as the smallest distance between pairs of members from the two clusters. Single linkage maximizes the connectedness of a cluster, but tends to lead to chain-like clusters containing dissimilar elements.

25.4.2.5 Complete Linkage

The distance between two clusters is calculated as the largest distance (dissimilarity) between any two members in the two clusters. Complete linkage minimizes the maximum within-cluster distance, and so tends to find compact clusters. However, it may overemphasize small differences between members.

25.4.2.6 Average Linkage

The distance between two clusters is calculated as the arithmetic mean of distances between all possible pairs from the two clusters. Average linkage is a compromise between single and complete linkage.

25.4.3 Missing Values Handling

For clustering on entities, a minimum of 50% of values for each entity in the selected entity list across conditions are required. Any entities with fewer values are excluded from clustering. To perform clustering, **GeneSpring** creates a separate entity list internally that contains valid entities based on this threshold. Clustering is only performed on that internal entity list.

For clustering on conditions, at least 50% of all entities associated with each condition should have valid values.

For clustering on both entities and conditions, the same conditions as for the individual clustering processes apply. **GeneSpring** first looks at the condition for entities, followed by the condition for conditions.

Chapter 26

Correlation Analysis

Correlation measures the strength and direction of the linear relationship between two quantitative variables. In biological systems, correlation analysis allows identification of co-regulated molecules such as genes and metabolites as well as to identify relationships between the samples in a study. The correlation analysis framework in **GeneSpring** supports pair-wise correlations on the entities or samples measured from a single technology platform or between entities measured from two different platforms.

This chapter describes in detail the different correlation analysis options available in **GeneSpring**. Correlation analysis in **GeneSpring** is supported for all the major analysis and experiment types in **GeneSpring** with the exception of: (1) Pathway experiments, (2) Copy Number experiments, and (3) Association experiments. Table 26.1 lists all the **GeneSpring** analysis types supported for performing Correlation Analysis.

Table 26.1: Analysis Types in **GeneSpring** for Correlation Analysis

Analysis Type	**Within Single Experiment**	**Across Two Experiments**
mRNA Expression	YES	YES
Exon Expression	YES	YES
miRNA	YES	YES
RTPCR	YES	YES
DNA-Seq	NO	NO
RNA-Seq	YES*	YES*
Small RNA-Seq	YES*	YES*
Metabolomics	YES	YES
Proteomics	YES	YES

* - To perform Correlation Analysis, RNA-Seq and small RNA-Seq experiments have to be imported from **Strand NGS** into **GeneSpring**. For these two NGS experiment types, Correlation Analysis is possible both in the imported experiment itself and as part of an MOA experiment (see How to Import a **Strand NGS** Experiment into **GeneSpring** on page 617).

GeneSpring offers two statistical methods for computing correlation coefficients: Pearson and Spearman. The results of Correlation Analysis are represented visually as a heatmap. Correlation coefficient values in the heatmap are in the range of +1 to -1. Positive correlation indicates entities are behaving similarly and negative correlation indicates opposite profiles, while proximity to 0 means that there is no linear relationship between the entity pair.

GeneSpring also provides the option to cluster entities in a correlation heatmap based on the correlation coefficients or filter entities in the heatmap based on available associated data.

26.1 Correlation Analysis in GeneSpring

This section describes in detail the required steps for each of the available correlation analyses. This section also describes the available options for analyzing the saved results in the form of a heatmap once it has been saved as part of the experiment.

26.1.1 How to Perform Correlation Analysis on Entities

Correlation analysis on entities is useful to reveal any potential linear relationships between the entities in the chosen entity list based on the entity profile defined by the chosen interpretation. The heatmap resulting from correlation analysis on entities can be used to identify clusters of entities that are co-regulated. The heatmap also illustrates any relationships between such entity clusters.

Use the following procedure to set up and run correlation analysis on entities.

1. Select **Correlation** from the **Analysis** section of the **Workflow** browser. The **Correlation** window opens. See "Correlation Window Features" on page 772.
2. Select **Entities** from the **Correlate on** drop-down list.
3. Click **Choose...** next to the **Entity List** field and select the entity list that contains the set of entities you want to analyze in the **Choose Entity List** dialog box. Click **OK**. See "Choose Entity List" on page 780.

NOTE Correlation Analysis is performed only on entities in the selected entity list. It is recommended (but not mandatory) to perform correlation on a set of entities filtered for reliability and/or statistically significant changes in expression by t-test or ANOVA. The selected entity list in the **Experiment Navigator** (the one that appears in bold) is selected by default.

---

4. Click **Choose...** next to the **Interpretation** field and select the interpretation that you want to analyze in the **Choose Interpretation** dialog box. Click **OK**. See "Choose Interpretation" on page 781.

NOTE Correlation Analysis on Entities can be performed with the **All Samples** interpretation, a non-averaged interpretation provided there are more than three samples and an averaged interpretation if it has more than three conditions and a minimum of two entities have values in each of these three samples or conditions.

NOTE Interpretations, other than the **All Samples** interpretation, have to be created before starting **Correlation** analysis (see "Create Interpretation") to be available for selection in this workflow step. The selected interpretation in the Experiment Navigator (the one that appears in bold) is selected by default. An interpretation should contain at least two samples or conditions to run correlation analysis.

---

5 Select one of the following from the **Similarity metric** drop-down list: **Pearson** or **Spearman**. See "Similarity Metric" on page 799 for more information on each method.

6 Click **OK** to perform correlation analysis and save the result (a correlation node) in the experiment as well as view the correlation heatmap in the desktop area.

NOTE A pre-requisite for computing the correlation coefficient for two entities in **GeneSpring** is the presence of three or more paired data points from the same samples or conditions in the selected interpretation.
For example, the correlation coefficient of two entities E_1, E_2 is computed for a non-averaged interpretation I_1 with conditions C_1, C_2, where, C_1 has samples S_1, S_2, and C_2 has sample S_3, when, both E_1 and E_2 have corresponding data points (E_1S_1, E_1S_2, E_1S_3, E_2S_1, E_2S_2, and E_2S_3) in all three samples. If either of the entities have missing data points in any of the samples, the correlation coefficient for E_1 and E_2 is not computed for the interpretation I_1. For an averaged interpretation I_2 with conditions C_1 and C_2, the correlation coefficient for entities E_1 and E_2 is not computed as the number of conditions is $<$ three.

The correlation analysis node is saved as an object in the experiment hierarchy below the selected input entity list.

26.1.2 How to Perform Correlation Analysis on Samples

Correlation analysis on samples is useful to reveal or confirm linear relationships between the samples of an experiment based on their expression/abundance profiles defined by the entity list and interpretation chosen for correlation analysis. The heatmap resulting from correlation analysis on samples can be used to illustrate the closeness between samples of a condition or the relationship between sample groups across conditions.

CAUTION The results of sample-sample correlation analysis are sensitive to baseline transformation performed during experiment creation. Create a new experiment with a modified baseline transformation parameter, or during experiment creation choose **Do not perform baseline** to change the input data for sample-sample correlation analysis.

Use the following procedure to set up and run correlation analysis on samples.

1. Select **Correlation** from the **Analysis** section of the **Workflow** browser. The **Correlation** window opens. See "Correlation Window Features" on page 772.
2. Select **Samples** from the **Correlate on** drop-down list.
3. Click **Choose...** next to the **Entity List** field and select the entity list that contains the set of entities you want to analyze in the **Choose Entity List** dialog box. Click **OK**. See "Choose Entity List" on page 780.

NOTE Correlation on samples is performed only on entities in the selected entity list. It is recommended (but not mandatory) to perform correlation on a set of entities filtered for reliability and/or statistically significant changes in expression by t-test or ANOVA. The selected entity list in the **Experiment Navigator** (the one that appears in bold) is selected by default.

4. Click **Choose...** next to the **Interpretation** field and select the interpretation that you want to analyze in the **Choose Interpretation** dialog box. Click **OK**. See "Choose Interpretation" on page 781.

NOTE Interpretations, other than the **All Samples** interpretation, have to be created before starting **Correlation** analysis (see "Create Interpretation") to be available for selection in this workflow step. The selected interpretation in the Experiment Navigator (the one that appears in bold) is selected by default. An interpretation should contain at least three samples or conditions to run correlation analysis.

5. Select one of the following from the **Similarity metric** drop-down list: **Pearson** or **Spearman**. See "Similarity Metric" on page 799.
6. Click **OK** to perform correlation analysis and save the result (a correlation node) in the experiment as well as view the correlation heatmap in the desktop area.

NOTE Correlation Analysis for samples is only allowed on interpretations where the selected entity list has at least three paired entity data points for the selected interpretation. Therefore, if you choose the **All Samples** or a non-averaged interpretation there have to be at least two samples present as well as a minimum of three entities for which data is available in each of these two samples. If you choose an averaged interpretation there have to be a minimum of two conditions as well as a minimum of three entities for which data is available in each of these two conditions.

The correlation analysis node is saved as an object in the experiment hierarchy below the selected input entity list.

26.1.3 How to Perform Correlation Analysis in a MOA experiment

Correlation Analysis is also one of the analysis types available in Multi-Omic Analysis (MOA) experiments (see "Multi-Omic Analysis in **GeneSpring** " on page 603) which can contain up to two experiments of different experiment types (e.g. transcriptomics and metabolomics). Correlation analysis in a MOA experiment is useful to reveal relationships among the entities of two separate experiments, typically of a different technology or experiment type. For example, MOA Correlation Analysis can be used to identify regulatory relationships between biological molecules from the same samples measured by two different technologies like microRNA and mRNA expression, or protein and small molecule abundance. In the resulting correlation heatmap, starting from the top left corner, the entities of Experiment 1 are listed first, followed by the entities in Experiment 2.

NOTE In a MOA experiment Correlation Analysis on Samples is currently not supported.

- Select Correlation Analysis from the Multi-Omic Analysis section of the Workflow browser.

The following sections describe in detail the different steps of the wizard-driven correlation analysis workflow in a MOA experiment.

26.1.3.1 Step 1: Specifying Interpretations, Entity Lists, and Similarity Metric

This section describes how to choose interpretations, entity lists, and the desired similarity metric in the first step of wizard-driven correlation analysis in a MOA experiment.

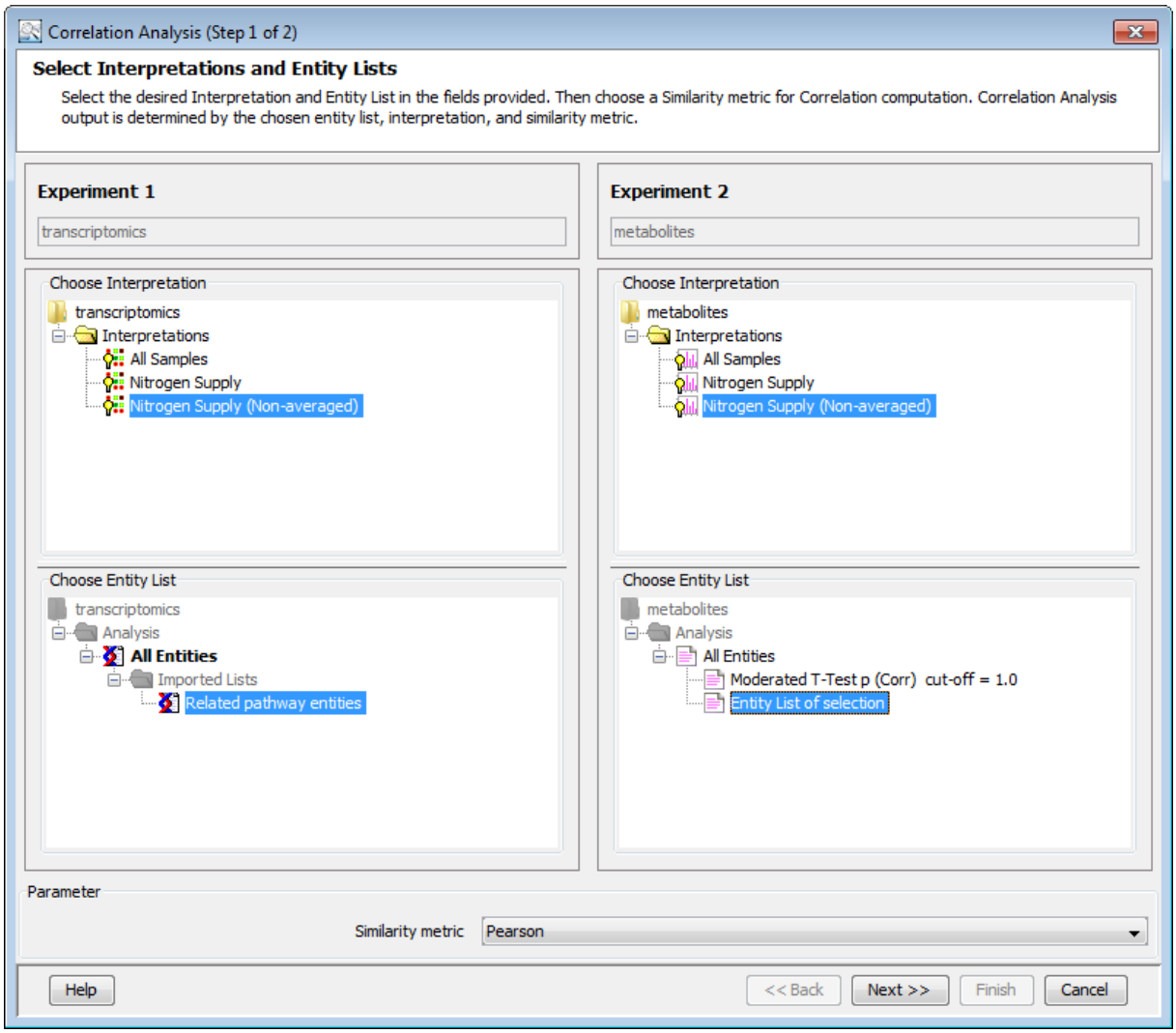


Figure 26.1: Correlation Analysis in a Multi-Omic Analysis experiment - Step 1 of 2

To choose an interpretation

An interpretation is required from each experiment to perform correlation analysis. Select an interpretation for Experiment 1 and Experiment 2 in the separate areas identified by the respective experiment names.

- Select the required interpretation from **Choose Interpretation**.

The chosen interpretation is highlighted.

NOTE

Interpretations, other than the All Samples interpretation, have to be created before starting correlation analysis (see Create Interpretation) to be available for selection in this workflow step. The selected interpretations must either both be averaged or non-averaged. An interpretation should contain at least three samples or conditions and a minimum of two entities with available values in each of these three samples/conditions to run correlation analysis.

To select an entity list

All available entity lists in the MOA experiments are displayed in separate areas identified by the respective experiment names in Step 1.

- Select the required entity list from **Choose Entity List**.

The chosen entity list is highlighted.

To select the Similarity metric

Select the desired method to determine the degree of similarity between the entities or samples for correlation analysis from these options: **Pearson** or **Spearman**.

- Select the desired option from the **Similarity metric** drop-down list.
- Click **Next** ≫ to proceed to Step 2.

For more information on each method see "Similarity Metric" on page 799.

26.1.3.2 Step 2: Sample or Condition Pairing

This section describes how to pair samples or conditions in the second step of the wizard-driven correlation analysis workflow in a MOA experiment.

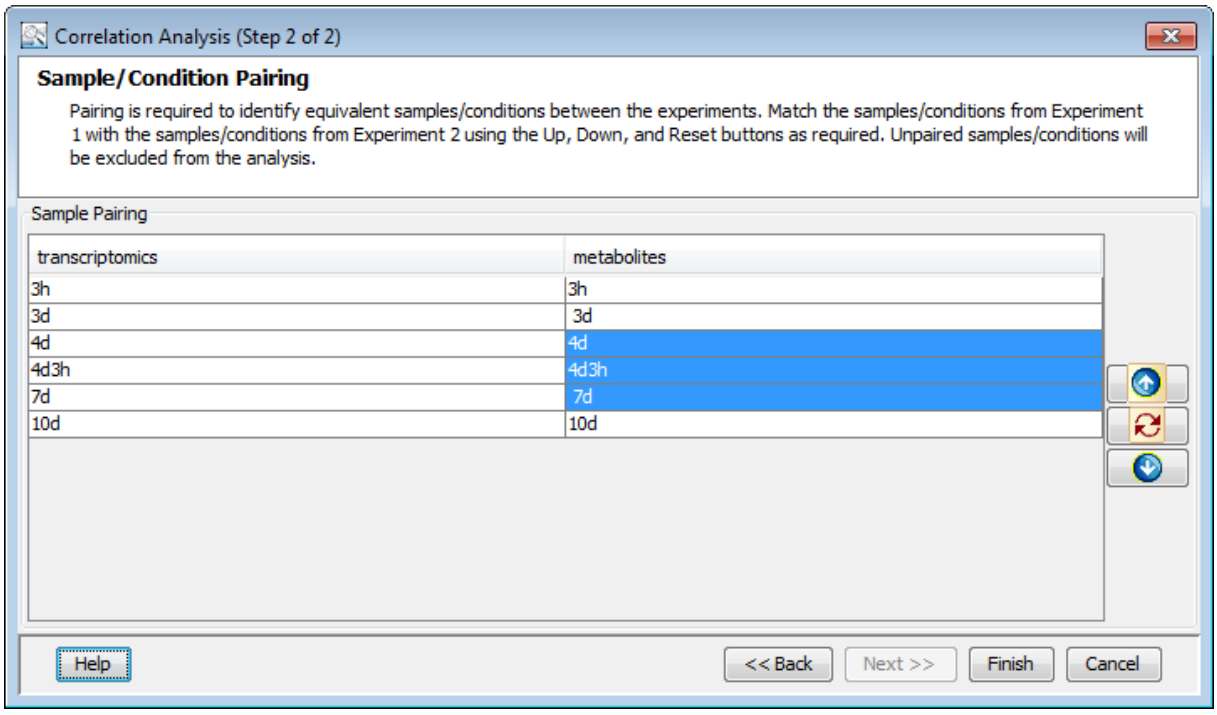


Figure 26.2: Correlation Analysis in a Multi-Omic Analysis experiment - Step 2 of 2

To pair samples or conditions

Samples or conditions from the two experiments have to be in pairs to perform correlation analysis. By default, **GeneSpring** displays the samples/conditions in the order in which they are present in the chosen interpretations. You can reorder the samples/conditions as required in this step.

1. Select a sample or condition to change their order. To select multiple samples/conditions and change their position in the order at the same time hold the <Ctrl> key and select the desired items.
2. Use the **Move up** , **Move down** , and **Reset** buttons to reorder the selected items, or refresh the order.
3. Repeat steps 1 and 2 until all the samples/conditions are paired up as required.
4. Click **Finish** to start correlation analysis.

NOTE

Correlation analysis is only performed where at least three pairs of samples or conditions are available. Pairing is required to identify equivalent samples or conditions between the experiments to perform correlation analysis. Unpaired samples or conditions from either of the experiments will be excluded from correlation analysis.

26.1.4 How to Explore the Correlation Heatmap View

Each correlation analysis you perform results in a correlation heatmap that is saved in the experiment hierarchy as a correlation node and which can be viewed and interactively in the desktop area as follows:

- To view a correlation heatmap in the desktop area:
  - Double-click the correlation analysis node in the Experiment Navigator.
  - The view is split into two panels vertically. The left panel shows the summary of the whole heatmap and the right panel shows a zoomed view of the selected area of the summary heatmap. By default the whole heatmap is selected in the summary view.
- To view the various parts of a correlation heatmap (see "The Correlation Heatmap View" on page 774:
  - Select or zoom into an area in the summary view to view it in the Detailed View.
  - Use the arrow keys to view neighboring areas from the selected heatmap area.
  - If you are using a computer mouse, moving the mouse scroll wheel will pan vertically by default. To pan horizontally, hold the <Ctrl> key and move the scroll wheel.
  - Left-click and drag out the mouse scroll in the detailed view for zoom in.
  - Left-click and drag in the mouse scroll in the detailed view for zoom out.
  - Click the **Redo** icon or **Undo** icon in the Correlation analysis heatmap Toolbar to change the summary view selection state.
- If clustering is performed on the correlation heatmap, to view the various parts of the clustered heatmap:
  - Click on a node to view the subtree or select **Column Tree Navigation Mode** or **Row Tree Navigation Mode** from the right-click options on the heatmap.
  - Press the <Shift> key and use the arrow keys to view child or parent nodes in a left hand tree or in a top tree.
  - Click the **Redo** icon or **Undo** icon in the Correlation Heatmap Toolbar to change the cluster tree selection state.
- Hover over a node, tree, or entity of interest to see the respective values (distance metric between nodes or correlation values) in the tool tip.
- Double-click on a heatmap cell to view the Correlation Coefficient scatter plot.

In a multi-omic correlation heatmap, use the **Show across experiment correlation view** icon and **Show whole correlation view** icon in the Correlation Heatmap Toolbar to toggle between the whole correlation view and the quadrant that contains the across experiment correlation result.

26.1.5 Working with a Correlation Heatmap View

This section describes the viewing and analysis options available in the Correlation Heatmap View including right-click menus and the functions accessible through the Correlation Heatmap View Toolbar.

26.1.5.1 To open a correlation node

- Double-click the correlation node name in the Experiment Navigator.
- Right-click over the correlation node name and select **Open Correlation Node**.

The correlation node opens in the Correlation Heatmap View in the main display area. The displayed entities or samples/conditions depend on the type of correlation analysis, the selected entity list, and interpretation as follows:

- **Correlation Analysis on Entities**: A correlation heatmap always displays entities of the entity list from which it was created.
- **Correlation Analysis on Samples/Conditions**: A correlation heatmap always displays samples/conditions of the interpretation from which it was created.
- **Correlation Analysis in a MOA experiment**: A correlation heatmap always displays entities present in the entities of both the experiments from which it was created.

NOTE Clicking on another entity list or interpretation in the Experiment Navigator after the correlation heatmap has opened does not result in any changes to the open correlation heatmap.

See the The Correlation Heatmap View on page 774 for more information on the individual components of the view.

26.1.5.2 To inspect a correlation node

- Right-click over the correlation node name and select **Inspect Correlation Node**.

A Correlation Node Inspector opens. The inspector holds all the available information associated with the creation of the correlation node, including its Name, Creation date, Last modified date, and Owner. The Notes contain all the correlation parameters that were used to create the correlation node.

You can edit the **Name** and **Notes**.

See "Correlation Node Inspector" on page 782.

26.1.5.3 To delete a correlation node

- Right-click the correlation node name in the Experiment Navigator and select **Delete Correlation Node**.

The correlation node is permanently deleted from **GeneSpring**.

26.1.5.4 To select entities on a correlation heatmap

For selecting entities on a correlation heatmap:

- Right-click on the heatmap and select **Selection Mode**.
- Left-click on the heatmap and drag across the desired section to select entities.

Selected entities are distinguished and highlighted by a green bar on the borders of the heatmap. A red bar on the border denotes the entities in the correlation analysis heatmap that match the entity list selected using the **Choose entity list to highlight matched entities** icon in the **Correlation Heatmap View Toolbar**. The entity list used to perform correlation analysis is selected by default to higlight entities in the heatmap with the red bar.

from the entity list on which correlation analysis was performed or the entity list selected using the **Choose entity list to highlight matched entities** icon in the **Correlation Heatmap View Toolbar**.

NOTE Selection Mode is not available for Correlation Analysis on Samples heatmaps.

26.1.5.5 To apply filters on a correlation heatmap

You can filter the entities displayed in a correlation heatmap on the available associated data in the entity list(s) selected for correlation analysis. Only if associated data is available for the entity list(s) on which correlation analysis was performed, the **Show filters** icon is enabled in the **Correlation Heatmap View Toolbar**.

1. Click icon in the **Correlation Heatmap View Toolbar** to open the filtering panel.
2. Reset the minimum and maximum values or select another filter type (check box or search field) from the available associated data columns for filtering. These filters will be applied to the correlation analysis heatmap view.
   - In a MOA correlation analysis heatmap, click on **Show filters** icon in the **Correlation Heatmap View Toolbar** to open the filtering panel.

- Filtering panel has two tabs, one for each experiment with their respective associated data filters.
- In a MOA correlation analysis heatmap, **Show filters** icon is enabled even when one of the two experiments have associated data columns.

3 In the filtering panel, click icon to open a **Column Selector** and select columns for filtering from the available associated data columns.

4 You can reset the filters applied by clicking the **Reset filter** icon in the filtering panel.

5 Click the **Hide filters** icon in the **Correlation Heatmap View Toolbar** to close the filtering panel.

NOTE This option is not available for Correlation Analysis on Samples heatmaps.

26.1.5.6 To hide or show the summary view of a correlation analysis heatmap

You can hide or show the summary view of a correlation heatmap. The summary view is useful to view the location of the area displayed by the detailed view within the entire heatmap when you have zoomed into a particular section of the heatmap.

- Click the **Hide summary view** icon in the **Correlation Heatmap View Toolbar** to close the summary view panel.
- Click the **Show summary view** icon in the **Correlation Heatmap View Toolbar** to open the summary view panel.

26.1.5.7 To toggle fullscreen mode of a correlation heatmap

You can view a correlation heatmap in fullscreen mode. The fullscreen view is useful to maximize the viewing area for large correlation heatmaps.

- Click the **Toggle fullscreen mode** icon in the **Correlation Heatmap View Toolbar** to expand the Correlation Heatmap View across the whole **GeneSpring** window and click the same icon again to return to the normal display mode limited to the **GeneSpring** desktop area.

26.1.5.8 To run clustering on a correlation heatmap

You can run clustering on a correlation analysis heatmap to organize genes/entities or samples/conditions in the dataset into clusters. Clusters can be based on the similarity of their correlation coefficient profiles using hierarchical clustering and data characterization techniques. The results of clustering analysis are represented visually in the form of cluster trees.

To run clustering on a correlation analysis heatmap, use the following procedure:

1. Click the **Run clustering** icon in the **Correlation Heatmap View Toolbar**. The **Clustering Inputs** dialog opens. See "Clustering Inputs" on page 783.
2. Select **Hierarchical** clustering algorithm (see "How to Perform Hierarchical Clustering Analysis" on page 671)
3. Select the desired **Distance metric** method to calculate the similarity between the entities or conditions for clustering from these options: **Eucledian**, **Squared Eucledian**, **Manhattan (Cityblock)**, **Maximum (Chebychev)**, **Minimum, Differential**, **Canberra**, **Harmonic**, **Pearson's Centered**, **Pearson's Uncentered (Cosine)**, **Pearson's Centered - Absolute**, or **Pearson's Uncentered - Absolute**.

   See "Distance Metric" on page 742 for the mathematical formulas for each of these options.
4. Select the desired method for representing the connection between clusters visually in the cluster tree from these options: **Ward's**, **Centroid**, **Median**, **Single**, **Complete**, **Average**.

   See "Linkage Rule" on page 744 for more details on each of these options.
5. Click **OK** to save the correlation analysis heatmap with clustering analysis cluster trees.

26.1.5.9 To remove clustering on a correlation analysis heatmap

- Click the **Remove clustering** icon in the **Correlation Heatmap View Toolbar**.

The cluster tree(s) are removed from the correlation heatmap and the entities are ordered as they are before **Run clustering**.

26.1.5.10 To save a filtered entities correlation analysis heatmap from a correlation analysis heatmap

To save a correlation heatmap that was filtered on associated data as a new correlation heatmap, use the following procedure:

1. Apply filters using the options from **Show filters** icon in the **Correlation Heatmap View Toolbar**.

2 Click the **Save filtered entities correlation** icon in the **Correlation Heatmap View Toolbar**.

The filtered entities correlation analysis heatmap is saved as a new correlation node below the parent correlation node in the Experiment Navigator.

NOTE Cluster trees of a filtered correlation analysis heatmap are not saved to the new filtered entities correlation analysis node. Open the filtered entities correlation analysis node and **Run clustering** for cluster trees. After filtering entities, **Run Clustering** applies the initial Clustering parameters selected by you to perform **Clustering Analysis**.

26.1.5.11 To show all entity labels in a correlation heatmap

- Click the **Show all entity labels** icon in the **Correlation Heatmap View Toolbar**.

The correlation heatmap display is automatically adjusted to display the entity labels of entities highlighted in the summary view in the detailed view. Left-click and drag over the summary view or detailed view of the correlation analysis heatmap to change the view field.

26.1.5.12 To fit all entities to view in a correlation analysis heatmap

- Click the **Fit all entities to view** icon in the **Correlation Heatmap View Toolbar**.

The correlation analysis heatmap display is automatically adjusted to fit all entities highlighted in the summary view into the detailed view. Left-click and drag over the summary view or detailed view of the correlation analysis heatmap to change the view field.

26.1.5.13 To undo or redo any actions in the summary view of a correlation analysis heatmap

You can undo or redo a zoom (left-click and drag) action in the summary view of a correlation analysis heatmap. This is useful to return the views to an earlier or later state for better visualization of the correlation analysis heatmap and the constituent entities, samples, or conditions.

- Click the **Undo** icon in the **Correlation Heatmap View Toolbar**.

The correlation analysis heatmap display is automatically adjusted to the previous state of the summary view. Click the **Undo** icon multiple times to view the actual launch zoom state in the summary view of the correlation analysis heatmap.

- Click the **Redo** icon in the **Correlation Heatmap View Toolbar**.
  The correlation analysis heatmap display is automatically adjusted to the later state of the summary view. Click the **Redo** icon multiple times to view the last action zoom state in the summary view of the correlation analysis heatmap.

26.1.5.14 To change the display settings of a correlation analysis heatmap

- Click the **Properties** icon in the **Correlation Heatmap View Toolbar**.
  The Properties dialog box opens. Change the color scheme of the correlation analysis heatmap, display or remove entity labels and rulers using the options provided in the Colors, Labels, Rulers, and Description tabs. See "Properties" on page 792.

26.1.5.15 To inspect selected entities from a correlation analysis heatmap

1. Left-click and drag over the heatmap area to select and zoom into an area.
2. Click the **Select all in zoomed view** icon in the **Correlation Heatmap View Toolbar**.
3. Click the **Inspect selected entity** icon in the main toolbar.

You can inspect all available details on the selected entities in the Entity Inspector that opens. This option is not available in an MOA experiment.

26.1.5.16 To create a new entity list from selected entities in a correlation heatmap

1. Left-click and drag over the heatmap area to select and zoom into an area.
2. Click the **Select all in zoomed view** icon in the **Correlation Heatmap View Toolbar**.
3. Click the **Create entity list from selection in view** icon in the **Correlation Heatmap View Toolbar**.
   An **Entitylist Inspector** opens.
4. Edit the default **Name** and **Notes** of the new entity list if desired.
5. Click **Configure Columns** to include or exclude any available Annotation columns.

6 Click **OK** to save the new entity list in your experiment.

For a correlation heatmap in a single experiment, the new entity list containing entities from the correlation heatmap view is saved in the Experiment Navigator as a new entity list.

For a correlation heatmap in a MOA experiment, the new entity lists containing entities from the correlation heatmap are saved in the MOA Experiment Navigator as new entity lists in the respective parent experiment folders.

26.1.5.17 To create a new entity list using the main toolbar option from selected entities in a correlation heatmap view

1 Left-click and drag over the heatmap area to select and zoom into an area.

2 Click the **Select all in zoomed view** icon in the **Correlation Heatmap View Toolbar**.

3 Click the **Create entity list from selection** icon in the main toolbar.
An **Entitylist Inspector** opens.

4 Edit the new entity list's default **Name** and **Notes** if desired.

5 Click **OK** to save the new entity list in your experiment.

The new entity list containing common entities from the correlation heatmap view and the active entity list is saved in the Experiment Navigator as a new entity list. This option is not available in MOA experiments.

26.1.5.18 To highlight matched entities from a selected entity list in a correlation heatmap view

In a correlation heatmap, you can highlight matched entities from a different entity list to the original entity list used for running **Correlation Analysis**. This is useful to match and visualize entities from a different entity list in comparison to the entities in the entity list used for **Correlation Analysis**.

1 Click the **Choose entity list to highlight matched entities** icon in the **Correlation Heatmap View Toolbar**.
An **Entity List Chooser** dialog opens.

2 Click **Choose...**
A **Choose Entity List** dialog opens.

3 Select a entity list to highlight entities in the correlation analysis heatmap.

4 Click **OK** to visualize the common entities between the selected entity list and the original entity list of correlation analysis.

You can change the transparency of the entities absent in the selected entity list and the correlation analysis entity list from the **Properties** ⟶ **Colors** tab. See "Properties" on page 792.

NOTE In a MOA Correlation Analysis experiment, you can choose one entity list for each of the experiments to highlight entities.

26.1.5.19 To view whole correlation analysis heatmap of a MOA experiment

In a MOA correlation heatmap view, you can view a whole heatmap of correlations across the two experiments in a single view. This is useful to visualize and understand the correlation of entities within an experiment as well as across experiments. See To view across experiments correlation analysis heatmap in a MOA experiment on page 764 to toggle between whole heatmap and across experiments correlation heatmap.

- On completion of Correlation Analysis in a MOA experiment, a heatmap with correlations across the two experiments is displayed.
- Click **Show whole correlation view** icon to display the whole correlation analysis heatmap for within and across experiments correlations.

26.1.5.20 To view across experiments correlation analysis heatmap in a MOA experiment

In a MOA correlation heatmap view, you can view a heatmap of correlations across the two experiments in a single view. This is useful to visualize and understand the correlation of entities across experiments. See To view whole correlation analysis heatmap of a MOA experiment on page 764 to toggle between across experiments correlation heatmap and whole heatmap.

- Click **Show across experiment correlation view** icon to display the correlation analysis heatmap for across experiments correlations.

26.1.5.21 To export a correlation heatmap as a report

You can save the visible correlation analysis heatmap as a report or download as a **.pdf* file.

Choose an option to export:

- **Summary View**: The current Summary View pane is exported.
- **Detailed View**: The current Detailed View pane is exported.
- **Combined View**: The current Combined View of the heatmap is exported.

NOTE If a report contains > 100 pages (default setting), it cannot be viewed in **GeneSpring**. You can download the report as a **.pdf* or modify report settings from **Tools** ⟶**Options** ⟶**Miscellaneous** ⟶**Report**.

To export a correlation analysis heatmap with entity labels:

1. Select an area in the **Summary View**.
2. Click the **Show all entity labels** icon.
3. Right-click on either of the panes, select **Export as Report** and choose one of the available options from, **Summary View**, **Detailed View**, or **Combined View**.
4. Modify the other parameters as required and click **Preview**.
5. Click **Save Report** to save the report in the experiment navigator. This report will include the **Summary View** pane, or **Detailed View** pane, or the **Combined View** along with the **Simple** or **Advanced** report parameters as selected.
6. If the Report has > 100 pages, download the report as a **.pdf* file using the **Download** icon.

NOTE To export a correlation analysis heatmap without entity labels, do not click on the **Show all entity labels** icon.

26.1.5.22 To export a correlation heatmap as an image

You can export a complete correlation heatmap as an image (*.tiff, *.png, *.jpg, *.jpeg, *.bmp, *.svg, *.eps).

Choose an option to export:

- **Summary View**: The current Summary View pane is exported.
- **Detailed View**: To export the complete Detailed View, uncheck the **Export only the visible region** and proceed. To export only the current Detailed View Pane, use the default options and proceed.
- **Combined View**: The current Combined View of the correlation heatmap is exported.

NOTE Exporting a complete **Detailed View** in the default *.tiff, *.svg, *or .eps format will result in a single image of the detailed correlation heatmap. If you select a big region or the complete Detailed View to export, selecting any other image format will result in a folder with a collection of images representing the Detailed View of the correlation analysis heatmap.

To export a correlation analysis heatmap with entity labels:

1. Select an area in the **Summary View**.
2. Click the **Show all entity labels** icon.
3. Right-click on either of the panes, select **Export as Image** and choose one of the available options from, **Summary View**, **Detailed View**, or **Combined View**.
4. Specify a location to save the image and choose an image format.
5. Clear the option **Export only the visible region** to export the complete **Detailed View** with entity labels to the specified location.
6. Use the default setting for **Export only the visible region** to export only the visible region. This image will include the **Summary View** pane, or **Detailed View** pane, or the **Combined View** as selected.
7. Modify the other parameters as required and click **OK**.

NOTE To export a correlation heatmap without entity labels, do not click on the **Show all entity labels** icon.

26.1.5.23 To open a correlation coefficient Scatter plot

Correlation coefficient scatter plot provides a graphical representation of the empirical data values that were used to calculate the correlation coefficient between an entity pair.

For opening a correlation coefficient scatter plot, use the following procedure:

1. Double-click on a correlation analysis heatmap cell.
   A **Correlation Coefficients** dialog opens.
   The regression fit and equation in the plot display the direction and strength of the dependency between a pair of entities or samples.

26.1.5.24 To export correlation coefficients of a correlation heatmap

You can export the correlation coefficients of a correlation analysis heatmap as a text file (*.txt, *.tsv, *.csv). For correlation analysis on entities heatmap, associated data, and annotations can also be exported.

For correlation analysis on entities heatmap, use the following procedure to export correlation coefficients:

1 Right-click on the correlation analysis heatmap and choose **Export Correlation Coefficients**.
   A **Export Correlation Coefficients** dialog opens.
2 Select the entities for which you want to export correlation coefficients from these options: **All**, **Filtered**, **Selected**.
3 Select which **Annotation** columns you want to export by moving the desired annotations from the **Available items** field to the **Selected items** field or vice versa using the Move Right and Move Left buttons.
4 Reorder the selected items, or refresh the order using the Up , Down , and Refresh Order buttons.
5 Click **OK**.
   A **Save** dialog box opens.
6 Provide a file name and determine where you want to save the list. Click **Save**.

NOTE In the option to select entities for which you want to export correlation coefficients, selecting;

- **All**: Includes all entities present in the correlation heatmap.
- **Filtered**: Includes only entities that satisfy the **Filters** settings (see To apply filters on a correlation heatmap) and are displayed in the correlation heatmap.
- **Selected**: Includes only entities that have been selected in the heatmap view. For this, the correlation coefficients for the selected entities against all the entities in the heatmap will be exported.

See "Export options" on page 789.

For a Correlation Analysis on Samples heatmap, use the following procedure to export correlation coefficients:

1 Right-click on the correlation analysis heatmap and choose **Export Correlation Coefficients**.
2 A **Save** dialog box opens.
3 Provide a file name and determine where you want to save the list. Click **Save**.

26.1.5.25 To export correlation coefficients of a correlation heatmap in a MOA experiment

You can export the correlation coefficients of a MOA correlation analysis heatmap as a text file (*.txt, *.tsv, *.csv). For a MOA correlation analysis heatmap, associated data, and annotations can also be exported.

In a MOA experiment, use the following procedure to export correlation coefficients:

1. Right-click on the correlation analysis heatmap and choose **Export Correlation Coefficients**.
   A **Export Correlation Coefficients** dialog opens.
2. Select the entities you want to export correlation coefficients, from these options: **All**, **Filtered**, **Selected**.
3. From each of the experiment tabs, select which **Annotation** columns you want to export by moving the desired annotations from the **Available items** field to the **Selected items** field or vice versa using the move right and move left buttons.
4. Reorder the selected items, or refresh the order using the up , down , and refresh order buttons.
5. Click **OK**.
   A **Save** dialog box opens.
6. Provide a file name and determine where you want to save the list. Click **Save**.

NOTE

In the option to select entities for which you want to export correlation coefficients, selecting;

- **All**: Includes all entities present in the correlation heatmap.
- **Filtered**: Includes only entities that satisfy the **Filters** settings (see To apply filters on a correlation heatmap) and are displayed in the correlation heatmap.
- **Selected**: Includes only entities that have been selected in the heatmap view. For this, the correlation coefficients for the selected entities against all the entities in the heatmap will be exported.

See "Export options" on page 789.

26.1.6 Correlation analysis result visualizations in GeneSpring

In **GeneSpring** correlation analysis always results in a heatmap to visualize the pair-wise relationship between the chosen variable: entities or samples. However, depending on whether correlation analysis was performed in a single experiment or in a MOA experiment and if clustering was performed on the correlation heatmap, some of the characteristics and functionalities of the heatmap vary.

The following sections describe the purpose and functionality of each of these result visualizations in **GeneSpring**: correlation heatmap, MOA correlation heatmap, and clustered correlation heatmap.

26.1.6.1 Correlation heatmap

Correlation analysis visualizes pair-wise relationships between entities or samples in an experiment. The result of correlation analysis is visualized as a heatmap drawn on the computed correlation coefficients.

Correlation coefficient values range from $+1$ to -1, where positive correlation signifies entities or samples with similar behavior and negative correlation indicating entities or samples with opposite behavior. Entities or samples with a 0 or near 0 correlation indicate no linear relationship between the entity or sample pair.

You can zoom in to a correlation heatmap and apply horizontal and vertical cell spacing to view the heatmap as individual color cells. Each color cell represents an entity-entity or a sample-sample correlation value. Double-click on a cell to view a correlation scatter plot showing the trend and distribution of entity values. On the scatter plot, the normalized intensity values of an entity are represented as points and the trendline shows the direction of correlation.

The correlation heatmap view in **GeneSpring** shows the same heatmap in two panels. The view is split into two panels horizontally. The left panel shows the whole heatmap and the right panel shows a zoomed view of the selected area of the whole heatmap. You can navigate along the selected area of the heatmap using the arrow keys to inspect certain cells more closely.

Entity or sample labels for each of the entities or samples are displayed beside and below the heatmap. By default, a Correlation Analysis on Entities heatmap has a red color bar along the right and bottom edges of the heatmap indicating the presence of the entity in the original correlation analysis entity list. An additional highlight color bar is shown to highlight matched entities. See "To highlight matched entities from a selected entity list in a correlation analysis heatmap view" on page 763. Heatmap colors and transparency settings for non-matched entities can be modified through the **Colors** tab of the **Properties** dialog box. See "Properties" on page 792.

Entity labels can reflect any one of the annotations available for the entities in the selected entity list. The type of annotation can be determined through the **Labels** tab of the **Properties** dialog box. See Properties on page 792.

NOTE For a Correlation on Samples heatmap, the highlight color bar is not available.

You will notice a red highlight bar along the right and bottom borders of the correlation heatmap. This bar represents or marks the entities that matched the original input entity list. It serves as a visual aid when you start contrasting this correlation result with other entity lists. Select an entity list using **Choose entity list to highlight matched entities** icon to make it the active entity list. This highlights the entities in the heatmap matching the active entity list and correlation analysis input entity list. This is useful when the heatmap contains a large number of entities.

The legend shows the color range of the heatmap and a description of parameters on which correlation analysis was performed.

Hovering over the cells of the heatmap shows the entity or sample pair correlation values in a tool tip. The tool tip on the cells also shows the complete entity or condition pair information.

See "How to Perform Correlation Analysis on Entities" on page 749 and "How to Perform Correlation Analysis on Samples" on page 750.

26.1.6.2 MOA Correlation heatmap

A MOA correlation heatmap is a concatenated view displaying entities from the first experiment with entities from the second experiment. This visualization method removes the process of normalizing values between two experiments. A MOA correlation heatmap is divided into four quadrants showing entity correlation between and across experiments.

You can zoom into a MOA correlation heatmap and apply horizontal and vertical cell spacing to view the heatmap as individual color cells. Each color cell represents an entity-entity correlation value. Double-click on a cell to view a correlation scatter plot showing the trend and distribution of entity values. On the scatter plot, the normalized intensity values of an entity are represented as points and the trendline shows the direction of correlation.

The Correlation Heatmap View in **GeneSpring** shows the same heatmap in two panels. The view is split into two panels horizontally. The left panel shows the whole heatmap and the right panel shows a zoomed in view of the selected area of the whole heatmap. You can navigate along the selected area of the heatmap using the arrow keys to inspect certain cells more closely.

Entity labels for each of the experiments' entities are displayed beside and below the heatmap in different colors. A MOA correlation on entities heatmap has a highlight color bar along the right and bottom edges of the heatmap indicating the presence of the entity in the original correlation analysis entity lists. An additional highlight color bar is shown to highlight matched entities. See "To highlight matched entities from a selected entity list in a correlation analysis heatmap view" on page 763. Heatmap colors and transparency settings for non-matched entities can be modified through the **Colors** tab of the **Properties** dialog box. See "Properties" on page 792.

Entity labels can reflect any one of the annotations available for the entities in the selected entity lists. The type of annotation can be determined through the **Labels** tab of the **Properties** dialog box. See "Properties" on page 792.

You will notice a red highlight bar along the right and bottom borders of the correlation heatmap. This bar represents or marks the entities that matched the original input entity lists. It serves as a visual aid when you start contrasting this correlation result with other entity lists. Select entity lists using **Choose entity lists to highlight matched entities** icon to make them the active entity lists. This highlights the entities in the heatmap matching the active entity lists and correlation analysis input entity lists. This is useful when the heatmap contains a large number of entities.

For example, consider a MOA experiment is created on two experiments A and B and their respective entity lists have m and n entities. If correlation analysis is performed using the entity lists with m and n entities, the resulting MOA correlation heatmap features:

- A concatenated heatmap is visualized with $m + n$ entities on both X- and Y-axis.
- The heatmap has four quadrants showing correlation of $m * m$, $n * n$, $m * n$, and $n * m$ entity values.
- You can click on the **Show across experiment correlation view** icon and visualize only the $m * n$ quadrant of the heatmap.
- The summary view displays a selection denoting the zoomed or selected state in the detailed view.

Also see "How to Perform Correlation Analysis in a MOA experiment on page 752".

26.1.6.3 Clustered Correlation heatmap

You can perform hierarchical clustering on a correlation heatmap to produce hierarchical cluster trees based on the correlation values of entities, samples, or entities of different experiments in MOA correlation.

In a MOA correlation heatmap, clustering of entities is based only on the correlation values of the entities across an experiment which is based on the pairing option selected in Step 2: Sample or Condition Pairing . This results in a subset of entities based on the profile of their relationship with entities of the second experiment and the resultant correlation values are distributed into clusters.

For example, in a case where expression data is being correlated with miRNA data, clustering based on across experiment correlation values will help in visualizing genes which show a similar or opposite linear relationship with a miRNA.

See "To run clustering on a correlation analysis heatmap" on page 760.

26.2 Correlation Analysis Screen Reference

This section describes the user interface elements of the Correlation Analysis window, as well as toolbar icons, dialog boxes, and right-click options associated with the workflow and its results.

26.2.1 Correlation Analysis Window Features

This window allows you to determine the inputs and settings for either correlation analysis on entities or on samples.

26.2.1.1 Correlate on drop-down list

Use this drop-down menu to select on what you want to perform correlation analysis:

- Entities (see How to Perform Correlation Analysis on Entities on page 749)
- Samples (see How to Perform Correlation Analysis on Samples on page 750)

26.2.1.2 Entity List Choose... button

The **Entity List** field displays the active entity list in the Experiment Navigator (appears in bold) by default. Click **Choose...** to change the entity list if required. Select the desired entity list in the **Choose EntityList** dialog box that opens. See Choose Entity List on page 780.

NOTE For best results, do not select the All Entities list for correlation analysis. Best results are obtained when the list of entities has been filtered to remove entities with measurements that are mostly in the noise range or that are mostly flagged absent.

26.2.1.3 Interpretation Choose... button

The **Interpretation** field displays the active interpretation in the Experiment Navigator by default. Click **Choose...** to change the interpretation if required. Select the desired interpretation in the **Choose Interpretation** dialog box that opens. See Choose Interpretation on page 781.

To perform correlation analysis the chosen interpretation in combination with the chosen entity list has to satisfy the minimum condition that there are at least three paired data points present in the selected data. This minimum requirement can be fulfilled differently depending on whether you perform correlation analysis on entities or samples:

- Entities: If you choose the **All Samples** or a non-averaged interpretation there has to be a minimum of three samples as well as at least two entities for which data is available in each of these three samples. If you choose an averaged interpretation there have to be a minimum of three conditions defined in the interpretation as well as at least two entities for which data is available in each of these conditions. In a MOA experiment, the chosen interpretations and entity lists have to fulfill these requirements for each input experiment.
- Samples: If you choose the **All Samples** or a non-averaged interpretation there have to be at least two samples present as well as a minimum of three entities for which data is available in each of these two samples. If you choose an averaged interpretation there have to be a minimum of two conditions as well as a minimum of three entities for which data is available in each of these two conditions.

26.2.1.4 Similarity metric drop-down list

Select the desired method to calculate the similarity between the entities or samples for correlation from these options: **Pearson** or **Spearman**.

See "Similarity Metric" on page 799 for more details on each method.

26.2.1.5 OK button

Click **OK** to save the result of correlation analysis as a correlation node in your experiment.

26.2.1.6 Cancel button

Closes the Correlation Analysis window. **GeneSpring** does not save the last settings that you specified. The next time you start correlation analysis, the default settings are displayed.

26.2.1.7 Help button

Click **Help** to find out more about correlation analysis and the available parameter settings in the Correlation Analysis window from the user manual.

26.2.2 The Correlation Heatmap View

The following sections describe all the display components and functional elements of the Correlation Heatmap View: view components, the correlation-specific toolbar as well as the available right-click options.

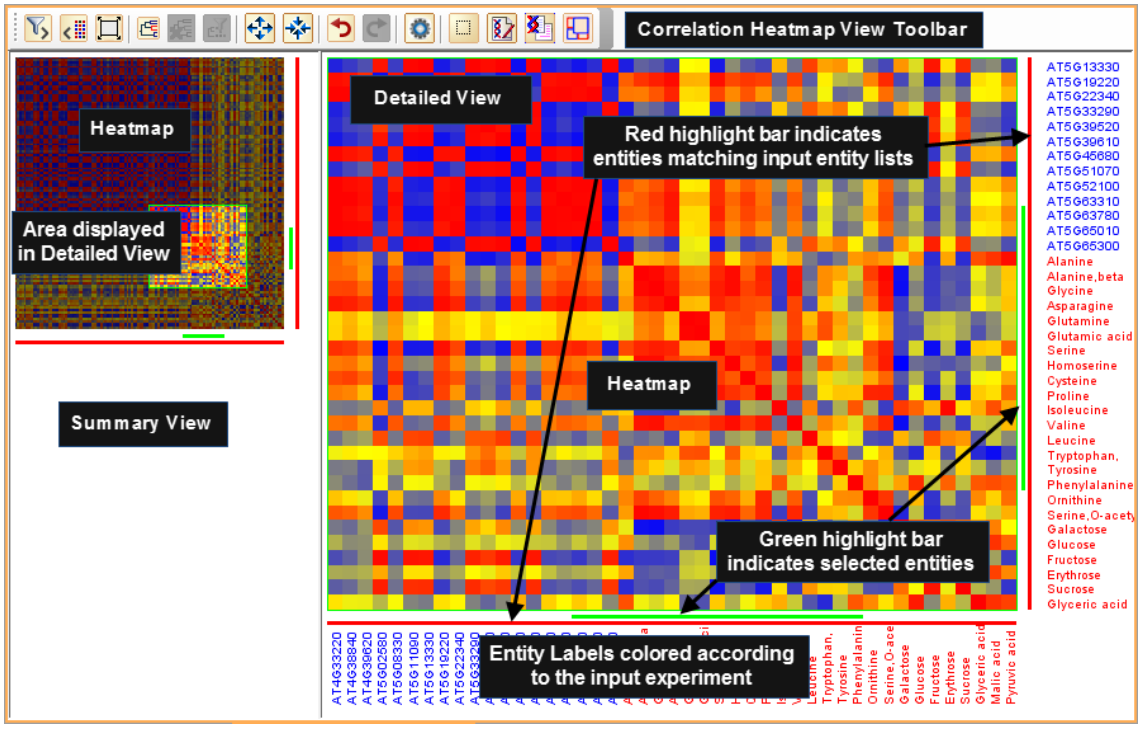


Figure 26.3: MOA Correlation View

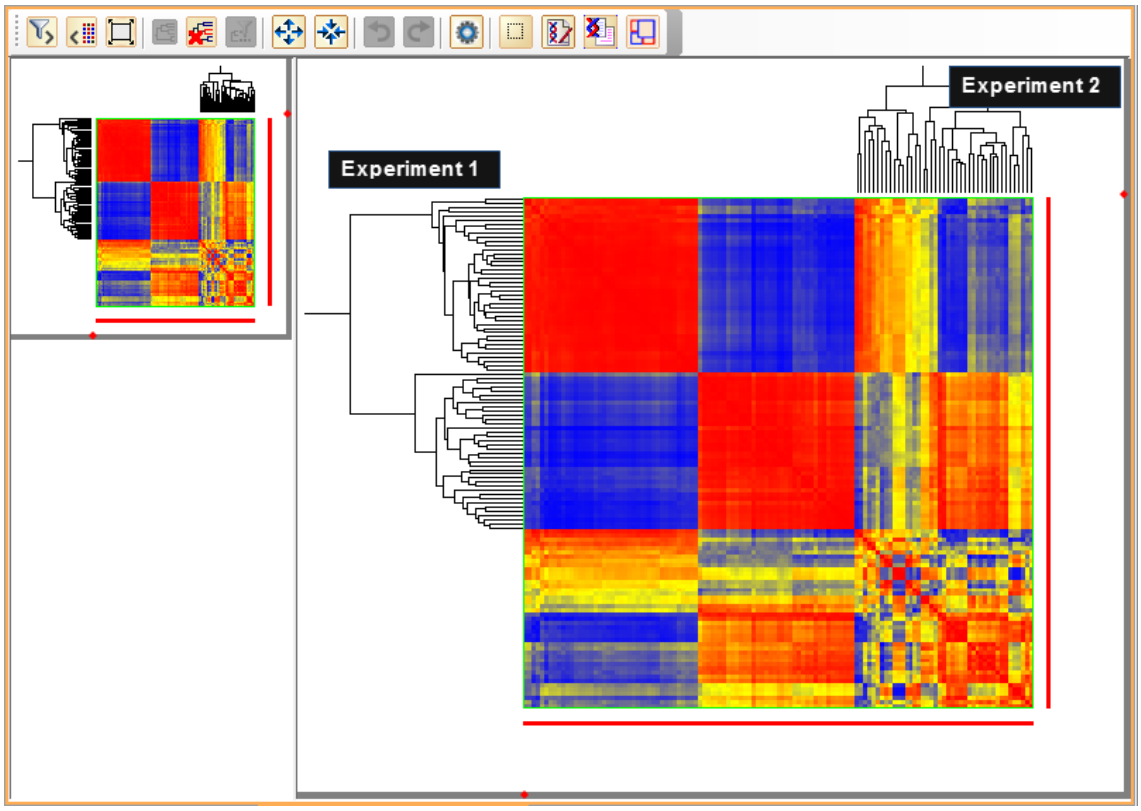


Figure 26.4: Clustered Correlation Heatmap view of an MOA experiment

26.2.2.1 Correlation Heatmap View Components

The Correlation Heatmap View consists of a **Summary View** pane and a **Detailed View** pane.

The **Summary View** always displays the entire correlation analysis heatmap. If clustering is performed on the correlation heatmap, the entire cluster tree is also shown. To explore a correlation heatmap, you can use any one of the options described under How to Explore a Correlation Analysis Result in the Heatmap View on page 756 and How to explore a cluster tree result" on page 679. The selected or zoomed in area in the **Summary View** is displayed in the **Detailed View**.

In the **Detailed View**, you can further zoom into the parts of the heatmap that you selected in the **Summary View**. At any given time, the area corresponding to the part of the heatmap that is displayed in the **Detailed View** is also highlighted in the **Summary View**.

26.2.2.2 Correlation Heatmap View Toolbar

The Correlation Heatmap View Toolbar provides access to visualization and analysis functions, including filter options, show/hide summary view, toggle full screen mode, run clustering, remove clustering, save filtered entities correlation heatmap, show all entity labels, fit all entities to view, undo, redo, display properties, select all in zoomed view, create entity list from selection in view, choose entity lists to highlight matched entities, and show complete entity correlation/show cross experiment entity correlation (only in MOA experiments). The availability of each function represented by an icon in the toolbar depends on the availability of associated data (for filtering) and prior actions carried out on the heatmap view. Table 26.2 details the icons, their utility, and availability.

Table 26.2: Function and Availability of Corrrelation Heatmap View Toolbar icons

Toolbar Icon	Function	Availability
	Show filters+	If associated data is available for filtering
	Hide filters+	If filters panel is already active
	No filters	If associated data is not available
	Hide summary view	If summary view is open
	Show summary view	If summary view is closed
	Toggle fullscreen mode	If correlation analysis heatmap view is in the main display area
	Run Clustering	If hierarchical clustering analysis was not performed
	Remove Clustering	If hierarchical clustering analysis was performed on the correlation heatmap
	Save filtered entities correlation+	If filtering was applied on the correlation heatmap
	Show all entity labels	Always available
	Fit all entities to view	If zoom is applied on the correlation analysis heatmap
	Undo	If an action is performed on the correlation analysis heatmap
	Redo	If multiple actions are performed on the correlation analysis heatmap
	Properties	Always available
	Select all in zoomed view+	If zoom is applied on the correlation analysis heatmap
	Create entity list from selection *+	If entities are selected in the correlation analysis heatmap
	Choose entity lists to highlight matched entities+	Always available
	Show whole correlation view#	On launch of a MOA correlation analysis heatmap view
	Show across experiment correlation view#	If the whole correlation view of a MOA experiment is active

+ - Available only for correlation analysis on entities and MOA experiment correlation analysis heatmap view.
* - See **NOTE**.
# - Available only in MOA experiment correlation analysis heatmap view.

NOTE If you select entities in the correlation heatmap and click the **Create entity list from selection** icon in the Main Toolbar, only those correlation heatmap entities that match with the active entity list are saved in a new entity list. If you click the **Create entity list from selection** icon in the Correlation Heatmap View Toolbar, all the selected entities from the correlation heatmap are saved.

26.2.2.3 Legend

The legend at the bottom right corner of the **GeneSpring** desktop area contains the color-coding information for the correlation heatmap and the legend for **Experiment Label Color** for a MOA correlation heatmap. It also contains a summary of the parameters used to create the heatmap. For more information on how to change the color settings see "To change the display settings of a correlation analysis heatmap" on page 762.

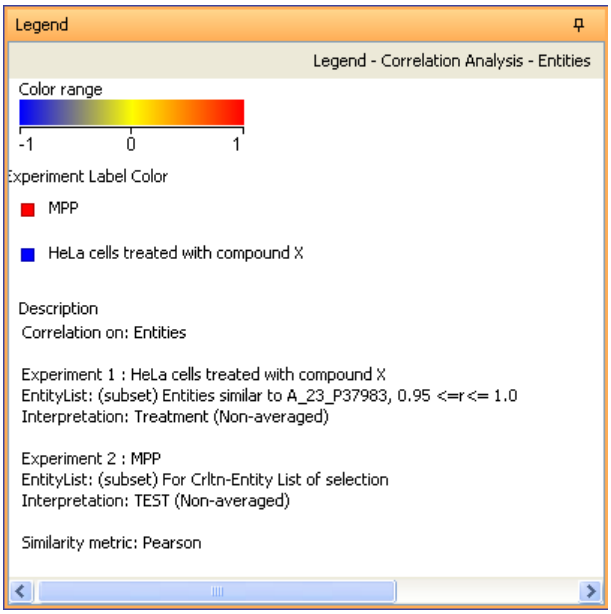


Figure 26.5: Legend of a MOA correlation analysis heatmap

26.2.2.4 Right-click options on a Correlation Heatmap View

The available right-click options on a correlation analysis heatmap are summarized in Table 26.3.

Table 26.3: Menu options available on right-click over a correlation analysis heatmap

Option	Description
Zoom Mode	Activates the zoom functions.
Selection Mode *	Activates the selection of entities on the correlation analysis heatmap.
Column Tree Navigation Mode	Available after clustering and activates the navigation functions for the column tree of entities.
Row Tree Navigation Mode	Available after clustering and activates the navigation functions for the row tree of entities.
Select All *	Selects all entities in the correlation analysis heatmap.
Invert Selection *	Inverts the entity selection.
Clear All Selections *	Deselects any selected entities.
Reset Zoom	Resets the zoomed level in the **Detailed View** to the selected view in the **Summary View**.
Copy View	Copies the current heatmap view to the system clipboard.
Export As • Report • Image • HTML	Allows you to export the **Summary**, **Detailed**, or **Combined View** as a report or in any one of these image formats: *.tiff, *.png, *.jpg, *.jpeg, *.bmp, *.svg., *.eps, or the current heatmap view in *.html format.
Export Correlation Coefficients	Exports the correlation coefficients of al the entities or samples to a *.txt, *.tsv, or *.csv file.
Properties	Allows you to change the display settings of the heatmap view.

* - Not available for a **Correlation Analysis on Samples** heatmap.

26.2.3 Dialog Boxes

This section describes all the dialog boxes that appear during Correlation Analysis.

26.2.3.1 Choose Entity List

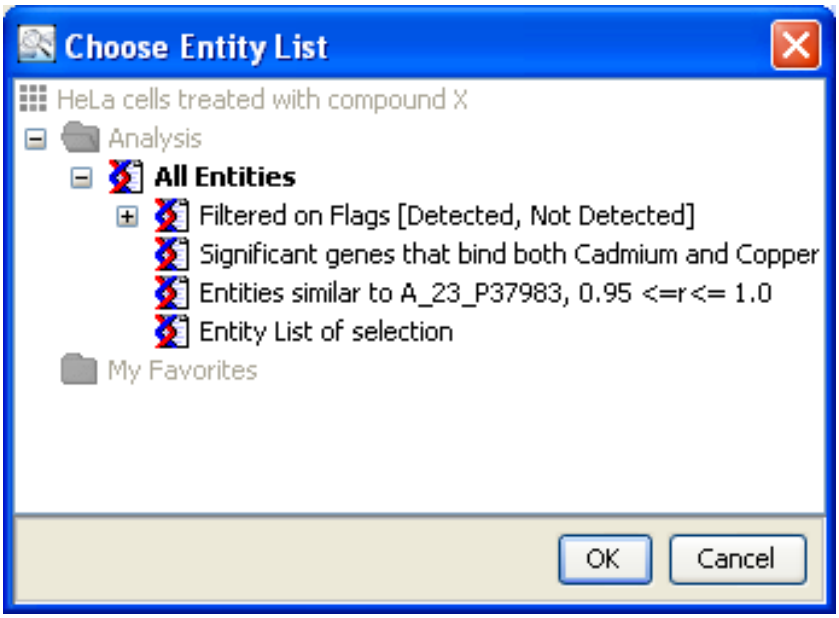


Figure 26.6: Choose Entity List dialog box

Purpose: Select the entity list that contains the entities on which you want to perform Correlation Analysis.

To open: Click **Choose...** in the Correlation Window. See "How to Perform Correlation Analysis on Entities" on page 749, "How to Perform Correlation Analysis on Samples" on page 750.

OK Confirms the selected entity list, closes the dialog and returns to the Correlation Analysis Window.

Cancel Aborts the selection process and returns to the Correlation Analysis Window.

26.2.3.2 Choose Interpretation

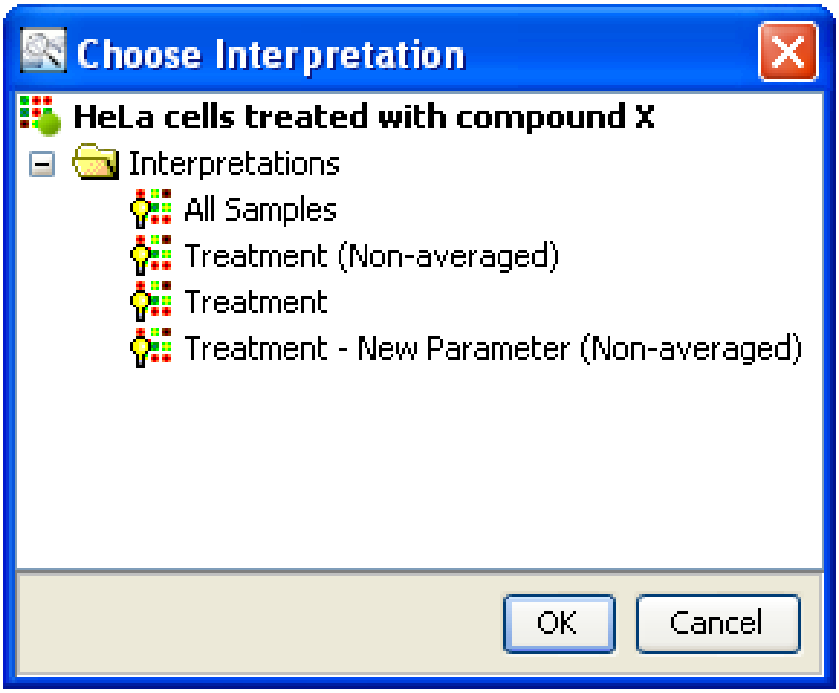


Figure 26.7: Choose Interpretation dialog box

Purpose: Select an interpretation based on which you want to perform Correlation Analysis from the available interpretations.

To open: Click **Choose...** in the Correlation Window. See "How to Perform Correlation Analysis on Entities" on page 749, "How to Perform Correlation Analysis on Samples" on page 750.

OK Confirms the selected interpretation, closes the dialog and returns to the Correlation Analysis Window.

Cancel Aborts the selection process and returns to the Correlation Analysis window.

26.2.3.3 Correlation Node Inspector

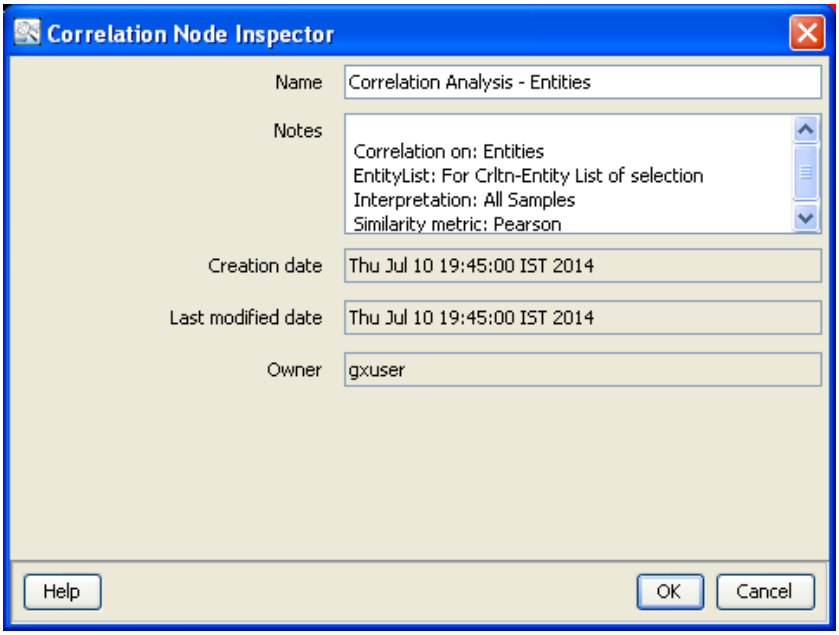


Figure 26.8: Correlation Node Inspector dialog box

Purpose: Review the details that are recorded with the result of Correlation Analysis. You can change the default Name and Notes.

To open: Right-click the correlation analysis node in the Experiment Navigator and select **Inspect Correlation Node**. See "To inspect a correlation node" on page 757.

Help Opens the **GeneSpring** manual in your default browser.

OK Closes the dialog and applies any changes to the correlation node.

Cancel Closes the dialog box. No changes are applied.

26.2.3.4 Clustering Inputs

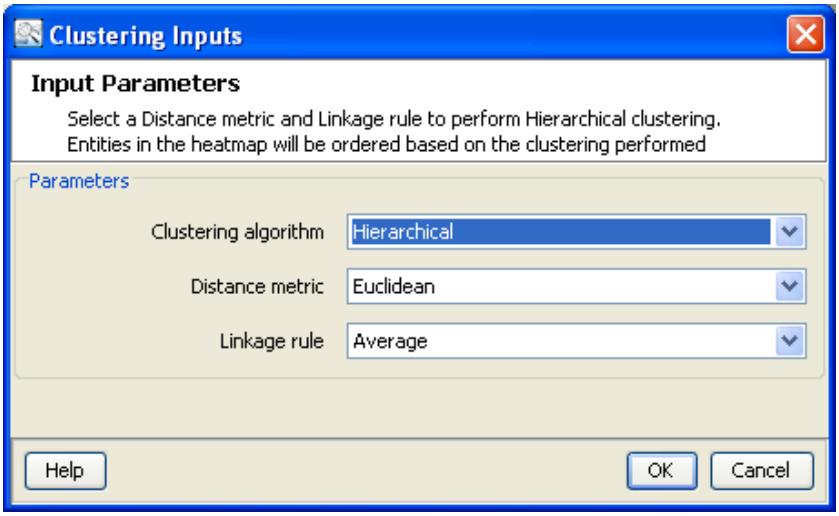


Figure 26.9: Correlation Analysis Clustering Inputs dialog box

Purpose: Select parameters to perform hierarchical clustering on the correlation analysis heatmap.

To open: Click the **Run clustering** icon from the correlation analysis heatmap toolbar. See "To run clustering on a correlation heatmap" on page 760.

Help Opens the **GeneSpring** manual in your default browser.

OK Closes the dialog and applies cluster trees to the correlation heatmap.

Cancel Closes the dialog box. No changes are applied.

26.2.3.5 Entitylist Inspector

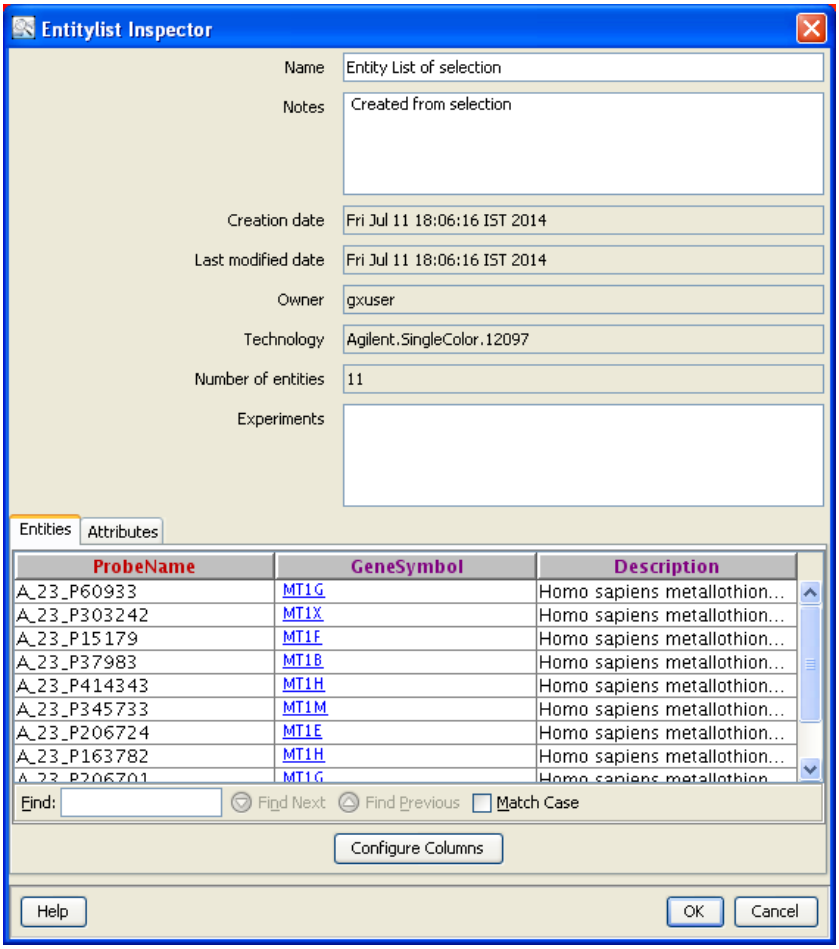


Figure 26.10: Correlation Analysis Create Entity List from selection in view dialog box

Purpose: Create an entity list from the selected entities in a correlation heatmap.

To open: Click **Create entity list from selection in view** icon from the correlation heatmap view toolbar. See "To create a new entity list from selected entities in a SEA correlation heatmap view" on page 762.

Help	Opens the **GeneSpring** manual in your default browser.
Configure Columns	Opens the **Select Annotation Columns** dialog.
OK	Closes the dialog and applies highlighting to the matched entities in the correlation heatmap.
Cancel	Closes the dialog box. No changes are applied.

26.2.3.6 Entity Lists Inspector

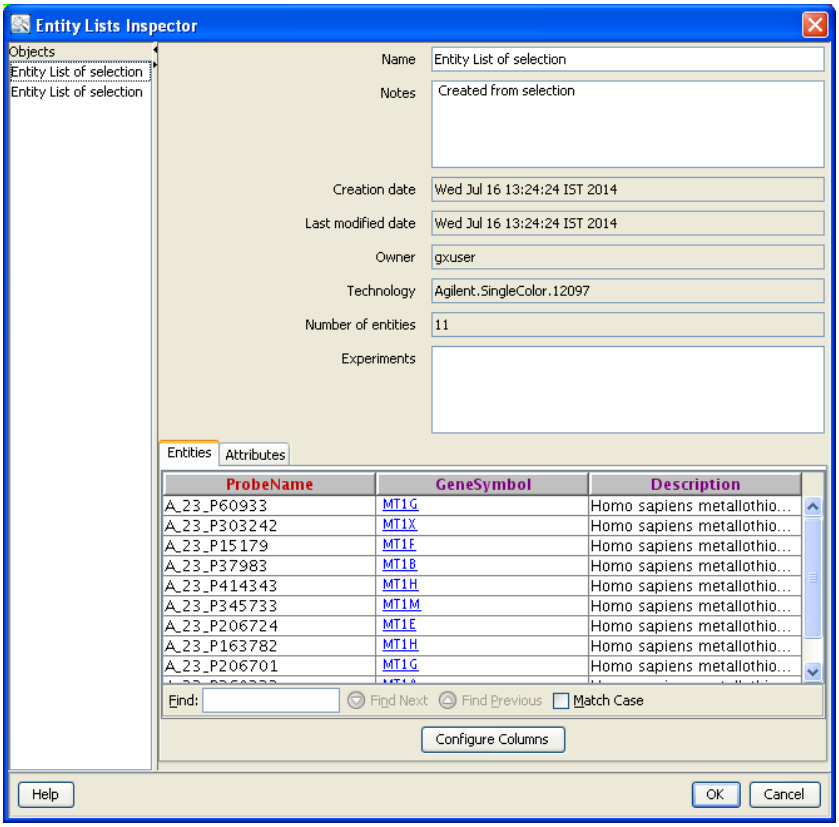


Figure 26.11: MOA Correlation Analysis Create Entity List from selection in view dialog box

Purpose: Create an entity list for each experiment from the selected experiment entities in a correlation heatmap.

To open: Click **Create entity list from selection in view** icon from the correlation heatmap view toolbar. See "To create a new entity list from selected entities in a MOA correlation heatmap view" on page 762.

Help	Opens the **GeneSpring** manual in your default browser.
Configure Columns	Opens the **Select Annotation Columns** dialog.
OK	Closes the dialog and applies highlighting to the matched entities in the correlation heatmap.
Cancel	Closes the dialog box. No changes are applied.

26.2.3.7 Select Annotation Columns

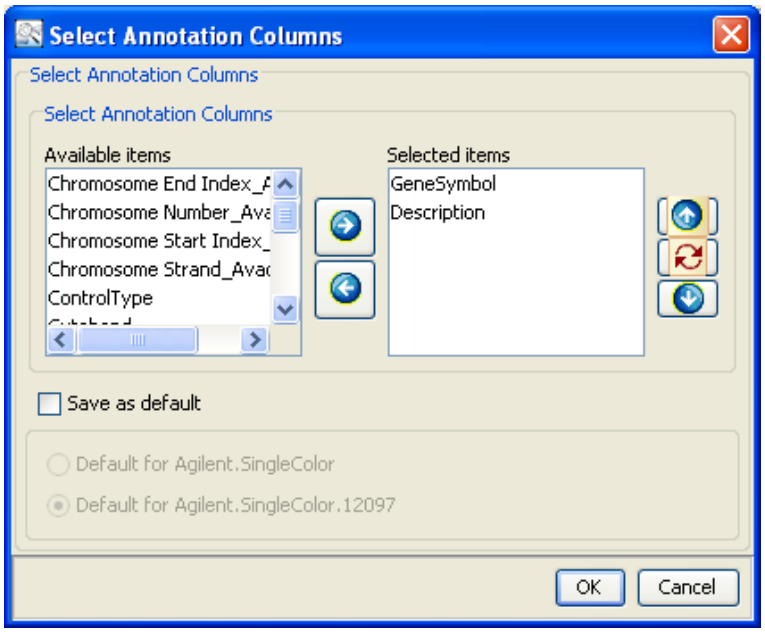


Figure 26.12: Select Annotation Columns dialog box

Purpose: Select available individual columns in the selected entity list to be included in the new entity list being created.

To open: Click **Configure Columns** in a **Entitylist Inspector** or a **Entity Lists Inspector** dialog. See "To create a new entity list from selected entities in a SEA correlation heatmap view" on page 762 and "To create a new entity list from selected entities in a MOA correlation heatmap view" on page 762.

Select Annotation Columns	Determine the columns you want to appear in the new entity list. Only columns listed under **Selected items** are displayed. You can also change the order in which the selected columns appear in the new entity list.
OK	Closes the dialog box. The selected columns and their associated data is included in the new entity list.
Cancel	Aborts the selection process and returns to the Entitylist Inspector or Entity Lists Inspector dialog box.

26.2.3.8 Entity List Chooser in a SEA experiment

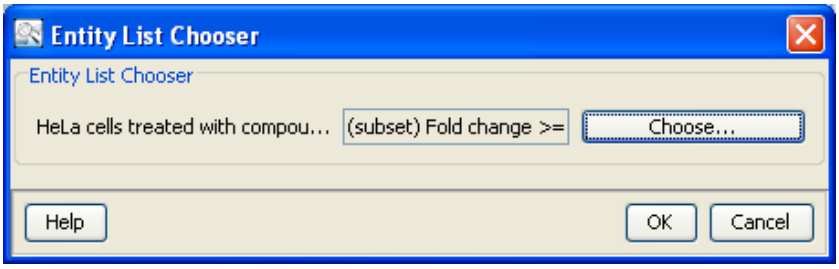


Figure 26.13: Correlation Analysis Entity List Chooser dialog box in a SEA experiment

Purpose: Select an entity list to match and highlight common entities with the entity list on which the correlation analysis heatmap is created.

To open: Click the **Choose entity list to highlight matched entities** icon from the correlation analysis heatmap toolbar. See "To highlight matched entities from a selected entity list in a correlation heatmap view" on page 762.

Help	Opens the **GeneSpring** manual in your default browser.
Choose...	Opens the **Choose Entity List** dialog.
OK	Closes the dialog and applies highlighting to the matched entities in the correlation heatmap.
Cancel	Closes the dialog box. No changes are applied.

26.2.3.9 Entity List Chooser in a MOA experiment

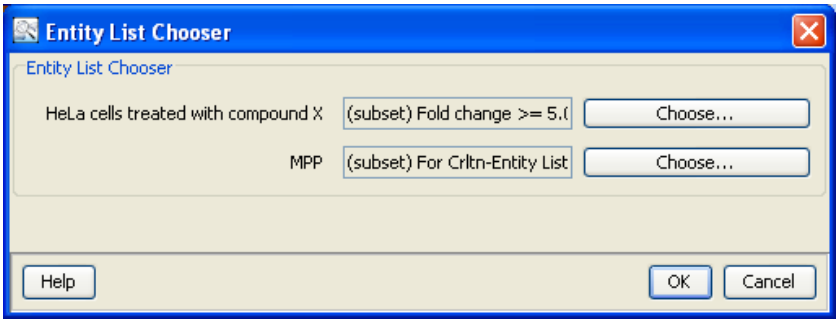


Figure 26.14: Correlation Analysis Entity List Chooser dialog box in a MOA experiment

Purpose: Select an entity list from each experiment to match and highlight common entities with the entity list on which the correlation heatmap is created.

To open: Click **Choose entity lists to highlight matched entities** icon from the correlation heatmap view toolbar. See "To highlight matched entities from a selected entity list in a correlation heatmap view" on page 762.

Help	Opens the **GeneSpring** manual in your default browser.
Choose...	Opens the **Choose Entity List** dialog.
OK	Closes the dialog and applies highlighting to the matched entities in the correlation heatmap.
Cancel	Closes the dialog box. No changes are applied.

26.2.3.10 Export options

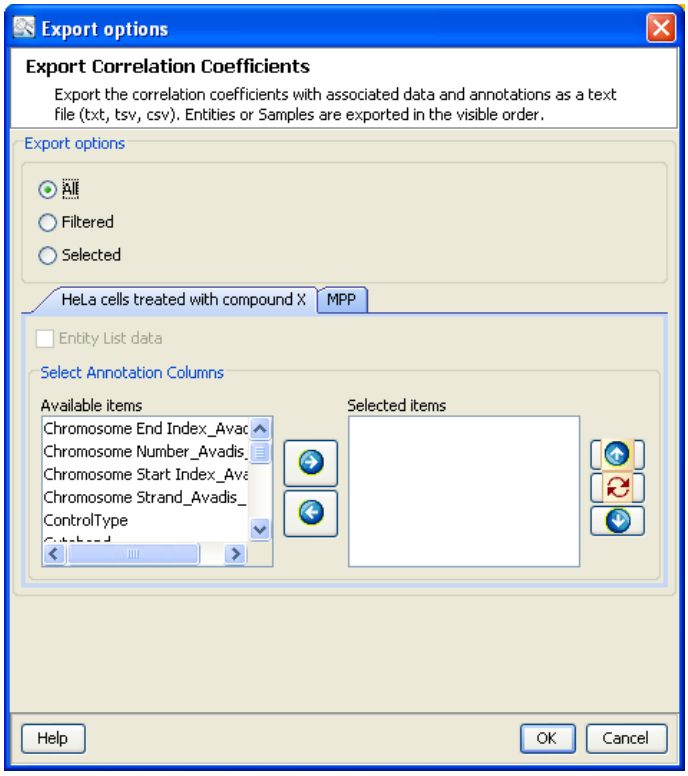


Figure 26.15: Export options dialog box to export correlation coefficients

Purpose: Export the correlation coefficients of the correlation heatmap in a *.txt, *.tsv, or, *.csv file. Select the data and annotation columns to be exported along with the correlation coefficients.

To open: Right-click on the correlation heatmap and select **Export Correlation Coefficients**. See "To export correlation coefficients of a correlation heatmap" on page 766 and "To export correlation coefficients of a correlation heatmap in a MOA experiment" on page 767.

Help Opens the **GeneSpring** manual in your default browser.

OK Closes the dialog and allows you to specify the file name and location to save the correlation coefficients in a *.txt, *.tsv, or, *.csv file.

Cancel Closes the dialog box and returns to the correlation heatmap view. The entities are not exported.

26.2.3.11 Report Exporter

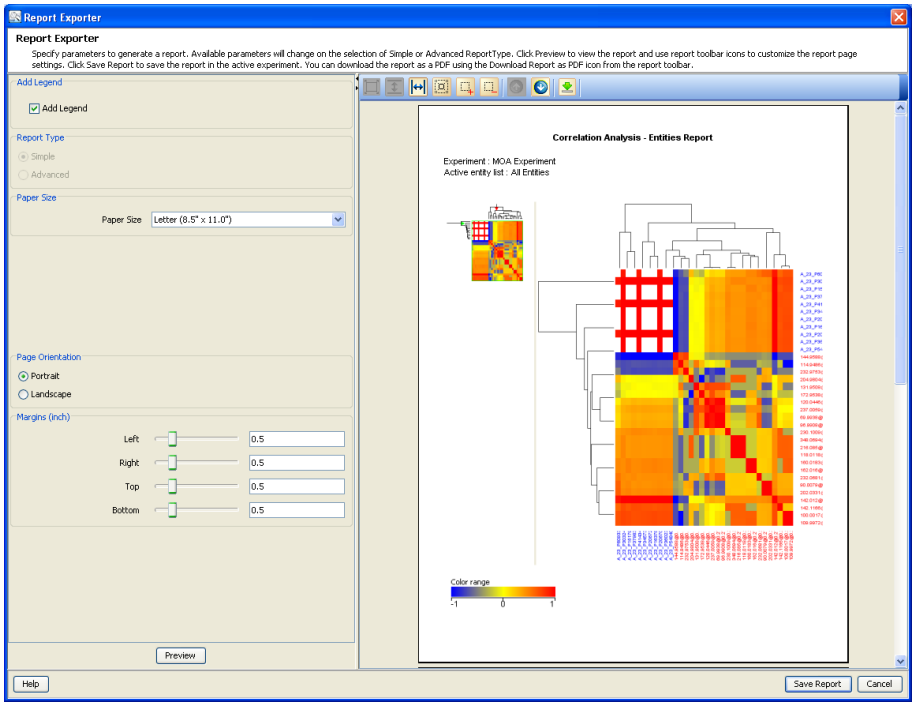


Figure 26.16: Export as Report options dialog box

Purpose: Export the correlation heatmap as a Report or download as a **.pdf* file.

To open: Right-click the correlation heatmap and select **Export as Report**. See "To export a correlation heatmap as a report" on page 764.

Help	Opens the **GeneSpring** manual in your default browser.
Save report	Closes the dialog and saves the correlation analysis report in the experiment navigator.
Cancel	Closes the dialog box and returns to the correlation heatmap view. The correlation analysis report is not exported.

26.2.3.12 Print Options

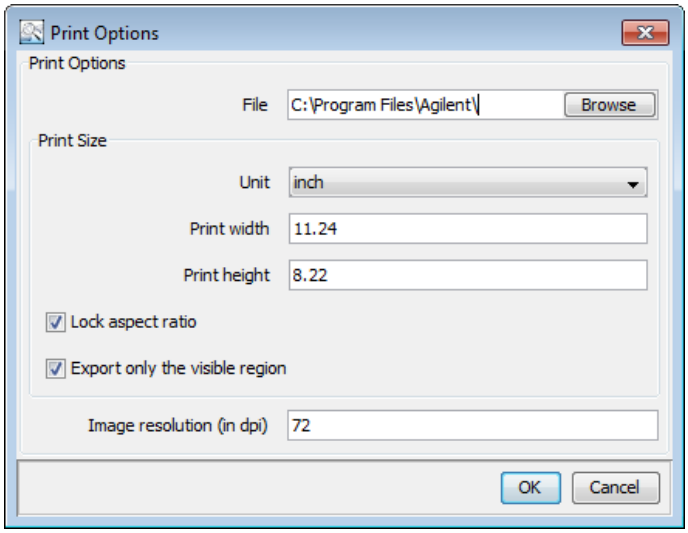


Figure 26.17: Export as Image options dialog box

Purpose: Export the correlation heatmap as an image (with or without entity labels).

To open: Right-click the correlation heatmap and select **Export as Image**. See "To export a correlation heatmap as an image" on page 765.

OK Closes the dialog and saves the correlation heatmap image at the specified location.

Cancel Closes the dialog box and returns to the correlation heatmap view. The correlation heatmap image is not exported.

26.2.3.13 Properties

Purpose: Inspect the display settings of the current correlation analysis heatmap. You can change the default color settings, labels to display, ruler settings, title, and description. The Properties dialog box has five tabs: **Colors**, **Labels**, **Rulers**, and **Description**. Each tab is described further in the following sections.

To open: Click the **Properties** icon in the **Correlation Heatmap View Toolbar** or right-click on the correlation heatmap view and select **Properties**. See "To change the display settings of a correlation heatmap" on page 762.

Help	Opens the **GeneSpring** manual in your default browser.
OK	Closes the dialog and applies any changes to the open correlation heatmap.
Cancel	Closes the dialog box.

Colors tab

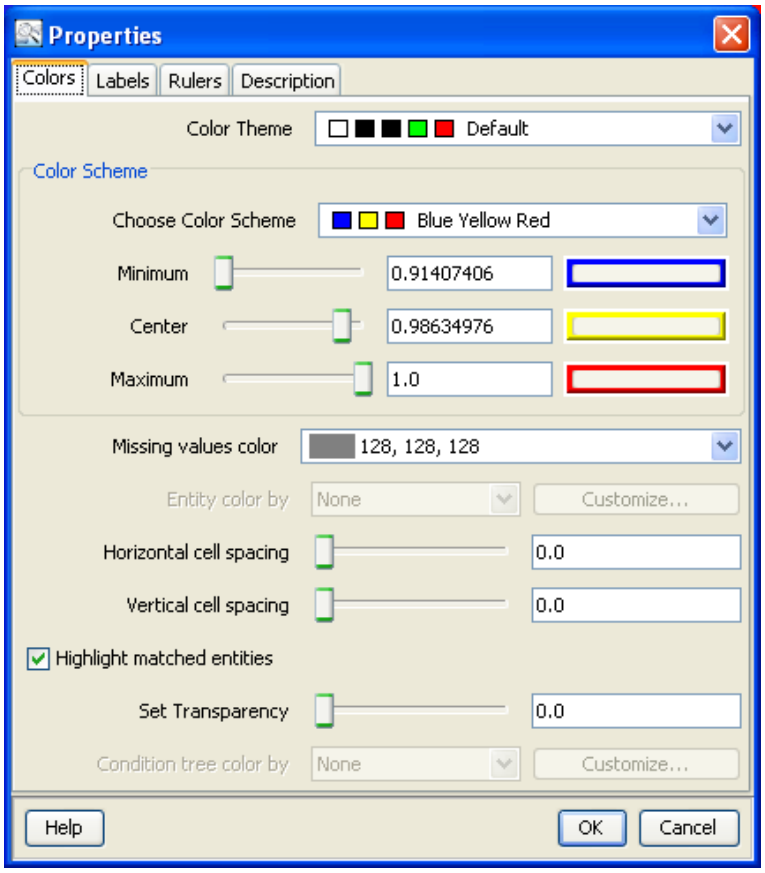


Figure 26.18: Properties dialog box - Colors tab

Customize the coloring options for the background, heatmap, and entity labels in the correlation analysis heatmap view.

Color Theme Choose from either **Default** or create a **Custom Color Theme** (see Create New Colors Theme on page 794) for the correlation analysis heatmap view.

Color Scheme Select one of the existing color schemes from the **Choose Color Scheme** drop-down list. To define a custom color scheme select the **Custom Color Scheme** option.

Specify the **Missing values color**.

Use the sliders to modify the value for **Horizontal cell spacing** and **Vertical cell spacing**. Cell spacing is available on the heatmap and you can view it only in the zoomed view of the correlation heatmap.

Use the slider to modify the value for **Set Transparency** and change the level of transparency of unmatched entities between a selected entity list and the correlation analysis entity list in the correlation heatmap.

Create New Color Theme

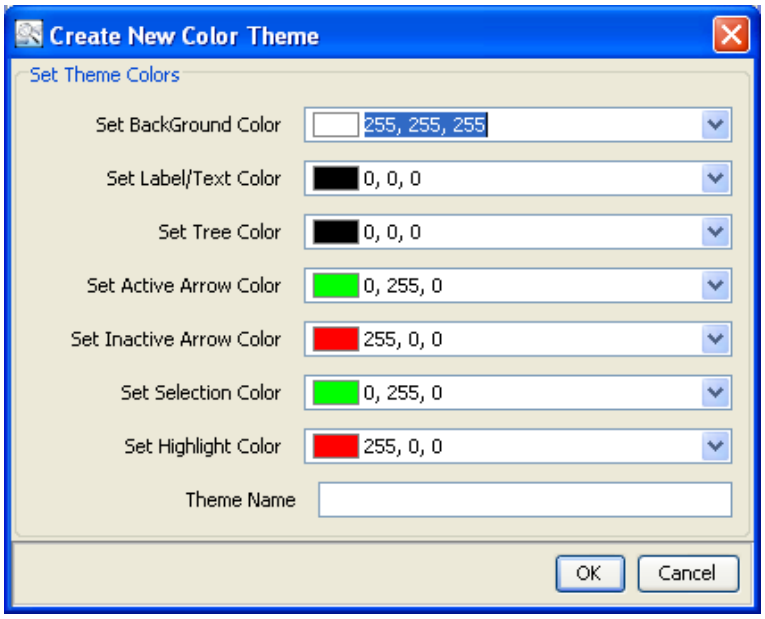


Figure 26.19: Properties dialog box - Colors tab - Create New Colors Theme

In a correlation heatmap, customize the colors for Background, Label or Text, Tree, Active and Inactive Arrows, Selection Bar, and Highlight Bar.

Set Color Choose a color of your choice from the the available colors for each of the available correlation heatmap components.

Theme Name Enter a name for the new color settings.

Labels tab

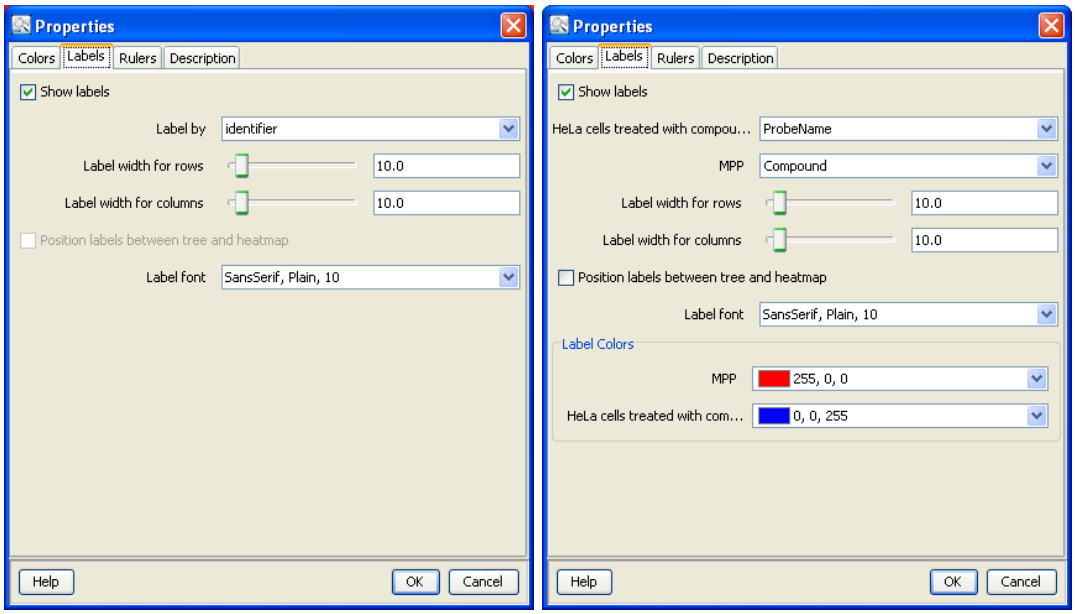


Figure 26.20: Properties dialog box - Labels tab

Select the display options for labels.

Show labels Choose to display or remove labels from the correlation analysis heatmap view.

Select from which annotation column the labels should be sourced from the **Label by** or **Experiment Name** drop-down list.

Determine the display width for the entity labels using the **Label width for rows** and **Label width for columns** slider or type the value and confirm by pressing <**Enter**>.

Position labels between tree and heatmap Select this option after clustering the correlation heatmap to display entity labels between the tree and heatmap, rather than below the heatmap.

Label font Select the font, size, and style for labels. The available fonts from the drop-down list are all the fonts installed on your system.

Label Colors Option available only in a MOA Correlation Analysis heatmap. Select one of the existing colors for each experiment from the drop-down list or use **More Colors** to select a color.

Rulers tab

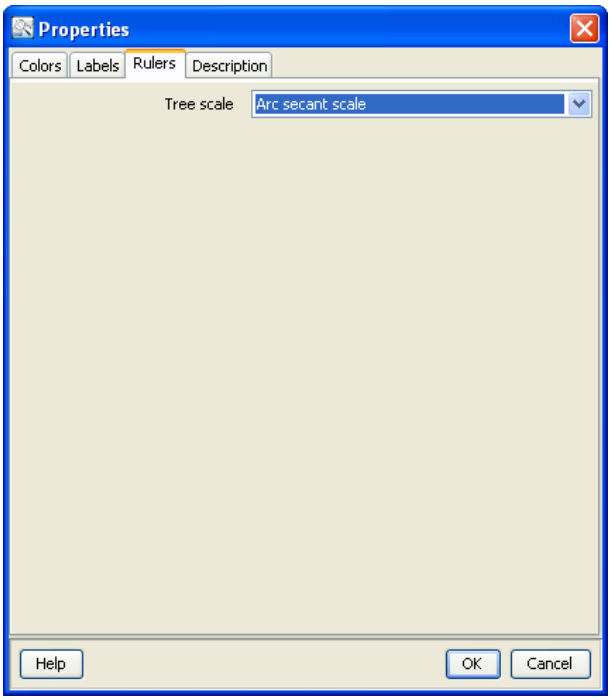


Figure 26.21: Properties dialog box - Rulers tab

Rulers tab is only available after performing clustering on the correlation analysis heatmap. Select the display settings for the cluster tree distance metric rulers.

Tree Scale Select the scale according to which the distance metric values are displayed from the **Tree Scale** drop-down list. Choose from these options: **Square root scale**, **Arc secant scale**, **True scale**, and **Log scale**.

Description tab

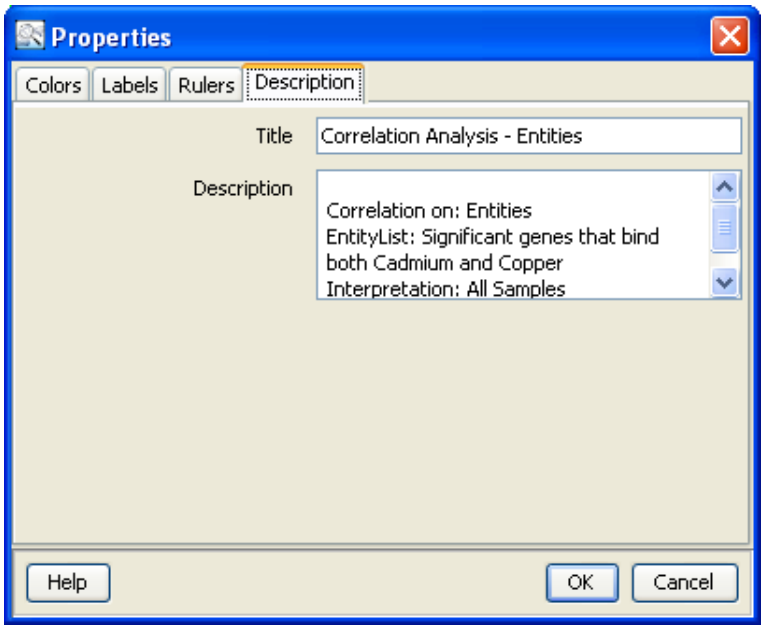


Figure 26.22: Properties dialog box - Description tab

Review the title of the correlation analysis node and a description of the parameters that were used to create it.

Title Contains the name of the correlation analysis node.

Description Contains information about all the parameters that were used to create the correlation analysis node.

NOTE You can make changes to both the **Title** and **Description** fields in this dialog. However, these changes are not saved when you close the correlation analysis heatmap.

26.3 Correlation Analysis - Statistical Algorithms

This section explains the statistical algorithms and options available for each algorithm involved in performing correlation analysis.

26.3.1 Correlation computation in GeneSpring

This section details on the computation of correlation coefficients in **GeneSpring**.

26.3.1.1 Entity-entity or sample-sample correlation in a SEA experiment

A table of expression values is created from the normalized intensities of entities and samples from the chosen Entity List and Interpretation. For correlation analysis on entities, correlation coefficients are computed between each pair of rows in the expression table. For correlation analysis on samples, correlation coefficients are computed from the column profiles in the expression table.

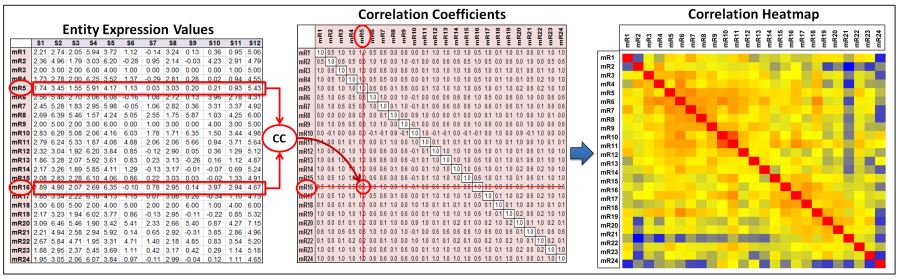


Figure 26.23: Computation of pair-wise correlation coefficients between entities in a SEA experiment

26.3.1.2 Entity-entity correlation in a MOA experiment

In a MOA experiment, an expression table is created individually for each of the experiments based on the selected Entity List and Interpretation for Correlation Analysis. The two expression tables are concatenated based on the sample pairing information specified in the Step 2: Sample or Condition Pairing (see Step 2: Sample or Condition Pairing on page 754) of MOA Correlation Analysis.

If m entities are present in the selected Entity List from experiment 1 and n entities are present in the selected Entity List from experiment 2, the concatenated expression table will have $m + n$ entities. Pair-wise entity correlation coefficients are computed between each pair of rows in the concatenated expression table.

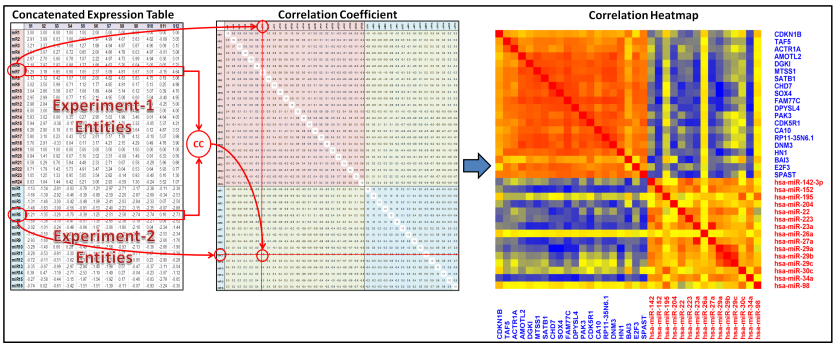


Figure 26.24: Computation of pair-wise correlation coefficients between entities of two different experiments in a MOA experiment

26.3.2 Similarity Metric

A correlation algorithm is required to calculate correlation co-efficient which determines the measure the similarity or difference between entities or samples. There are several similarity metrics that are available to determine the similarity between individual entities or samples while computing the similarity scores. The available metric options in **GeneSpring** are explained in this section along with their respective computation formulae.

See "How to Perform Correlation Analysis on Entities" on page 749, "How to Perform Correlation Analysis on Samples" on page 750, "How to Perform Correlation Analysis in a MOA experiment" on page 752.

26.3.2.1 Pearson

Pearson similarity metric is widely used for finding the strength in linear relationships within normally distributed data. If your data is not normally distributed it is advised to use the **Spearman** method [62].

Formula

$$r = \frac{\Sigma_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\Sigma_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\Sigma_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

26.3.2.2 Spearman

Spearman similarity metric is a rank based algorithm and is more tolerant towards outliers present in data sets [63].

Formula

$$\rho = 1 - \frac{6\Sigma d_i^2}{n(n^2 - 1)}$$

where,

$$d_i = x_i - y_i$$

Chapter 27

Class Prediction: Learning and Predicting Outcomes

27.1 General Principles of Building a Prediction Model

Classification algorithms in **GeneSpring** are a set of powerful tools that allow researchers to exploit microarray data for building prediction models. These tools stretch the use of microarray technology into the arena of diagnostics and understanding the genetic basis of complex diseases. Classification predicts the class label of an input object. The input data set is divided into a **Training Set** and a **Validation Set**; the Training set is used for creating a function for prediction of unknown class labels.

The Validation set is used to confirm and validate the initial analysis.

Prediction models in **GeneSpring** are built based on the expression profile of conditions. This model is then used to predict the condition class of an unknown sample. For example, given gene expression data for different kinds of cancer samples, a model which can predict the cancer type for an new sample can be learnt .

GeneSpring provides five powerful machine learning algorithms for building the classification models - Decision Tree (DT), Neural Network (NN), Support Vector Machine (SVM), Naive Bayesian (NB) and PLSD. Models built with these algorithms can then be used to classify samples or genes into discrete classes based on its gene expression.

The models built by these algorithms range from visually intuitive ones (as with Decision Trees) to abstract ones (as for Support Vector Machines). Together, these methods constitute a comprehensive toolset for learning, classification and prediction.

GeneSpring provides a workflow link to build a model and predict the sample from gene expression data.

27.2 Prediction Pipeline

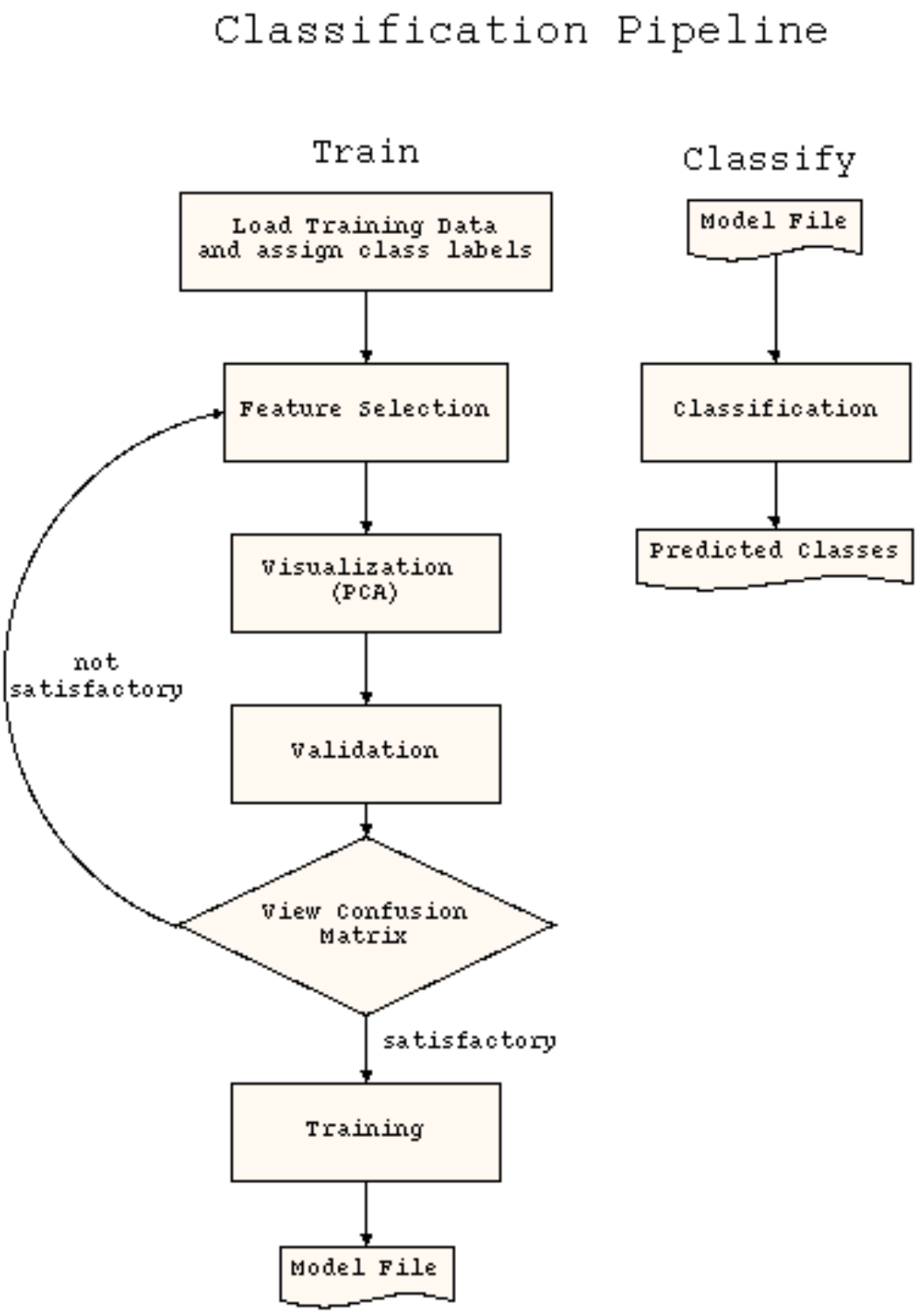


Figure 27.1: Classification Pipeline

The problem statement for building a prediction model is to build a robust model to predict phenotypic samples from gene expression data. The model is built with the dependent variable being the sample type and the independent variables being the genes and their expression values corresponding to the sample. With the above example, given the gene expression profiles of the different types of cancerous tissue, you want to build a robust model to predict the nature of the sample from the model. The model should not over-fit the data used for building the model.

Once the model has been validated, it can be saved and used to predict the outcome of a new sample from gene expression data of the sample. See Figure 27.1.

Note: All classification algorithms in **GeneSpring** for prediction of discrete classes (i.e. SVM, NN, NB,DT and PLSD) allow for validation, training and classification.

27.2.1 Validate

Validation helps to choose the right set of features or entity lists, an appropriate algorithm and associated parameters for a particular dataset. Validation is also an important tool to avoid over-fitting models on training data as over-fitting will give low accuracy on validation. Validation can be run on the same dataset using various algorithms and altering the parameters of each algorithm. The results of validation, presented in the Confusion Matrix (a matrix which gives the accuracy of prediction of each class), are examined to choose the best algorithm and parameters for the classification model.

Two types of validation have been implemented in **GeneSpring**.

Leave One Out: All data with the exception of one row is used to train the learning algorithm. The model thus learnt is used to classify the remaining row. The process is repeated for every row in the dataset and a Confusion Matrix is generated.

N-fold: The classes in the input data are randomly divided into N equal parts; N-1 parts are used for training, and the remaining one part is used for testing. The process repeats N times, with a different part being used for testing in every iteration. Thus each row is used at least once in training and once in testing, and a Confusion Matrix is generated. This whole process can then be repeated as many times as specified by the number of repeats.

The default values of three-fold validation and one repeat should suffice for most approximate analysis. If greater confidence in the classification model is desired, the Confusion Matrix of a 10-fold validation with three repeats needs to be examined. However, such trials would run the classification algorithm 30 times and may require considerable computing time with large datasets.

27.2.2 Prediction Model

Once the results of validation are satisfactory, as viewed from the confusion matrix of the validation process, a prediction model can be built and saved. The results of training yield a Model, a Report,

a Confusion Matrix and a plot of the Lorenz Curve. These views will be described in detail later.

27.3 Running Class Prediction in GeneSpring

Class prediction can be invoked from the workflow browser of the tool. There are two steps in class prediction; building prediction models and running prediction. Each of these takes you through a wizard collecting inputs providing visual outputs for examination and finally saving the results of building and running prediction models.

27.3.1 Build Prediction Model

The *Build Prediction Model* workflow link launches a wizard with six steps for building a prediction model.

Input Parameters The first step of building prediction models is to collect the required inputs. The prediction model is run on an entity list and an interpretation. The model is built to predict the interpretation based upon the expression values in the entity list. The entity list should thus be a filtered and analyzed entity list of genes that are significant to the interpretation. Normally these entity lists that are filtered and significant at a chosen p-value between the conditions in the interpretation. Thus the entity list is the set of features that are significant for the interpretation. See Figure 27.2.

In the first step, the entity list, the interpretation and the class prediction algorithm are chose. By default, the entity list is the active entity list in the experiment. To change the entity list, click on the *Choose* button and select an entity list from the tree of entity list shown in the experiment. The default interpretation is the active interpretation in the dataset. To build a prediction model on another interpretation in the experiment, click on *Choose* and select another interpretation from the interpretation tree shown in the active experiment. Choose the prediction model from the drop-down list and click *Next*.

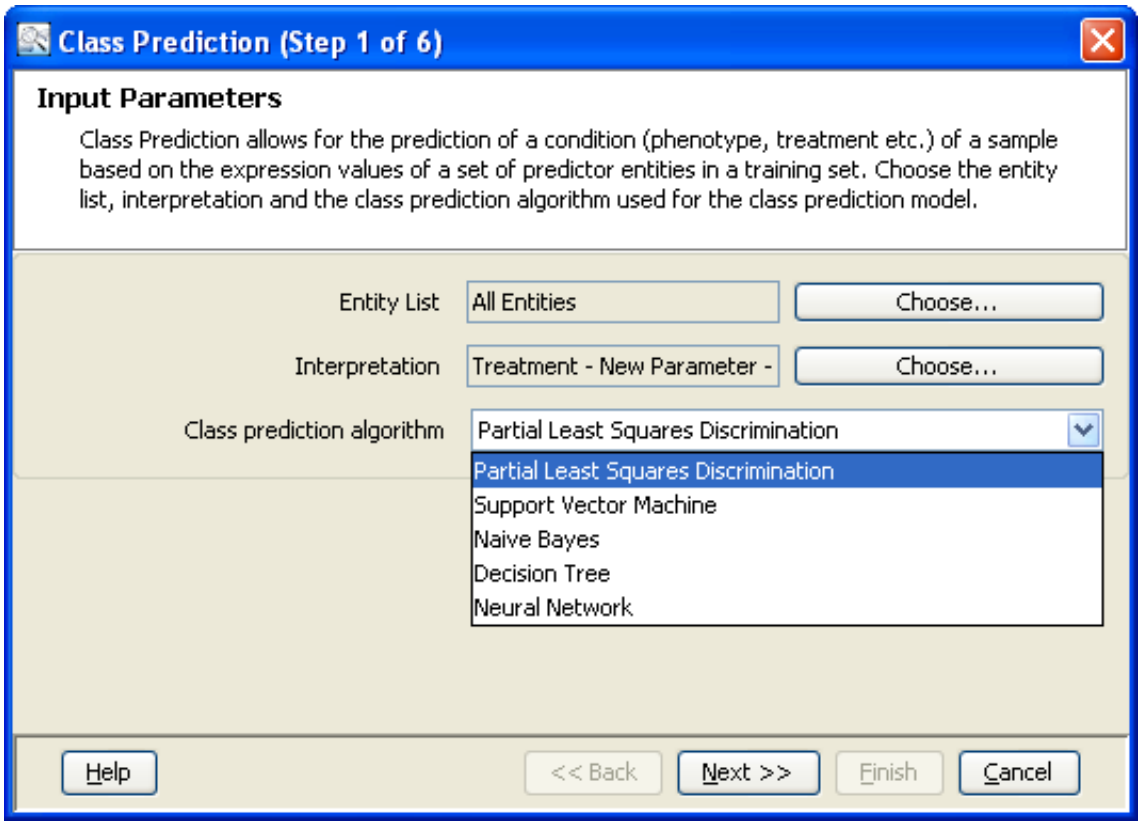


Figure 27.2: Build Prediction Model: Input Parameters

Validation Parameters The second step in building a prediction model is to choose the model parameters and the validation parameters. Here, the model specific parameters will be displayed and the validation type and parameters for validation can be chosen. For details on the model parameters refer the sections on Partial Least Squares Discrimination (PLSD) Support Vector Machine (SVM), Naive Bayesian (NB), Decision Tree (DT), and Neural Network (NN). For details on the validation parameters see the section on Validate. See Figure 27.3.

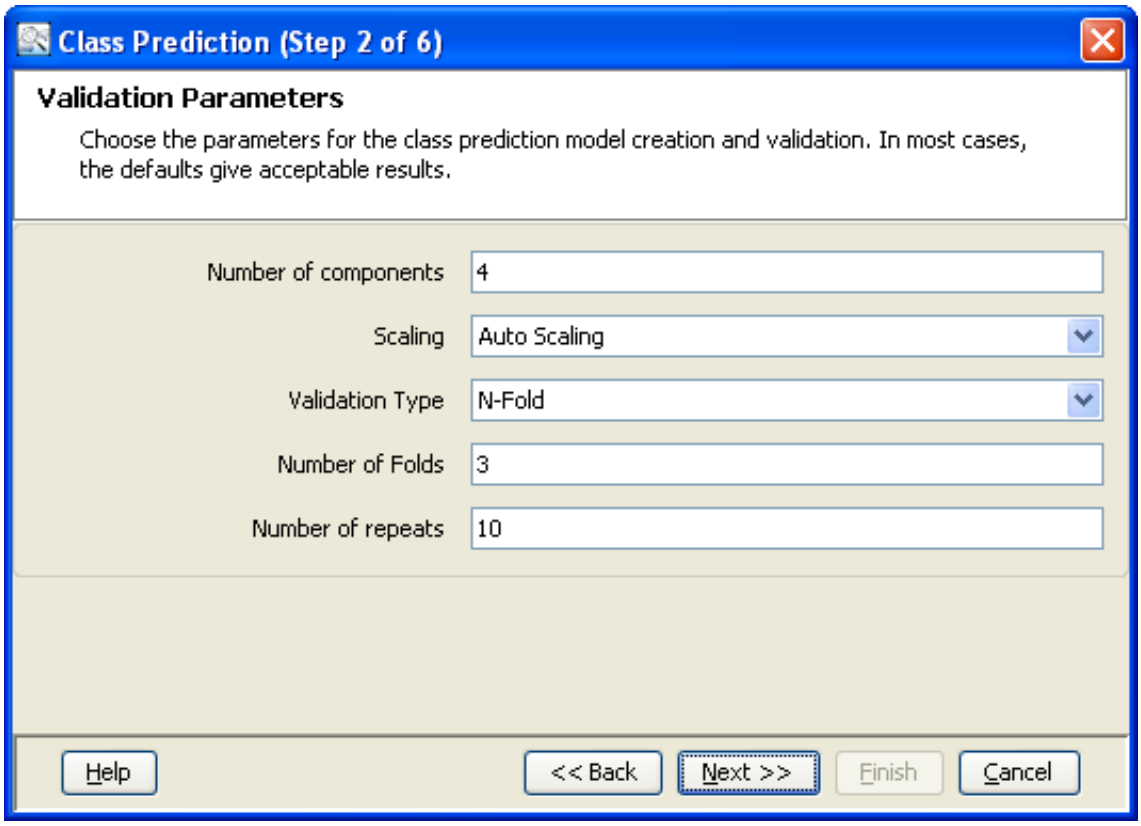


Figure 27.3: Build Prediction Model: Validation Parameters for PLSD

R2Q2 Output This step is only displayed if **Partial Least Squares Discrimination** class prediction algorithm was selected in **Step 1 of 6** of the **Build Prediction Model** workflow. The dialog contains a **R2X Cumulative Values** histogram and **R2Y Cumulative and Q2Y Comparison** histograms. Here, R2 values show the accuracy of the PLSDA model fit and Q2 values show the accuracy of prediction of the model. From these R2-Q2 plots, select the **Number of Components** for class prediction model creation and validation.

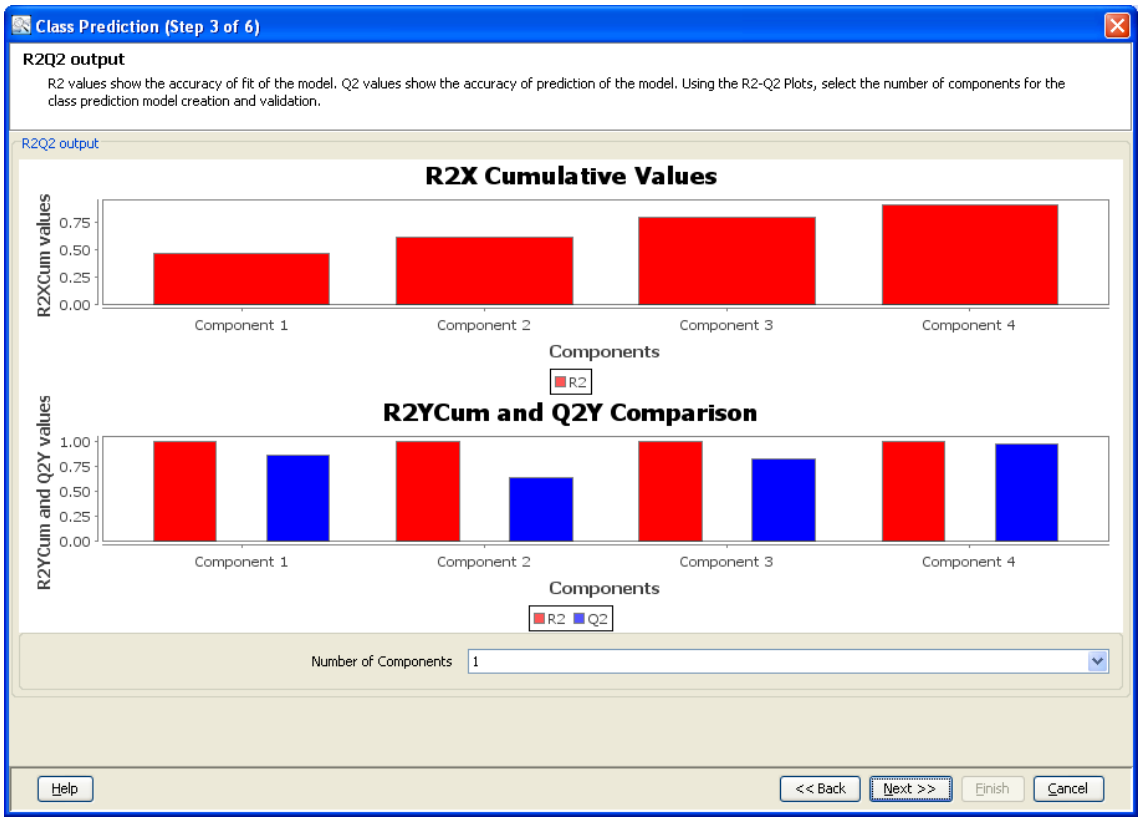


Figure 27.4: Build Prediction Model: R2Q2 output for PLSD

Validation Algorithm Outputs The next step in building prediction algorithms is to examine the validation algorithm outputs. These are a confusion matrix and a prediction report table. The confusion matrix gives the efficacy of the prediction model and the report gives details of the prediction of each condition. For more details, see the section on Viewing Classification Results. If the results are satisfactory, click *Next* or click *Back* to choose a different different model or a different set of parameters. Clicking *Next* will build the prediction model. See Figure 27.5.

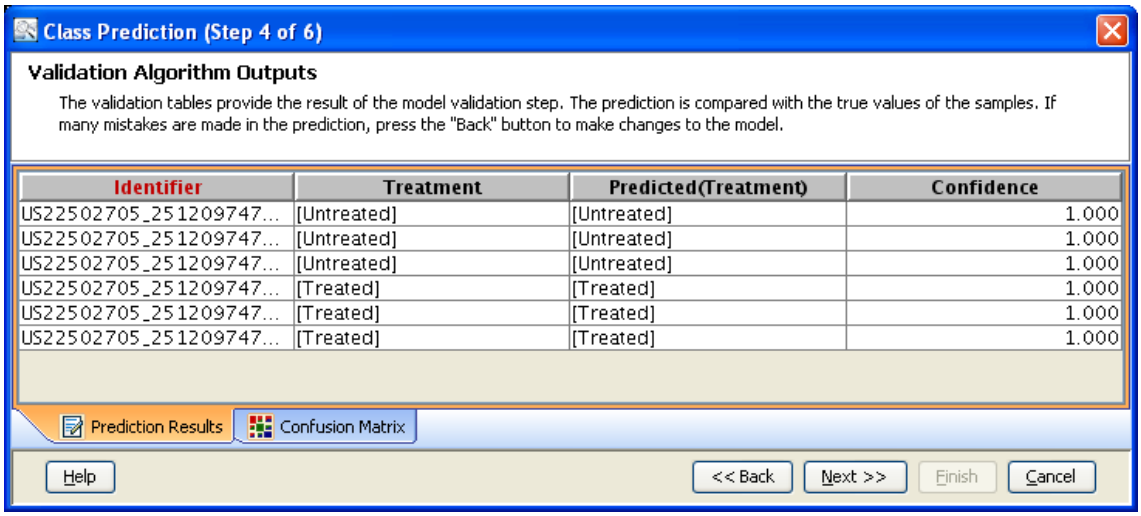


Figure 27.5: Build Prediction Model: Validation Output

Training Algorithm Output The next step provides the output of the training algorithm. It provides a confusion matrix for the training model on the whole entity list, report table, the lorenz curve showing the efficacy of classification and prediction model. Wherever appropriate, a visual output of the classification model is presented. For more details refer to the section on Viewing Classification Results. For details on the model for each algorithm, go to the appropriate section of Partial Least Squares Discrimination, Decision Tree (DT), Neural Network (NN), Support Vector Machine (SVM), and Naive Bayesian (NB). If you want to rerun the model and change the parameters, click *Back*. Click *Next* to save the model. See Figure 27.6.

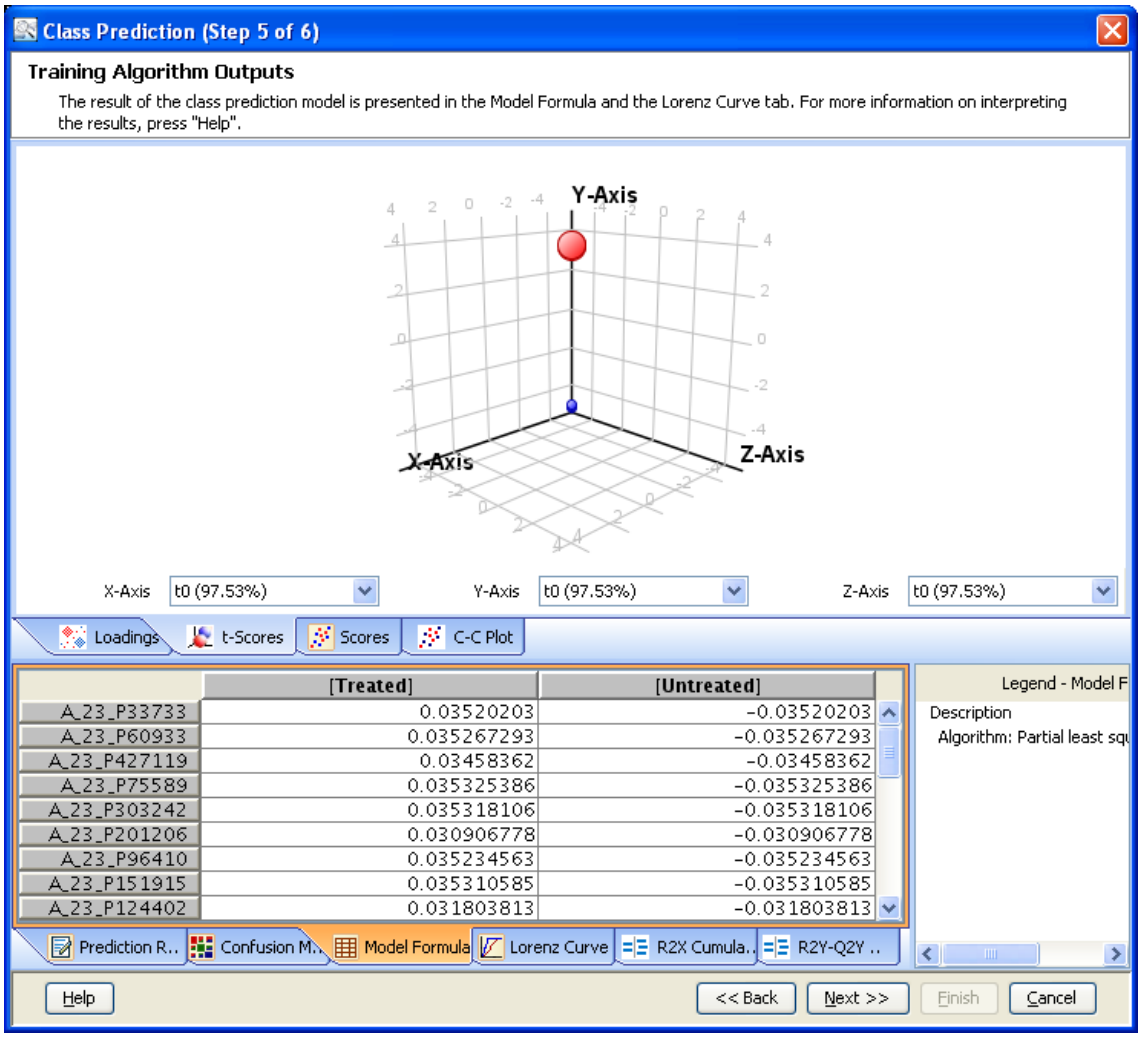


Figure 27.6: Build Prediction Model: Training Output

Class Prediction Model Object The last step of building the prediction model is to save the class prediction model object in the tool. The view shows the model object with a default name and the notes showing the details of the prediction model and the parameters used. The view also shows a set of system generated fields that are stored with the model. You can change the name of the model and add additional notes in the text box provided. All these fields will be stored as annotations of the model can be searched and selected. Clicking *Finish* will save the model in the tool and show it in the *Analysis* tree of the experiment navigator. A right click on the model in the navigator will show options to inspect the model, copy it or remove it. Additionally, the entity list that was actually used in building the model can be created by clicking *Expand as Entity List*. This utility is useful to get that subset of the original entity list that actually goes into the model; this is especially true for decision trees where the final model is most likely to use a subset of the original entities.
The saved model can be used in any other experiment of the same technology in the tool. See Figure 27.7.

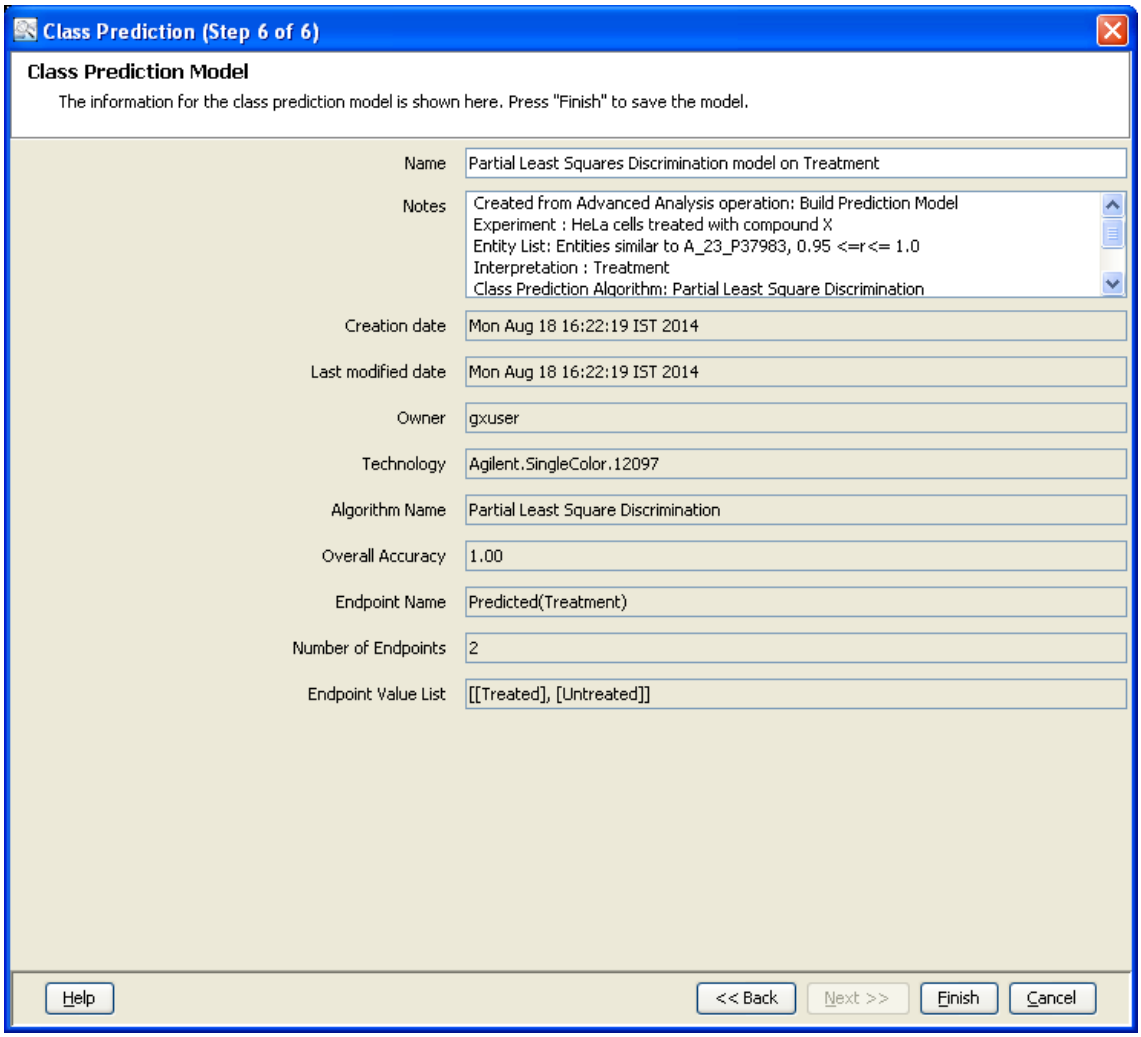
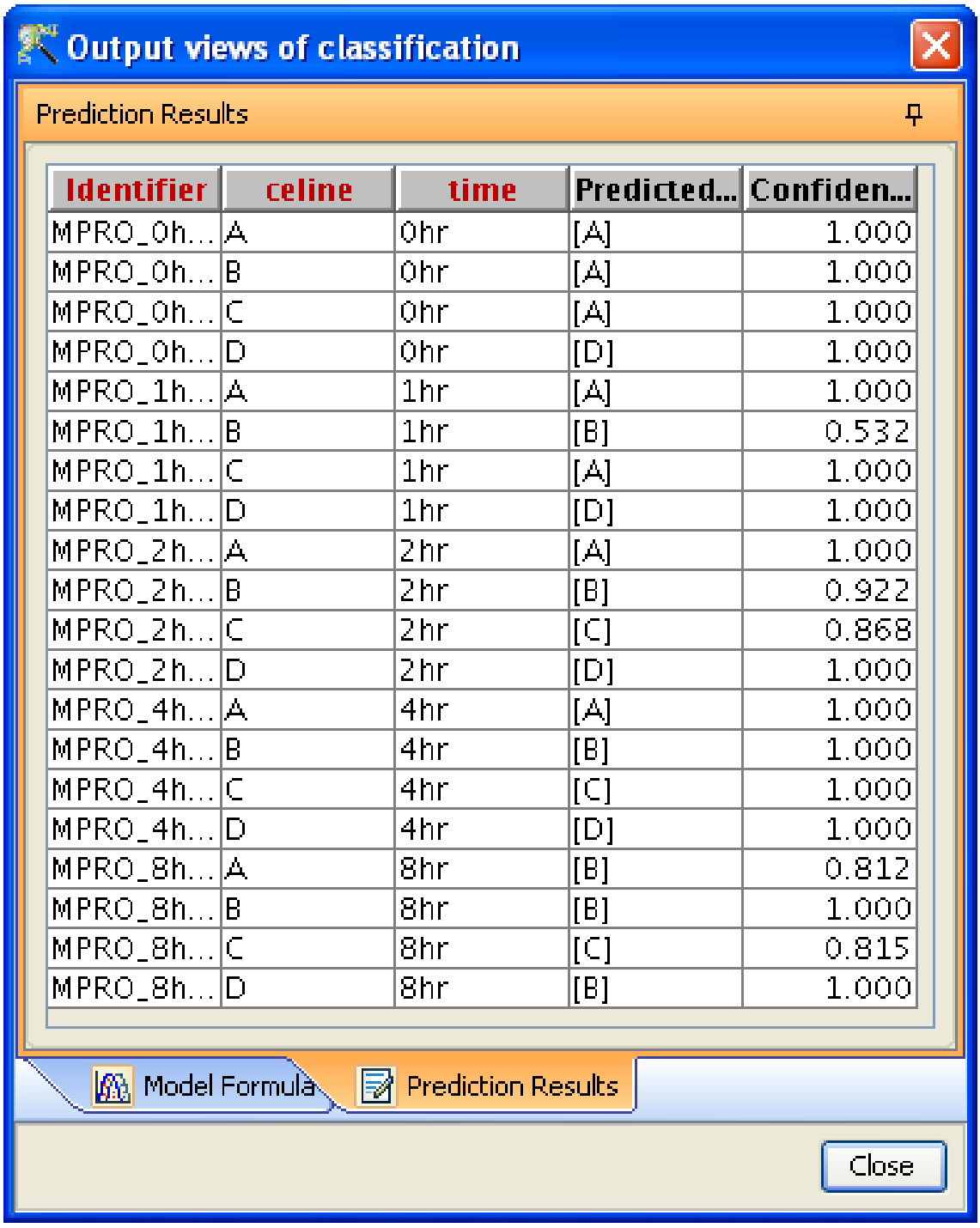


Figure 27.7: Build Prediction Model: Model Object

27.3.2 Run Prediction

The *Run Prediction* workflow link is used to run a prediction model in an experiment. Clicking on this link will show all the models in the tool that have been created on the same technology. select a model and click *OK*. This will run the prediction model on the current experiment and output the results in a table. The model will take the entities in the technology used to model, run the model on all the samples in the experiment and predict the outcome for each sample in the experiment. The predicted results will be shown in the table along with a confidence measure appropriate to the model. For details on the prediction results and the confidence measures of prediction, see the appropriate sections of Partial Least Squares Discrimination, Decision Tree (DT), Neural Network (NN), Support Vector Machine (SVM), and Naive Bayesian (NB). See Figure 27.8.



Identifier	celine	time	Predicted...	Confiden...
MPRO_0h...	A	0hr	[A]	1.000
MPRO_0h...	B	0hr	[A]	1.000
MPRO_0h...	C	0hr	[A]	1.000
MPRO_0h...	D	0hr	[D]	1.000
MPRO_1h...	A	1hr	[A]	1.000
MPRO_1h...	B	1hr	[B]	0.532
MPRO_1h...	C	1hr	[A]	1.000
MPRO_1h...	D	1hr	[D]	1.000
MPRO_2h...	A	2hr	[A]	1.000
MPRO_2h...	B	2hr	[B]	0.922
MPRO_2h...	C	2hr	[C]	0.868
MPRO_2h...	D	2hr	[D]	1.000
MPRO_4h...	A	4hr	[A]	1.000
MPRO_4h...	B	4hr	[B]	1.000
MPRO_4h...	C	4hr	[C]	1.000
MPRO_4h...	D	4hr	[D]	1.000
MPRO_8h...	A	8hr	[B]	0.812
MPRO_8h...	B	8hr	[B]	1.000
MPRO_8h...	C	8hr	[C]	0.815
MPRO_8h...	D	8hr	[B]	1.000

Figure 27.8: Run Prediction: Prediction Output

Note: A prediction model created on a technology can be used only in experiments of the same technology.

27.4 Partial Least Squares Discrimination

PLSD is an extension of the PLSR (Partial Least Squares Regression) - a PLS version of LDA (Linear Discriminant Analysis). PLSD is useful when you need to predict a set of variables, and identify them as functional classes from a large number of independent variables (i.e., predictors).

Notions and Notations

The PLSD model is developed from a training set of N observations (objects, cases, compounds, etc.) with K **X**-variables denoted by $x_k(k = 1, \ldots, K)$, and M **Y**-classes $y_m(m = 1, 2, \ldots, M)$. These training data form the two matrices X and Y of dimensions (NxK) and (NxM), respectively.

Later, predictions for new observations are made based on their **X**-data. This gives predicted t-scores, loadings, and prediction results with confidence intervals.

27.4.1 PLSD Model and Parameters

The goal of PLS regression is to predict **Y** from **X** and to describe their common structure.

PLS regression decomposes both **X** and **Y** as a product of a common set of orthogonal factors and specific loadings. So, the independent variables are decomposed as $X = TP^T$ with $T^TT = I$; where **I** and **P** are the identity and loading matrices. Likewise, **Y** is estimated as $\hat{Y} = TBC^T$; where **B** is a diagonal matrix with ‘‘regression weights" as the diagonal elements and **C** is the "weight matrix" of the dependent variables. The columns of **T** are the latent vectors.

The dependent variables are predicted using the multivariate regression formula $\hat{Y} = TBC^T = XB_{PLS}$; where $B_{PLS} = (P^{T+})BC^T$ and P^{T+} is the Moore-Penrose pseudoinverse of P^T.

Step 1 of 6: Input Parameters Select the entity list and the interpretation along with the algorithm (PLSD).

Step 2 of 6: Validation Parameters Select the Model Parameters from the Validation Parameters dialog, and then click **Next**. Refer to Table 27.1 for details.

Step 3 of 6: R2Q2 Output Using the R2-Q2 Plots, select the number of components to be included for class prediction model creation.

Step 4 of 6: Validation Algorithm Outputs The results of validation with PLSD are displayed in the dialog. They consist of the Confusion Matrix and the Prediction Results. If the validations results are good, proceed for training in step 4 or go back and redo validation with different parameter settings.

Step 5 of 6: Training Algorithm Outputs The results of model building with PLSD are displayed in the view.

Step 6 of 6: Class Prediction You can edit the Name and Notes. The dialog informs about the Creation and Last modified date, Owner, Technology, Algorithm Name, Overall Accuracy,

Parameter	Additional Information
Number of Components	Number of components to decompose to; default value is 4.
Scaling	**Auto Scaling:** Select the Auto Scaling option from the Scaling drop-down list. 1. Subtracts the mean μ_i from each m_{ij}: $m_{ij} = m_{ij}$ - μ_i. 2. Scales down the value by a factor equal to the standard deviation σ_i: $m_{ij} = \frac{(m_{ij}-\mu_i)}{\sigma_i}$. **Pareto:** Select the Pareto Scaling option from the Scaling drop-down list. It scales down the value by a factor equal to the square root of the standard deviation σ_i: $m_{ij} = \frac{m_{ij}}{\sqrt{\sigma_i}}$. **No Scaling:** You can select No Scaling option to skip scaling.
Validation Type	Only N-Fold validation is supported.
Number of Folds	Sets the number of folds; default value of 3 folds is good to go with.
Number of Repeats	Sets the number of repeats; default value of 10 is good to go with.

Table 27.1: Validation Parameters

Endpoint Name, Number of Endpoints, and Endpoint Value List. Clicking *Finish* will add a node called 'Partial Least Squares Discrimination Model' in the experiment navigator and exit the wizard.

27.5 Support Vector Machines

Support Vector Machines (SVM) attempts to separate conditions or samples into classes by imagining these to be points in space and then determining a separating plane which separates the two classes of points.

While there could be several such separating planes, the algorithm finds a good separator which maximizes the separation between the classes of points. The power of SVMs stems from the fact that before this separating plane is determined, the points are transformed using a so called kernel function so that separation by planes post application of the kernel function actually corresponds

to separation by more complicated surfaces on the original set of points. In other words, SVMs effectively separate point sets using non-linear functions and can therefore separate out intertwined sets of points.

The **GeneSpring** implementation of SVMs, uses a unique and fast algorithm for convergence based on the Sequential Minimal Optimization method. It supports three types of kernel transformations - Linear, Polynomial and Gaussian. In all these kernel functions, it so turns out that only the dot product (or inner product) of the rows (or conditions) is important and that the rows (or conditions) themselves do not matter, and therefore the description of the kernel function choices below is in terms of dot products of rows, where the dot product between rows a and b is denoted by $x(a).x(b)$.

The Linear Kernel is represented by the inner product given by the equation $x(a).x(b)$.

The Polynomial Kernel is represented by a function of the inner product given by the equation $(k_1[x(a).x(b)] + k_2)^p$, where p is a positive integer.

The Gaussian Kernel is given by the equation $e^{-(\frac{x(a)-x(b)}{\sigma})^2}$

Polynomial and Gaussian kernels can separate intertwined datasets but at the risk of over-fitting. Linear kernels cannot separate intertwined datasets but are less prone to over-fitting and therefore, more generalizable.

An SVM model consists of a set of support vectors and associated weights called Lagrange Multipliers, along with a description of the kernel function parameters. Support vectors are those points which lie on (actually, very close to) the separating plane itself. Since small perturbations in the separating plane could cause these points to switch sides, the number of support vectors is an indication of the robustness of the model; the more this number, the less robust the model. The separating plane itself is expressible by combining support vectors using weights called Lagrange Multipliers.

For points which are not support vectors, the distance from the separating plane is a measure of the belongingness of the point to its appropriate class. When training is performed to build a model, these belongingness numbers are also output. The higher the belongingness for a point, the more the confidence in its classification.

27.5.1 SVM ModelParameters

The parameters for building a *SVM Model* are detailed below:

Kernel Type Available options in the dropdown menu are - Linear, Polynomial, and Gaussian. The default is Linear.

Max Number of Iterations A multiplier to the number of conditions needs to be specified here. The default multiplier is 100. Increasing the number of iterations might improve convergence, but will take more time for computations. Typically, start with the default number of iterations and work upwards watching any changes in accuracy.

Cost This is the cost or penalty for misclassification. The default is 100. Increasing this parameter has the tendency to reduce the error in classification at the cost of generalization. More precisely, increasing this may lead to a completely different separating plane which has either more support vectors or less physical separation between classes but fewer misclassifications.

Ratio This is the ratio of the cost of misclassification for one class to the cost of the misclassification for the other class. The default ratio is 1.0. If this ratio is set to a value r, then the cost of misclassification for the class corresponding to the first row is set to the cost of misclassification specified, and the cost of misclassification for the other class is set to r times this value. Changing this ratio will penalize misclassification more for one class than the other. This is useful in situations where, for example, false positives can be tolerated while false negatives cannot. Then setting the ratio appropriately will have a tendency to control the number of false negatives at the expense of possibly increased false positives. This is also useful in situations where the classes have very different sizes. In such situations, it may be useful to penalize classifications much more for the smaller class than the bigger class.

Kernel Parameter (1) This is the first kernel parameter k1 for polynomial kernels and can be specified only when the polynomial kernel is chosen. Default if 0.1.

Kernel parameter (2) This is the second kernel parameter k2 for polynomial kernels. Default is set to 1. It is preferable to keep this parameter non-zero.

Exponent This is the exponent of the polynomial for a polynomial kernel (p). The default value is 2. A larger exponent increases the power of the separation plane to separate intertwined datasets at the expense of potential over-fitting.

Sigma This is a parameter for the Gaussian kernel. The default value is set to 1.0. Typically, there is an optimum value of sigma such that going below this value decreases both misclassification and generalization and going above this value increases misclassification. This optimum value of sigma should be close to the average nearest neighbor distance between points.

Validation Type Choose one of the two types from the dropdown menu - Leave One Out, N-Fold. The default is N fold validation.

Number of Folds If N-Fold is chosen, specify the number of folds. The default value is 3.

Number of Repeats The default value is 10.

The results of validation with SVM are displayed in the dialog. The Support Vector Machine view appears under the current spreadsheet and the results of validation are listed under it. They consist of the Confusion Matrix and the Lorenz Curve. The Confusion Matrix displays the parameters used for validation. If the validations results are good then these parameters can be used for training.

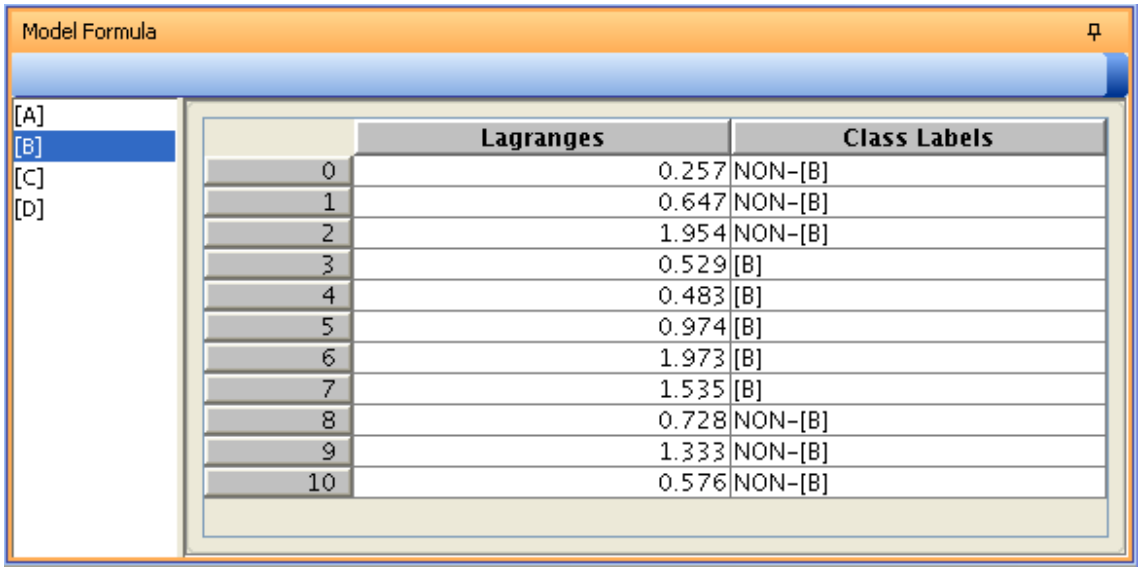
The results of training with SVM are displayed in the dialog. They consist of the SVM model, a Report, a Confusion Matrix, and a Lorenz Curve, all of which will be described later.

27.5.1.1 Support Vector Machine Model

For Support Vector Machine training, the model output contains the following training parameters in addition to the model parameters: See Figure 27.9.

The top panel contains the Offset which is the distance of the separating hyperplane from the origin in addition to the input model parameters.

The lower panel contains the Support Vectors, with three columns corresponding to row identifiers(if marked)/row indices, Lagranges and Class Labels. These are input points, which determine the separating surface between two classes. For support vectors, the value of Lagrange Multipliers is non-zero and for other points it is zero. If there are too many support vectors, the SVM model has over-fit the data and may not be generalizable.



	Lagranges	Class Labels
0	0.257	NON-[B]
1	0.647	NON-[B]
2	1.954	NON-[B]
3	0.529	[B]
4	0.483	[B]
5	0.974	[B]
6	1.973	[B]
7	1.535	[B]
8	0.728	NON-[B]
9	1.333	NON-[B]
10	0.576	NON-[B]

Figure 27.9: Model Parameters for Support Vector Machines

27.6 Naive Bayesian

Bayesian classifiers are parameter based statistical classifiers. They are multi-class classifiers and can handle continuous and categorical variables. They predict the probability that a sample belongs to a certain class. The Naive Bayesian classifier assumes that the effect of an attribute on a given class is independent of the value of other attributes. This assumption is called the *class conditional independence*. The Naive Bayesian model is built based on the probability distribution function of the training data along each feature. The model is then used to classify a data point based on the learnt probability density functions for each class.

Each row in the data is presented as an n dimensional feature vector, $X = (x_1, x_2, \ldots, x_n)$. If there are m classes, $C_1, C_2, \ldots, C_m$. Given an unknown data sample X the classifier predicts that X belongs to the class having the highest posterior probability, conditioned on X. The Naive Bayesian assigns X to class C_i if and only if

$P(C_i|X) > P(C_j|X)$ for $1 <= j <= m, j \neq i$

Applying bayesian rule, and given the assumption of *class conditional independence*, the probability

can be computed as

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$

The Probabilities $P(x_1|C_i, P(x_2|C_i), \ldots, P(x_n|Ci$ is estimated from the training samples and forms the Naive Bayesian Model.

27.6.1 Naive Bayesian Model Parameters

The parameters for building a *Naive Bayesian Model* are detailed below:

Validation Type Choose one of the two types from the dropdown menu - Leave One Out, N-Fold. The default is N fold validation.

Number of Folds If N-Fold is chosen, specify the number of folds. The default value is 3.

Number of Repeats The default value is 10.

The results of validation with Naive Bayesian are displayed in the dialog. They consist of the Confusion Matrix, Validation Report and the Lorenz Curve. The Confusion Matrix displays the parameters used for validation. If the validations results are good these parameters can be used to train and build a model.

The results of the model with are displayed in the dialog. They consist of the NB Model Formula, a Report, a Confusion Matrix, and a Lorenz Curve, all of which will be described later.

27.6.2 Naive Bayesian Model View

For Naive Bayesian training, the model output contains the row identifier(if marked)/row index on the left panel and the Naive Bayesian Model parameters in the right panel. The Model parameters consist of the Class Distribution for each class in the training data and parameters for each feature or column. For continuous features the parameters are the mean and standard deviation for the particular class and for categorical variables these are the proportion of each category in the particular class. See Figure 27.10.

To View Classification Clicking on a row identifier/index highlights the classified class of the sample. It shows the computed posterior probability for the selected sample. The row will be classified into that class which shows the largest posterior probability.

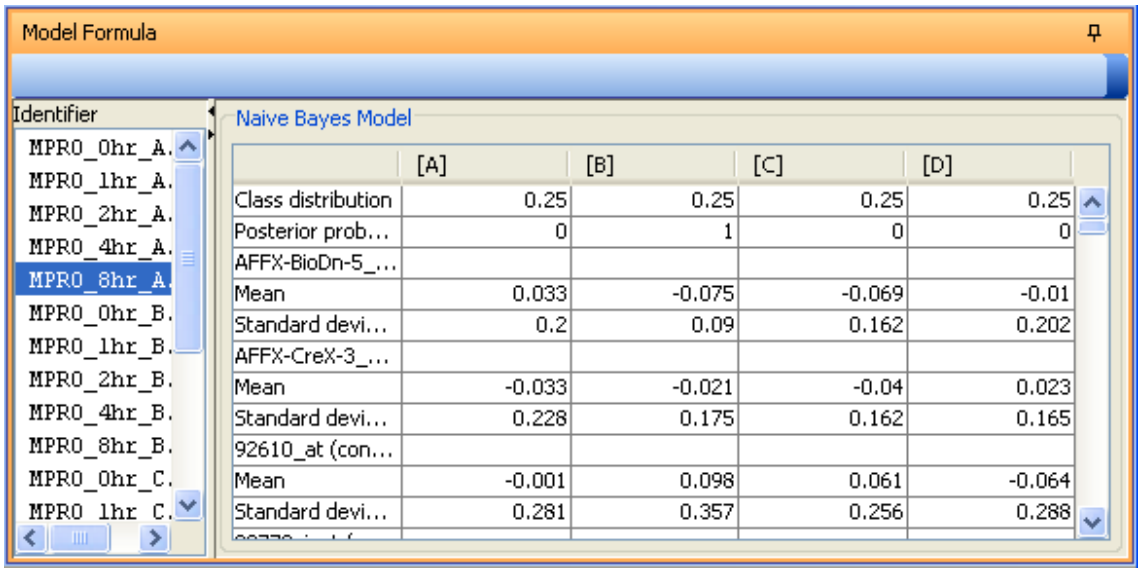


Figure 27.10: Model Parameters for Naive Bayesian Model

27.7 Decision Trees

A Decision Tree is best illustrated by an example. Consider three samples belonging to classes A,B,C, respectively, which need to be classified, and suppose the rows corresponding to these samples have values shown below:

	Feature 1	Feature 2	Feature 3	Class Label
Sample 1	4	6	7	A
Sample 2	0	12	9	B
Sample 3	0	5	7	C

Table 27.2: Decision Tree Table

Then the following sequence of Decisions classifies the samples - if feature 1 is at least 4 then the sample is of type A, and otherwise, if feature 2 is bigger than 10 then the sample is of Type B and if feature 2 is smaller than 10 then the sample is of type C. This sequence of if-then-otherwise decisions can be arranged as a tree. This tree is called a decision tree.

GeneSpring implements Axis Parallel Decision Trees. In an axis parallel tree, decisions at each step are made using one single feature of the many features present, e.g. a decision of the form if feature 2 is less than 10.

The decision points in a decision tree are called internal nodes. A sample gets classified by following the appropriate path down the decision tree. All samples which follow the same path down the

tree are said to be at the same leaf. The tree building process continues until each leaf has purity above a certain specified threshold, i.e., of all samples which are associated with this leaf, at least a certain fraction comes from one class. Once the tree building process is done, a pruning process is used to prune off portions of the tree to reduce chances of over-fitting.

Axis parallel decision trees can handle multiple class problems. Both varieties of decision trees produce intuitively appealing and visualizable classifiers.

27.7.1 Decision Tree Model Parameters

The parameters for building a *Decision Tree Model* are detailed below:

Pruning Method The options available in the drop-down menu are - Minimum Error, Pessimistic Error, and No Pruning. The default is Minimum Error. The No Pruning option will improve accuracy at the cost of potential over-fitting.

Goodness Function Two functions are available from the dropdown menu - Gini Function and Information Gain. This is implemented only for the Axis Parallel decision trees. The default is Gini Function.

Allowable Leaf Impurity Percentage (Global or Local) If this number is chosen to be x with the global option and the total number of rows is y, then tree building stops with each leaf having at most x*y/100 rows of a class different from the majority class for that leaf. And if this number is chosen to be x with the local option, then tree building stops with at most x% of the rows in each leaf having a class different from the majority class for that leaf. The default value is 1% and Global. Decreasing this number will improve accuracy at the cost of over-fitting.

Validation Type Choose one of the two types from the dropdown menu - Leave One Out, N-Fold. The default is N fold.

Number of Folds If N-Fold is chosen , specify the number of folds. The default value is 3.

Number of Repeats The default value is 10.

Attribute Fraction at nodes This parameter refers to the fraction of attributes or entities in the input entity list that will be considered while building the decision tree. The default value, 1.0, implies that ALL i.e, 100% of the entities will be considered; if the value is less than that, say 0.6, only 60% of the entities (randomly chosen) will be used to create the tree. Only values between 0.0 and 1.0 should be input to this parameter.

The results of validation with Decision Trees are displayed in the dialog. They consist of the Confusion Matrix and the Lorenz Curve. The Confusion Matrix displays the parameters used for validation. If the validations results are good these parameters can be used for training.

The results of model building with Decision Tree are displayed in the view. These consists of Decision Tree model, a Report, a Confusion Matrix, and a Lorenz Curve, all of which will be described later.

27.7.2 Decision Tree Model

GeneSpring implements the axis parallel decision trees.

The Decision Tree Model shows the learnt decision tree and the corresponding table. The left panel lists the row identifiers(if marked)/row indices of the dataset. The right panel shows the collapsed view of the tree. Clicking on the Expand/Collapse Tree icon in the toolbar can expand it. The leaf nodes are marked with the Class Label and the intermediate nodes in the Axis Parallel case show the Split Attribute.

To Expand the tree Click on an internal node (marked in brown) to expand the tree below it. The tree can be expanded until all the leaf nodes (marked in green) are visible. The table on the right gives information associated with each node.

The table shows the Split Value for the internal nodes. When a candidate for classification is propagated through the decision tree, its value for the particular split attribute decides its path. For values below the split attribute value, the feature goes to the left node, and for values above the split attribute, it moves to the right node. For the leaf nodes, the table shows the predicted Class Label. It also shows the distribution of features in each class at every node, in the last two columns. See Figure 27.11.

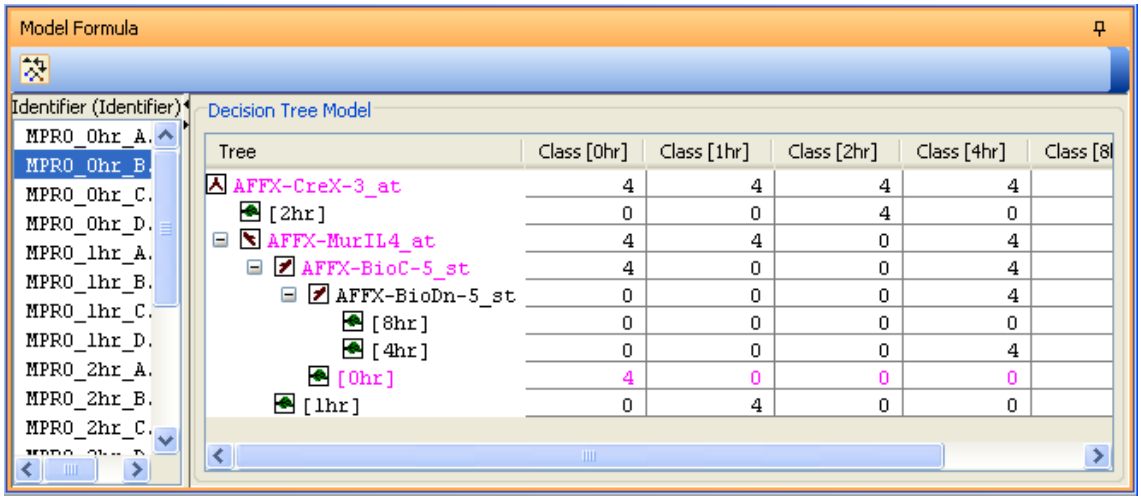


Figure 27.11: Axis Parallel Decision Tree Model

To View Classification Click on an identifier to view the propagation of the feature through the decision tree and its predicted Class Label.

Expand/Collapse Tree: This is a toggle to expand or collapse the decision tree.

27.8 Neural Network

Neural Networks can handle multi-class problems, where there are more than two classes in the data. The Neural Network implementation in **GeneSpring** is the multi-layer perception trained using the back-propagation algorithm. It consists of layers of neurons. The first is called the input layer and features for a row to be classified are fed into this layer. The last is the output layer which has an output node for each class in the dataset. Each neuron in an intermediate layer is interconnected with all the neurons in the adjacent layers.

The strength of the interconnections between adjacent layers is given by a set of weights which are continuously modified during the training stage using an iterative process. The rate of modification is determined by a constant called the learning rate. The certainty of convergence improves as the learning rate becomes smaller. However, the time taken for convergence typically increases when this happens. The momentum rate determines the effect of weight modification due to the previous iteration on the weight modification in the current iteration. It can be used to help avoid local minima to some extent. However, very large momentum rates can also push the neural network away from convergence.

The performance of the neural network also depends to a large extent on the number of hidden layers (the layers in between the input and output layers) and the number of neurons in the hidden layers. Neural networks which use linear functions do not need any hidden layers. Nonlinear functions need at least one hidden layer. There is no clear rule to determine the number of hidden layers or the number of neurons in each hidden layer. Having too many hidden layers may affect the rate of convergence adversely. Too many neurons in the hidden layer may lead to over-fitting, while with too few neurons the network may not learn.

27.8.1 Neural Network Model Parameters

The parameters for building a *Neural Network Model* are detailed below:

Number of Layers Specify the number of hidden layers, from layer 0 to layer 9. A value of '0' would mean 'no hidden layers'. In this case, the Neural Network behaves like a linear classifier. In **GeneSpring**, the default number of layers are 3.

Set Neurons This specifies the number of neurons in each layer. The default value is 15 neurons for each layer. Vary this parameter along with the number of layers.

Choose an optimal number of layers, which yield the best validation accuracy. Normally, up to 3 hidden layers are sufficient.

Number of Iterations The default is 100 iterations. This is normally adequate for convergence.

Learning Rate The default is a learning rate of 0.7. Decreasing this would improve chances of convergence but increase time for convergence.

Momentum The default is a 0.3.

Validation Type Choose one of the two types from the dropdown menu - Leave One Out, N-Fold. The default is N fold validation in **GeneSpring** .

Number of Folds If N-Fold is chosen, specify the number of folds. The default value is 3.

Number of Repeats The default value is 10.

The results of validation with Neural Network are displayed in the dialog. They consist of the Confusion Matrix and the Lorenz Curve. The Confusion Matrix displays the parameters used for validation. If the validations results are good these parameters can be used for training.

The results of training with Neural Network are displayed in the view. They consist of the Neural Network model, a Report, a Confusion Matrix, and a Lorenz Curve, all of which will be described later.

27.8.2 Neural Network Model

The Neural Network Model displays a graphical representation of the learnt model. There are two parts to the view. The left panel contains the row identifier(if marked)/row index list. The panel on the right contains a representation of the model neural network. The first layer, displayed on the left, is the input layer. It has one neuron for each feature in the dataset represented by a square. The last layer, displayed on the right, is the output layer. It has one neuron for each class in the dataset represented by a circle. The hidden layers are between the input and output layers, and the number of neurons in each hidden layer is user specified. Each layer is connected to every neuron in the previous layer by arcs. The values on the arcs are the weights for that particular linkage. Each neuron (other than those in the input layer) has a bias, represented by a vertical line into it. See Figure 27.12.

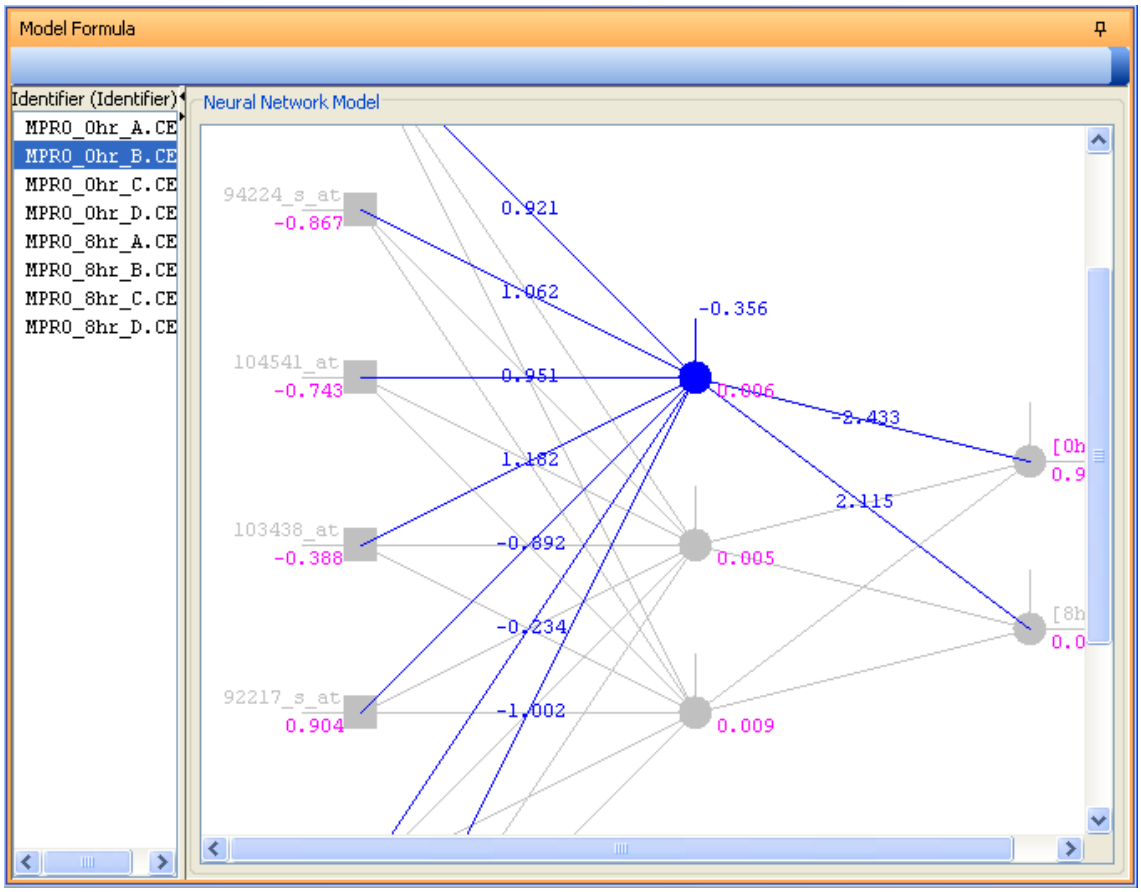


Figure 27.12: Neural Network Model

To View Linkages Click on a particular neuron to highlight all its linkages in blue. The weight of each linkage is displayed on the respective linkage line. Click outside the diagram to remove highlights.

To View Classification Click on an id to view the propagation of the feature through the network and its predicted Class Label. The values adjacent to each neuron represent its activation value subjected to that particular input.

27.9 Viewing Classification Results

The results of classification consist of the following views - The Classification Report, and if Class Labels are present in this dataset, the Confusion Matrix and the Lorenz Curve as well. These views provide an intuitive feel for the results of classification, help to understand the strengths and weaknesses of models, and can be used to tune the model for a particular problem. For example, a classification model may be required to work very accurately for one class, while allowing a greater degree of error on another class. The graphical views help tweak the model parameters to achieve this.

27.9.1 Confusion Matrix

A Confusion Matrix presents results of classification algorithms, along with the input parameters. It is common to all classification algorithms in **GeneSpring** - classification.SVM, Neural Network, Naive Bayesian Classifier, and Decision Tree, appears as follows:

The Confusion Matrix is a table with the true class in rows and the predicted class in columns. The diagonal elements represent correctly classified experiments, and cross diagonal elements represent misclassified experiments. The table also shows the learning accuracy of the model as the percentage of correctly classified experiments in a given class divided by the total number of experiments in that class. The average accuracy of the model is also given. See Figure 27.13.

Confusion Matrix

	[0hr] (Predi...	[1hr] (Predi...	[2hr] (Predi...	[4hr] (Pred...	[8hr] (Predi...	Accuracy
(True) [0hr]	3	1	0	0	0	75.000
(True) [1hr]	1	3	0	0	0	75.000
(True) [2hr]	0	1	3	0	0	75.000
(True) [4hr]	1	0	0	2	1	50.000
(True) [8hr]	0	0	0	1	3	75.000
Overall Accuracy						70.000

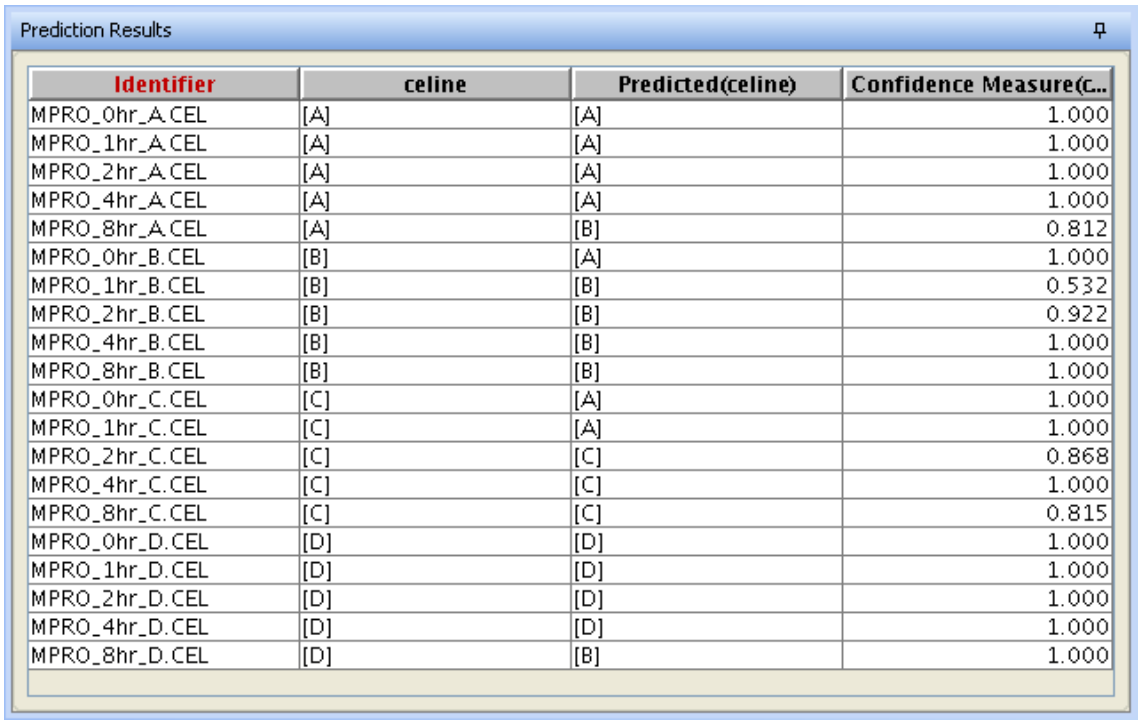
Figure 27.13: Confusion Matrix for Training with Decision Tree

- For validation, the output shows a cumulative Confusion Matrix, which is the sum of confusion matrices for individual runs of the learning algorithm.
- For training, the output shows a Confusion Matrix of the experiments using the model that has been learnt.
- For classification, a Confusion Matrix is produced after classification with the learnt model only if class labels are present in the input data.

27.9.2 Classification Report

This report presents the results of classification. It is common to the three classification algorithms - Support Vector Machine, Neural Network, and Decision Tree.

The report table gives the identifiers; the true Class Labels (if they exist), the predicted Class Labels and class belongingness measure. The class belongingness measure represents the strength of the prediction of belonging to the particular class. See Figure 27.14.



Prediction Results

Identifier	celine	Predicted(celine)	Confidence Measure(C...
MPRO_0hr_A.CEL	[A]	[A]	1.000
MPRO_1hr_A.CEL	[A]	[A]	1.000
MPRO_2hr_A.CEL	[A]	[A]	1.000
MPRO_4hr_A.CEL	[A]	[A]	1.000
MPRO_8hr_A.CEL	[A]	[B]	0.812
MPRO_0hr_B.CEL	[B]	[A]	1.000
MPRO_1hr_B.CEL	[B]	[B]	0.532
MPRO_2hr_B.CEL	[B]	[B]	0.922
MPRO_4hr_B.CEL	[B]	[B]	1.000
MPRO_8hr_B.CEL	[B]	[B]	1.000
MPRO_0hr_C.CEL	[C]	[A]	1.000
MPRO_1hr_C.CEL	[C]	[A]	1.000
MPRO_2hr_C.CEL	[C]	[C]	0.868
MPRO_4hr_C.CEL	[C]	[C]	1.000
MPRO_8hr_C.CEL	[C]	[C]	0.815
MPRO_0hr_D.CEL	[D]	[D]	1.000
MPRO_1hr_D.CEL	[D]	[D]	1.000
MPRO_2hr_D.CEL	[D]	[D]	1.000
MPRO_4hr_D.CEL	[D]	[D]	1.000
MPRO_8hr_D.CEL	[D]	[B]	1.000

Figure 27.14: Decision Tree Classification Report

27.9.3 Lorenz Curve

Predictive classification in **GeneSpring** is accompanied by a class belongingness measure, which ranges from 0 to 1. The Lorenz Curve is used to visualize the ordering of this measure for a particular class.

The items are ordered with the predicted class being sorted from 1 to 0 and the other classes being sorted from 0 to 1 for each class. The Lorenz Curve plots the fraction of items of a particular class encountered (Y-axis) against the total item count (X-axis).

For a given class, the following intercepts on the X-axis have particular significance:

The light red traces the number of items predicted to belong to the selected class.

Classification Quality The point where the red curve reaches its maximum value (Y=1) indicates the number of items which would be predicted to be in a particular selected class if all the items actually belonging to this class need to be classified correctly.

Consider a dataset with two classes A and B. All points are sorted in decreasing order of their belongingness to A. The fraction of items classified as A is plotted against the number of items, as all points in the sort are traversed. The deviation of the curve from the ideal indicates the quality

of classification. An ideal classifier would get all points in A first (linear slope to 1) followed by all items in B (flat thereafter). The Lorenz Curve thus provides further insight into the classification results produced by **GeneSpring**. The main advantage of this curve is that in situations where the overall classification accuracy is not very high, one may still be able to correctly classify a certain fraction of the items in a class with very few false positives; the Lorenz Curve allows visual identification of this fraction (essentially the point where the red line starts departing substantially from the steady slope line to Y=1). See Figure 27.15.

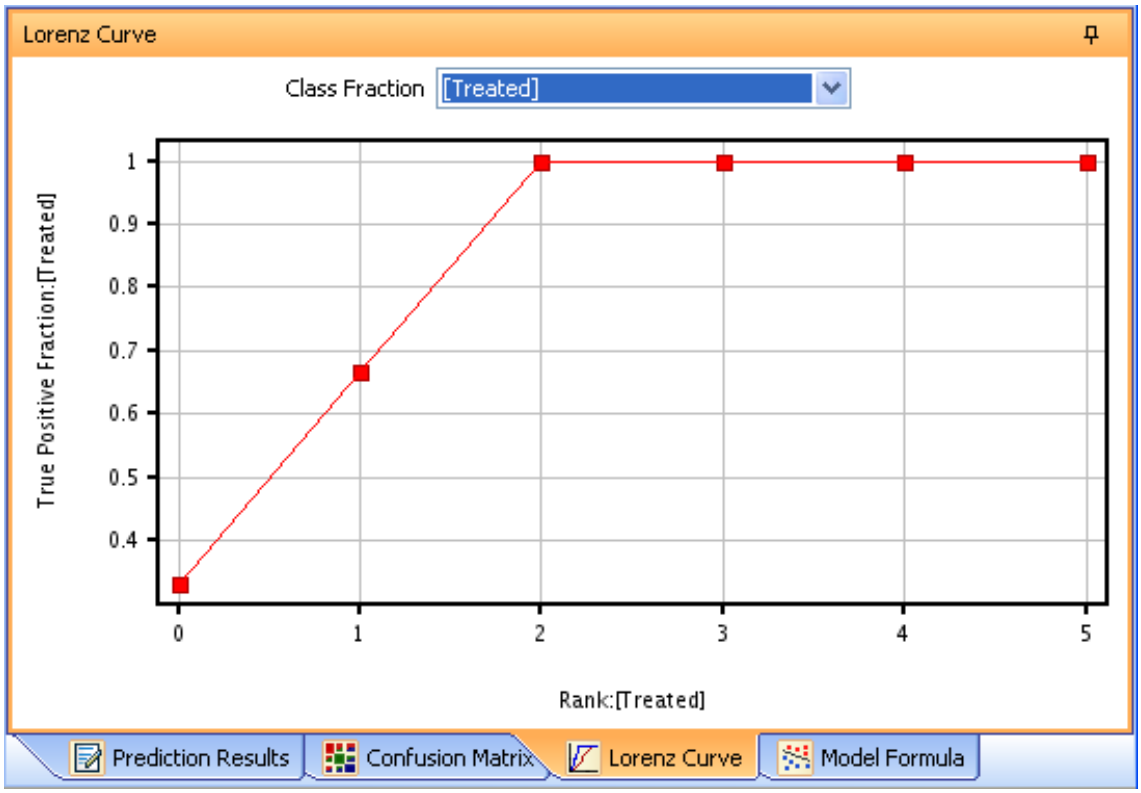


Figure 27.15: Lorenz Curve for Neural Network Training

27.9.3.1 Lorenz Curve Operations

The Lorenz Curve view is a lassoed view and is synchronized with all other lassoed views open in the desktop. It supports all selection and zoom operations like the scatter plot.

Chapter 28

Gene Ontology Analysis

28.1 Working with Gene Ontology Terms

The Gene Ontology™(GO) Consortium maintains a database of controlled vocabularies for the description of molecular functions, biological processes and cellular components of gene products. These GO terms are represented as a Directed Acyclic Graph (DAG) structure. Detailed documentation for the GO is available at the Gene Ontology homepage (http://geneontology.org). A gene product can have one or more molecular functions, be used in one or more biological processes, and may be associated with one or more cellular components. The DAG structure ensures that a gene with a particular GO term also has several other ancestor GO terms implicitly.

In **GeneSpring**, the technology associated with an experiment provides GO terms associated with the entities in the experiment. For Affymetrix, Agilent and Illumina technologies, GO terms are packaged with **GeneSpring**. For custom technologies, GO terms must be imported and marked while creating custom technology for using the GO analysis. For further details, refer to Step 9 of Technology Creation in the generic chapters.

GeneSpring is packaged with the GO terms and their DAG relationships as provided by the GO Ontology Consortium on their website (http://geneontology.org). These ontology files will be updated periodically and these updates will be available via *Annotations* ⟶ *Update Technology Annotations* ⟶ *From Agilent Server*. Locate and click on *GOData* when the *Automatic Software Update* window appears. Click *Update* button. It is necessary to have an active internet connection to avail this feature.

Users can also update the ontology files directly from GO consortium website using a script. In order to execute the script, do the following:

- Download the OBO file from GO consortium, at (http://geneontology.org/GO.downloads.ontology.shtml).
- Open the Script Editor in **GeneSpring** from *Tools* ⟶ *Script Editor*

- Copy the following script in the Script Editor and select Run icon.
  script.marray.genespring.updateGODataFromFile.run()

- Select the downloaded OBO file as an input and select **Update** to update GO data.

Custom GO annotation file (mapping of probe-ID to GO terms) from any source can also be imported in **GeneSpring** . This can be done while creating a generic single or two color technology, or while updating an existing technology using ***Annotations*** $\longrightarrow$ ***Update Technology Annotations***. For carrying out GO analysis, the custom annotation file can either contain a single column with all the GO IDs in it, separated by a separator or it can contain separate columns for the different GO processes. Some of the GO formats supported by **GeneSpring** is given below (not inclusive of all):

- GO:0000012
- go:012
- 12
- GO:0000012(single strand break repair);GO:0000910 (cytokinesis);GO:0006260 (DNA replication); GO:000626 (DNA ligation);GO:0006281(DNA repair);GO:0006310 (DNA recombination);GO:0008150 (biological_process)

In case of multiple columns, while each column can be in a different format, multiple formats within a column is not supported.

The single column with multiple GO IDs should be marked as *Gene Ontology accession* from the drop-down menu. Instead if columns containing individual GO processes(Biological Process, Cellular Component and Molecular Function) are present, they should be marked accordingly in the dropdown menu.

28.2 Introduction to GO Analysis in GeneSpring

GeneSpring has a fully-featured gene ontology analysis module that allows exploring gene ontology terms associated with the entities of interest. **GeneSpring** allows the user to visualize and query the *GO Tree* dynamically, to view GO terms at any level as a *Pie Chart*, to dynamically drill into the pie, to navigate through different levels of the GO tree, to compute enrichment scores for GO terms based upon a set of selected entities, and to use enrichment scores and FDR corrected p-values to filter the selected set of entities. The results of GO analysis can then provide insights into the biology of the system being studied.

In the normal flow of gene expression analysis, GO analysis is performed after identifying a set of entities that are of interest, either from statistical tests or from already identified gene lists. You can select a set of entities in the dataset and launch GO analysis from the *results Interpretation* section on the workflow panel.

Note: To perform GO Analysis, GO terms associated with the entities should be available. These are derived from the technology of the experiment. For Affymetrix, Agilent and Illumina technologies, **GeneSpring** packages the GO Terms associated with the entities. For custom technologies, GO terms must be imported and marked while creating custom technology for using the GO analysis.

This chapter details GO Analysis, the algorithms to compute enrichment scores, the different views launched by the GO analysis and methods to explore the results of GO analysis.

28.3 GO Analysis

GO Analysis can be accessed from most of the workflows in **GeneSpring** . Clicking on the *GO Analysis* link in the *Results Interpretations* section on the workflow panel will launch a wizard that will guide you through collecting the inputs for the analysis and creating an entity list with the significant entities.

Input Parameters The input parameter for GO analysis is any entity list in the current active experiment. By default, the active entity list in the current experiment is shown as the chosen entity list. Clicking on *Choose* will show a tree of entity lists in the current experiment. You can choose any of the entity lists and launch GO Analysis. See Figure 28.1.

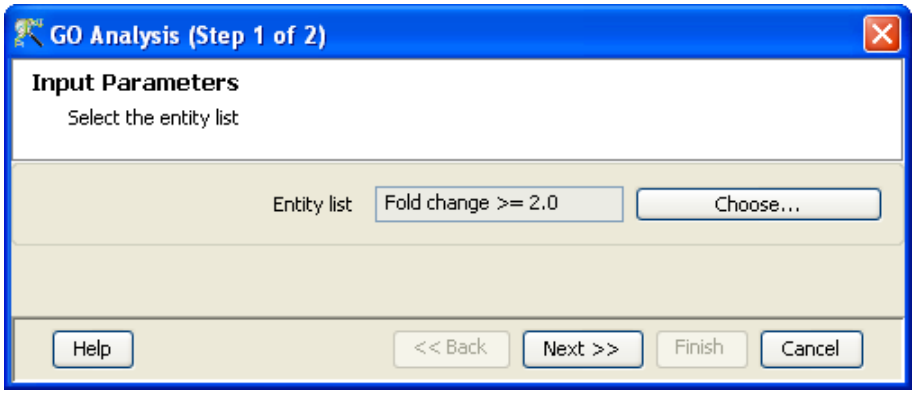


Figure 28.1: Input Parameters

Output Views The results of GO Analysis are shown in the view. Depending upon the input entity list, GO terms that are enriched with a p-value cut-off of 0.1 are shown. If no entities satisfy the cut-off, click on the *Change cutoff* button and change the cut-off from the slider or in the text box. This will dynamically update the views accordingly

The output views show a pie chart, a spreadsheet with the GO terms that satisfy the p-value cut-off and a GO Tree. These GO terms have been derived from only those GO IDs that were part of the input entity list and obeying the cut-off p-value. The spreadsheet has GO accession numbers, their description, uncorrected p-values obtained from the hypergeometric test (described later), Benjamini-Yekutieli corrected p-values to take care of multiple GO term testing and correlation between them (the p-value cutoff is based on the corrected

values), along with frequencies of occurrence of GO terms in the input entity list and in the All Entities list.

All the views are interactive and are dynamically linked. Thus clicking on the pie chart will select the GO Term in the GO tree and will show the corresponding entities associated with the GO terms. Clicking on a GO term on the spreadsheet will highlight the corresponding term in the GO Tree and show the corresponding entities. For details on the views and navigation see the section on GO Analysis Views. See Figure 28.2.

Click *Finish* to save the entity lists corresponding to relevant GO terms. This will create the appropriate entity lists in a folder called GO Analysis. You can also manually select a set of GO terms and save entity lists corresponding these GO terms using the Save Custom button.

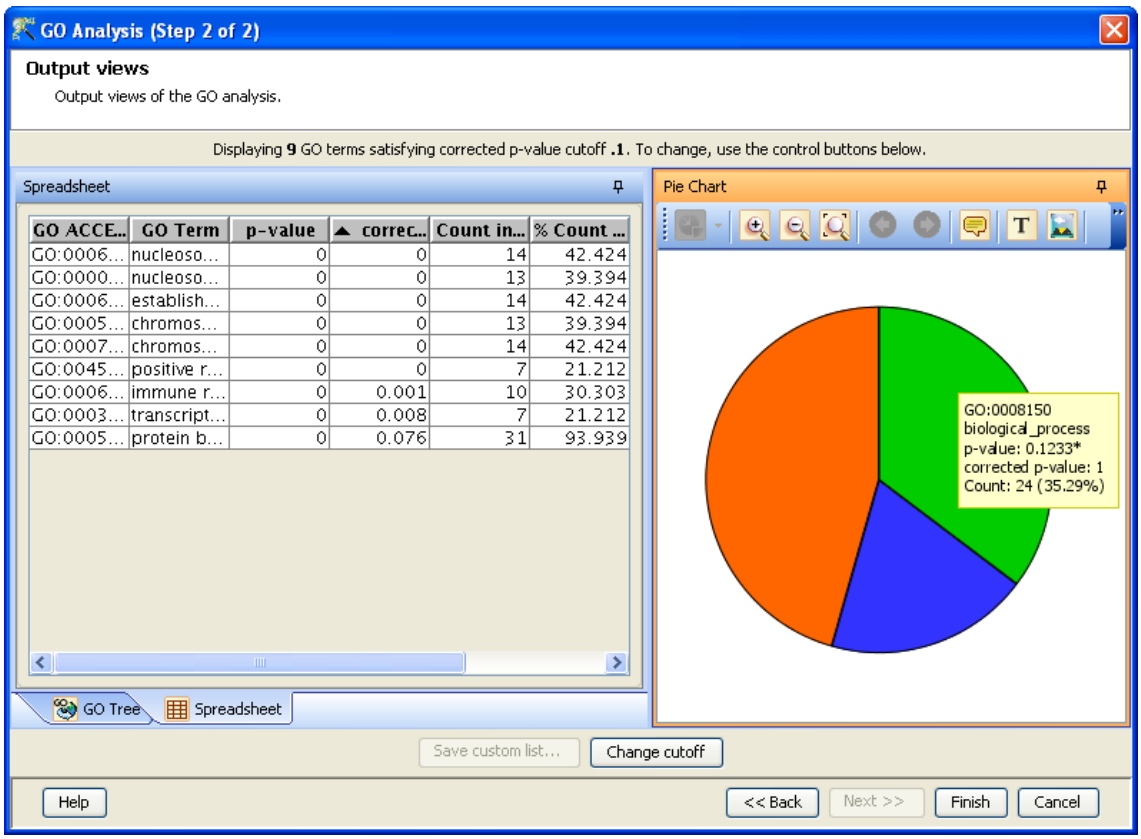


Figure 28.2: Output Views of GO Analysis

The p-value for each GO term reflects the enrichment in frequency of that GO term in the input entity list relative to the All Entities list. The p-value for a GO term g is determined by the following quantities:

- Number of entities in the input entity list which have the term g or any descendant term.
- The number of entities in the All Entities list which have the term g or any descendant term.
- The total number of entities in the input entity list, and
- The total number of entities in the All Entities list.

Note that **GeneSpring** takes GO terms from Biological Processes, Molecular functions and Cellular components together. For details on the computation of the enrichment score or p-value see below.

Note the following rules govern which GO terms are displayed in the various output views.

- All GO terms associated with the input entity list are taken.
- For each term, all ancestors are considered and the cut off is now applied on both the GO term as well as its ancestors.
- The GO terms which pass the cut off are now shown in the output table view.
- The `Count in Selection` shown in the table for a GO term g is obtained as follows: take all descendant GO terms of g and add up genes in the input list associated with these terms. To this, genes in the input list associated with g itself are added. This is the count in selection.
- When saving entity lists for GO terms, what is saved are exactly those entities which contribute to the count above.
- The `Count in Total` is obtained as follows: just add up all genes in *All Entities* associated with this term or any descendant terms.

Note: Previously, only those GO terms which were associated directly with the entity list and also satisfied the p-value cut-off were selected. Then, all the parents which satisfied the p-value cut-off were also pooled in. Now, we initially don't consider the p-value cut-off. We take all the GO-Terms directly associated with the entity list and then trace all the parents. Once we have this whole set, we then apply the p-value cut-off.
This was done as the previous system caused parent entities to show 'inconsistent' behavior. For example, if the p-value cut-off was changed from 0.05 to 0.1, certain new entities with p-value < 0.5 would come in (which should have been coming earlier with cut-of 0.5).

28.4 GO Analysis Views

28.4.1 GO Spreadsheet

The GO Spreadsheet shows the following for each Go Term displayed.

p-value: The probability of obtaining the specified GO accession number from a list of random entities. Less the p-value more significant is the GO accession number.

corrected p-value: Since multiple GO accession number are tested for their significance, hence a multiple testing correction is performed. The GO spreadsheet is sorted based on the corrected p-values.

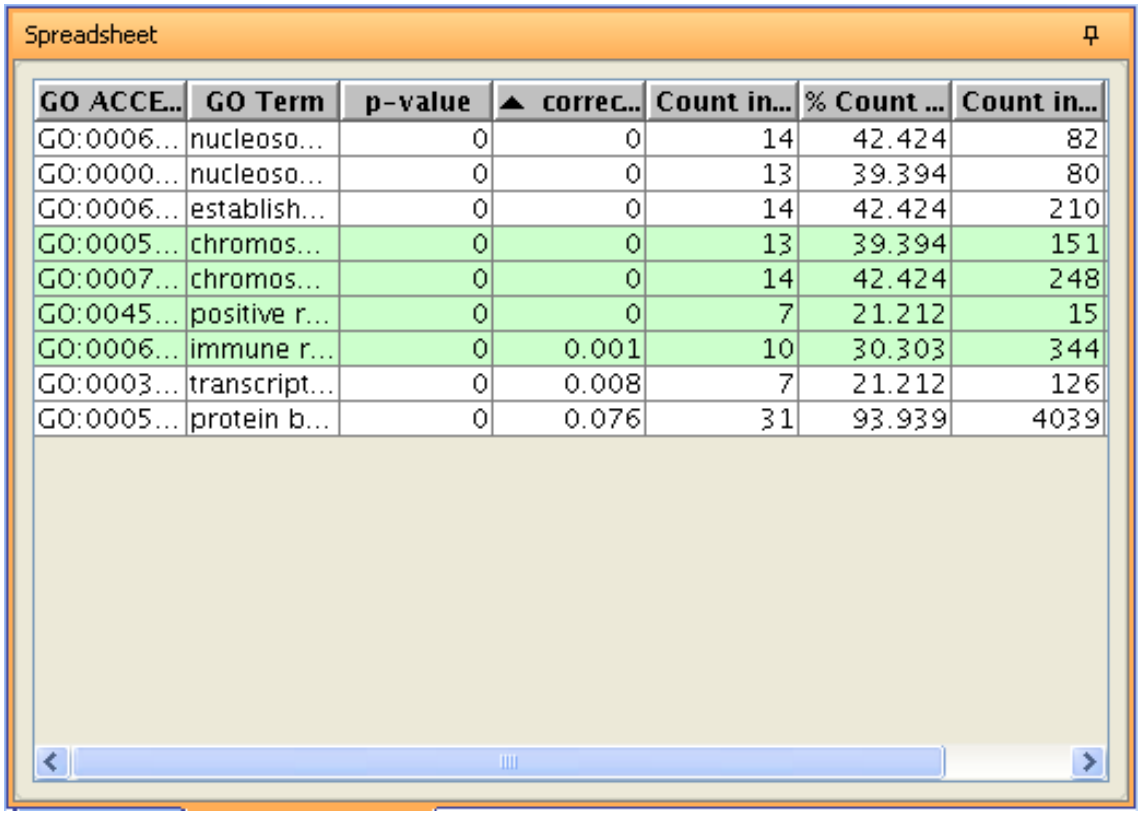
Count in Selection: This refers to the number of genes in the selected entity (for example, from T-test) list which have that particular GO term or its descendants.

% **Count in Selection:** This refers to the percentage of genes in the input entity list which have that GO term.

Count in Total: This refers to the number of genes in *All Entities* which have that GO term or its descendants.

% **Count in Total:** This refers to the percentage of genes in the All Entities list which have that GO term.

The Selection of GO terms in this table will select the corresponding GO terms in the *GO Tree* view and will show the entities associated with the GO term. See Figure 28.3.



GO ACCE...	GO Term	p-value	▲ correc...	Count in...	% Count ...	Count in...
GO:0006...	nucleoso...	0	0	14	42.424	82
GO:0000...	nucleoso...	0	0	13	39.394	80
GO:0006...	establish...	0	0	14	42.424	210
GO:0005...	chromos...	0	0	13	39.394	151
GO:0007...	chromos...	0	0	14	42.424	248
GO:0045...	positive r...	0	0	7	21.212	15
GO:0006...	immune r...	0	0.001	10	30.303	344
GO:0003...	transcript...	0	0.008	7	21.212	126
GO:0005...	protein b...	0	0.076	31	93.939	4039

Figure 28.3: Spreadsheet view of GO Terms

28.4.2 The GO Tree View

The *GO Tree* view is a tree representation of the GO Directed Acyclic Graph (DAG) as a tree view with all GO Terms and their children. Thus there could be GO terms that occur along multiple paths of the GO tree. The GO tree is represented on the left panel of the view. The panel to the right of the GO tree shows the list of entities in the experiment that corresponds to the selected GO term(s). The selection operation is detailed below. See Figure 28.4.

The GO tree is always launched expanded up to three levels. The GO tree shows the GO terms

along with their enrichment p-value in brackets. The GO tree shows only those GO terms along with their full path that satisfy the specified p-value cut-off. GO terms that satisfy the specified p-value cut-off are shown in blue, while others that are on the path and do not satisfy the cut-off are shown in black.

Note that the final leaf node along any path will always have GO term with a p-value that is below the specified cut-off and shown in blue. Also note that along an extended path of the tree there could be multiple GO terms that satisfy the p-value cut-off.

The *GO Tree* provides a link between the GO terms and the entities in the experiment. Operations on the GO Tree are detailed below:

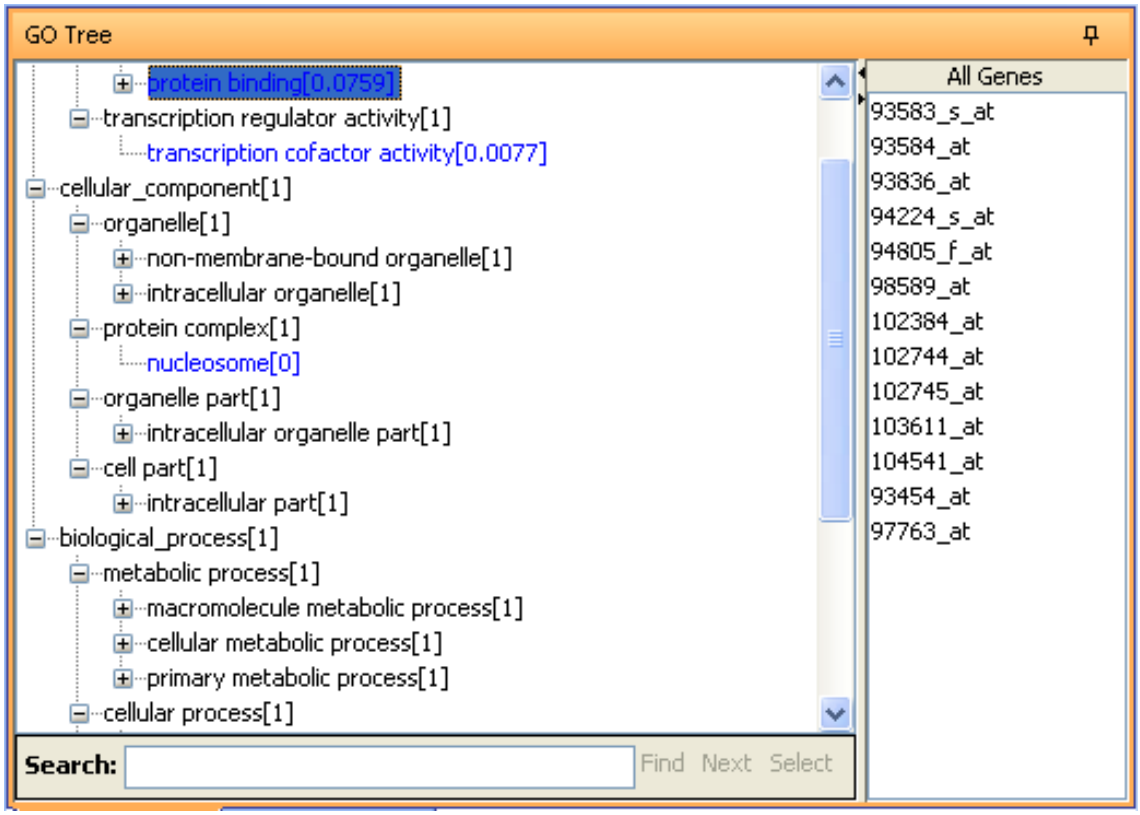


Figure 28.4: The GO Tree View

Expand and Collapse the GO tree: The GO tree can be expanded or collapsed by clicking on the root nodes.

GO Tree Labels: The GO tree is labelled with GO terms as default. You can change the GO tree to be labelled by either the GO Accession; the GO terms; or both from the right-click properties dialog.

p-value and Count: The number in the bracket corresponding to a GO term shows the p-value or enrichment value of the GO term. You can display the p-value, the actual counts of both the p-value and the actual counts for the GO term from the right-click properties dialog.

The counts show two values. The first value shows the number of entities in the entity list contributing to any significant GO term in the hierarchy. The second count value shows the number of entities that contribute any significant GO term in the hierarchy in the experiment.

Select Genes: Clicking on a GO term in the tree will select the entities in the entity list that contributed to any significant GO term in the hierarchy.

You can choose multiple GO terms in the tree and and see *All Genes* that contributed to any significant GO term in the hierarchies. This will show a union of all the entities corresponding to the selected GO terms. Or you can choose multiple GO terms in the tree and select the *Common Genes* that contributed to any significant GO term in the hierarchies. This will show an intersection of the entities corresponding to the selected GO terms. See Figure 28.5.

Selecting *Show All Genes* or *Show Common Genes* can be chosen from the right-click *Properties* menu of the GO tree.

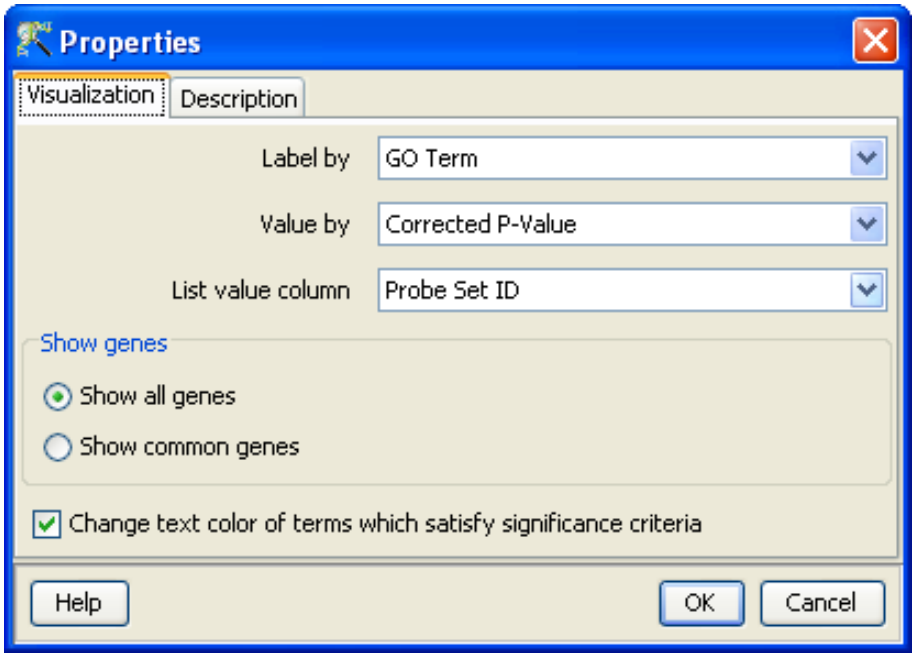


Figure 28.5: Properties of GO Tree View

28.4.3 The Pie Chart

The pie chart view shows a pie of the GO terms with the number of entities that contribute to the any significant GO term in the hierarchy. When the pie chart is launched, it is launched with the top level GO terms of *Molecular Function*, *Biological Process* and *Cellular Component*. The slices of the pie is drawn with the number of entities in each of the three terms that contribute to any significant GO terms in whole hierarchy of GO terms. See Figure 28.6.

The pie chart view is rich with functionality. It allows you to drill into the pie and reach any level of the GO tree, and navigate through the different drill levels. You can select the entities corresponding to the pies or the GO terms in any view. The pie chart allows you to zoom in and out of view, fit the pie chart to view, enable and delete callouts for the slices, add text and images

to the view and create publication quality outputs. The functionality of the pie chart is detailed below:

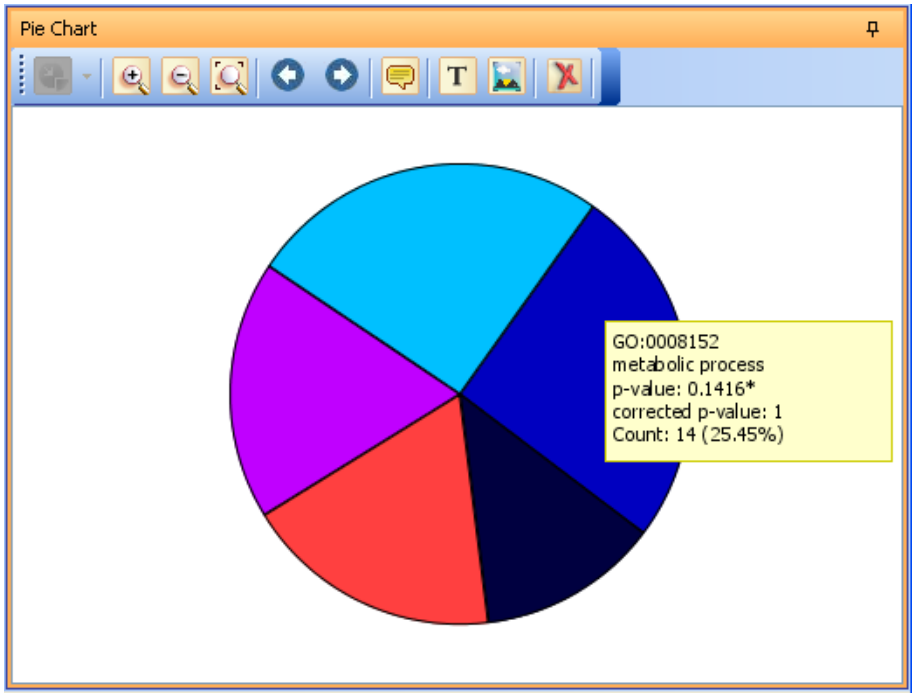


Figure 28.6: Pie Chart View

Default launch The pie chart by default is launched with the three top level GO terms of *Molecular Function*, *Biological Process* and *Cellular Component*.

Selecting Slices of the Pie To select a slice of a pie, click on the slice of interest. To add to the selection *Shift + Left-click* on the pies of interest. All the selected pies will be shown with a yellow border. You can also select slices by clicking and dragging the mouse over the canvas. A selection rectangle will be shown and all the slices within the selection rectangle will be selected.

Drill into pie To drill into a GO term and traverse down the hierarchy, select the pie or pies of interest by clicking on it. Click the Drill Selected Pie icon on the toolbar. This will execute one of the four selected options that are chosen in the drop-down list of the Drill Selected Pie icon. Double-click on any pie has exactly the same effect as drilling down the slice according to the chosen option.

Drill Pie One-Level This option will replace the current pie chart with a new pie chart, with GO terms one level below the GO terms of the selected slices. For example, if *Molecular Function* is selected, and the *Drill Pie One-Level* option is chosen, then the current top level pie will be replaced a pie with the first level children of *Molecular Function*. This is the default option.

Drill Pie All-Levels This option will replace the current pie chart with a new pie chart, with all the GO terms of the selected slices(s) below the GO terms of the selected slice(s). This pie chart cannot be drilled down further since it has been expanded to the last level.

Expand Slice One-Level This option will expand the selected slice(s) with GO terms one level below the GO terms of the selected slices. The other unselected slices, their GO terms, and their counts will remain unaffected. However, the slice sectors may change depending upon the counts of the individual slices

Expand Slice All-Levels This option will expand the selected slice(s) with all the GO terms of the selected slice(s) below the GO term of the selected slice(s). The other unselected slices, their GO terms, and their counts will remain unaffected. However, the slice sectors may change depending upon the counts of the individual slices

Zoom and fit to view To zoom in, zoom out or fit the pie chart view to the displayed canvas, click on the zoom in icon zoom out icon and Fit to view icon respectively.

Navigating through pies In the course of exploring the GO Analysis pie chart, you may be drilled into different levels of selected slices using different drill methods detailed above. You can navigate between the different drilled states of the pie chart by clicking on the Back icon and Forward icon respectively. These icons will be enabled or disabled appropriately depending upon the current state of the pie chart.

The pie chart can only remember a single path from the original top level pie to the current state. Thus, for example, if you had drilled into one slice, then went back, choose another slice to drill into then the previous drilled path will not be maintained.

Callouts for slices The slices of the pie chart denote different GO terms. If you hover the mouse on the slice the tool-tip shows the associated GO ID; the GO term; the p-value of the GO term; and the count of the number of entities contributing to any significant GO term in the hierarchy. Note that GO terms could be present even if they did not pass the specified cut-off because a GO term that was lower in the hierarchy satisfied the p-value cut-off. We use an asterisk (*) in the p-value to indicate this.

You can create a callout for selected slices by selecting the slices of interest and clicking on the Show Callouts icon on the tool bar. This will create a callout with the GO ID; the GO term; the p-value of the GO term; and the count of the number of entities contributing to any significant GO term in the hierarchy. The callouts can be selected, moved, and resized. To delete a callout, select the callout and click the Delete icon icon.

Add text and Image Texts can be added to the pie chart wherever required. To add text to the pie chart, click on the Switch Text Mode icon. This will change the cursor. You can click on the canvas of the pie chart and add text. Click on the icon again to toggle back to the selection mode. To add an image to the pie chart, click on the Insert Image icon. This will pop-up a file chooser. Choose the required image and add it to the pie chart.

Right-click menu on the pie chart The right click menu on the pie chart has options to print the pie chart to a browser, export the pie chart as an image to any desired resolution; and access the properties of the pie chart. The properties options of the pie chart allows you to change the properties of the view as detailed below: See Figure 28.7.

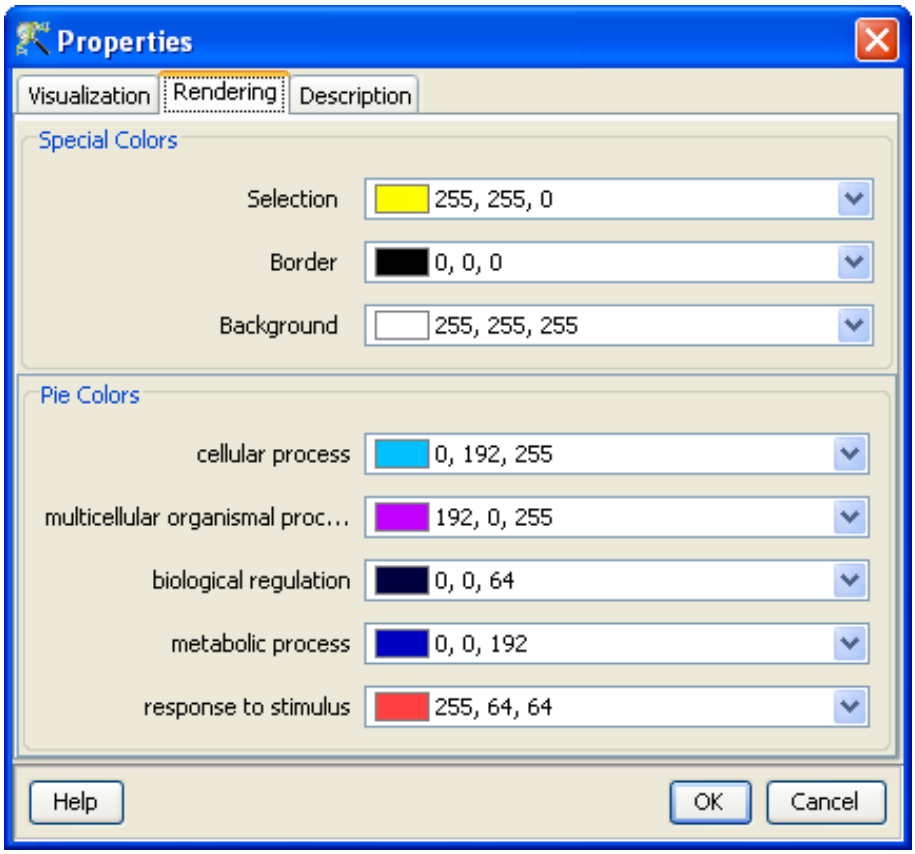


Figure 28.7: Pie Chart Properties

Visualization The *Visualization* tab of the properties dialog allow you to change the height of the pie chart from 0 to 100. the default is set at 100, when the pie chart is represented as a circle. The height can be decreased to make the pie chart an ellipse.

The *Minimum row count* of the pie chart can be changed. The default is set to 1. If the count, or number of entities is less than that specified in this dialog, the slice will not be displayed. This can be used to filter out GO terms with only a small number of entities.

Rendering The selection color, the border color, the background color, and the color of the slices of the pie can be changed.

Description You can add any description to the pie chart from the *Description* tab.

28.5 GO Enrichment Score Computation

Suppose we have selected a subset of significant entities from a larger set and we want to classify these entities according to their ontological category. The aim is to see which ontological categories are important with respect to the significant entities. Are these the categories with the maximum number of significant entities, or are these the categories with maximum enrichment? Formally stated, consider a particular GO term G. Suppose we start with an array of n entities, m of which have this GO term G. We then identify x of the n entities as being significant, via a t-test, for instance. Suppose y of these x entities have GO term G. The question now is whether there is enrichment for G, i.e., is y/x significantly larger than m/n. How do we measure this significance?

In most arrays each probeset is associated with single or multiple GO terms. Since some genes (Entrez-ids) are represented by multiple probesets, therefore GO term enrichment calculation gets biased toward genes having multiple probesets. Hence for unbiased calculation, multiple probesets corresponding to the same Entrez id are collapsed before running the GO analysis. A union of GO terms corresponding to multiple probesets for the same Entrez id is used for collapsed probeset. The following rule sets are followed for systematically condensing the probesets:

- If the entity has a single Entrez ID then take associated GO terms and associate it with this Entrez ID.
- If an entity has multiple Entrez IDs then if the Entrez ID has occurred previously and has an associated GO term, these are removed from the list. Each remaining Entrez ID get is then associated with GO terms.
- If an Entrez ID occurs only in groups, then its associated GO terms are found by taking an intersection of the GO terms in all the rows(entities) where it is present.

GeneSpring computes a p-value to quantify the above significance. This p-value is the probability that a random subset of x entities drawn from the total set of n entities will have y or more entities containing the GO term G. This probability is described by a standard hypergeometric distribution (given n balls, m white, n-m black, choose x balls at random, what is the probability of getting y or more white balls). **GeneSpring** uses the hypergeometric formula from first principles to compute this probability.

Since very often large number of hypothesis will be tested, some form of correction is required. However, there is no simple or straight forward way to do that. The different hypotheses are not independent by virtue of the way that GO is structured and even with this difficulty addressed, we are most interested in patterns of p-values that correspond to a structure in GO rather than single p-values exceeding some fixed threshold. In **GeneSpring** we have addressed the first issue using the Benjamini-Yekutieli correction [60, 3], which takes into account the dependency among the GO terms.

Finally, one interprets the p-value as follows. A small p-value means that a random subset is unlikely to match the actually observed incidence rate y/x of GO term G, amongst the x significant entities. Consequently, a low p-value implies that G is enriched (relative to a random subset of x entities) in the set of x significant entities.

NOTE: The GO analysis implementation in **GeneSpring** considers Molecular Function, and Cellular Component all together. Further, ‘‘part-of’’ relations in the ontology are ignored.

Chapter 29

Gene Set Enrichment Analysis

29.1 Introduction to GSEA

Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether an *a priori* defined set of genes shows statistically significant differences between two phenotypes. Traditional analysis of expression profiles in a microarray experiment involves applying statistical analysis to identify genes that are differentially expressed. In many cases, few genes pass the statistical significance criterion. When a larger number of genes qualify, there is often a lack of unifying biological theme, which makes the biological interpretation difficult. GSEA overcomes these analytical difficulties by focussing on gene sets rather than individual genes. It uses the ranked gene list to identify the gene sets that are significantly differentially expressed between two phenotypes.

GSEA analysis in **GeneSpring** is based on the GSEA implementation by the Broad Institute (http://www.broad.mit.edu/gsea) The current chapter details the GSEA Analysis, the algorithms to compute enrichment scores and methods to explore the results of GSEA analysis in **GeneSpring** .

29.2 Gene sets

A gene set from the Broad Institute is a group of genes, based on prior biological knowledge, that share a common function, chromosomal location or regulation. In **GeneSpring**, gene sets can also be defined as any entity lists created in the application that are used for GSEA.

The Broad Institute (http://www.broad.mit.edu/index.html) maintains a collection of gene sets. **GeneSpring** supports the import of MIT-Harvard-Broad gene sets in the following file formats:

- **txt/csv:** First line is header information and the remaining lines are genes.

- **grp:** Gene set file format where each gene is in a new line
- **gmt:** Gene Matrix Transposed file format where each row represents a gene set
- **xml:** Molecular signature database file format (msigdb_*.xml)

A detailed description of the file formats can be found at http://www.broad.mit.edu/cancer/software/gsea/wiki/index.php/Data_formats. The Broad gene sets can be found at http://www.broad.mit.edu/gsea/msigdb/msigdb_index.html. Each individual gene set can be viewed, downloaded and imported into **GeneSpring** . Alternatively, after registering with the web-site, one can download the entire collection.

Once Broad gene sets have been downloaded, they can be imported into **GeneSpring**. To import the Broad gene sets, click on the Import BROAD GSEA Gene sets link from the ***Tools*** section of the menu.

Importing gene sets in .grp, .gmt, or .xml formats into **GeneSpring** converts them into **GeneSpring** Entity Lists which can be searched from Search $\longrightarrow$ Entity Lists .

Note:To perform GSEA, Gene Symbol mark is essential and this is derived from the technology of the experiment. For Affymetrix, Agilent, and Illumina technologies, **GeneSpring** packages the Gene Symbol ID marks. For custom technologies, Gene Symbol must be imported and marked while creating a custom technology for performing the GSEA.

Note: If you would like to import all the gene sets in a single file, import *msigdb_v4.0.xml* all gene sets file.

29.3 Performing GSEA in GeneSpring

GSEA can be accessed from most of the workflows in **GeneSpring** . Clicking on the *GSEA* link in the *Result Interpretations* section of the Workflow panel will launch a wizard that will guide you through GSEA in **GeneSpring**.

Step 1: Input Parameters The input parameters for GSEA analysis is an entity list and an interpretation in the current active experiment. By default, the active entity list and the active interpretation in the experiment are selected. Clicking on the **Choose** option will show a tree of entity lists or interpretations in the experiment. You can choose any of the entity lists and interpretation from the tree as inputs to the GSEA Analysis. See Figure 29.1

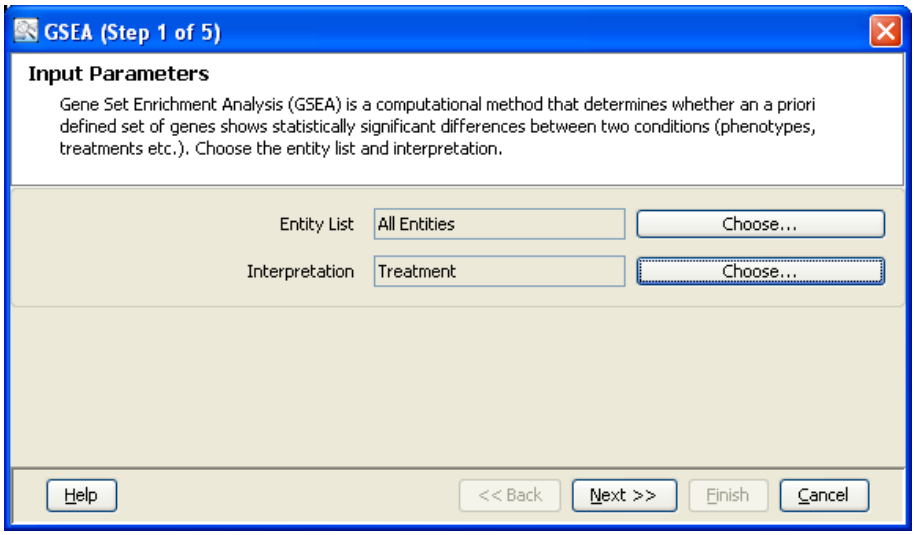


Figure 29.1: Step1: Input Parameters

Step2: Pairing Options In the Pairing Options page, you can explicitly select pairs of conditions for GSEA, or, you can select all the conditions in the interpretation against a single control condition. If you choose pairs of conditions, the table shows all the pairs. Choose the pairs of conditions to test by checking off the corresponding boxes. If you choose all conditions against control, select the condition to use as the control from the drop-down menu. See Figure 29.2

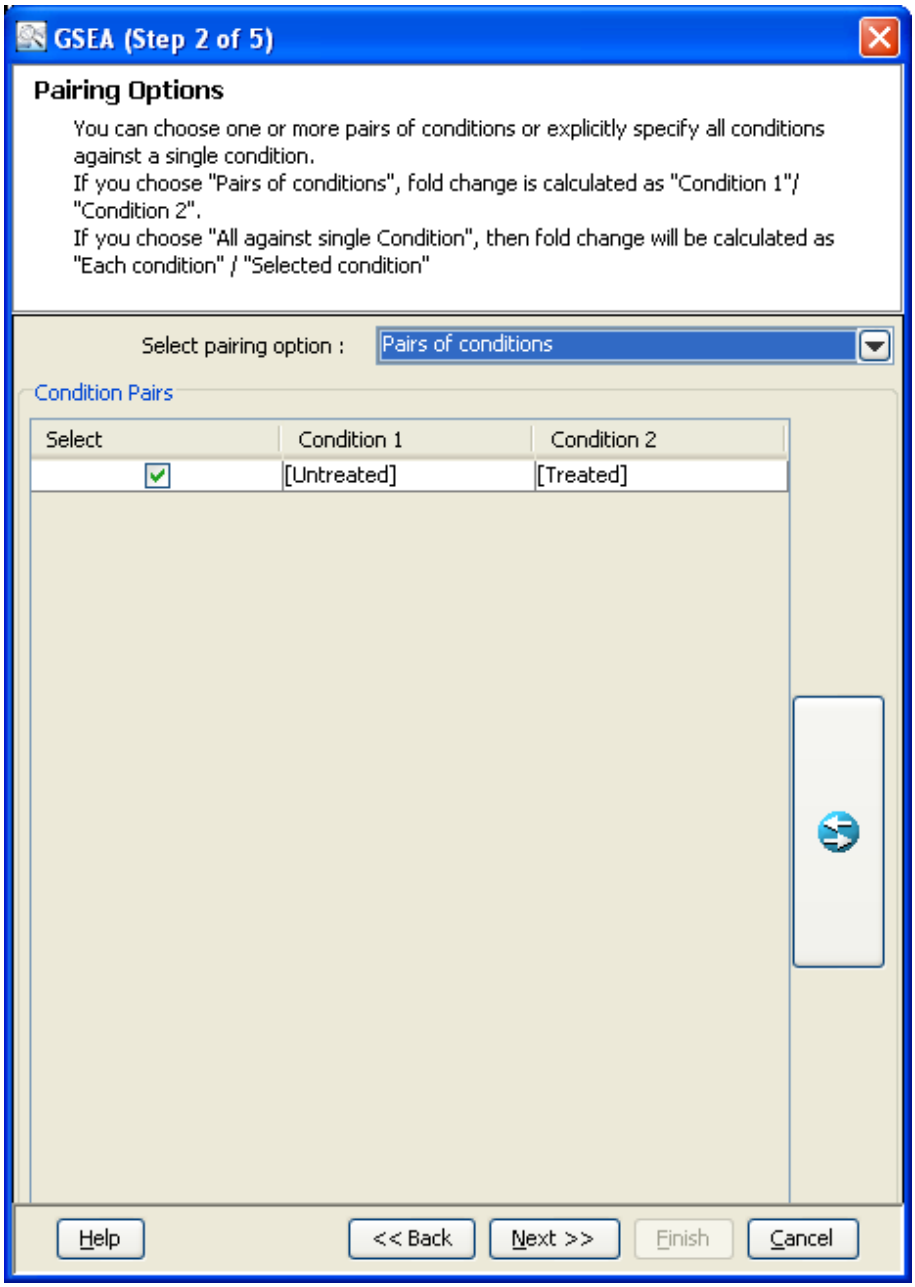


Figure 29.2: Step 2: Pairing Options

Step 3: Choose Gene Sets In the **Choose Gene Sets** options page, select Simple Search and choose one or more of the Broad gene sets that have been imported. See Figure 29.3.

You can also specify the minimum number of genes that must match between the gene set and the input entity list for GSEA in order for the gene set to be considered in the analysis. The default is set at 15 genes. Thus, if a gene set has less than 15 genes matching the entity list, then this gene set will not be considered. The default number of permutations used for analysis is set at 100.

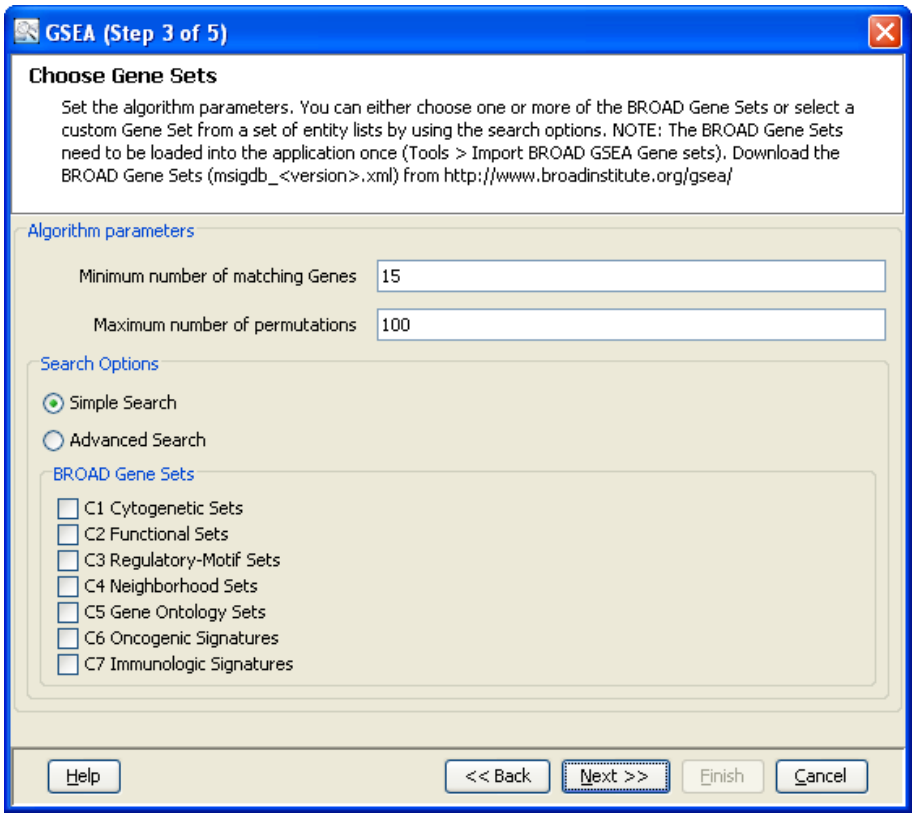


Figure 29.3: Step 3: Choose Gene Sets

Step 4: Choose Gene Sets Click on the Advanced Search radio button, search for the entity lists of interest, and select the ones to be used as gene sets for GSEA. Select custom gene sets from entity lists that you have created in **GeneSpring**. See Figure 29.4.

You can also specify the minimum number of genes that must match between the gene set and the input entity list for GSEA in order for the gene set to be considered in the analysis. The default is set at 15 genes. Thus, if a gene set has less than 15 genes matching the entity list, then this gene set will not be considered. The default number of permutations used for analysis is set at 100.

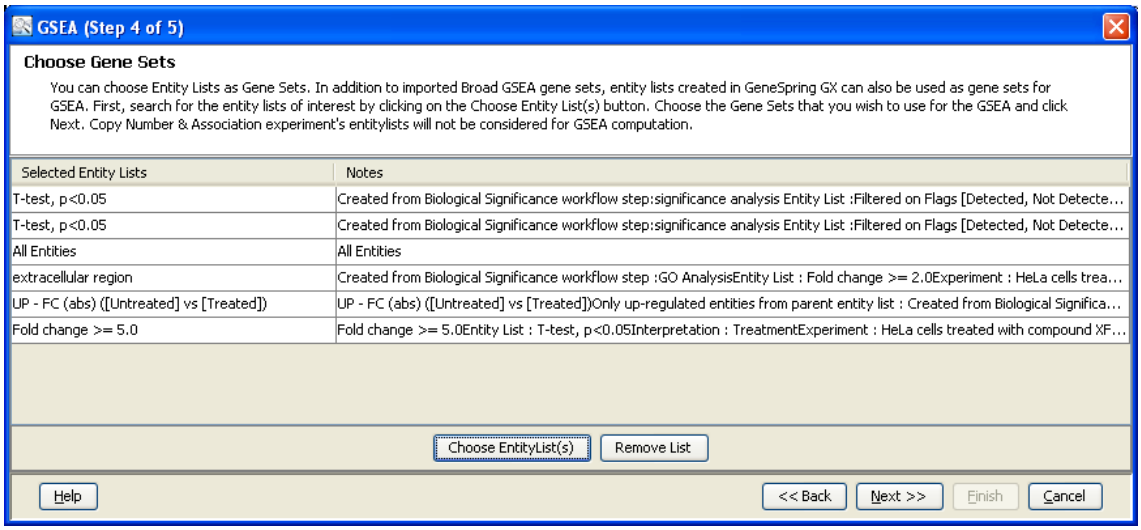


Figure 29.4: Step 4: Choose Entity Lists as Gene Sets

Step5: Results from GSEA The **Gene Sets satisfying minimum Gene requirement** spreadsheet shows the gene sets with q values below the specified cut-off. The **Gene Sets with fewer than minimum necessary matches** spreadsheet shows the gene sets satisfying the minimum number of matching genes specified in step 3. You can change the q-value cut-off by clicking on the Change q-value cut-off button and entering a new value. See Figure 29.5.

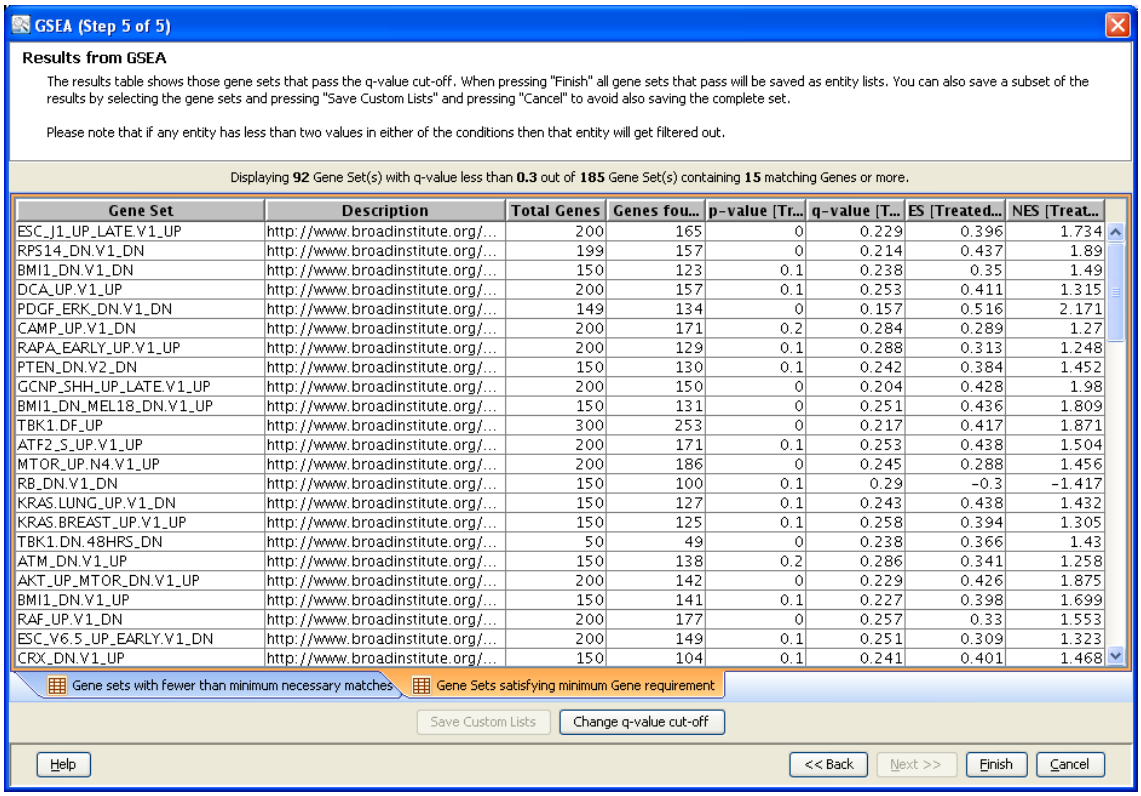


Figure 29.5: Step 5: Results

GSEA results spreadsheet reports the following columns of values:

- **Gene Sets:** List of gene sets that pass the threshold criterion.
- **Description:** User supplied description associated with the gene set.
- **Total Genes:** Total number of genes in the gene set.
- **Genes Found:** Number of genes in the gene set that are also present in the dataset on which analysis is performed.
- **p value:** Nominal p-value (from null-distribution of the gene-set).
- **q value:** False Discovery Rate q-value.
- **ES value:** Enrichment score of the gene set for the indicated pairs of conditions.
- **NES value:** Normalized enrichment score of the gene set for the indicated pairs of conditions.

Last four columns are repeated when multiple pairs of conditions are selected for analysis.

Gene sets with q-values below the cut-off can be saved to the Navigator. Click **Finish** to save all the gene sets within the **Gene Sets satisfying minimum Gene requirement** spreadsheet.

To save a subset of these gene sets, select the gene sets of interest and click Save Custom Lists. These gene sets will be automatically translated to the technology of the experiment and saved as entity lists in a GSEA folder within the Navigator. The saved entity lists are named according their respective gene set names.

29.4 GSEA Computation

GSEA analysis works on a ranked list of genes to compute the enrichment scores for gene sets. Correlation value for each gene is computed as the ratio between difference in mean expression values and the sum of standard deviations within each phenotype. Correlation values is used rank the genes in the dataset. Enrichment score for a gene set is defined as the maximum deviation from zero, when walking through the ranked list of genes, incrementing the running score for a hit, and decrementing for a miss. A hit is observed when the gene in the ranked list is present in the gene set, and a miss otherwise.

Thus analysis is restricted to log summarized datasets. If a gene has multiple probes in the dataset, the probe with maximum inter quartile expression range value is considered to compute the mean. Inter quantile range is immune to baseline transformation and hence GSEA results on baseline transformed data and no baseline transformed data remains same. GSEA algorithm and computation of associated metric is detailed in the paper http://www.pnas.org/content/102/43/15545.full.pdf. The permutative procedure described in the paper is used to compute the *p-values* and *q-values*. Number of permutations can be configured from *Tools* $\longrightarrow$ *Options* $\longrightarrow$ *Data Analysis Algorithms* $\longrightarrow$ *GSEA* of the menu bar.

Pseudo code for GSEA in **GeneSpring**

Create a collapsed with rows with unique genes for a given pair of phenotypes

For each row in the dataset
$m_1 = mean(phenotype_1)$
$m_2 = mean(phenotype_2)$
$stdev_1 = variance(phenotype_1)$
$stdev_2 = variance(phenotype_2)$
$correlationValue = \frac{m_1 - m_2}{Sqrt(stdev_1 + stdev_2)}$ *Note: stdev is forced to a minimum value of 0.2*abs(mean_i)*

Sort correlation values in increasing order to get the RankedGeneList

For each gene set

```
runningScore = 0
R = number of rows in dataset
g = number of genes in gene set
correlationTotal = sum (correlationValues for genes in gene set)
for gene in RankedGeneList
  if gene present in gene set
```

```
    runningScore = runningScore + (correlationValue of gene)/(correlationTotal)
  else
    runningScore = runningScore - 1/(R-g)
ES = max(abs(runningScore))
```

Compute ES_* for each random permutation of phenotype ordering Divide the enrichment scores obtained into $ES_*(>= 0)$ and $ES_*(< 0)$

If $ES > 0$
$p - value = \frac{NumberOfPermutationsWhereES>ES_*andES_*>=0}{NumberOfPermutationsWhereES_*>=0}$
$NES = ES/max(ES_*(> 0))$
$NES_* = NES/max(ES_*(> 0))$

For a given gene set with NES_i
$ratio1 = \frac{NumberOfRandomPermutationsForAllGeneSetsWhereNES<NES_*andNES_*>=0}{NumberOfRandomPermutationsForAllGeneSetsWhereNES_*>=0}$
$ratio2 = \frac{NumberOfGenesetsWhereNES_i<NESandNES>=0}{NumberofGeneSetsWhereNES>0}$
$q - value = \frac{ratio1}{ratio2}$

The process is repeated for gene sets with $ES < 0$

29.5 Import BROAD GSEA Genesets

GSEA can be performed using the seven genesets which are available from the Broad Institutes website (http://www.broad.mit.edu/gsea/). These genesets can be downloaded and imported into the **GeneSpring** to perform GSEA. Clicking on this option allows the user to navigate to the appropriate folder where the genesets are stored and select the set of interest. The files should be present either in .xml or .grp or .gmt formats. Remember to provide 'Read' permissions to these gene lists after import, if you intend having others use these lists to run GSEA.

Chapter 30

Gene Set Analysis

30.1 Introduction to GSA

Gene Set Analysis (GSA) is a computational method that determines whether an *a priori* defined set of genes shows statistically significant differences between two phenotypes. Traditional analysis of expression profiles in a microarray experiment involves applying statistical analysis to identify genes that are differentially expressed. In many cases, few genes pass the statistical significance criterion. When a larger number of genes qualify, there is often a lack of unifying biological theme, which makes the biological interpretation difficult. GSA overcomes these analytical difficulties by focussing on gene sets rather than individual genes. The implementation uses the S-Score, computed as a cumulative sum of t-Scores across the phenotypes for genes in a gene set for significance testing.

GSA in **GeneSpring** is based on the GSA implementation by the Department of Statistics (http://www-stat.stanford.edu/). The chapter details the GSA Analysis, the algorithms to computeS-Scores and methods to explore the results of GSA analysis in **GeneSpring** .

30.2 Gene sets

A gene set from the Broad Institute is a group of genes, based on prior biological knowledge, that share a common biological function, chromosomal location or regulation. In **GeneSpring**, gene sets can also be defined as any entity lists created in the application that are used for GSA.

The Broad Institute (http://www.broad.mit.edu/index.html) maintains a collection of gene sets. **GeneSpring** supports the import of MIT-Harvard-Broad gene sets in the following file formats:

- **txt/csv:** First line is header information and the remaining lines are genes.

- **grp:** Gene set file format where each gene is in a new line
- **gmt:** Gene Matrix Transposed file format where each row represents a gene set
- **xml:** Molecular signature database file format (msigdb_*.xml)

A detailed description of the file formats can be found at http://www.broad.mit.edu/cancer/software/gsea/wiki/index.php/Data_formats. The Broad gene sets can be found at http://www.broad.mit.edu/gsea/msigdb/msigdb_index.html. Each individual gene set can be viewed, downloaded and imported into **GeneSpring** . Alternatively, after registering with the website, one can download the entire collection. Links to other gene sets can be found at (http://www-stat.stanford.edu/~tibs/GSA/).

Once Broad gene sets have been downloaded, they can be imported into **GeneSpring**. To import the Broad gene sets, click on the *Import BROAD GSEA Gene sets* link from the ***Tools*** section of the menu bar.

Importing gene sets in .grp, .gmt or .xml formats into **GeneSpring** converts them into **GeneSpring** Gene Lists which are automatically marked as Gene Symbol. (Note that importing the msigdb_v2.xml into **GeneSpring** takes around 10 minutes as the XML file is parsed)

Note: To perform GSA, the Entrez ID or Gene Symbol mark is essential. These are derived from the technology of the experiment. For Affymetrix, Agilent and Illumina technologies, **GeneSpring** packages the Entrez ID and Gene Symbol IDs marks. For custom technologies, Entrez ID or Gene Symbol must be imported and marked while creating custom technology for using the GSA.

30.3 Performing GSA in GeneSpring

GSA can be accessed from most of the workflows in **GeneSpring** . Clicking on the *GSA* link in the *Result Interpretations* section of the Workflow panel will launch a wizard that will guide you through GSA in **GeneSpring**.

Input Parameters The input parameters for GSA analysis is an entity list and an interpretation in the current active experiment. By default, the active entity list and the active interpretation in the experiment are selected. Clicking on the *Choose* option will show a tree of entity lists or interpretations in the experiment. You can choose any of the entity lists and interpretation from the tree as inputs to the GSA Analysis. See Figure 30.1.

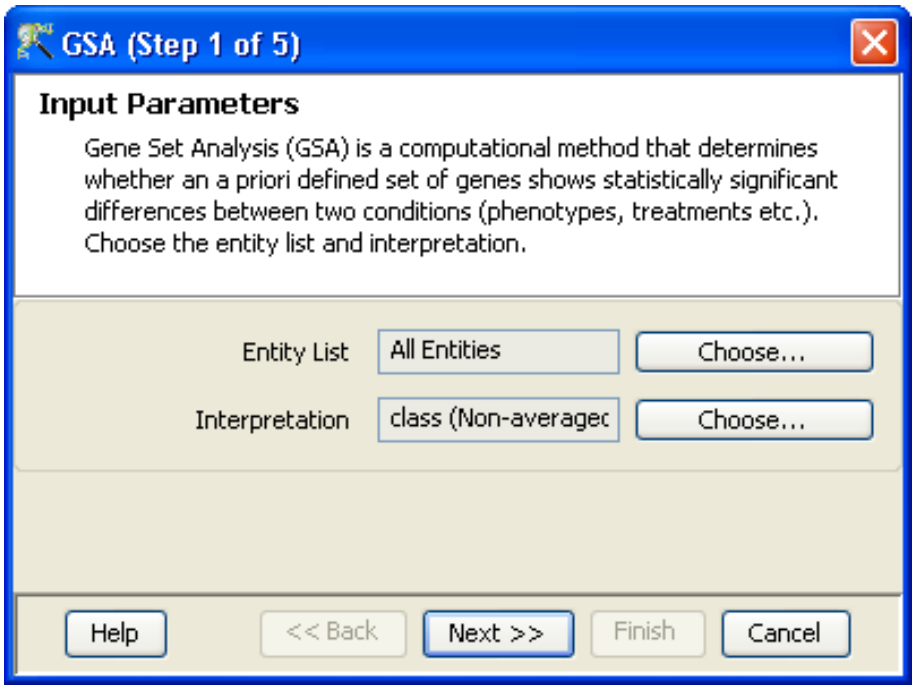


Figure 30.1: Input Parameters

Pairing Options In the Pairing Options page, you can explicitly select pairs of conditions for GSA, or, you can select all the conditions in the interpretation against a single control condition. If you choose pairs of conditions, the table shows all the pairs. Choose the pairs of conditions to test by checking off the corresponding boxes. If you choose all conditions against control, select the condition to use as the control from the drop-down menu. See Figure 30.2.

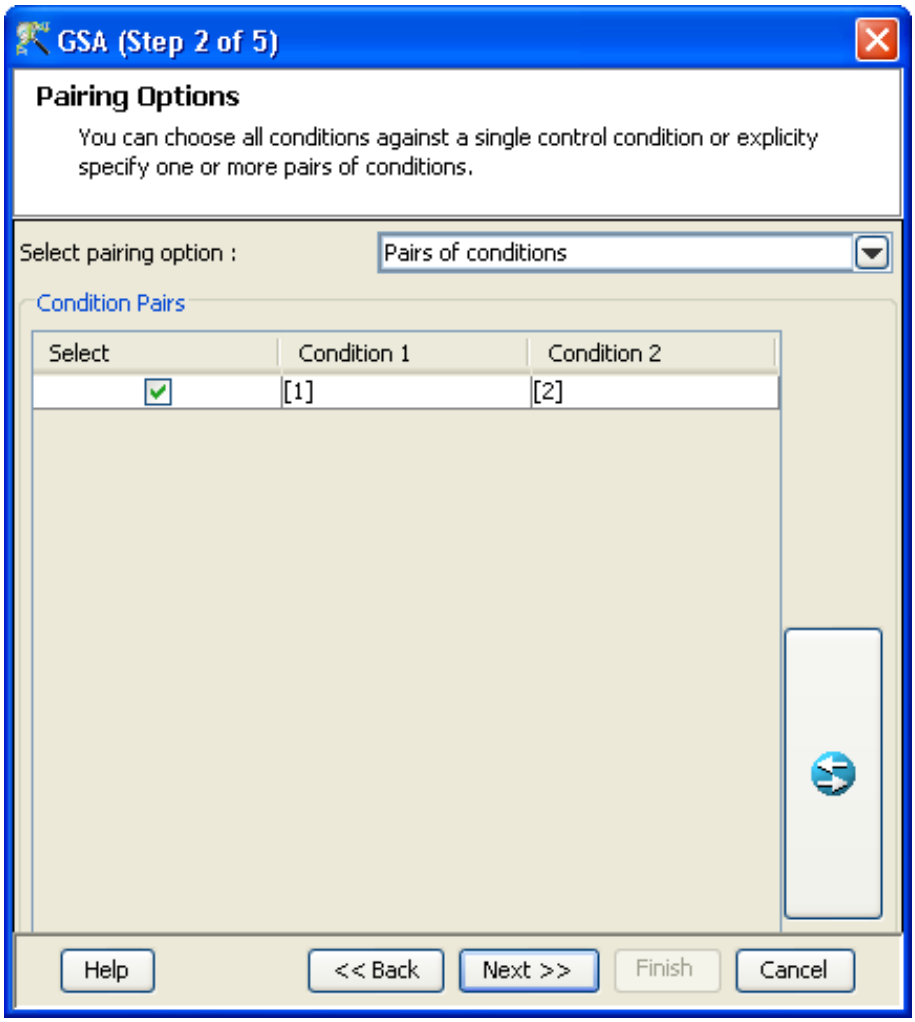


Figure 30.2: Pairing Options

Choose Gene Sets In the *Choose Gene Sets* options page, you can choose one or more of the BROAD gene sets that have been imported. See Figure 30.3. Alternatively, you can select custom gene sets from entity lists that you have created in **GeneSpring**. To do this, click on the Advanced Search radio button, search for the entity lists of interest, and select the ones to be used as gene sets for GSA. See Figure 30.4.

You can also specify the minimum number of genes that must match between the gene set and the input entity list for GSA in order for the gene set to be considered in the analysis. The default is set at 15 genes. Thus, if a gene set has less than 15 genes matching the entity list, then this gene set will not be considered. The default number of permutations used for analysis is set at 100.

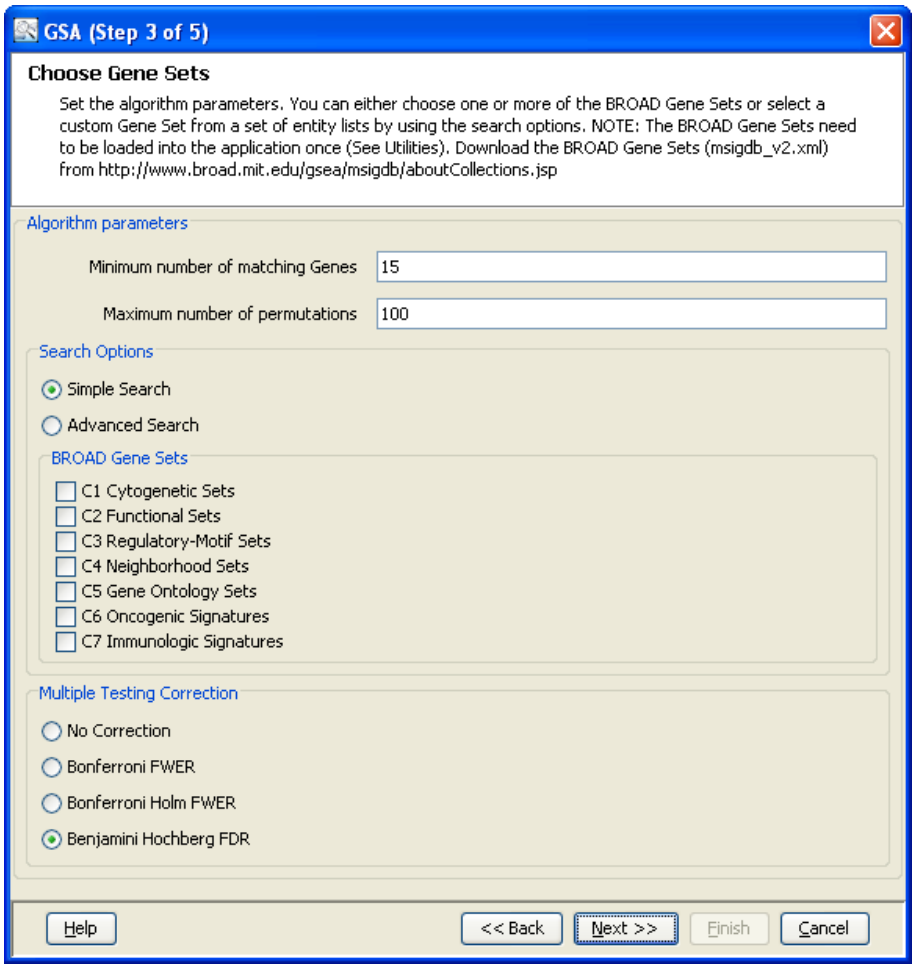


Figure 30.3: Choose Gene Sets

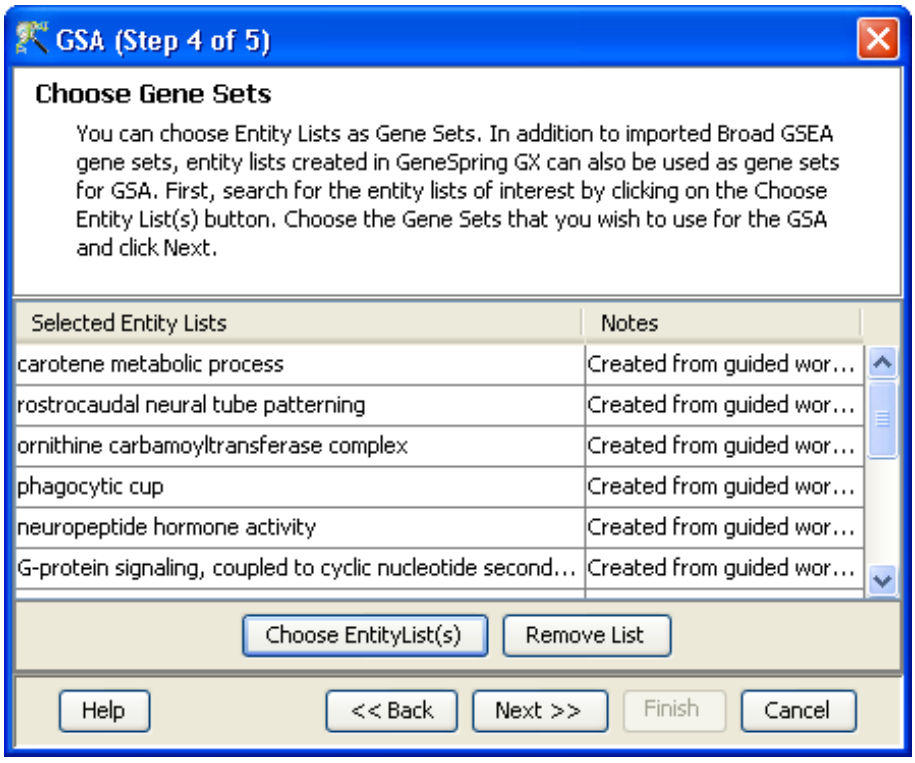


Figure 30.4: Choose Gene Sets

Results from GSA The *Gene Sets satisfying minimum Gene requirement* spreadsheet shows the gene sets with p-values below the specified cut-off. The *Gene Sets with fewer than minimum necessary matches* spreadsheet shows the gene sets with p-values above the specified cut-off. You can change the p-value cut-off by clicking on the Change p-value cut-off button and entering a new value. See Figure 30.5.

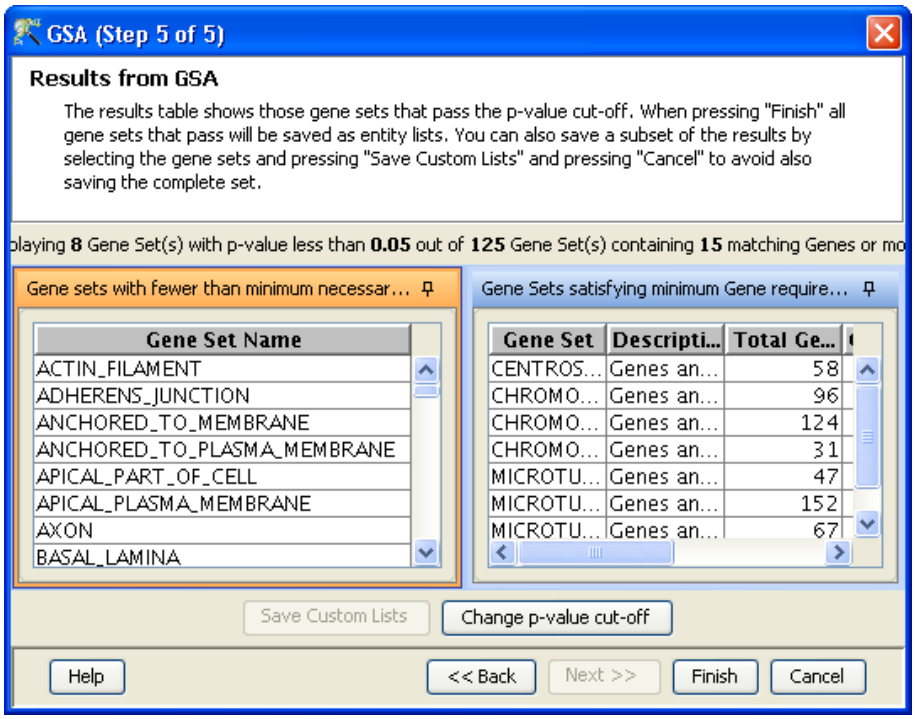


Figure 30.5: Choose Gene Sets

GSA results spreadsheet reports the following columns of values:

- **Gene Sets:** List of gene sets that pass the threshold criterion.
- **Details:** User supplied description associated with the gene set.
- **Total Genes:** Total number of genes in the gene set.
- **Genes Found:** Number of gene in the gene set that are also present in the dataset on which analysis is performed.
- **S-Score:** Enrichment score of the gene set for the indicated pairs of conditions.
- **p-value:** Nominal p-value (from null-distribution of the gene-set)
- **Corrected p-value:** False Discovery Rate p-value, if p-value correction is requested

Columns except the first four are repeated when multiple pairs of conditions are selected for analysis.

Gene sets with Corrected p-values below the cut-off can be saved to the Navigator. Click *Finish* to save all the gene sets within the *Gene Sets satisfying minimum Gene requirement* spreadsheet. To save a subset of these gene sets, select the gene sets of interest and click Save Custom Lists. These gene sets will be automatically translated to the technology of the experiment and saved as entity lists in a GSA folder within the Navigator. The saved entity lists are named according their respective gene set names.

30.4 GSA Computation

Algorithm and computation of associated metric is detailed in the paper http://www-stat.stanford.edu/~tibs/ftp/GSA.pdf.

GSA analysis works on the t-Score computed for a list of genes, and uses a sum of t-Scores to compute the s-Score for each for gene set. t-Score in **GeneSpring** is computed for each gene assuming unequal variance between the conditions. The max-mean statistic is used to compute S-Scores for each gene set. The permutative procedure described in the paper is used to compute the *p-values*. S-Score for a gene set is compared with scores generated from random permutations of samples between the phenotypes to determine the level of significance. Number of permutations can be configured at *Tools* $\longrightarrow$ *Options* $\longrightarrow$ *Data Analysis Algorithms* $\longrightarrow$ *GSEA* of the menu bar.

Analysis is restricted to log summarized datasets. If a gene has multiple probes in the dataset, the probe with maximum inter-quartile range across the expression values (with no more than 25% of the values missing) is used to represent the gene. Inter-quantile range is immune to baseline transformation and hence GSA results on data with and without baseline transformation remain the same.

Pseudo code for GSA in **GeneSpring**
$R = NumberOfRowsInDataset$

Create a collapsed with rows with unique genes for a given pair of phenotypes

For each row in the dataset
$n_1 = NumberOfSamplesOfPhenotype_1$
$n_2 = NumberOfSamplesOfPhenotype_2$
$m_1 = mean(phenotype_1)$
$m_2 = mean(phenotype_2)$
$var_1 = variance(phenotype_1)$
$var_2 = variance(phenotype_2)$
$tScore = \frac{m_1 - m_2}{\sqrt{(\frac{var_1}{n_1} + \frac{var_2}{n_2})}}$

Separate the positive and negative t-scores into two arrays
$t(>0) = ArrayOfPositivetScores$
$t(<0) = ArrayOfNegativetScores$
$mean_+ = \frac{sum(t(>0))}{R}$
$mean_- = \frac{sum(-t(<0))}{R}$
$stdev_+ = \sqrt{(\frac{sum(t(>0)^2) - R * mean_+}{R})}$
$stdev_+ = \sqrt{(\frac{sum(t(>0)^2) - R * mean_-}{R})}$

For each gene set
$g = NumberOfGenesInGeneSet$
$sScore_+ = \frac{\frac{sum(t(>0)ForGenesInGeneSet)}{g} - mean_+}{stdev_+}$
$sScore_- = \frac{\frac{sum(-t(<0)ForGenesInGeneSet)}{g} - mean_-}{stdev_-}$

$sScoreForGeneSet = max(sScore_+, sScore_-)$ Compute $sScore_*$ for each random permutation of phenotype ordering
$p - value = \frac{NumberOfPermutationsWhere(abs(sScore)<abs(sScore_*))}{TotalNumberOfPermutations}$

Chapter 31

Pathway Analysis

One of the key advantages of **GeneSpring** is that it not only enables you to analyze data from different high-throughput technologies like genomics, transcriptomics, proteomics, and metabolomics in the same tool, it also allows you to compare data from these different experiment types in the same project.

Analysis of any 'omics' data in **GeneSpring** typically results in a list of entities (genes or proteins, small molecules) that are significantly different in the experimental conditions of interest. Pathway analysis in **GeneSpring** provides the necessary biological context for a functional analysis of these entities to better understand their role in a biological process.

In **GeneSpring**, the Pathway Analysis module supports analysis on well-studied, curated pathways, while the NLP Network Discovery component drives discovery by creating networks around the entities of your interest using a powerful Natural Language Processing (NLP) algorithm that extracts information from published literature.

Multi-Omic Analysis in **GeneSpring** now allows you to overlay data from two different experiments on the same pathway simultaneously and thus enables an integrated analysis of data from different experiment types. Read more about how this can help with your research under What is Multi-Omic Pathway Analysis (MOA)? on page 896.

You can interpret your experimental data through the following pathway analysis features in **GeneSpring**:

- Import curated pathways:
  - directly from the WikiPathways portal (`http://www.wikipathways.org`),
  - derived from the BioCyc Pathway/Genome databases (`http://www.biocyc.org/`) provided by the Bioinformatics Research Group at SRI International® (used under license) [9],

Figure 31.1: BioCyc Database Collection

  - derived from the KEGG (Kyoto Encyclopedia of Genes and Genomes) Pathway Databases (`http://www.genome.jp/kegg/pathway.html`) developed by Kanehisa Laboratories® (can be used with a valid end user license) [37],
  - pathways from other sources in BioPAX (Level 2 and Level 3), GPML, or Text format,

  or create interaction networks from a database of biological and chemical entities, relationships between entities, and properties of these entities and relationships derived from our proprietary NLP algorithm.

  Determine which of the created or imported pathways have significant overlap with a specified list of entities from one (Single Experiment Analysis) or two experiments of the same or differing experiment types (Multi-Omic Analysis). Find out more about this under Pathway Analysis in **GeneSpring** from page 896.

- View and investigate these pathways and interaction networks in an interactive pathway viewer and overlay your experimental data on these pathways.
- Access other popular pathway analysis tools like IPA, Metacore, and Cytoscape from within **GeneSpring**, allowing you to export data to these tools. In the case of IPA, you can also import entity lists resulting from pathway analysis in that tool into **GeneSpring**. See Analyzing Pathways using External Tools Accessible from **GeneSpring** on page 1015 on how to use this feature.
- Create networks based on information in PubMed abstracts and identify interactions associated with Medical Subject Headings (MeSH) terms using the in-built NLP feature. This is an alternate way of creating pathways based on terms/concepts instead of entities (see NLP Networks on page 1037). Once you have created and saved such networks in **GeneSpring**, you can also overlay data from your experiments on these networks when looking for significant pathways and networks.

Using these pathway analysis features in **GeneSpring**, you can answer questions such as: What well-known biological pathways and processes are significantly represented by genes of my interest? What other entities and processes have been reported in literature to be affected by these genes/proteins/metabolites? Is there a pattern in the expression of connected genes across different experimental conditions or different entity types as measured by experiments under similar conditions? For instance, overlaying data on a signaling pathway can provide an understanding of the cause-effect relationships between the genes or proteins of interest and insights into the mechanism of a specific condition under study.

Data from different experiment types, say, a Transcriptomics and Metabolomics experiment on a metabolic pathway may reveal that some of the metabolite abundances in a pathway closely mimic the expression of enzymes catalyzing the

inter-conversion of these metabolites. Such reactions in the pathway could be enzyme limited reactions - the rate-limiting steps - in the pathways. The absence of a correlation on the other hand may suggest a different regulatory mechanism for the enzyme.

Connected genes, proteins or metabolites in a pathway can also suggest a design for your next set of experiments to support your hypothesis. Pathway Analysis and Network Discovery thus provide a unifying biological theme for the entities resulting from analysis in **GeneSpring**.

31.1 Getting Started

To begin Pathway Analysis in **GeneSpring**, you require the following:

1. A valid license for the Pathway Architect module (If you would like to use curated pathways).
2. Pathways from sources of interest for the organism under study.
3. Supporting databases to perform pathway analysis for a single organism or across different organisms.

This section describes the licensing process for the pathway analysis module in **GeneSpring**. It also describes how you can import pathways from external sources into **GeneSpring**, how to acquire the required supporting databases to perform pathway analysis with a single or mulitiple organisms and how to work with organisms for which such supporting databases are not available. In addition, details on all the elements and organization of the **GeneSpring** Interaction Databases, as well as information on how to work with and update pathway interactions in the interaction database are also provided in this section.

31.1.1 Pathway Module Licensing

Any pathway features in **GeneSpring** that use curated pathways are licensed separately and can only be accessed with a valid Pathway Architect module license.

If you have a desktop license for **GeneSpring** that does not include pathway features, contact **GeneSpring** support to get a Pathway module license order-id. Then use **Help** $\longrightarrow$ **License Manager** $\longrightarrow$ **Change** and provide this order-id in the specified field. All pathway and network analysis features will be available after authentication with the **GeneSpring** license server (you have to be connected to the internet to complete this step). Once the license is entered, **GeneSpring** automatically closes and has to be restarted.

If you have a **WorkGroupPlus** license that does not include pathway features, contact your Systems Administrator and request them to contact **GeneSpring** support to get a pathway module license order-id; this order-id can be used to upgrade the **GeneSpring** floating license server to include the pathway module. Once this is done, the next time you launch the **GeneSpring** client, all pathway-related functionality will be enabled.

31.1.2 KEGG Pathways Licensing

A valid KEGG end user license is required to import KEGG pathways into **GeneSpring**. KEGG ftp website login credentials are compulsorily validated when you try to Import KEGG pathways from the Agilent Server (see To import pathways from KEGG through the main menu on 868 and on page 906). If the automatic validation is unsuccessful, **GeneSpring** prompts you to enter valid credentials to access the KEGG ftp website. After successful validation, **GeneSpring** proceeds with the download of KEGG pathways from the Agilent Server.

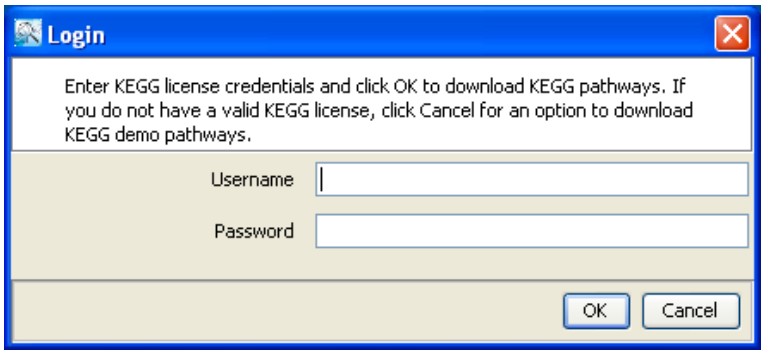


Figure 31.2: Validation to import KEGG pathways

If you require access to the KEGG pathway download from a network location that is not linked to your KEGG account, please contact the **GeneSpring** support team (`http://genespring-support.com/support`).

If you do not have a KEGG pathway license but would like to try out the KEGG pathway features in **GeneSpring**, we provide an option to download ten demo KEGG pathways for seven organisms after KEGG license authentication has failed.

31.1.3 Importing Pathways into GeneSpring

Your **GeneSpring** installation contains 21 pathways. One of these pathways was created using the NLP Network Discovery workflow in **GeneSpring** with the **cadmium ion binding** entity list as input and can be viewed in the HeLa cells demo experiment. The remaining set of 20 pathways is included in **GeneSpring**, courtesy of the Computational Biology Center at Memorial Sloan-Kettering Cancer Center, Gary Bader's lab at the University of Toronto for the 'Cancer Cell Map', the Pandey Lab at Johns Hopkins University and the Institute of Bioinformatics (Bangalore, India). The pathway diagrams were loaded through the BioPAX exchange format, an open platform for the distribution of network and pathway information.

In addition, you can migrate pathways from earlier versions of **GeneSpring**. All these pathways are available to perform pathway analysis in **GeneSpring** through either Single Experiment Analysis (SEA) or Multi-Omic Analysis (MOA) as described in Pathway Analysis in **GeneSpring** on page 896.

Additional pathways can be imported into **GeneSpring** directly from the WikiPathways portal (`http://www.wikipathways.org`), from the Agilent Server (BioCyc pathways and KEGG pathways), or from any source in the BioPAX (Level 2 and Level 3), GPML and text formats. To import pathways in the BioPAX (Level 2) or text format and Pathway Architect pathways, you have to download an organism's Interaction Database first.

31.1.3.1 To migrate pathways from earlier versions

GeneSpring 12.0 and higher versions support viewing, searching, and analysis of NLP-derived pathways from earlier versions. When updating your installation of **GeneSpring** to version 12.0 or higher, these pathways are migrated into the

version and their **Source** is automatically labeled as "Legacy". Note that the original **Source** label is overwritten in this process. If you want to retain the original source information you can do so by manually adding it to the **Notes** field when inspecting the pathway.

31.1.3.2 To import pathways from WikiPathways through the main menu

Organism specific pathways available at http://www.wikipathways.org/ can be imported into **GeneSpring** through **Tools** ⟶ **Import Pathways from WikiPathways** via the main menu. Use this option to import all pathways for all organisms, or all pathways for one selected organism at a time (available from a drop-down menu).

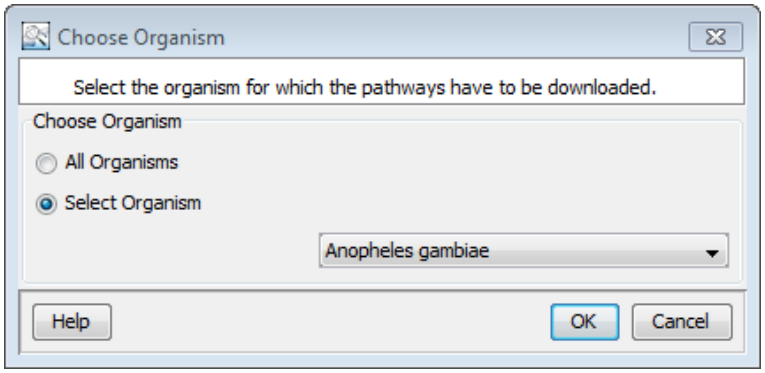


Figure 31.3: Select Organism for WikiPathways download

When **GeneSpring** has downloaded all the pathways, the number of pathways with an Analysis collection tag and those without is displayed in the **Import Statistics** dialog. Please note, that the same pathway may be present both under Analysis and Non-Analysis pathways, differentiated only by their version numbers.

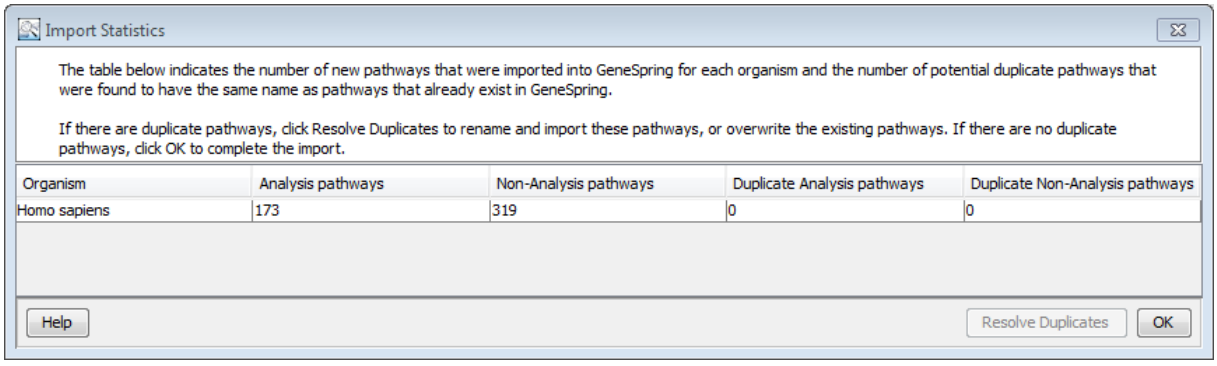


Figure 31.4: Import Statistics dialog - no duplicates

When downloading pathways from WikiPathways, **GeneSpring** detects potential duplicates (pathways that already exist in **GeneSpring** with the same name and are associated with the same organism). If such potential duplicates are present, the number of duplicates is indicated in the last two columns.

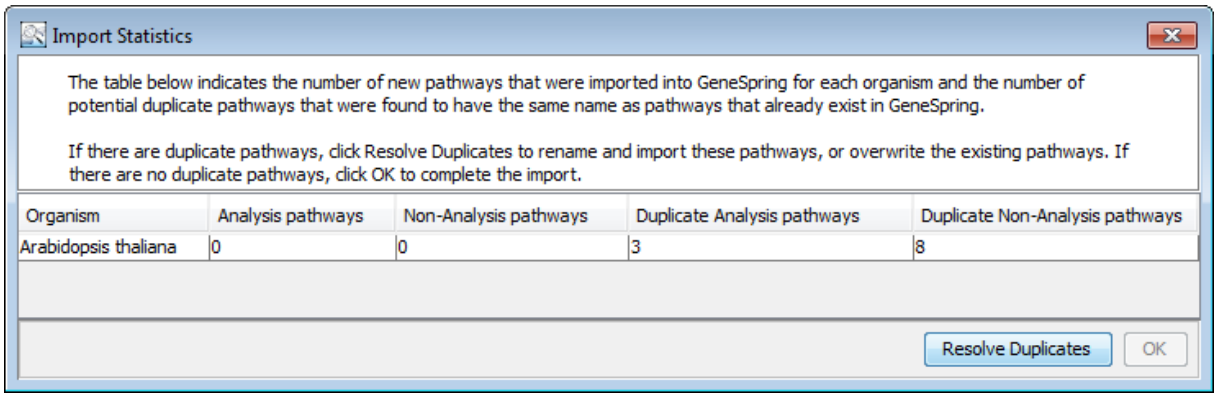


Figure 31.5: Import Statistics dialog - duplicates are present

Click **Resolve Duplicates** to import these duplicate pathways with different names, overwrite the existing pathways with these pathways or not import these pathways at all.

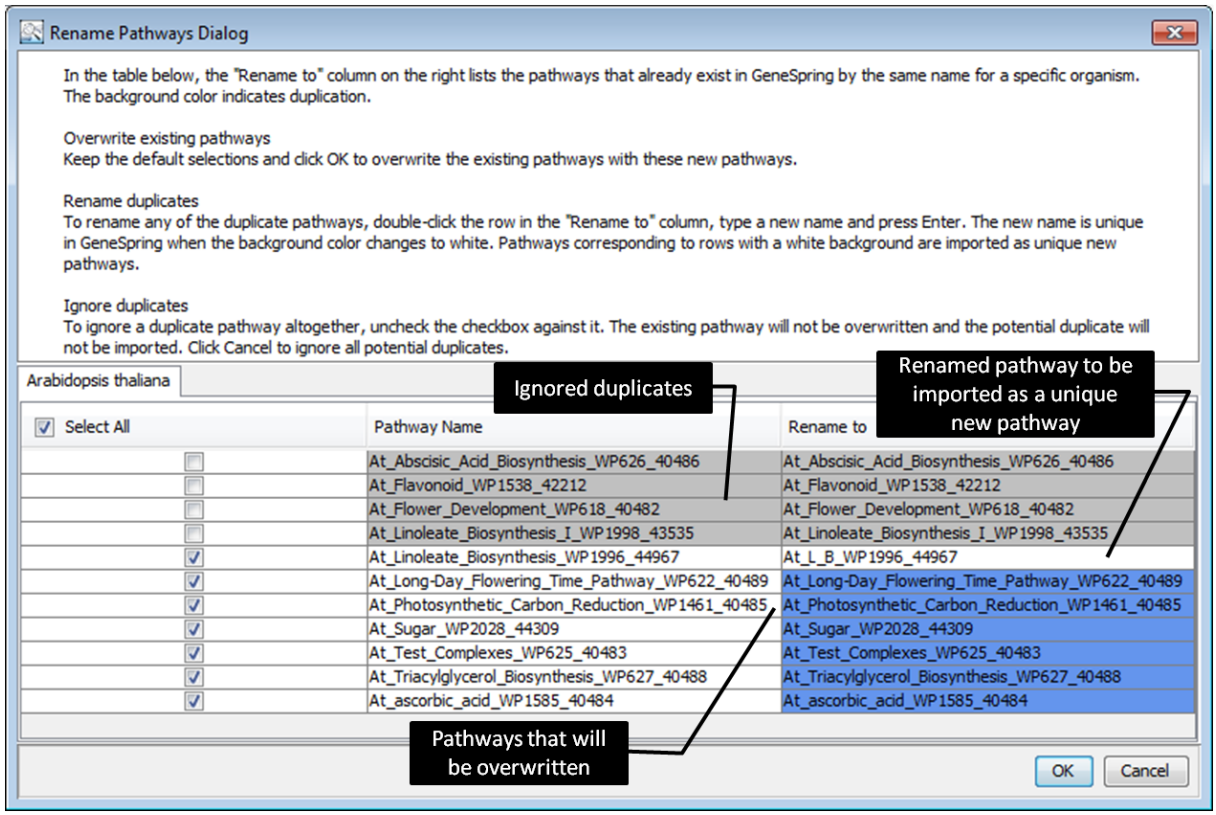


Figure 31.6: Rename Pathways Dialog

By default, all pathways are selected and on clicking **Import** the existing pathways are overwritten with the just downloaded pathways. The background color of the cells in the **Rename to** column of the **Rename Duplicates** dialog indicates

that this pathway name already exists for pathways associated with the same organism and with WikiPathways as pathway source. If there are duplicates for several organisms, a separate spreadsheet is created for each organism. These can be accessed by clicking on the spreadsheet tab with the corresponding organism name.

To rename any of the duplicate pathways and import them as new pathways, double-click the corresponding row in the **Rename to** column and type the new name. Click **Enter** to confirm. If the background color stays white, **GeneSpring** has identified that the new pathway name is unique for pathways associated with the same organism and with WikiPathways as pathway source. Otherwise the background color will reappear and the pathway with that name, which already exists in **GeneSpring**, will be overwritten when clicking **Import**.

If you do not want to overwrite any existing pathways or rename the new pathways, click **Cancel**, or unselect all pathways and click **Import**.

WikiPathways offers downloadable pathways in multiple formats (see `http://wikipathways.org/index.php/Download_Pathways`). However, for best viewing and rendering of these pathways, we recommend the import of pathways from WikiPathways in the GPML format only. Previously downloaded pathways from WikiPathways (in the GPML format) can be imported into **GeneSpring** through **Tools** ⟶ **Import Pathways** ⟶ **GPML**. (see To import pathways from file on page 870). However, it is worth noting here, that **GeneSpring** is not able to extract from this file whether the pathway was tagged as an Analysis Collection pathway on the WikiPathways portal or not.

You can also import pathways from WikiPathways directly from the SEA or MOA workflows (see To download WikiPathways in an SEA or MOA workflow on page 905).

NOTE A **GeneSpring WorkGroupPlus** user cannot download the WikiPathways of an organism if the download is being accessed by another user.

31.1.3.3 To import pathways from BioCyc through the main menu

GeneSpring provides an option to import pathways for a number of organisms from the BioCyc Database Collection [9] (`http://biocyc.org/`) directly from the Agilent server or an Update File provided by the **GeneSpring** Support Team. To import these pathways:

1. Go to **Tools** ⟶ **Import Pathways from BioCyc** ⟶ **From Agilent Server**.

   The **Automatic Software Update** dialog opens.
2. Select the organisms for which you want to import available BioCyc pathways. Click **Update**.

   An Information dialog informs you of the size requirements of the pathway download. Click **OK**. **GeneSpring** confirms completion of the download with another Information dialog. Click **OK**.

GeneSpring imports all the pathways for the selected organisms. The progress of the import process is indicated by a progress bar. The pathways can now be used for SEA or MOA by selecting BioCyc from the **Curated pathways** section in **Step 1: Specifying Experiments, Organisms, and Pathway Sources** of the wizard-driven SEA and MOA workflows.

NOTE

If you publish research results in scientific journals based on the use of data from any of these BioCyc pathways, reference the results using the BioCyc and EcoCyc (for E.coli pathways) citations provided in the bibliography [9].

BioCyc pathway updates

When updates become available for the BioCyc Database Collection, these updates are also provided on the Agilent server after some time. Go to **Tools** $\longrightarrow$ **Import Pathways from BioCyc** $\longrightarrow$ **From Agilent Server** to check for any available BioCyc pathway updates. The import procedure for these updates is the same as described above. These updated versions do not overwrite previously downloaded BioCyc pathways but are stored in **GeneSpring** in addition to the earlier versions. If you have downloaded multiple versions, go to **Tools** $\longrightarrow$ **Options** $\longrightarrow$ **Pathway** $\longrightarrow$ **BioCyc version for analysis** before starting a SEA or MOA workflow to specify which BioCyc pathways you want to use for analysis from one of these three options: **Latest version only**, **Older versions only**, or **All versions**. By default, **GeneSpring** uses the **Latest version only** setting for pathway analysis. You can identify the version of a particular BioCyc pathway (that was previously downloaded from the Agilent Server) in the pathway's Inspector.

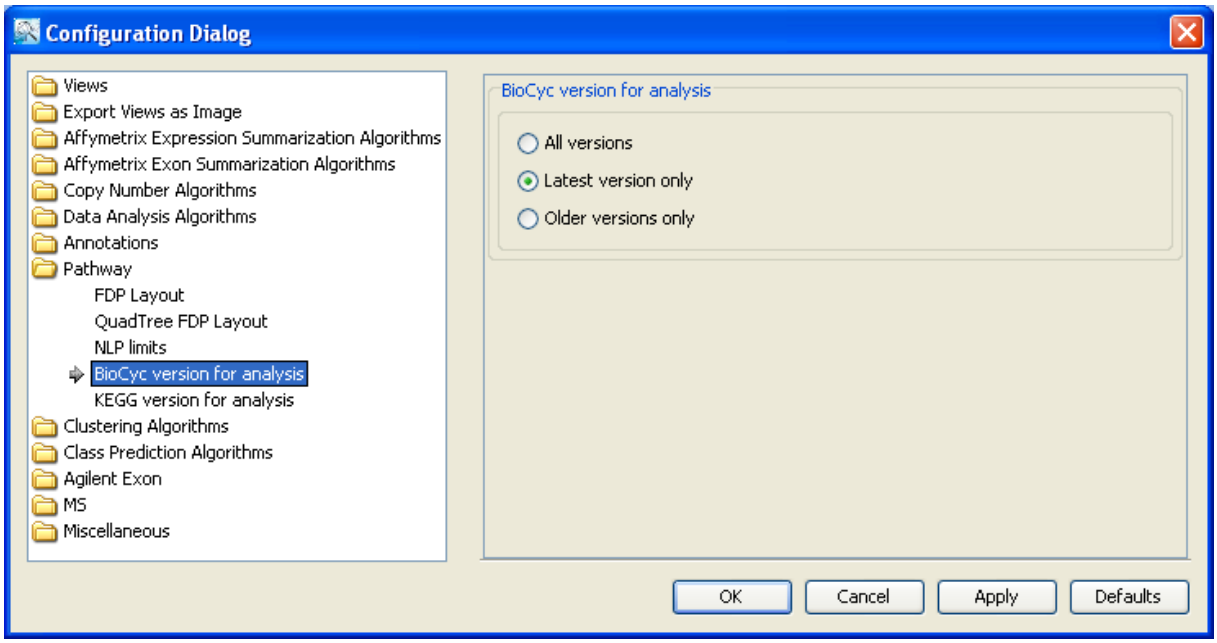


Figure 31.7: Options dialog to select BioCyc pathway version to use for analysis

31.1.3.4 To import pathways from KEGG through the main menu

GeneSpring also has an option to import pathways for a number of organisms from the KEGG Database Collection [37] (`http://www.genome.jp/kegg/pathway.html`) directly from the Agilent Server or an Update File provided by the **GeneSpring** Support Team. Before proceeding with the pathway import, **GeneSpring** validates your KEGG ftp website login details in the background (see KEGG Pathways Licensing on page 862).

The first time you import KEGG pathways, **GeneSpring** also downloads the KEGG Infrastructure Database from the Agilent Server. If you are using an Update File to import KEGG pathways into **GeneSpring** for the first time, select the KEGG Infrastructure Database as well as the organisms for which you want to import pathways in the **Automatic Software Update** dialog before clicking **Update**.

If you do not have a KEGG license, **GeneSpring** provides the option to download 10 demo pathways for seven organisms. Follow the same steps as detailed below. When KEGG license authentication fails, the error message allows you to choose to download the demo pathways instead.

NOTE

KEGG pathways can only be imported into **GeneSpring** from the Agilent server using the procedure detailed below. KEGG pathways downloaded directly from the KEGG Database Collection website to your system cannot be imported into **GeneSpring**.

To import KEGG pathways:

1. Go to **Tools** ⟶ **Import Pathways from KEGG** ⟶ **From Agilent Server**.

   The **Automatic Software Update** dialog opens.

2. Select the organisms for which you want to import available KEGG pathways. Click **Update**.

   An Information dialog informs you of the size requirements of the pathway download. Click **OK**. **GeneSpring** confirms completion of the download with another Information dialog. Click **OK**.

GeneSpring imports all the pathways for the selected organisms. The progress of the import process is indicated by a progress bar. The pathways can now be used for SEA or MOA by selecting KEGG from the **Curated pathways** section in **Step 1: Specifying Experiments, Organisms, and Pathway Sources** of the wizard-driven SEA and MOA workflows.

NOTE If you publish research results in scientific journals based on the use of data from any of these KEGG pathways, reference the results using the KEGG citations provided in the bibliography [37].

KEGG pathway updates

When updates become available for the KEGG pathways, these updates are also provided on the Agilent server after some time. Go to **Tools** ⟶ **Import Pathways from KEGG** ⟶ **From Agilent Server** to check for any available KEGG pathway updates. The import procedure for these updates is the same as described above. These updated versions do not overwrite previously downloaded KEGG pathways but are stored in **GeneSpring** in addition to the earlier versions. If you have downloaded multiple versions, go to **Tools** ⟶ **Options** ⟶ **Pathway** ⟶ **KEGG version for analysis** before starting a SEA or MOA workflow to specify which KEGG pathways you want to use for analysis from one of these three options: **Latest version only**, **Older versions only**, or **All versions**. By default, **GeneSpring** uses the **Latest version only** setting for pathway analysis. You can identify the version of a particular KEGG pathway (that was previously downloaded from the Agilent Server) in the pathway's Inspector.

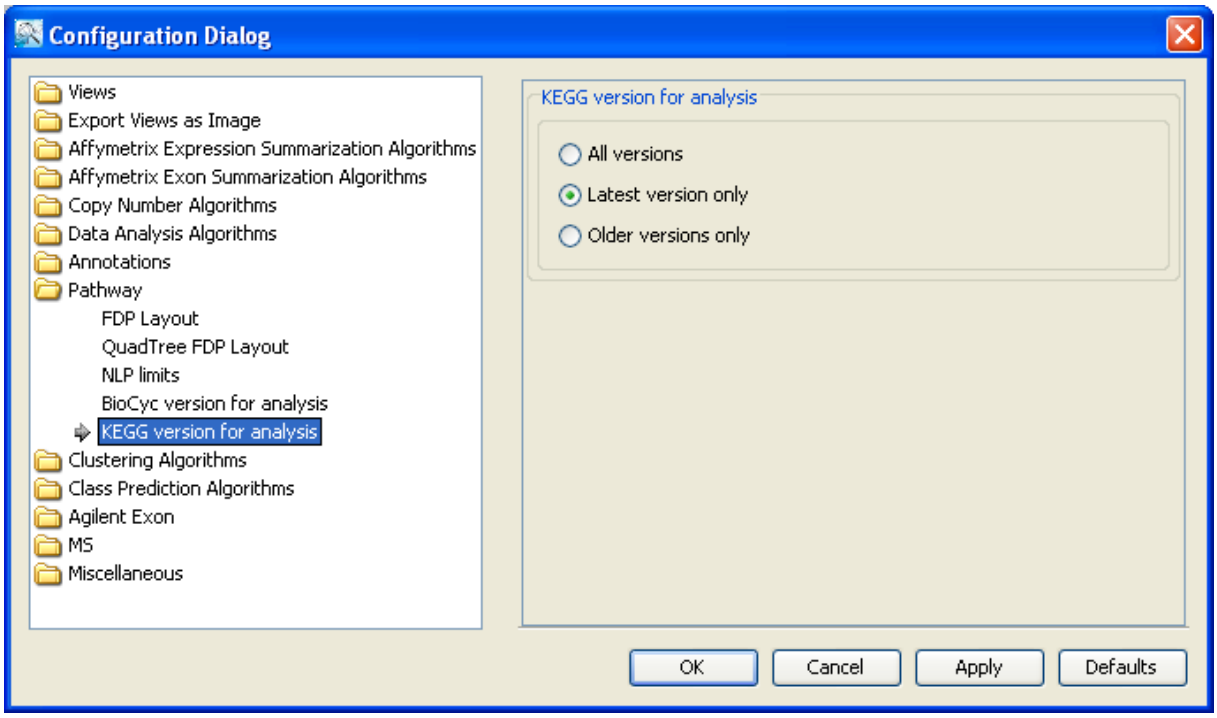


Figure 31.8: Options dialog to select KEGG pathway version to use for analysis

31.1.3.5 To import pathways from file

GeneSpring can import pathways from a file (saved locally on a computer) through **Tools** ⟶ **Import Pathway from File** from the main menu.

Pathway files in the following three file formats can be imported into **GeneSpring**:

- BioPAX (.owl file format)
  - Level 2
  - Level 3
- Text (.txt, .tsv file formats)
- GPML file format

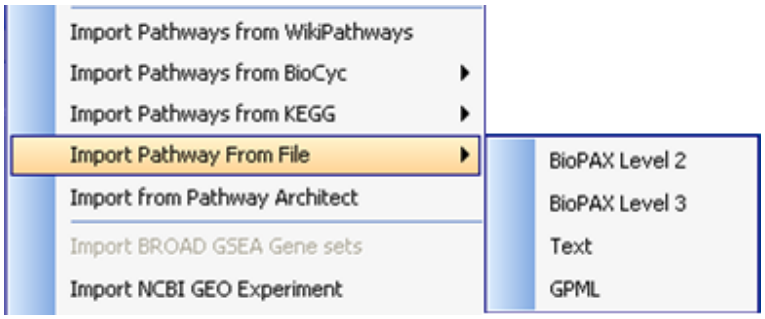


Figure 31.9: Import Pathways From File options

NOTE

When you import pathways in the BioPAX (Level 2) and Text format, **GeneSpring** recognizes information about the entities and relations in the imported pathway and adds any new information to its internal database. These new entities and relations are then available during future NLP Network Discovery or Extract Relations via NLP workflows. This is not the case for pathways imported in the BioPAX (Level 3) and GPML format. In addition, GPML pathways are imported into **GeneSpring** in such a way that they will look exactly the same as in other pathway tools or on the WikiPathways portal. BioPAX pathways (both Level 2 and 3) on the other hand look different compared to other tools because of the way entities and relations are visualized in **GeneSpring** (see Representation of Entities and Relations in Networks and Pathways).

Import BioPAX files

Pathways in the .owl file format (BioPAX Level 2 and Level 3) can be imported using this option. To import BioPAX Level 2 pathways choose **Tools** ⟶ **Import Pathway from File** ⟶ **BioPAX Level 2**. To import BioPAX Level 3 pathways choose **Tools** ⟶ **Import Pathway from File** ⟶ **BioPAX Level 3**.

In the **Import BioPAX Files** window, click **Select BioPAX .owl files...** to browse for and select the BioPAX files that contain the pathways you want to import. Select one or more .owl files and click **Open** to bring the pathways into the **Import BioPAX Files** dialog.

The import wizard automatically parses BioPAX files and displays information about the pathways contained in the selected files in the following columns:

- **Pathway Name:** Name of the imported pathway.
- **Organism:** Pathway organism identified from the file or assigned by you.
- **File Name:** Name of the .owl file containing the pathway.

If the .owl file does not contain information about the organism (**Organism** column reads **None**), you have to assign an organism to each pathway. Select all the pathways for which you want to assign a particular organism and click **Assign Organism...**. Select the appropriate organism from the drop-down list.

For BioPAX Level 2 pathways, the only available organisms from this drop-down list are those for which the Interaction Database was downloaded previously (see To download an organism's Interaction Database on page 890).

- If you are unsure about which organism the pathway should be associated with, do not select the pathway for import in this session. Download the appropriate organism's Interaction Database (if it is available from the Agilent Server) and then import the pathway file again.
- If the **Organism** field contains information but the Interaction Database is not currently downloaded, close this dialog and go to Annotations ⟶ Update Pathway Interactions ⟶ From Agilent Server to download the corresponding database, if the given organism's database is available on the Agilent Server, and then import the pathway files again.

For BioPAX Level 3 pathways, none of these Interaction Database related restrictions apply.

Select the **Pathway provider** from the drop-down list. If the provider for your pathways is not available from the list, select **Custom...** and type the provider name.

If you do not want to import any of the pathways listed in this dialog, select those pathways and click **Remove**.

When you have provided all the necessary organism and pathway provider information and reviewed the list of pathways in the window, click **OK** to import the pathways into **GeneSpring**. Once the import is completed, the pathways are available for search and analysis.

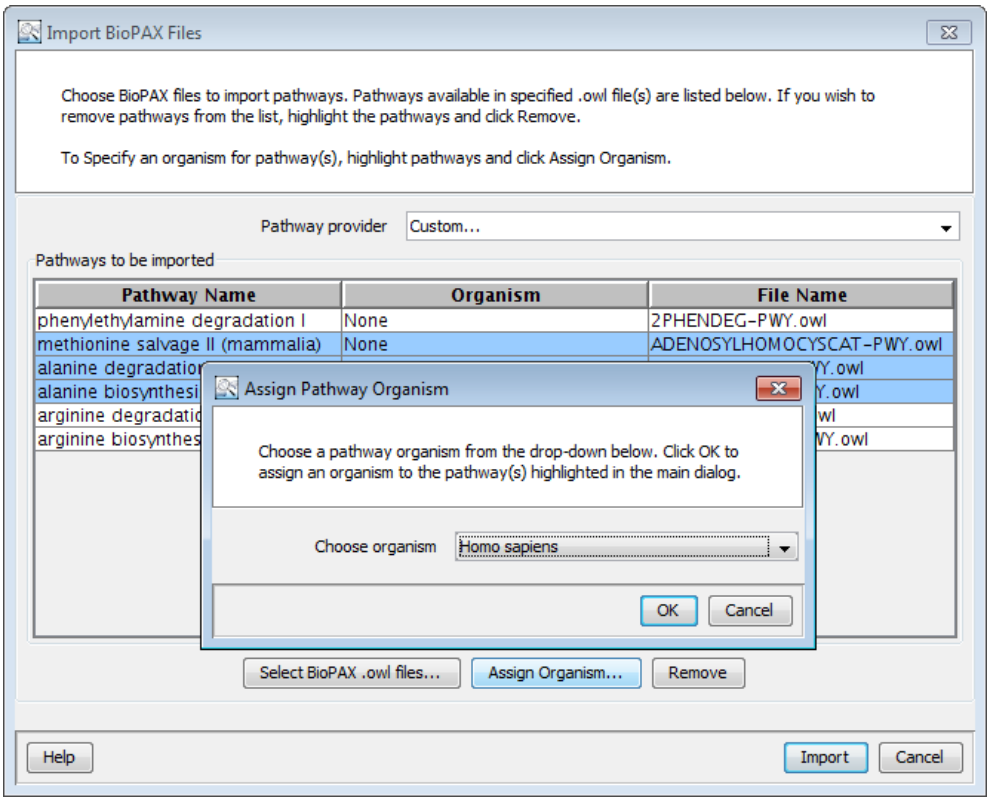


Figure 31.10: Import BioPAX pathways

A number of online resources from where BioPAX pathways are available to download are listed here:

- **Pathway Commons** (http://www.pathwaycommons.org)
- **Reactome** (http://www.reactome.org)
- **BioCyc** (http://biocyc.org)
- **BioModels** (http://www.ebi.ac.uk/biomodels/)
- **Nature/NCI Pathway Interaction Database** (http://pid.nci.nih.gov/)
- **Cancer Cell Map** - Available from the Pathway Commons portal (http://www.pathwaycommons.org/pc/dbSnapshot.do?snapshot_id=8)
- **Netpath - A curated resource of signal transduction pathways in humans** (http://www.netpath.org/)
- **ConsensusPathDB - A database integrating human functional interaction networks** (http://cpdb.molgen.mpg.de/)
- **PANTHER** (http://www.pantherdb.org/pathway/pathwayList.jsp)

Some of the resources are not available free of charge and licenses have to be purchased for acquiring the pathways.

Importing Text files

GeneSpring recognizes a specific format for importing pathways in a Text format and the pathway files have to be in that exact format for import. A single file or multiple files can be imported using this option.

Interaction import format

Interactions can be imported via a tab delimited file (**.txt**, **.tsv**) where each row has three parts:

- **Part I:** In this part you have to specify the overall interaction details in the first seven columns. **GeneSpring** can only process certain values for the information you provide in Columns 2-5. These values are specified in the Interaction Databases section on page 1040.
  - Column 1: **Organism** specified using Taxonomy ID (if the pathway is applicable to all taxa, '0' is used)
  - Column 2: **Interaction Type**
  - Column 3: **Interaction Mechanism**
  - Column 4: **Location**
  - Column 5: **Tissue**
  - Column 6: **Score** is a number between 0 (lowest) and 10 (highest); NLP derived interactions can have a maximum score of 9
  - Column 7: **Degree** is the number of participants in the interaction and has to be an Integer > 1.
- **Part II:** This part specifies the participants. Each participant's specification is featured in five columns. **GeneSpring** can only process certain values for the information you provide in the last four of these columns. These are specified in the Interaction Databases section on page 1040.

  The columns for the i^{th} participant are:
  - column $7+(i-1)*5+1$: **Participant Id**. This can be the urn or object ID (from the Query DB).
  - column $7+(i-1)*5+2$: **Participant Type**. Must be one of
    - Protein
    - Small Molecule
    - Enzyme
    - Complex
    - Process
    - Function
    - Family
  - column $7+(i-1)*5+3$: **Participant Role**
  - column $7+(i-1)*5+4$: **Participant Location**
  - column $7+(i-1)*5+5$: **Participant State**
- **Part III:** This part specifies the reference properties. There is no restriction on the number of columns which can be added in this section. They must take the form "k:v" where k is the reference property name in the database and v is the value associated with this reference property. So this can be "Year:1993", "PMID:23445", "Source:BIND", "SourceID:12434", etc.

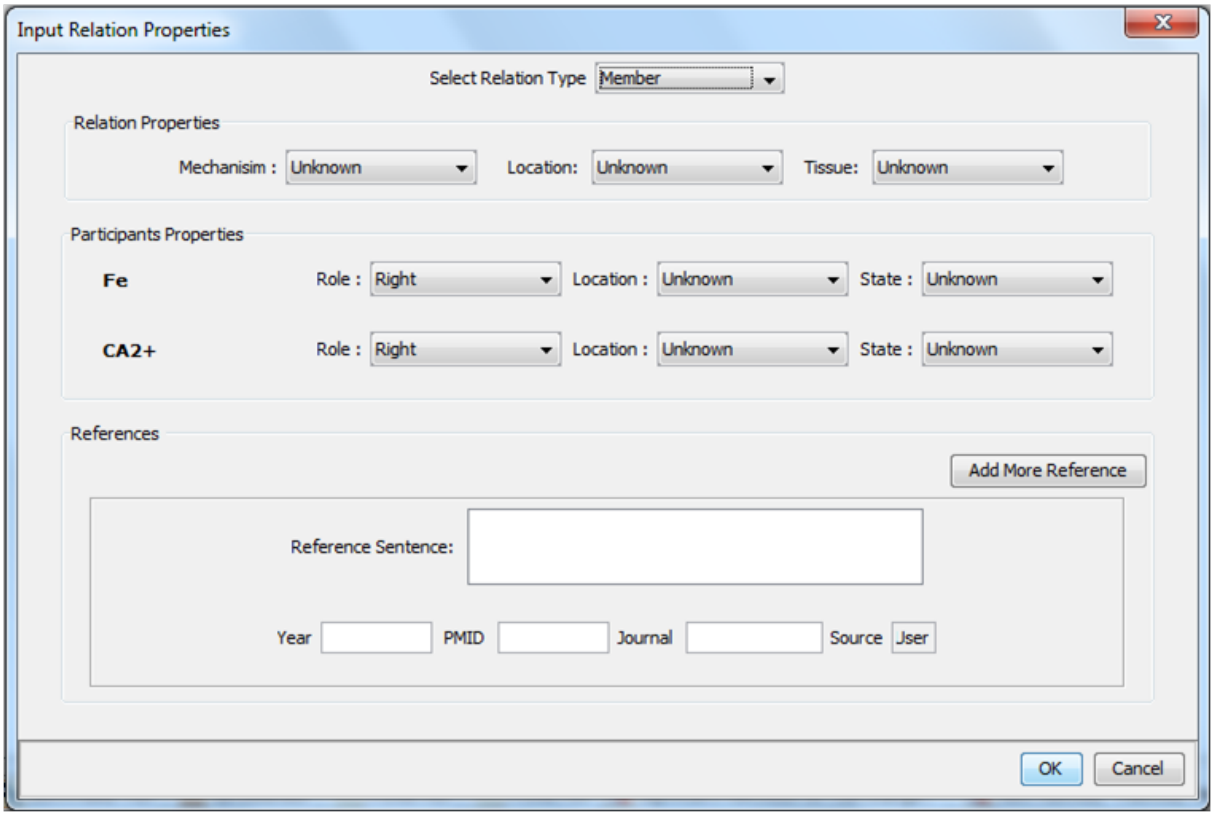


Figure 31.11: Input Relation Properties dialog

For example, a line in the import file can look like this:

```
10090 Transport Unknown Cytoplasm Melanocyte 10 2
Atp7A (EZ ID :11977) Protein (+)Regulator Unknown
Unknown  Cu (CAS ID:133353-46-5) Small Molecule
(+)Target Unknown Unknown PMID:18650808
Sentence:Copper is supplied to melanosomes by
ATP7A, a cohort of which localizes to melanosomes
in a biogenesis of lysosome-related organelles
complex-1 (BLOC-1)-dependent manner
Year:2008 Journal: Nature
```

Import GPML files

A single pathway or multiple pathway files in .gpml format can be imported into **GeneSpring** using this option.

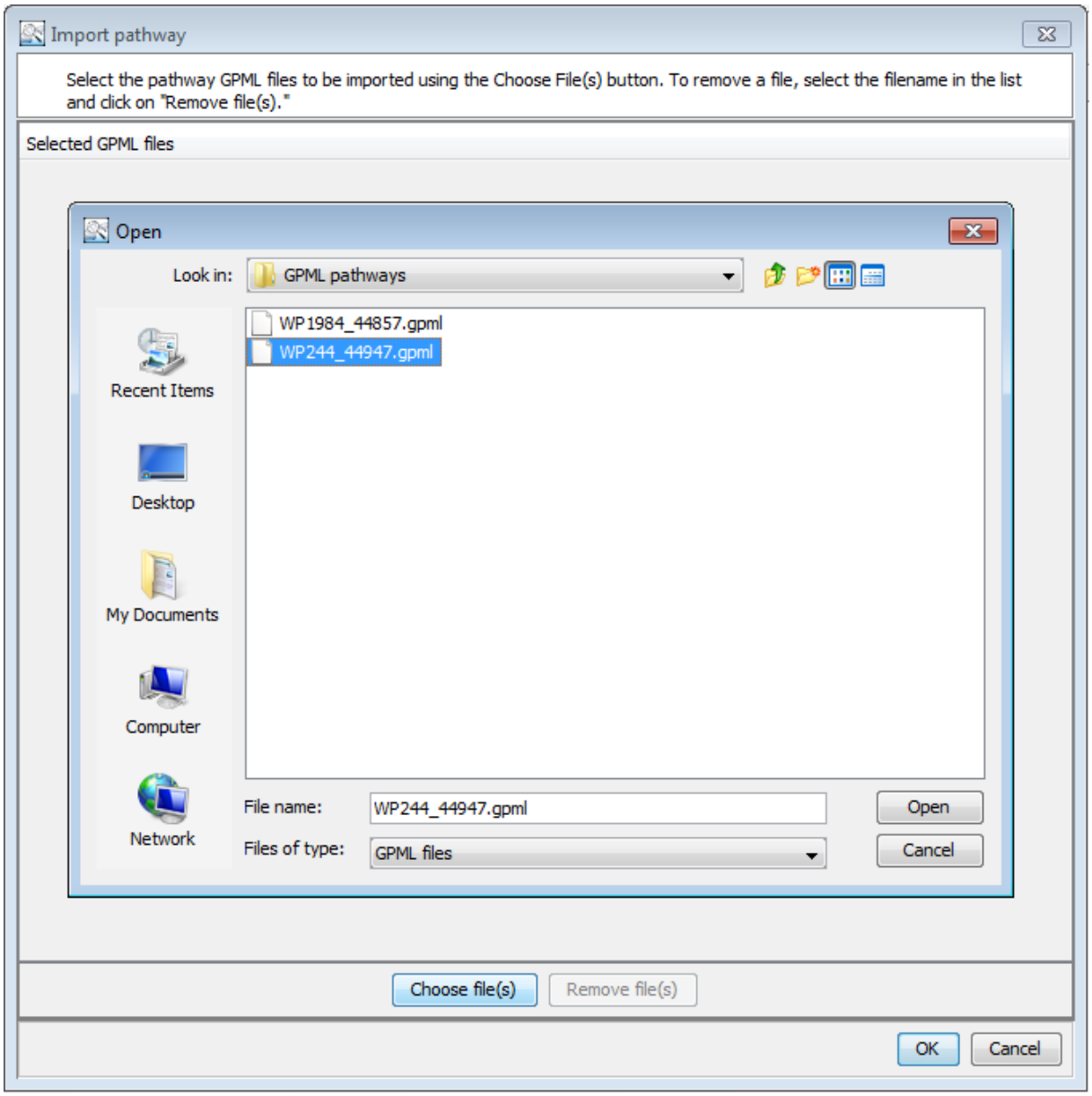


Figure 31.12: Choose pathway files in the GPML format

After selecting the desired files, an organism and the pathway provider has to be specified for the completion of the import.

The organism can be specified (if **GeneSpring** does not automatically identify it) or changed from a drop-down list that contains organism names for which WikiPathways provides downloads. Click the row representing the pathway for which you want to specify an organism to open the **Choose Organism** dialog. Use the **Enter New** option from the **Select Organism** drop-down in this dialog to add a new pathway organism. To specify the organism for multiple files at once, hold **Ctrl** and select all the desired rows and click **Set Organism**.

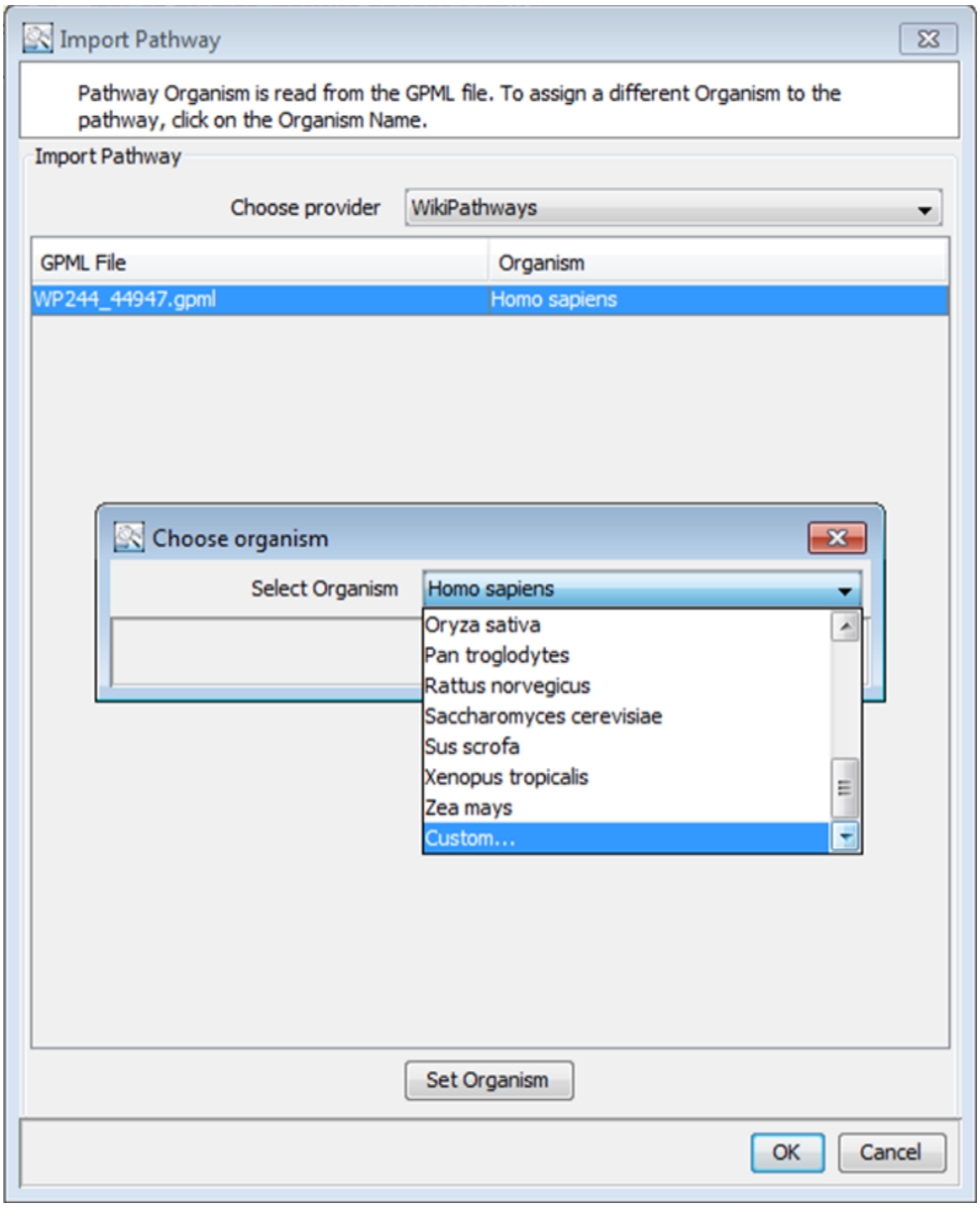


Figure 31.13: Set the organism for the imported pathway

For example, if a GPML file is imported from a custom file or exported from another source, **GeneSpring** is not always able to identify the pathway organism and provider. In this case the **Organism** and **Pathway provider** has to be specified. To specify a provider other than WikiPathways or **GeneSpring**, select **Custom...** from the **Choose provider** drop-down and type the name of the provider. Click **OK** to complete the import. When importing pathways, **GeneSpring** looks for duplicate pathways that already exist in **GeneSpring** on the basis of the file name, the organism, and the provider. In the **Rename To** column of the **Rename Pathways Dialog**, a blue background color indicates a duplicate of a pathway and an orange background color indicates the duplicate of a shared pathway. If duplicates are identified, **GeneSpring** allows you to rename the pathway, overwrite the existing pathway with the same name, or abort the import

of the duplicate pathway. This process is described in more detail in the section on how To import pathways from WikiPathways through the main menu on page 864.

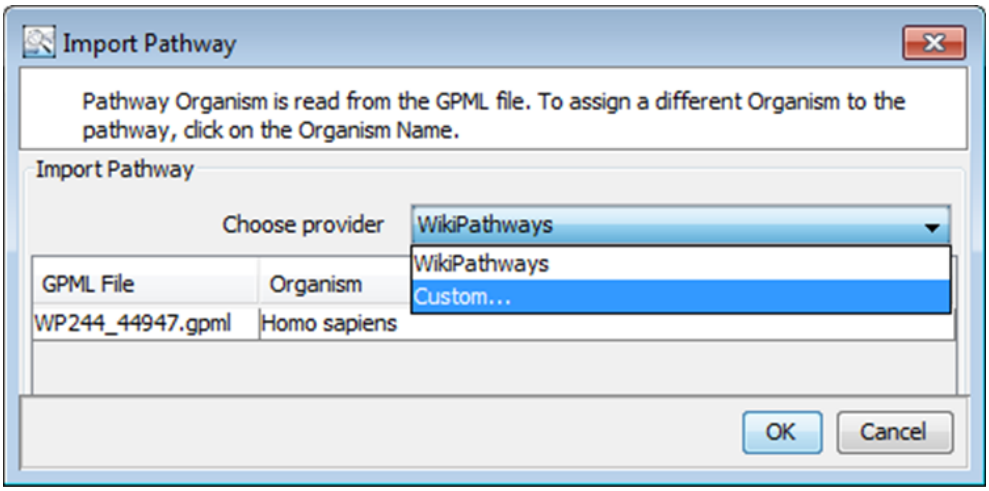


Figure 31.14: Choose the pathway provider for the imported pathway

NOTE

When importing pathways in the GPML format, **GeneSpring** checks the existing pathway database for duplicates based on the exact pathway name and the same pathway organism. You have the option to either rename a duplicate pathway and import it as a new pathway, or replace the existing pathway in **GeneSpring** with the same name. This process is described in To import pathways from WikiPathways through the main menu. However, **GeneSpring** cannot distinguish between different pathway organisms when preparing a file containing multiple pathways for import. Therefore, if the file contains multiple pathways with the same name, but associated with different organisms, only one of them will be imported. To avoid this, import pathways for different organisms separately.

31.1.3.6 To import pathways from Pathway Architect

Pathways in the .bin or .xml format from the Pathway Architect tool can be imported into **GeneSpring** using **Tools** ⟶ **Import from Pathway Architect**. Prior to this import, the HomoloGene Groups, Interaction Database, and BridgeDb [57] database for the organism of interest have to be downloaded (see Supporting Databases for Pathway Analysis on page 887).

31.1.4 Creating and Editing Pathways

In addition to creating interaction networks using one of the three Natural Language Processing (NLP) Networks features, NLP Network Discovery, MeSH Network Builder, and Extract Relations via NLP, you can also create your own

pathways/interaction networks or edit previously imported/created pathways (only applicable to pathways in the Text and BioPAX Level 2 format and pathways imported from Pathway Architect) or interaction networks using the **Create New Pathway** feature in **GeneSpring**.

This section explains how you can create new pathways, edit existing pathways (Text and BioPAX Level 2 format and Pathway Architect pathways only), add entities and interactions, remove unlinked entities, copy and paste selected pathway elements, change the pathway layout and finally save the new pathway or interaction network in **GeneSpring** .

Many of the features explained in this section can also be used in the workflows of the three **NLP Networks** features to edit the resulting interaction network before it is saved as a pathway list.

31.1.4.1 To create a new pathway

- Click the **Create new pathway** icon in the main tool bar.

A new pathway viewer opens. All the required functionality to add entities and relations, change pathway layout, copy and paste selected entities, or to undo or redo a previous action is accessible from the Pathway View Toolbar and is described in the following sections of this part of the manual. Read To save a new/edited pathway for important information about saving and naming your new pathway.

Figure 31.15: Pathway View Toolbar with enabled icons for creating a new pathway

31.1.4.2 To edit a pathway

In **GeneSpring**, you can edit all pathways and interaction networks that were imported from Pathway Architect, in the BioPAX (Level 2 only) and Text formats, or created with one of **GeneSpring**'s NLP Network features (NLP Network Discovery, MeSH Network Builder, and Extract Relations via NLP).

1. Open the pathway/interaction network that you want to edit in The Pathway View.
2. In the Pathway Viewer panel, select all the elements of the pathway/interaction network that you want to edit. This could be the entire pathway, or just a selection of entities and relations.
3. Click the **Copy Selection** icon in the Pathway View Toolbar.
4. Click the **Create new pathway** icon in the main toolbar.
   A new pathway viewer opens.
5. Click the **Paste Selection** icon in the Pathway View Toolbar.
   The selected entities and relations are pasted in the new Pathway View and can now be edited.

6 Make all the desired changes to this new pathway. You can access all the required functionality to add entities and relations, change pathway layout, copy and paste selected entities, or to undo or redo a previous action from the **Pathway View Toolbar**. The corresponding steps are described in the following sections of this part of the manual.

7 Click the **Save Pathway** icon to save the pathway as a pathway list only when you have completed all your changes. Please note that you cannot make any further changes to a pathway, once it has been saved.

A Pathway List Inspector opens. Read To save a new/edited pathway for important information about saving and naming your new pathway.

NOTE The **Create new pathway** icon is not available in the main toolbar of an MOA experiment. To edit a pathway from a pathway list in an MOA experiment, follow steps 1 and 2. Then open one of the input experiments of the MOA experiment and open a new pathway viewer as described in step 3. Follow the remaining steps to complete the new pathway.

31.1.4.3 To add a new entity to the pathway via the Entities Legend

Choose this option to add a new entity to the pathway, if it is likely that the entity is already present in one of the Interaction Databases in **GeneSpring**. This would be the case for any entities for which interactions with other entities have been sufficiently described in PubMed abstracts before the last Interaction Database update. These databases are frequently updated and you are notified on starting up **GeneSpring** if an update is available to download from the Agilent Server.

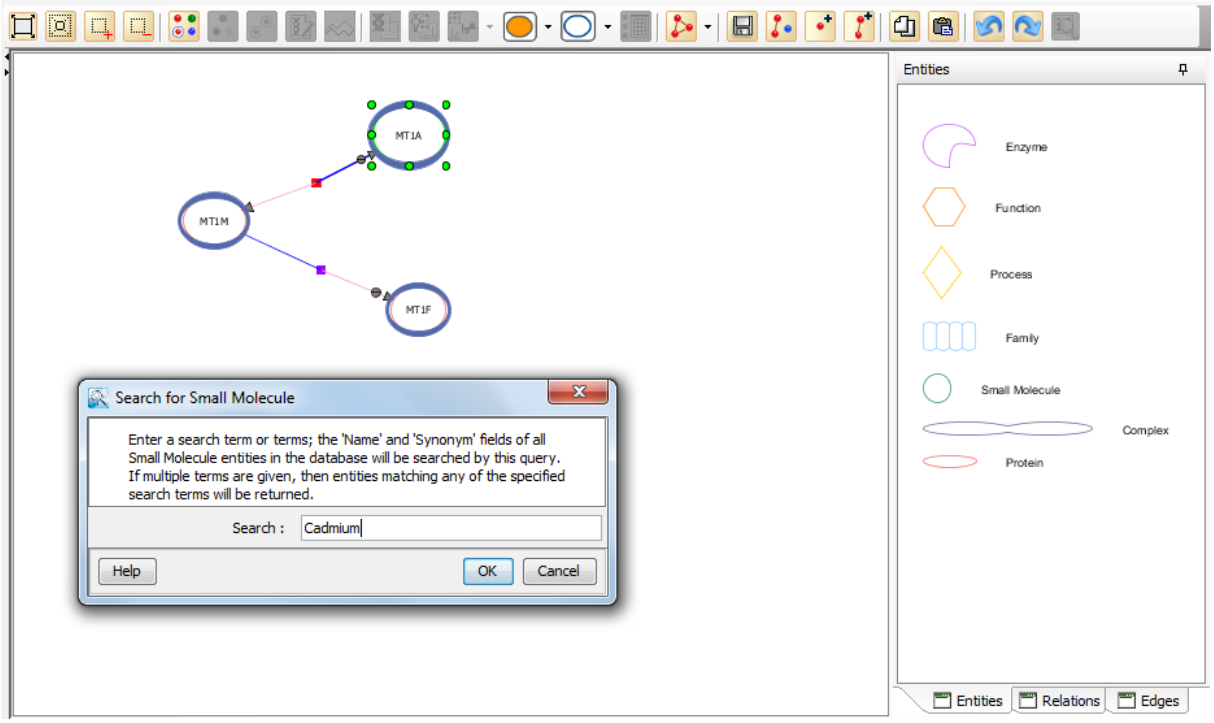


Figure 31.16: Entities Legend visible in the pathway viewer

1 Click or mouse over the **Entities** tab visible at the right edge of the pathway viewer to bring the Entities Legend into view.

2 Drag the symbol for the desired entity type into the pathway viewer.
A **Search for 'Entity Type'** dialog opens.

3 Type the name most commonly used for the desired entity into the **Search** field. Click **OK**.
GeneSpring searches and retrieves matching entities along with their properties from the Interaction Database. Their names, known synonyms, and the number of known connections with other entities are displayed in a **Search Results** dialog.

4 Select the desired entity or multiple entities and click **OK**.

The selected entity/entities are added to the pathway viewer. If **GeneSpring** could not find the desired entity in one of the Interaction Databases, or none of the entities displayed in the Search Result dialog represent the entity you want to add to this new pathway, follow the instructions under To add a new entity to the pathway via the toolbar to add this entity to the pathway.

31.1.4.4 To add a new entity to the pathway via the toolbar

There are two ways to add entities to a pathway, depending on whether they are present in one of the Interaction Databases included in **GeneSpring** or not. Select this option if **GeneSpring** could not identify the desired entity when you were trying To add a new entity to the pathway via the Entities Legend.

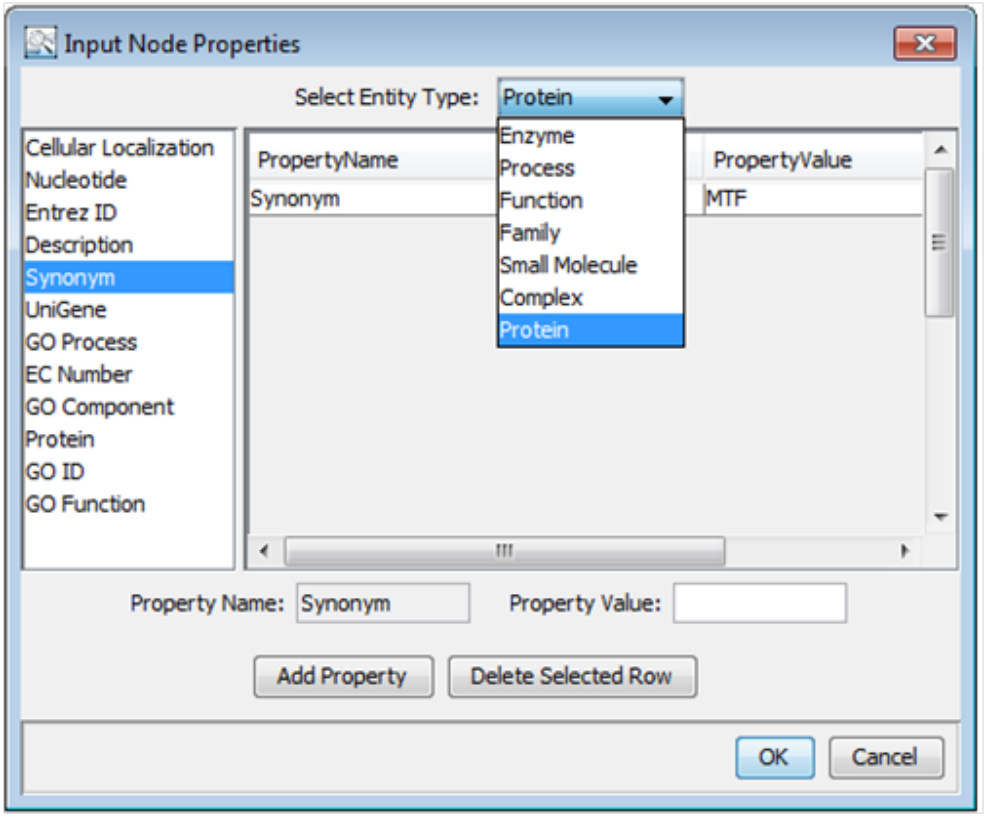


Figure 31.17: Input Node Properties dialog box

1 Click the **Add entity** icon from the Pathway View Toolbar. The Input Node Properties dialog opens.

2 Select the desired **Entity Type** from the drop-down and specify the **Property Name** by clicking on the available options for the selected **Entity Type** that are listed on the left side of the dialog. Then type the **Property Value**. Click **Add Property**.

   Each property is added to the spreadsheet view as a new row. Repeat this step to add as much information about the new entity as is available to you. The **Property Name** 'Synonym' is required for all entity types in order to complete this step. Provide the name that is most commonly used to refer to the entity that you want to add to the pathway, to avoid duplication of entities at a later stage. You can also add multiple values for all Property Names that are not unique annotations (eg. Entrez Gene IDs, or UniGene IDs). For example, for the **Entity Type** 'Process', and the **Property Name** 'Synonym', possible **Property Values** could be translation, protein translation, protein formation, protein synthesis, etc.

3 Click **OK** to add the entity to the pathway viewer.

At this point, **GeneSpring** compares the value you entered for the **Property Name** 'Synonym' with the entities present in the stored Interaction Databases. If the entity is already present in one of the databases, you cannot add it to the pathway using this method. See To add a new entity to the pathway via the Entities Legend for adding pathway entities that already exist in one of the Interaction Databases.

NOTE

New entities added to an interaction network using this method are saved in **GeneSpring** and become available to be retrieved during any of the NLP Networks workflows to create new interaction networks.

31.1.4.5 To add an interaction between two entities in the pathway

After adding desired entities (see To add a new entity to the pathway via the Entities Legend on page 880 and To add a new entity to the pathway via the toolbar on page 881), you can add the interactions between those entities to the pathway.

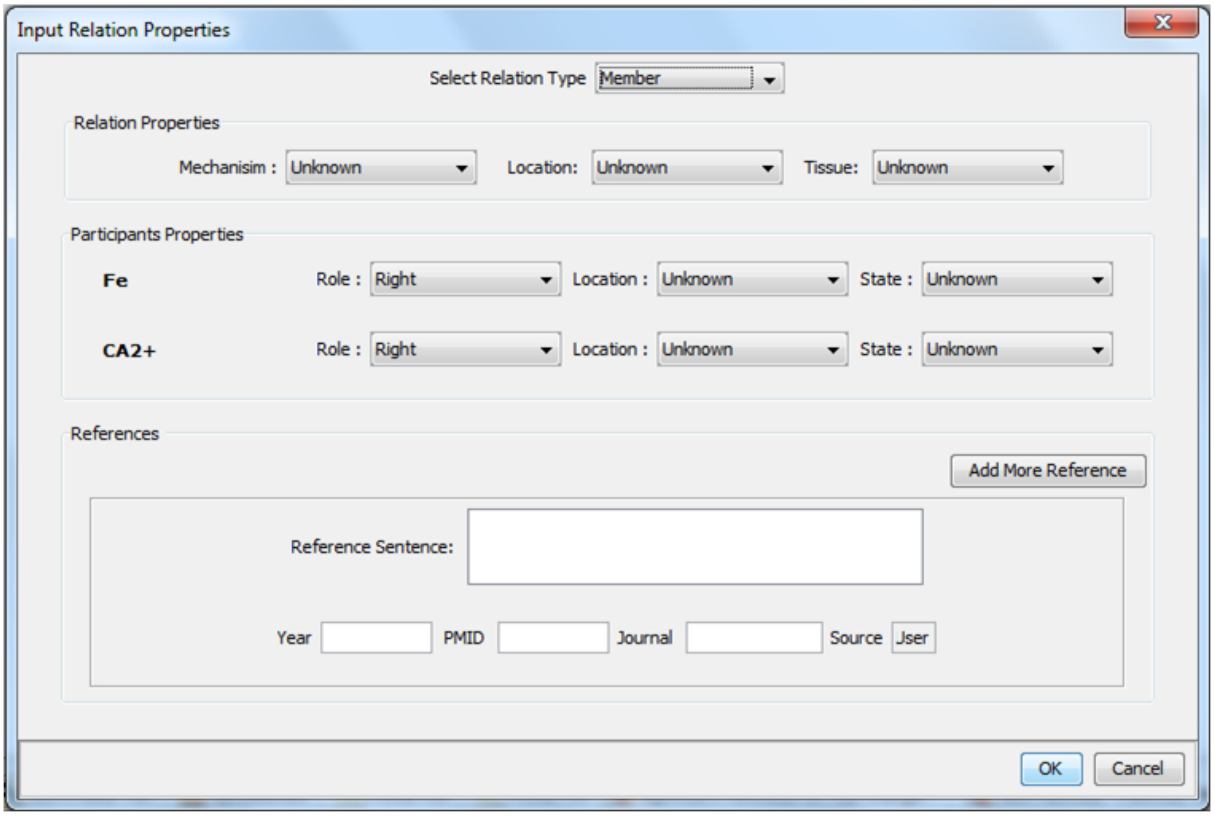


Figure 31.18: Input Relation Properties dialog

There are two ways to add interactions to a pathway, depending on whether the interaction is already present in the Interaction Database or not. To see whether the desired interaction is already present, select two entities and follow the process described in To add a new entity to the pathway via the Entities Legend replacing the **Entities** tab with the **Relations** tab. For new interactions, which are not present in the Interaction Database, use the method described here.

1 Select the two entities from the pathway viewer, for which you want to add a relation, by clicking the first entity and holding **Ctrl** and clicking the second entity.

2 Click the **Add interaction** icon from the Pathway View Toolbar. The Input Relation Properties dialog opens.

3 Select the desired **Relation Type**, **Relation Properties** (depending on the relation type), the **Participants Properties** from the respective drop-down menus, and provide **References** for the new relation. You can find more information about the available relation types in Relations on page 1045. Click **OK**.

The selected relation is added to the new pathway. Repeat steps 1 and 2 to add further new relations to the pathway.

NOTE **GeneSpring**'s Interaction Databases consist of a list of known interactions between entities. Any interactions between newly added entities or new interactions between entities that are not already present in an Interaction Database are saved and become available to be retrieved during any of the NLP Networks workflows to create new interaction networks.

31.1.4.6 To remove unlinked entities from the view

- Click the **Remove unlinked entities from the view** icon from the Pathway View Toolbar.

All entities that are not linked to another entity through a relation are removed from the pathway viewer. This functionality is particularly useful when editing pathways during the NLP Network Discovery workflow, since **GeneSpring** will not always be able to find relations for all the input entities in the stored Interaction Databases.

31.1.4.7 To copy selected entities and relations

Go to To copy selected entities and relations on page 924.

31.1.4.8 To paste selected entities and relations

The paste option is only available in a pathway viewer that was opened to create a new pathway.

- Click the **Paste Selection** icon from the Pathway View Toolbar.

The copied items are pasted into the current Pathway Viewer. If you clicked the **Copy Selection** icon after selecting entities and relations in another pathway (NLP networks, Hand created pathways and BioPax Level 2 format only) or the current pathway viewer, these entities and relations are pasted into the current pathway viewer and can be edited here.

31.1.4.9 To change pathway layout

See To change pathway layout on page 923.

31.1.4.10 To save a new/edited pathway

When you are happy with your new pathway you can save it as a pathway list. Please note, once you save a pathway it cannot be edited anymore.

1. Click the **Save Pathway** icon from the Pathway View Toolbar. Confirm that you want to save the pathway by clicking **Yes** in the Confirm dialog that opens. A Pathway List Inspector opens.
2. Rename the pathway list from the default **Name** 'New Pathway' and edit the **Notes** so that you can easily recognize it when you review your experiment at a later date.
3. Double-click the only row in the spreadsheet view at the bottom of the Pathway List Inspector, which contains the information associated with the new pathway, including the pathway organism, to change the name of the pathway from the default. This is important in order to distinguish it from other new pathways, for example when the pathway (as opposed to the pathway list) is identified as a matched pathway as a result of Single Experiment and Multi-Omic Analysis workflows. A Multiple Objects Inspector opens.
4. Name the pathway and fill in the Notes section. The **Pathway Source** is set to 'Hand created' by default. You can change the default to any of the other options provided in the drop-down. **GeneSpring** will only include this pathway in an SEA or MOA when the corresponding Pathway Source is selected during the respective workflows. Click **OK** to save these details.
5. Click **OK** in the **Pathway List Inspector** to save the edited pathway as a pathway list.

The new pathway list is saved in the Experiment Navigator as part of the active experiment, at the bottom of the **Analysis** folder. By default, any new pathway list is associated with the **All Entities** entity list, and the **All Samples** interpretation of the active experiment.

NOTE Entities and relations that are added via the toolbar are permanently stored in **GeneSpring** and will be queried during the NLP Networks workflows (NLP Network Discovery, MeSH Net- work Builder, and Extract Relations via NLP) to create new interaction networks.

31.1.5 Exporting and Transferring Pathways

Exporting Pathways

Exporting pathway images is a common requirement of the analysis. In **GeneSpring**, three options are provided for exporting pathway images:

Report: You can save an image of the pathway as a report in the experiment navigator or download the image of the pathway as a *.pdf* file as it is displayed in the Pathway Viewer. See How to Create a Report to customize settings for the report.

Image: You can save an image of the pathway as it is displayed in the Pathway Viewer as a *.png* or *.jpg*. The resolution and image size are customizable.

Navigable HTML: You can save a navigable HTML image that can be opened using an internet browser. This format allows you to click on any node or relation and view the associated properties.

Select the **Export As** option from the right-click menu on the Pathway Viewer panel to choose either one of these options.

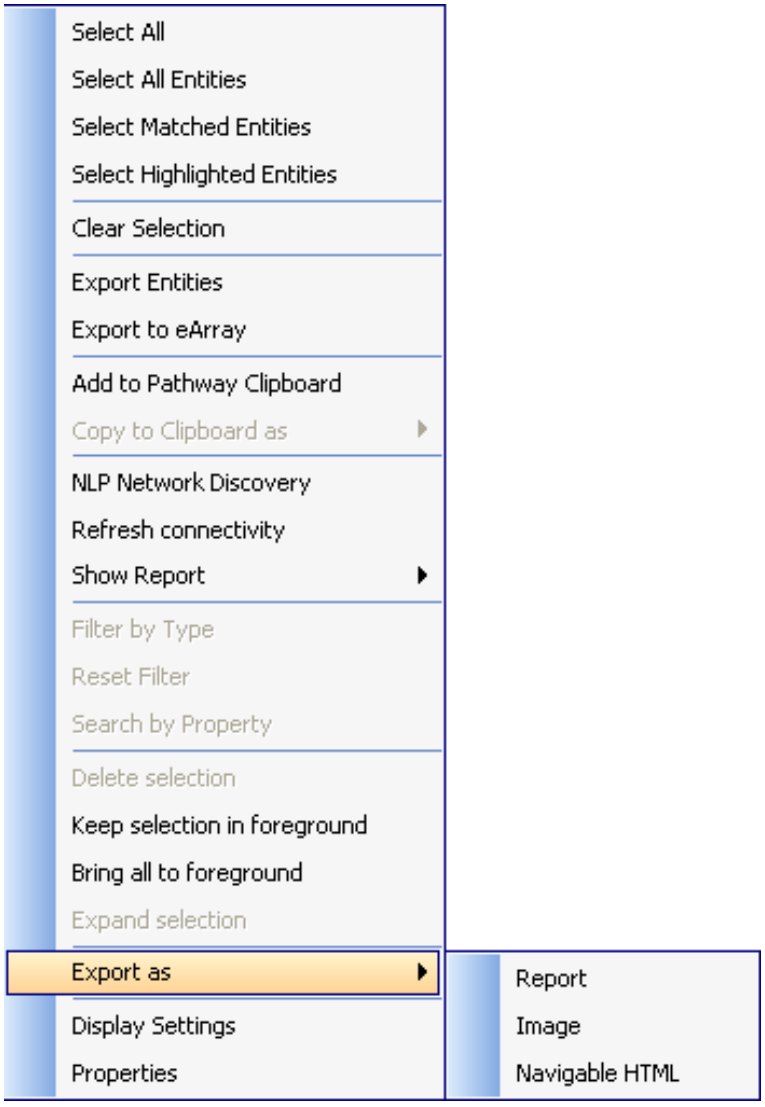


Figure 31.19: Export options for Pathways

For a given pathway the list of participating entities and relations can be exported as follows:

- Right-click a pathway image. Select **Show Report** ⟶ **Entities Table** (or **Selected Entities Table**) to generate a list of the entities. Right-click the spread sheet to export it as a *.txt* file.
- Right-click a pathway image. **Select Show Report** ⟶ **Relations Table** (or **Selected Relations Table**) to generate a list of the relations. Right-click the spreadsheet to export it as a *.txt* file.

You can customize the column selections by right clicking the spreadsheet, selecting Properties and configuring the columns.

Transferring Pathways

Pathways can be transferred from one installation of **GeneSpring** to another as part of an exported project. The exported project contains all data in the selected experiments, including pathways, within a single file, except for data relating to entities and interactions that were added by you when editing or creating new pathways and which were saved in the user space of the Interaction Database. These entities and interactions will still appear in a pathway when it is opened in the pathway view in the other installation but will no longer be associated with their original properties as specified by you and will be indicated as obsolete when updating pathways as this information is not contained in the system space of the Interaction Database.

To export the current project, select **File** ⟶ **Export Project** from the main menu. The project is exported as a *.tar* file, which can be imported into another installation of **GeneSpring** by selecting **File** ⟶ **Import Project** from the main menu. Importing the project will also import all pathways that were saved within the experiments. If the other installation has a more up-to-date version of relevant pathway Interaction Databases, **GeneSpring** provides an option to update the entities and relations in these pathways accordingly (see To download an organism's Interaction Database on page 890).

NOTE

You require a valid end user license to access KEGG pathways in **GeneSpring**. If an imported project has KEGG pathways, you cannot open these pathways until **GeneSpring** has validated the required KEGG end user license. A working internet connection is required to validate the KEGG end user license (see KEGG Pathways Licensing on page 862).

31.1.6 Supporting Databases for Pathway Analysis

Pathway Analysis in **GeneSpring** is supported by organism specific Interaction Databases, BridgeDb databases and HomoloGene annotations.

The Interaction Database specific to the pathway organism has to be downloaded from the Agilent Server to work with pathways that are not in the KEGG, GPML or BioPAX Level 3 format, for instance, to create new pathways, to import pathways in the BioPAX Level 2 or Text format into **GeneSpring**, or to perform Single Experiment Analysis (SEA) or Multi-Omic Analysis (MOA) using these pathways.

Pathways acquired from different sources may refer to the same entity using synonymous names and/or identifiers from different biological databases. Organism specific BridgeDb and ID Mapper databases and a common metabolite ID Mapper database provide the mapping for the same entity across different biological databases. Refer to the BridgeDb - ID Mapping section of this manual for more details on identifier mapping in Pathway Analysis.

In **GeneSpring**, it is possible to overlay experimental data on pathways from organisms different than the experiment organism. This is useful, for instance, in cases where pathways specific to the organism of interest are sparse or not readily available, but pathways are available for closely related well-studied organisms. Comparison of genes across organisms is made possible by HomoloGene annotations from the NCBI database to identify homologous genes across the different organisms under study. Refer to the HomoloGene Translation for Gene/Protein Identifiers section of this manual for more details on HomoloGene annotation mapping in Pathway Analysis.

Details on how to download each of these supporting databases into **GeneSpring** and working with organisms for which these databases are not available are also outlined in this section.

31.1.6.1 To download HomoloGene groups for gene / protein identifiers

To carry out pathway analysis across two or more organisms of interest, including organisms other than your experiment's organism, all the participating organisms' HomoloGene groups have to be downloaded using the Annotations Manager in **GeneSpring**. This information is required for HomoloGene Translation for Gene/Protein Identifiers particularly during SEA and MOA.

Select **Annotations** $\longrightarrow$ **Annotations Manager** $\longrightarrow$ **List** from the main menu. All the organisms for which you can download annotations, including HomoloGene groups, from the Agilent Server are listed on the left side of the dialog. Select only the **HomoloGene Groups (Homology)** option for the desired organisms. Click **Update**.

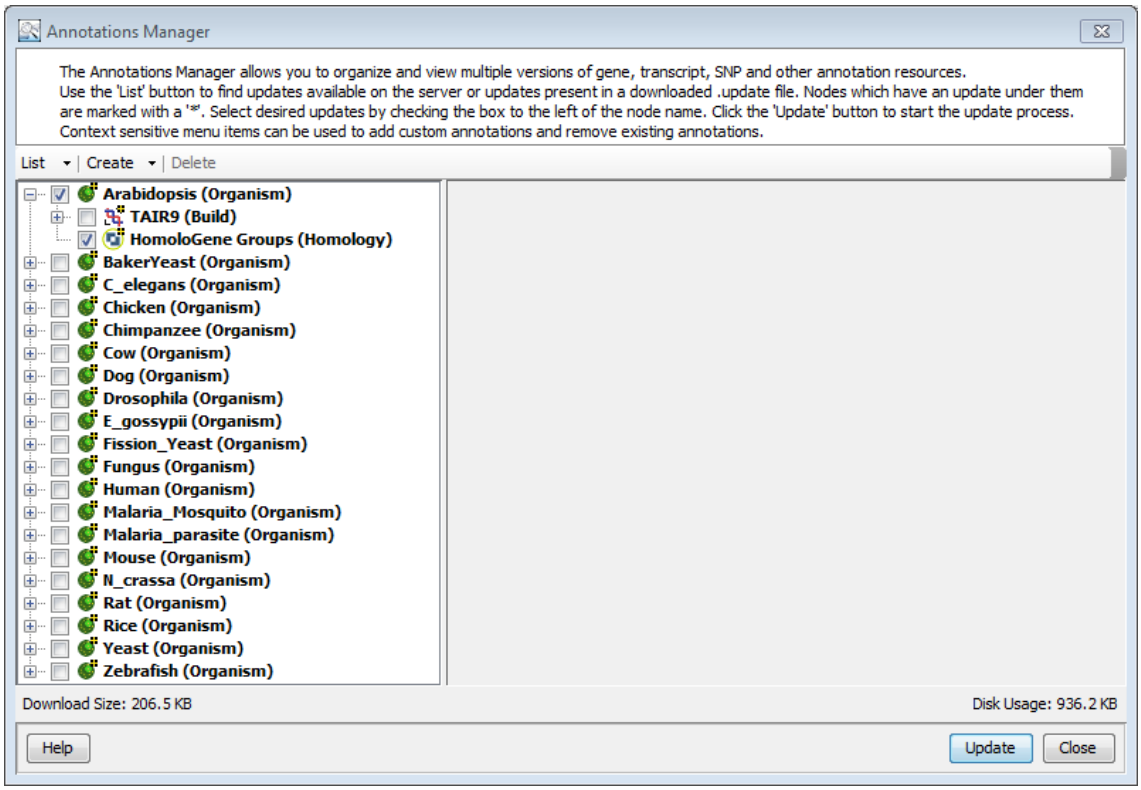


Figure 31.20: Annotations available from the Annotations Manager

When the download is complete, only the organisms for which you downloaded HomoloGene groups are listed. Click **Close**.

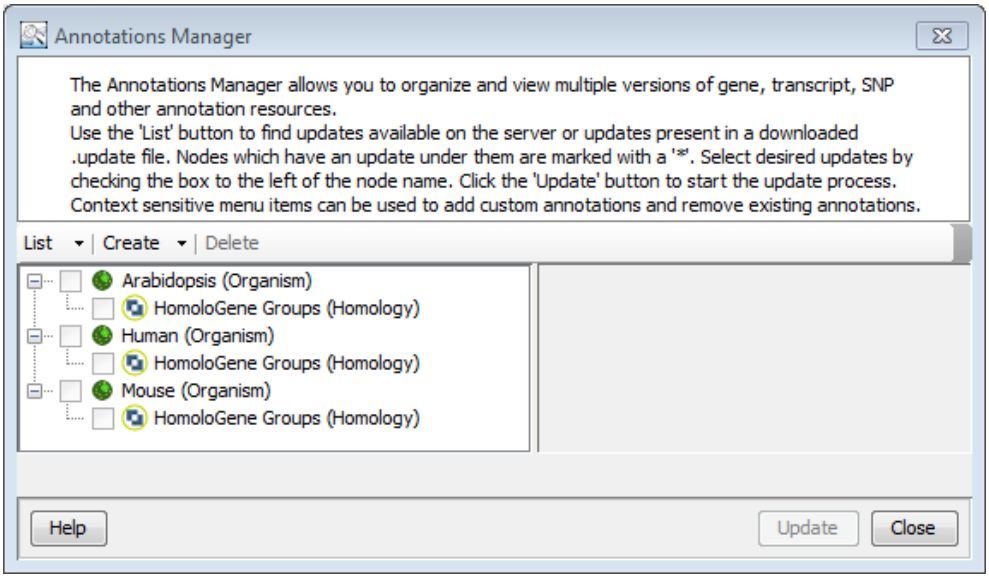


Figure 31.21: Downloaded annotations as shown in the Annotations Manager

31.1.6.2 To download an organism's Interaction Database

Before you can use any of the pre-installed 21 human pathways, or import new pathways in the BioPAX Level 2 or Text format (see Importing Pathways into **GeneSpring** on page 863), you have to download the corresponding organism's Interaction Database from the Agilent Server (if you are connected to the internet) or a **GeneSpring** update file available from **GeneSpring** customer support.

In the MOA and SEA workflow, you are notified if the required Interaction Databases have not been downloaded yet for the chosen organisms if you have selected any of the pathway sources that may contain pathways in the BioPAX Level 2 or Text format or pathways that you created in **GeneSpring**. In this case you have to cancel the current workflow, follow the instructions provided here to update the Interaction Databases and restart the MOA or SEA workflow once the update is complete.

Go to **Annotations** $\longrightarrow$ **Update Pathway Interactions** and select the respective option to download the required databases. The first time you access this feature **GeneSpring** prompts you to download the Pathway Infrastructure Database, which is required regardless of which organism you are working with. Click **Yes**.

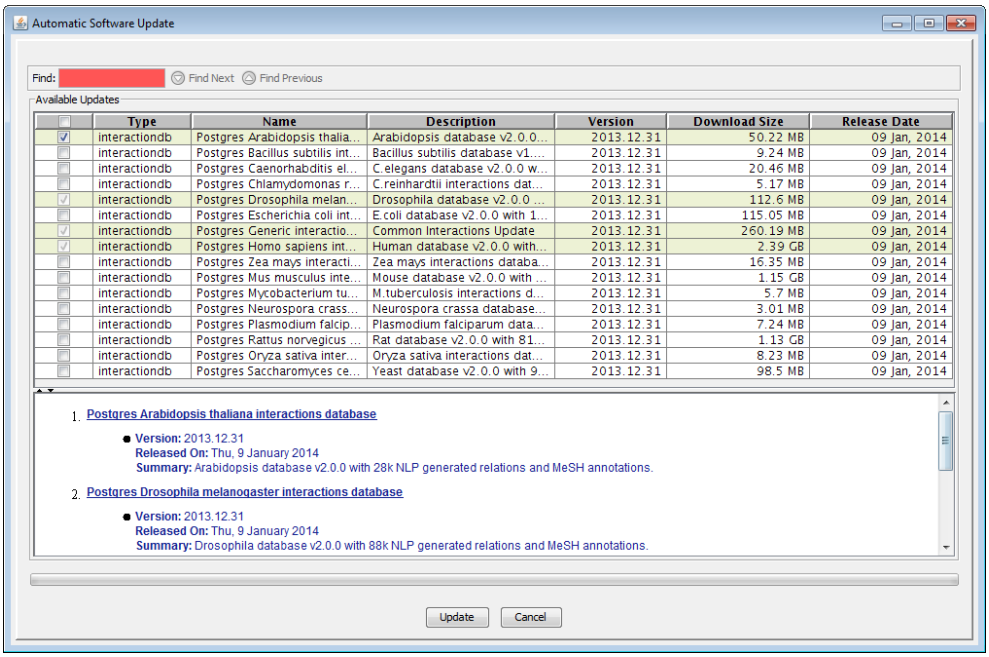


Figure 31.22: Download Pathway Infrastructure Database

After the Pathway Infrastructure Database has downloaded, all the organisms for which you can download an Interaction Database are listed in the Automatic Software Update dialog. Select the databases for the required organisms. Click **Update**.

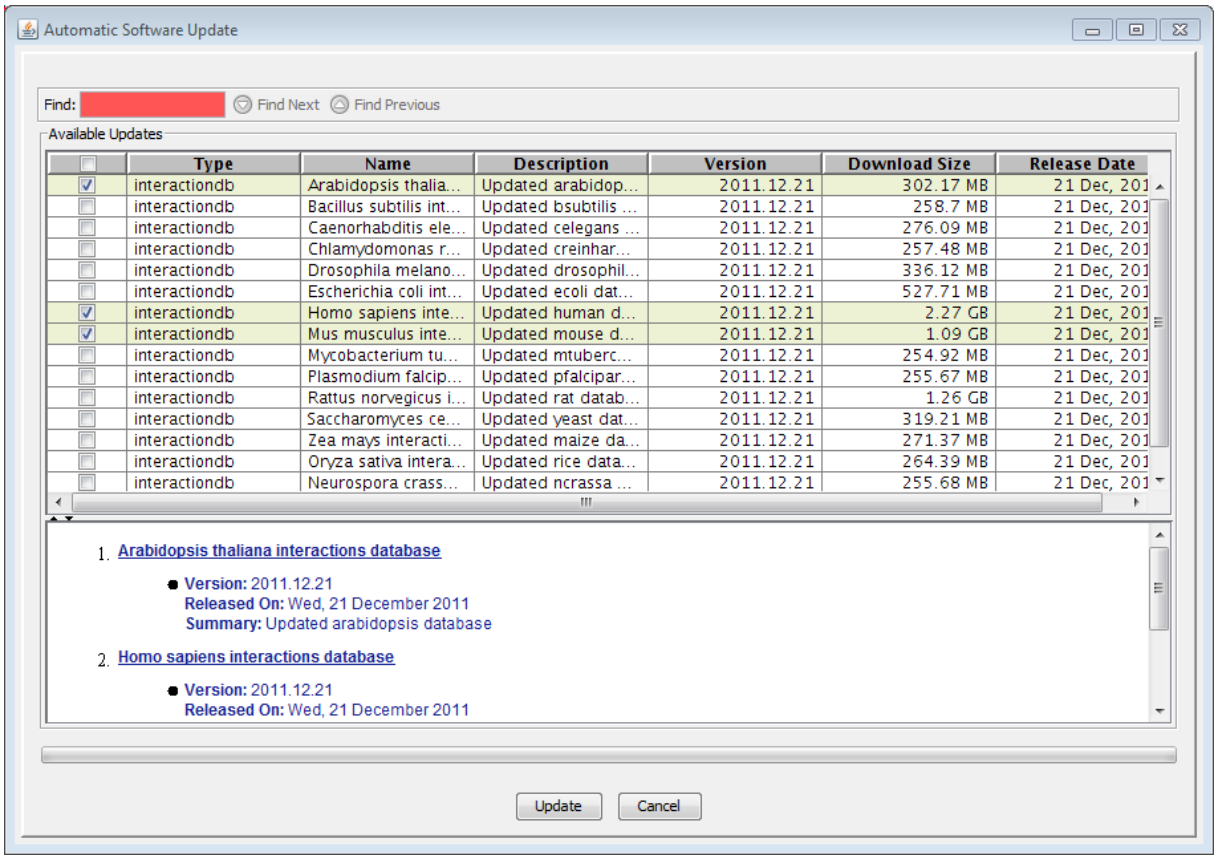


Figure 31.23: Automatic Software Update dialog for Interaction Databases

You can contact Agilent to inquire about building a new pathway interaction database for an organism that does not appear on the list in the Automatic Software Update dialog.

The **GeneSpring** Interaction Database is frequently updated with relations derived from the latest PubMed abstracts. When updates are available, **GeneSpring** notifies you with a message the next time you launch the application. For more information on the contents of the Interaction Databases go to Interaction Databases on page 1040.

Before updating Interaction Databases please ensure that you have enough disk space in your **GeneSpring** installation folder (for example the Human Interaction Database requires around 6GB of disk space). **GeneSpring** calculates and displays the required space for the selected organisms before you start the download. If you do not have enough disk space in the **GeneSpring** installation folder you can change the repository path as described in How to change the repository path section on page 68.

Once the updates are complete, go to **Annotations** $\longrightarrow$ **Pathway Database Statistics** to confirm that the pathway databases of interest are now listed in the left panel, with the associated statistics on the right. Once you have downloaded an organism's Interaction Database the organism is removed from the Automatic Software Update dialog and only reappears when an update becomes available.

Update existing pathways

After updating an Interaction Database, **GeneSpring** indicates if there are existing pathways that are based on this Interaction Database in your **GeneSpring** installation for which entities and/or relations have become obsolete or were updated as a result of the database update in the **Update Pathways** dialog (see Figure 31.24).

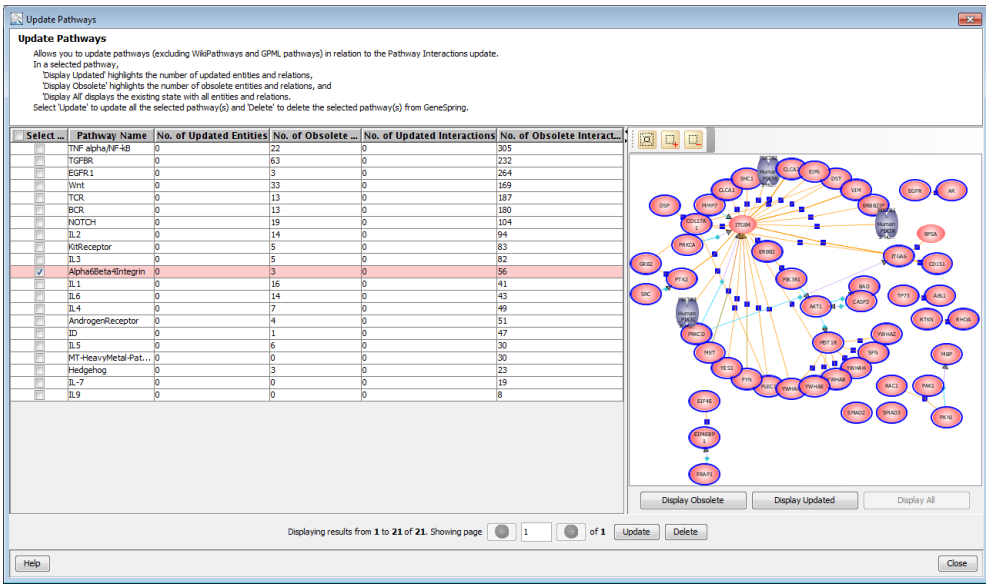


Figure 31.24: Update Pathways dialog

Click on each of the listed pathway names in the panel on the left to view the respective pathways in the pathway viewer on the right. Click **Display Obsolete** to see which entities and relations are no longer present in the new Interaction Database and have therefore become obsolete (see Figure 31.25). Click **Display Updated** to see which entities and relations were updated in the new Interaction Database. Click **Display All** to see the all the entities of the old version of the pathway.

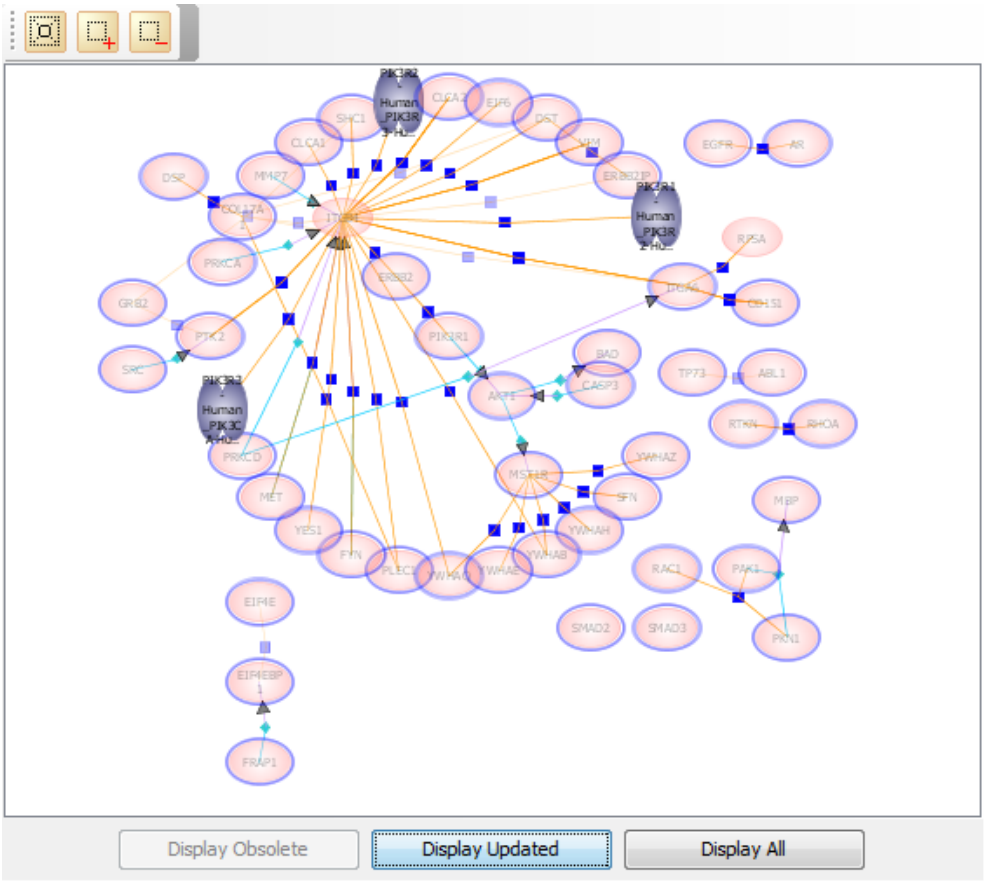


Figure 31.25: View obsolete entities and relations

You can choose to either update or delete pathways in the **Update Pathways** dialog by selecting the corresponding pathways in the panel on the left. If you do not want to decide at this point in time whether to update or delete any of the listed pathways, you can return to the **Update Pathways** dialog via **Annotations** ⟶ **Update Pathways** in the main menu.

WorkGroupPlus Mode

In **WorkGroupPlus** mode, the **Annotations** ⟶ **Update Pathway Interactions** menu item is disabled. Contact your Systems Administrator to update the Pathway Interactions Server with the appropriate organism databases. The Pathway Interactions Server is a centralized server which stores the organism databases.

Only an administrator can update organism databases via a web page. Once the administrator has updated the required databases, all **GeneSpring** clients can use all the pathway analysis features described in this chapter. Note that entities, relationships and their properties in the Pathway Interaction Server do not have any user access controls; all users can query these and write back new entities and relationships. Pathways derived from these entities and relationships however are governed by the usual **WorkGroupPlus** user access rules. A key feature of the Pathway Interactions Server is the speed with which pathway analysis queries are processed compared to the desktop version.

31.1.6.3 To download an organism's BridgeDb or ID Mapper database

Your **GeneSpring** installation includes the metabolite ID Mapper database, which is required to map small molecules regardless of the chosen organism, and the human BridgeDb database. To download BridgeDb or custom ID Mapper databases for other organisms, click **Annotations** ⟶ **Update BridgeDb** and follow the applicable workflow to download the required databases from the Agilent Server (an active internet connection is required to access the server) or a **GeneSpring** update file available from **GeneSpring** customer support. The organisms for which BridgeDb or ID Mapper databases are available for download from the Agilent Server are listed in the Automatic Software Update dialog. You are notified if updates are available for any of the databases you have previously downloaded.

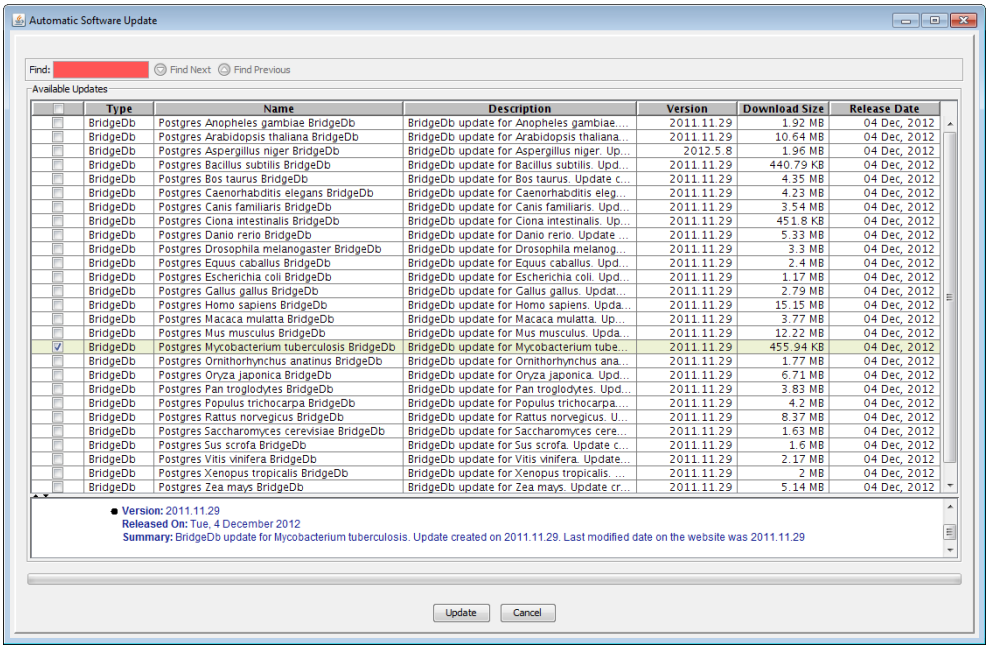


Figure 31.26: Automatic Software Update dialog for BridgeDb and ID Mapper databases

Select the databases for the required organisms.

In the MOA and SEA workflow, you are notified with a **Confirm** dialog if BridgeDb or an ID Mapper has not been downloaded yet for the chosen organisms. Selecting **Yes** in this **Confirm** dialog automatically downloads the chosen organisms' databases in the workflow itself.

See BridgeDb - ID Mapping for a description of how the BridgeDb algorithm is implemented in **GeneSpring**.

31.1.6.4 Supported organisms and annotations

Each of the supporting databases described in the earlier sections (HomoloGene annotations, Interaction Databases, BridgeDb, and ID Mapper databases) are available for specific organisms. As a result pathway analysis in **GeneSpring** is also limited by this list of organisms.

To create new pathways, import pathways in the BioPAX Level 2 or Text format and edit or view pathways for organisms other than the ones which are already

available in **GeneSpring**, you can create a new organism via **Annotations** ⟶ **Create Pathway Organism**. Provide the exact scientific name, common name and taxonomy identifier in the dialog that opens. Although **GeneSpring** does not perform correctness checks on these terms, it is useful to provide exact and valid names for smooth functioning later. For example, use "9606" as taxonomy identifier, "Homo sapiens" (note the space between the two words and the small letter which begins the second word) as scientific name, and "Human" as common name. Valid taxonomy ids can be found at the NCBI Taxonomy site (`http://www.ncbi.nlm.nih.gov/taxonomy`).

NOTE

Any analysis (SEA or MOA) and data overlay on these types of pathways is only possible if the corresponding Interaction Database (see To download an organism's Interaction Database) and BridgeDb database (see To download an organism's BridgeDb or ID Mapper database) for this organism are available.

In **WorkGroupPlus** mode, only an administrator can create a new organism; this can be done via the **GeneSpring** client with **Annotations** ⟶ **Create Pathway Organism**.

31.2 Pathway Analysis in GeneSpring

This section provides an overview of how to use the Single Experiment Analysis (SEA) workflow (previously Find Significant Pathways) and the new Multi-Omic Analysis (MOA) workflow to find pathways containing entities from one or two experiments of the same or different experiment type, respectively. MOA is a new feature in **GeneSpring** that allows you to compare experimental results across experiment types and organisms by identifying pathways that contain one or more entities that are also present in your experiment. In the resulting new MOA experiment or pathway list (resulting from SEA) you can then isolate significant pathways by determining a p-value cut-off and the minimum number of matched entities.

It also provides additional information on MOA, as well as the prerequisites for performing SEA or MOA.

31.2.1 What is Multi-Omic Pathway Analysis (MOA)?

Multi-omic pathway analysis is a new feature in **GeneSpring** that provides a means for pathway analysis at metabolite, transcript, protein, and peptide levels. This new analysis and interpretation option, called Multi-Omic Analysis (MOA), allows you to integrate your results across differing experiment types and multiple organisms and to visualize experimental data overlaid on curated pathways or literature derived networks. For example, you can now identify significant pathways for data from any combination of genomics, transcriptomics, proteomics, and metabolomics experiments, or you could compare results in a model organism that is different from your experiment organism.

After identifying a list of entities that is of particular interest to your research, **GeneSpring** has so far allowed you to create interaction networks from these entities and look for pathways that are significant in your research context. In this latest version of **GeneSpring** we have integrated support for the curated/canonical pathways from the WikiPathways portal, BioCyc genome databases, KEGG database as well as pathways from other sources in the .gpml format. The curated/canonical pathway view is a powerful way to view biological pathways that biologists are typically accustomed to viewing and complements the literature derived network views derived from Natural Language Processing (NLP) or Medical Subject Headings (MeSH terms).

A number of different kinds of analysis become possible as a result, for example:

- Exploring the 'neighborhood' around a metabolite of interest in a pathway provides information about which enzymes interact with the metabolite and which other metabolites are involved in those interactions. These interacting metabolites and enzymes (and the genes that encode those enzymes) are likely to be ideal candidates for the next set of ''omics" experiments. The interpretation of their measurement profiles in the context of the pathway(s) can help one synthesize a hypothesis for the observed behavioral pattern of a system, for example a cell, tissue or biofluid.
- Transcript, protein, peptide and/or metabolite levels are used to identify significant pathways by applying statistical tests. Understanding of these pathways is enhanced by overlaying the data observed in a heatmap table

onto nodes on a selected pathway diagram that highlights the elements of the pathway of interest.

- Pathway information enhances preliminary compound identification in MS as the formulas corresponding to metabolites found in matched pathways can be selected and exported to create a new database. This database can subsequently be queried in a "targeted" fashion against the acquired MS data to find metabolites from the proposed pathway.

NOTE An alternative way to perform Multi-Omic Pathway Analysis is to create an MOA experiment using the **New MOA experiment** icon in the main toolbar, select the two desired experiments and selecting Pathway Analysis from the Workflow Browser.

31.2.2 What is Single Experiment Analysis (SEA)?

Single Experiment Analysis (SEA) replaces the Find Significant Pathways (FSP) workflow step and is enhanced with some of the features that were introduced in this version of **GeneSpring** to enable Multi-Omic Analysis (MOA). As the name implies, the SEA only identifies matching pathways for one experiment, compared to two experiments in an MOA. But unlike in the previous FSP workflow, you can now choose a differing organism for pathway analysis from the organism associated with your chosen experiment. In addition, the curated pathways options (for example WikiPathways, BioCyC pathways, KEGG pathways if you have a valid end user license, or BioPAX pathways) as well as NLP and MeSH created pathways can also be individually selected as sources for pathway analysis.

31.2.3 Prerequisites for Multi-Omic Pathway Analysis

This section describes the steps that are necessary before starting the wizard-driven MOA and SEA workflows. This includes creating interpretations to group large numbers of samples into experimental conditions, importing pathways into **GeneSpring** and adding experiments, if you want to want to use them in an MOA workflow.

31.2.3.1 To create an Interpretation

You can overlay experimental data from the experiment in the pathway viewer in the form of quilt plots and heatstrips, either for all samples individually, or grouped by experimental conditions in the form of an interpretation. The **All Samples** interpretation is always available by default, as it is created automatically when samples are imported into **GeneSpring**. Interpretations based on experimental conditions have to be created before starting either the MOA or SEA workflow. Follow the instructions in the Create Interpretation section on page 240 to create a new interpretation.

If an experiment has a particularly high number of samples (more than ten samples), grouping samples allows you to better visualize experimental data in the pathway viewer. However, non-averaged interpretations can only be selected if there is only one sample assigned to each condition.

31.2.3.2 To import known pathways into GeneSpring

The SEA and MOA workflow identifies matching pathways only from a list of pathways that were created in **GeneSpring** or previously imported into **GeneSpring** from sources like the WikiPathway portal or from file in BioPAX, GPML, or Text formats. You can find all the necessary information on the prerequisites for working with pathways in **GeneSpring**, how to import known pathways, compatible pathway formats, and a list of pathway sources in the dedicated sections on Supporting Databases for Pathway Analysis on page 887 and Importing Pathways into **GeneSpring** on page 863.

31.2.3.3 To add an existing experiment to a project

The experiments that you would like to use as input experiments for MOA have to be in the same project from which you launch the MOA workflow. If you would like to perform MOA with an experiment that is currently in a different project, you can add it to the open project by clicking the **Add Experiment** icon, or via the main menu item **Project** $\longrightarrow$ **Add Experiment**.

31.2.4 Starting and Navigating the SEA and MOA Workflows

This section describes how to start and navigate the wizard-driven SEA and MOA workflows.

31.2.4.1 To start the MOA workflow

- Click **Multi-Omic Analysis** in the Workflow Browser under **Pathway Analysis** to start the wizard-driven MOA workflow.

After selecting the two input experiments in the **New Multi-Omic Experiment** window, the MOA workflow opens in a new window.

31.2.4.2 To start the SEA workflow

- Click **Single Experiment Analysis** in the Workflow Browser under **Pathway Analysis** to start the wizard-driven SEA workflow.

The SEA workflow opens in a new window.

31.2.4.3 To navigate the wizard-driven workflows

The wizard-driven SEA and MOA workflows contain four steps. You can move forward or backward in the workflows to make necessary changes in all four steps until you click **Finish** in the last step.

- Click **Next** ≫ to proceed to the next step in the workflow.
- Click ≪ **Back** to return to the previous step in the workflow.
- Click **Cancel** or to exit the workflow at any time.
- Click **Finish** in Step 4 to complete the workflow.

31.3 Single Experiment and Multi-Omic Analysis

This section describes in detail the different steps of the wizard- driven Single Experiment Analysis (SEA) and Multi-Omic Analysis (MOA) workflows. The steps for both workflows are

1. choosing the experiments, a pathway organism, and pathway sources,
2. specifying Interpretations and Entity Lists,
3. reviewing the analysis results, and
4. saving the result as a new pathway list or MOA experiment.

There are some differences between the SEA and MOA workflows, notably the fact that SEA requires only a single experiment, while MOA requires two experiments. This and other differences are mentioned separately in the appropriate sections.

This section also contains detailed information on the available options for further analyzing the saved pathways once the workflow is completed.

31.3.1 Step 1: Specifying Experiments, Organisms, and Pathway Sources

This section describes how to choose experiments, select a pathway organism and pathway sources in the wizard-driven SEA and MOA workflows (see Multi-Omic Analysis (Step 1 of 4)).

31.3.1.1 To choose an experiment

In SEA, by default the active experiment is selected for analysis.

Using MOA, it is possible to visualize overlapping entities from two experiments of either the same or different analysis type, compared to just one experiment in the SEA workflow. See What is Multi-Omic Pathway Analysis (MOA)? on page 896 for more information on how MOA can support your multi-omic research.

When you start the MOA workflow, **GeneSpring** first prompts you to select the two input experiments from the list of available experiments in the project and creates an MOA experiment containing all the entity lists and interpretations from the two selected experiments.

1. Select the two input experiments in the **New Multi-Omic Experiment** dialog box. You can also provide a Name and Description for the new MOA experiment that **GeneSpring** is creating from your selection.
2. Click **OK**.

GeneSpring creates a new MOA experiment from the selected experiments. The selected experiment names are reflected in the Experiment 1 and Experiment 2 fields of Step 1 of the Multi-Omic Analysis workflow, which opens automatically after the MOA experiment is created.

NOTE Only those experiments in the open project that match one of the supported experiment/technology types are available for selection in the **New Multi-Omic Experiment** window.

NOTE **GeneSpring** lists only those experiments that have a valid **GeneSpring** module license and are of a supported type and technology.

The order in which Experiment 1 and Experiment 2 appear in the list (and subsequently in the workflow window) is not relevant for finding matched pathways. To find matched pathways from just one experiment, choose **Single Experiment Analysis** from the **Pathway Analysis** section in the Workflow Browser.

The MOA (and standard SEA) workflow can analyze data from the following analysis types:

- Gene expression analysis
- Exon analysis
- Pathway experiments
- Mass Profiler Professional (MPP) experiments
- Strand NGS (MOA only)
- miRNA
- RTPCR

The following exceptions apply to some experiment types:

- For miRNA experiments you can only select entity lists resulting from a TargetScan in Step 2 of the Pathway Analysis workflow. Since the entity lists in miRNA experiments do not contain any associated data, the option to select an interpretation is not available. Hence, the pathways resulting from an SEA in a miRNA experiment will display matched entities but no Heatstrips or Quilt Plots will be visible for these matched entities.
- Similarly, for Whole Genome and target enriched DNA-Seq and Methyl-Seq experiments imported from Strand NGS, only matching entities are displayed and no data is overlaid as heatstrips or quilt plots.

GeneSpring does not support SEA or MOA workflows for Copy Number and Association Analysis experiments.

31.3.1.2 To choose a pathway organism

In the SEA and MOA workflows it is possible to choose an organism for finding matched pathways that is different from the organisms of the selected experiments. Selecting a different organism for searching pathways is useful when the experiment organism is less or not sufficiently described in literature, or when you want to observe the effects of one organism's pathogen/metabolite in another organism. By default, the drop-down displays the organism associated with Experiment 1.

- Select the pathway organism from the **Choose Pathway Organism** drop-down.

The list of organisms available from this drop-down depends on:

- the pathway organisms for which you have previously downloaded pathways or the Interactions Databases available from the Agilent Server (see Supporting Databases for Pathway Analysis on page 887),
- custom-created pathway organisms (see Supported organisms and annotations on page 894)
- the list of organisms for which pathways are available on the WikiPathways portal.

When different organisms are selected, **GeneSpring** requires both HomoloGene Translation for Gene/Protein Identifiers and BridgeDb - ID Mapping to map matching entities from the different organisms and with differing annotations. HomoloGene is used to identify homologous genes in these organisms. HomoloGene mapping between two organisms is done using Entrez Gene IDs. In cases where the same entity is identified by different annotations, for example Entrez Gene ID and RefSeq Transcript ID, the entities are mapped using BridgeDb.

In the absence of the HomoloGene or BridgeDb database for one or several selected organisms (experiment and pathway organisms) an SEA or MOA workflow can only proceed under certain circumstances that depend on the involved experiments' type. See Figure 31.27.

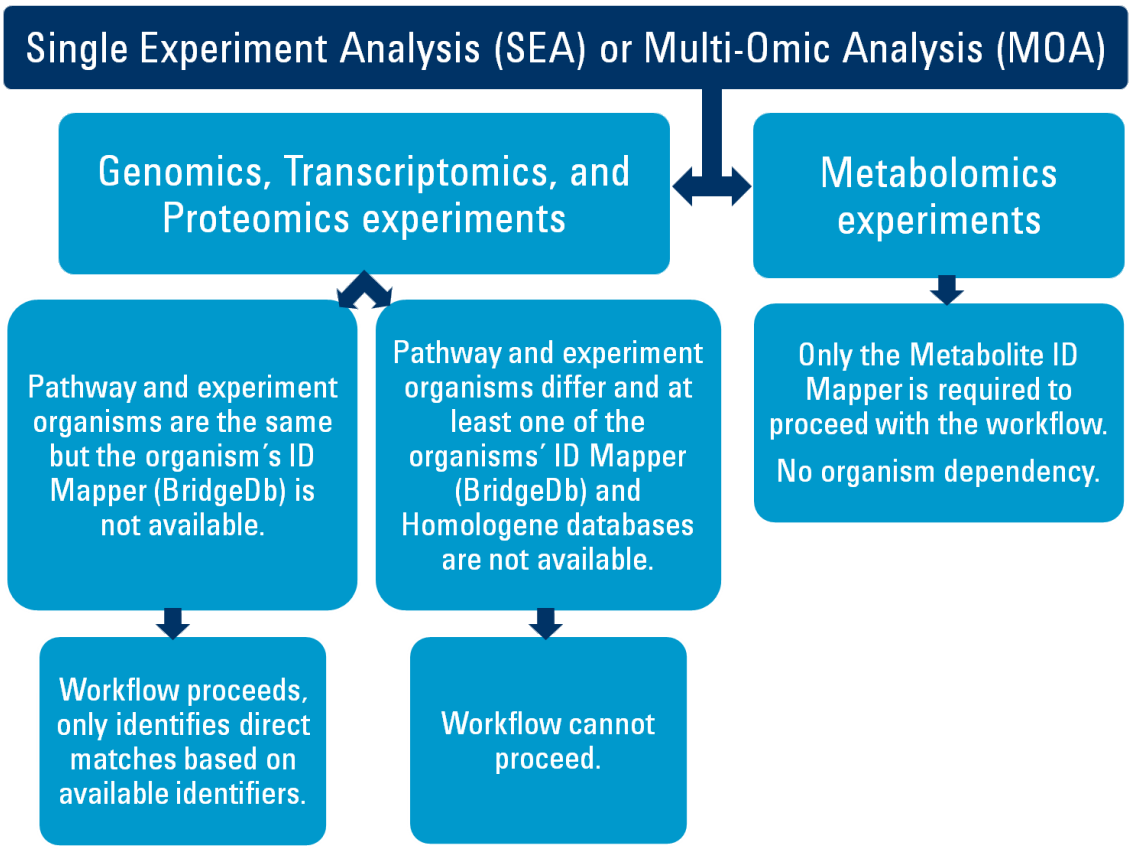


Figure 31.27: SEA/MOA workflow behavior based on experiment types and non-availability of at least one of the organisms' BridgeDb or HomoloGene database

31.3.1.3 To select pathway sources

You can select pathways from two Pathway Sources - **Curated pathways** and **Literature Derived Networks** - and their associated sources and formats.

1. Select **Pathway Sources**.
   The respective fields **Curated pathways only** and **Literature Derived Networks only** become active.
2. Select the required pathway sources from the provided lists.

GeneSpring searches only the selected pathway sources to identify matching pathways, which become available for selection in Step 3: Multi-Omic Analysis Results.

The following pathway sources are available under **Curated pathways**:

- WikiPathways - Analysis
- WikiPathways - Reactome
- WikiPathways - GenMAPP
- WikiPathways - Other
- BioCyc

- KEGG
- BioPAX (Imported)
- GPML (Imported)
- Hand created
- Legacy

The WikiPathways-Reactome and WikiPathways-GenMAPP options in the SEA/MOA workflow refer to the source of the pathway as it appears in the "Data-Source" field present within the GPML file. These source tags differ from the "GenMAPP-Approved" and "Reactome-Approved" curation tags that are available on the WikiPathways portal.

NOTE The WikiPathways options are selected in the list by default, to allow you to download pathways from the selected pathway organism from WikiPathways directly in both the MOA and SEA workflows (see To download WikiPathways in an SEA or MOA workflow).

By default, the BioCyc category includes the latest version only of BioCyc pathways that you have previously downloaded from the Agilent Server. You can change this setting in **Tools** ⟶ **Options** ⟶ **Pathway** ⟶ **BioCyc version for analysis** to use older versions only, or all versions; also see BioCyc pathway updates on page 867. Any pathways you imported from file for which you selected BioCyc as the provider are only included if you select the all versions option in this **Options** dialog.

If you select BioCyc when the number of pathways is '0', **GeneSpring** checks the availability of BioCyc pathways for the selected pathway organism on the Agilent Server when you click **Next** >>. If pathways are available to download, you have the option to download them directly in both the MOA or SEA workflows (see To download BioCyc pathways in an SEA or MOA workflow).

By default, the KEGG category includes the latest version only of KEGG pathways that you have previously downloaded from the Agilent Server. You can change this setting in **Tools** ⟶ **Options** ⟶ **Pathway** ⟶ **KEGG version for analysis** to use older versions only, or all versions; also see KEGG pathway updates on page 869. Any pathways you imported from file for which you selected KEGG as the provider are only included if you select the all versions option in this **Options** dialog.

If you select KEGG when the number of pathways is '0', **GeneSpring** checks the availability of KEGG pathways (if you have a valid KEGG end user license) for the selected pathway organism on the Agilent Server when you click **Next** >> (see KEGG Pathways Licensing on page 862). If pathways are available to download, you have the option to download them directly in both the MOA or SEA workflows (see To download KEGG pathways in an SEA or MOA workflow).

Other available pathway sources and formats are described in more detail in Importing Pathways into **GeneSpring** on page 863.

The 21 human pathways that are packaged with your **GeneSpring** installation and any pathways that you migrated from previous versions of **GeneSpring** are

categorized as **Legacy**. Pathways that you created in **GeneSpring** (see Creating and Editing Pathways on page 878) are categorized as **Hand created**.

The sources available under **Literature Derived Networks** are:

- NLP
- MeSH term

These include all interaction networks you imported or created using either one of the NLP Network Discovery, MeSH Network Builder, or Extract Relations via NLP workflows in **GeneSpring**.

The number of pathways previously imported into **GeneSpring** for each of the sources for the selected pathway organism is displayed in brackets next to the source name and changes accordingly when you choose a different pathway organism. The number represents all the pathways that were previously imported into **GeneSpring** (see Importing Pathways into **GeneSpring** on page 863).

31.3.1.4 To download WikiPathways in an SEA or MOA workflow

There are two ways to import pathways from the WikiPathways portal into **GeneSpring**. You can import all pathways for one organism or all available organisms before you start an SEA or MOA workflow by selecting **Tools** $\longrightarrow$ **Import Pathways from WikiPathways** from the main menu (see To import pathways from WikiPathways through the main menu on page 864).

Or you can download pathways from WikiPathways only for the pathway organism that you chose in Step 1 of the SEA or MOA, directly in the SEA or MOA workflow. However, this option is only available if pathways from WikiPathways for that organism have not previously been downloaded or imported in the GPML format.

1. Select any of the WikiPathways sources from **Curated pathways** and any of the other sources you want to include for analysis.
2. Click **Next** >>.
   A Confirm WikiPathways Download dialog informs you that no pathways were found for the chosen organism and provides the option to download available pathways.
3. Click **Yes** to download pathways from WikiPathways for the chosen organism.

The available pathways from WikiPathways are imported into **GeneSpring** in the background and are included in the analysis as you complete the SEA or MOA workflow.

NOTE The WikiPathways portal allows you to download pathways either from the "Analysis Collection" only, or "All Pathways". Go to To import pathways from WikiPathways through the main menu for more detailed information on how each of these collections are downloaded and stored in **GeneSpring**.

If BridgeDb databases have not yet been downloaded for the chosen organism, **GeneSpring** gives you the option to download the corresponding database here.

31.3.1.5 To download BioCyc pathways in an SEA or MOA workflow

There are two ways to import BioCyc pathways into **GeneSpring**. You can import all pathways for one organism or all available organisms before you start an SEA or MOA workflow by selecting **Tools** ⟶ **Import Pathways from BioCyc** from the main menu (see To import pathways from BioCyc through the main menu on page 866).

Or you can download BioCyc pathways only for the pathway organism that you chose in Step 1 of the SEA or MOA, directly in the SEA or MOA workflow. However, this option is only available if BioCyc pathways for that organism have not previously been downloaded or imported in the GPML format.

1. Select BioCyc pathways from **Curated pathways** and any of the other sources you want to include for analysis.
2. Click **Next** >>.
   A Confirm BioCyc Pathways Download dialog informs you that no pathways were found for the chosen organism and provides the option to download available pathways.
3. Click **Yes** to download BioCyc pathways for the chosen organism.

The available BioCyc pathways are imported into **GeneSpring** in the background and are included in the analysis as you complete the SEA or MOA workflow.

If BridgeDb or ID Mapper databases have not yet been downloaded for the chosen organism, **GeneSpring** gives you the option to download the corresponding database here.

NOTE If you publish research results in scientific journals based on the use of data from any of these BioCyc pathways, reference the results using the BioCyc and EcoCyc (for E.coli pathways) citations provided in the bibliography [9].

31.3.1.6 To download KEGG pathways in an SEA or MOA workflow

There are two ways to import KEGG pathways into **GeneSpring**. You can import all pathways for one organism or all available organisms before you start an SEA or MOA workflow by selecting **Tools** ⟶ **Import Pathways from KEGG** from the main menu (see To import pathways from KEGG through the main menu on page 868).

Or you can download KEGG pathways only for the pathway organism that you chose in Step 1 of the SEA or MOA, directly in the SEA or MOA workflow. However, this option is only available if KEGG pathways for that organism have not previously been downloaded or imported. Before proceeding with the pathway import, **GeneSpring** validates your KEGG ftp website login details in the background (see KEGG Pathways Licensing on page 862).

1. Select KEGG pathways from **Curated pathways** and any of the other sources you want to include for analysis.

2 Click **Next >>**.

A Confirm KEGG Pathways Download dialog informs you that no pathways were found for the chosen organism and provides the option to download available pathways.

3 Click **Yes** to download KEGG pathways for the chosen organism.

The available KEGG pathways are imported into **GeneSpring** in the background and are included in the analysis as you complete the SEA or MOA workflow.

If BridgeDb or ID Mapper databases have not yet been downloaded for the chosen organism, **GeneSpring** gives you the option to download the corresponding database here.

NOTE If you publish research results in scientific journals based on the use of data from any of these KEGG pathways, reference the results using the KEGG citations provided in the bibliography [37].

31.3.2 Step 2: Selecting Entity Lists and Interpretations

This section describes how to select interpretations, entity lists, and annotations in the second step of the SEA and MOA workflows (see Multi-Omic Analysis (Step 2 of 4)).

31.3.2.1 To select an interpretation

An interpretation is required from each experiment to perform SEA or MOA. You can select an interpretation for Experiment 1 and Experiment 2 (in MOA only) in the separate areas identified by the respective experiment names.

- Select the required interpretation from **Choose Interpretation**.

The chosen interpretation is highlighted.

NOTE Interpretations, other than the All Samples interpretation, have to be created before starting the MOA workflow (see Create Interpretation) to be available for selection in this workflow step. Only averaged interpretations or non-averaged interpretations with one sample in each condition are available for selection in the MOA workflow. An interpretation is not required for an SEA workflow with a miRNA experiment as input experiment.

31.3.2.2 To select an entity list

All available entity lists in the chosen experiments selected in Step 1 are displayed in separate areas identified by the respective experiment names in Step 2.

- Select the required entity list from **Choose Entity List**.

The chosen entity list is highlighted.

31.3.2.3 To select annotations

Entities from the selected entity lists and the pathways from the selected organism are matched based on their annotation identifiers. If the selected pathway organism differs from the experiment organisms, matching is accomplished by identifying homologous genes based on Entrez Gene IDs using HomoloGene Translation for Gene/Protein Identifiers. When the pathway and experiment organism are the same, annotation identifiers are matched using BridgeDb - ID Mapping.

GeneSpring first tries to find direct matches between the pathway entities and the entities in the selected entity list. A direct match can only occur when entities from both pathways and entity lists have identifiers from the same annotation. When identifiers from differing annotations have to be matched, the BridgeDb algorithm looks for a match in the order in which the annotations are displayed in this step. The first matching annotation and corresponding identifier are then displayed in the Heatmap Pane of The Pathway View.

The available annotations for the corresponding experiment type are listed under **Annotations**. By default, all listed annotations are selected. Please note that for "Generic" experiments, **GeneSpring** is only able to recognize all the annotations in a technology if the **Column Mark** was set correctly during experiment creation (see Update Technology Annotations on page 83).

You have to select at least one of the annotations to proceed with the SEA or MOA workflow.

- Select required annotations.
- Unselect unwanted annotations.

The list of annotations available for selection is fixed for each omic type: Genomics, Proteomics, Transcriptomics, and Metabolomics. They are listed in Table 31.1.

Table 31.1: Available Annotations by Omic Type

Genomics/ Transcriptomics	**Proteomics**	**Metabolomics**
RefSeq Protein ID	Swiss-Prot	KEGG ID
RefSeq Transcript ID	NCBI GI number	ChEBI ID
Entrez Gene ID		HMP ID
Ensembl Gene ID		CAS Number
GenBank Accession		
UniGene ID		
TIGR ID		

The annotations for each omic type were carefully chosen and placed in this order based on their popularity for the given experiment type and that they are supported by BridgeDb. If your pathway analysis does not identify any pathways, this could be due to a number of reasons, including:

- BridgeDb mapping is not supported between the relevant annotations,
- only few or no pathways are available from the specified Pathway Sources for the selected organism,
- **GeneSpring** could not recognize annotation identifiers, in particular in pathways downloaded from WikiPathways, because of a mismatch between the label added by the creators of the pathway and the data sources **GeneSpring** recognizes (for example if an annotation type is misspelled, see Supported Data Sources). This is also the reason behind instances in which **GeneSpring** declares that an imported pathway has '0' pathway entities (see Step 3: Multi-Omic Analysis Results).

31.3.3 Step 3: Multi-Omic Analysis Results

In this step all the queried pathways for the chosen pathway organism are displayed as a table with the associated metrics that were computed based on the chosen settings in the previous two steps (see Multi-Omic Analysis (Step 3 of 4)).

Between Step 2 and Step 3, **GeneSpring** computes p-values for all the matched pathways to determine their significance (see To filter pathway lists by p-value

cut-off). The **p-values**, number of **Matched Entities** between the input entity list and the pathway and **Pathway Entities of Experiment Type** are listed in their respective columns for each input entity list. No p-values are calculated for Metabolomics experiments (see p-value Computation).

The **Matched entities** columns contain the number of entities in the pathway that matched with the input entity list using the selected annotations during the SEA or MOA workflow.

The **Pathway entities of Experiment Type** columns contain the number of entities in the pathway that are of the experiment type: the number of genes and proteins in the pathway if the experiment type is Genomics/Proteomics/Transcriptomics and the number of metabolites if the experiment type is Metabolomics.

To inspect the pathway results before proceeding with the workflow, you can sort the results table by any one of these columns, or custom save a selection of these pathways as an MOA experiment in the MOA workflow and a new pathway list in the chosen experiment in the SEA workflow.

NOTE

For a small number of pathways from WikiPathways, **GeneSpring** will not be able to recognize any of the pathway entities and will report the number of **Pathway Entities of Experiment Type** as '0', despite the fact that the visual representation of the pathway in the Pathway View contains a number of entities. This is caused by a mismatch between the way the creator of the pathway labeled the entities and the data sources **GeneSpring** recognizes (see Supported Data Sources). **GeneSpring** is capable of recognizing a number of permutations of common ways to spell certain annotation types, but not all. We recommend that you inspect any of these pathways more closely, if you think they should have had at least some matches.

31.3.3.1 To sort pathways by column content

- Click any of the column headers to sort the pathway list from A-Z, from Z-A and to go back to the original order

An arrow will appear next to the column name indicating the sort direction.

31.3.3.2 To custom save selected pathways

You can select pathways from the displayed list of pathways and custom save this selection in both the MOA and the SEA workflows.

1. Select the row that contains the pathway of interest. To select multiple rows, hold **Ctrl** and select the additional rows.

   The selected rows are highlighted.
2. Click **Custom Save**.

   A Pathway List Inspector opens.
3. Rename the pathway list, if required. Click **OK**.

In an MOA workflow, the pathway list containing the selected pathways is saved in the MOA experiment created at the beginning of the workflow. This is a new experiment type, specifically created to hold the results of Multi-Omic Pathway as well as Correlation Analysis.

An MOA experiment contains all the entity lists and interpretations of the source experiments selected in Step 1 and the pathway list resulting from the analysis when completing the workflow, or the selected pathways when using the **Custom Save** option. It does not contain samples. After creation of an MOA experiment, it is a standalone item and is not connected to the parent experiments. A limited selection of analysis options is available from the Workflow Browser and main Toolbar in this type of experiment (see Working in an MOA Experiment on page 931).

In an SEA workflow, the pathway list containing the selected pathways is saved as a new pathway list in the source experiment under the entity list chosen for SEA.

31.3.4 Step 4: Saving Pathway List

This section describes the available options to review the list of pathways identified in Step 3: Multi-Omic Analysis Results before saving the result in a Pathway Analysis folder of the MOA experiment created at the beginning of the MOA workflow, or a new pathway list in the SEA workflow (see Multi-Omic Analysis (Step 4 of 4)).

31.3.4.1 To rename pathway list

By default, the pathway list is named after the two input experiments: "MOA-[name of Experiment 1]-[name of Experiment 2]".

- Delete the default name and type the desired name under **Name**.

31.3.4.2 To inspect and edit Notes

Notes contains information about the MOA settings for finding matched pathways. You can edit the default **Notes**.

- Scroll to view the entire contents of **Notes**.
- Delete or add information in **Notes** as desired.

31.3.4.3 To view the list of pathways to be saved

All the pathways that were displayed in the previous step are listed as a table at the bottom of the view.

- Scroll to view all the pathways (when the scroll bar is present).

31.3.4.4 To inspect pathways

You can inspect additional details of the listed pathways in the table:

- Double-click the row containing the chosen pathway. A Multiple Objects Inspector opens. It contains all the information related to the chosen pathway.

31.3.4.5 To save the pathway list

Once you are satisfied with the **Name**, **Notes**, you can complete the workflow.

- Click **Finish**.

In the MOA workflow, the pathway list is saved in the new MOA experiment that was created in the same project as the two source experiments at the beginning of the MOA workflow. The new MOA experiment contains all the interpretations and entity lists of the two source experiments. The pathway list is saved in a dedicated **Pathway Analysis** folder.

By default the new MOA experiment opens The Pathway View in the main **GeneSpring** Desktop. The first pathway listed in the pathway list panel is displayed in the Pathway Viewer. In the Heatmap Pane section of the view, the tab displaying the data for the experiment with the most matching entities between a pathway and the experiment's **All Entities** entity list is displayed.

The entity lists and new pathway list can be viewed and further analyzed as described in Working with Pathway Lists on page 912 and Working in an MOA Experiment on page 931.

In the SEA workflow the list of pathways is saved as a pathway list under the source entity list in the chosen experiment, rather than a separate experiment. By default, the new pathway list is displayed in The Pathway View when the SEA workflow is completed.

NOTE In **GeneSpring WorkGroupPlus** mode, to share a SEA pathway list, you have to share the SEA experiment from which the SEA pathway list was generated. To share a MOA pathway list, you have to share the MOA experiment from which the MOA pathway list was generated.

NOTE In **GeneSpring WorkGroupPlus** mode, to share a MOA pathway list from MOA experiments of versions earlier to **GeneSpring** version 13.0, share the project containing the experiments from which the MOA pathway list was generated.

31.3.5 Working with Pathway Lists

This section describes the viewing and analysis options that are common to pathway lists that were created from a Single Experiment (SEA) or Multi-Omic Analysis (MOA) workflow. These options include all the functions available in The Pathway View of a pathway list through right-click menus and the Pathway View Toolbar. Some of the toolbar options are only available for pathways that are not in the GPML format. To see these options, refer to the Pathway View Toolbar section on page 940.

31.3.5.1 To open a pathway list in the Pathway View

- Double-click the pathway list name in the **Experiment Navigator** or right-click over the pathway list name and select **Open Pathway List**.

The pathway list opens in the Pathway View. See the reference section of The Pathway View on page 940 for more information on its components.

31.3.5.2 To inspect a pathway list

- Right-click the pathway list name in the **Experiment Navigator** and select **Inspect Pathway List**.

The Pathway List Inspector opens. It contains all the information, like **Name** of the pathway list, **Notes**, and the list of pathways associated with this pathway list.

31.3.5.3 To remove a pathway list

- Right-click the pathway list name in the **Experiment Navigator** and select **Remove Pathway List**.

The pathway list is removed from the experiment, but remains available to be searched and added to this or other experiments at a later date.

31.3.5.4 To delete a pathway list

- Right-click the pathway list name in the **Experiment Navigator** and select **Delete Pathway List**.

The pathway list is permanently deleted from **GeneSpring**.

31.3.5.5 To display a pathway in the pathway viewer

By default, the first pathway listed in the Pathway list panel on the left side of The Pathway View is displayed in the Pathway Viewer. Use this panel to select other pathways that you want to bring into view.

- Double-click the row in the pathway list panel that corresponds to the selected pathway.

The selected pathway is displayed in the Pathway Viewer and the Heatmap Pane spreadsheet contains any experiment data associated with this pathway.

In **GeneSpring**, pathways are visualized depending on the file format in which they were imported. Pathways imported in the GPML format are rendered exactly as on WikiPathways. However, the entities and relations of pathways imported from Pathway Architect, a Text file or in the BioPAX format are visualized according to the legends for **Entities**, **Relations**, and **Edges** accessible from the right edge of the pathway viewer.

For all pathway formats, matching entities between the pathway and the input entity list that was selected for creating the pathway list are colored differently from the default setting. To switch off the default coloring of matching entities, use the **Select Entity Background Color to display** icon drop-down

options. In MOA experiments, the same toolbar icon drop-down list allows you to display only matching entities for one of the experiments. To switch off entity fill coloring, or toggle between experiments, use the **Select Entity Background Color to display** icon drop-down options. Experiment data for these matching entities is displayed in the Pathway Viewer in the form of Heatstrips or Quilt Plots. The Heatmap Pane below the Pathway Viewer contains a list of all pathway entities and experiment data for all pathway entities that match the technology, which corresponds to the **All Entities** entity list of the input experiment. For pathway lists resulting from an MOA workflow, the Heatmap Pane consists of two tabs, one for each input experiment.

You can visualize matching entities between the pathway and any entity list of any open experiment, by clicking on their names in the Experiment Navigator. These entities are referred to as highlighted entities and appear with a blue border in the Pathway Viewer. To switch off the default highlighting of matching entities with the active entity list in the Experiment Navigator use the **Select Highlight display** icon drop-down options.

All the color and display settings mentioned in this section are the default settings and can be modified as per your requirement (see To view and modify display settings on page 923).

You can overlay numerical as well as categorical data associated with the entities in the selected entity list onto the corresponding entities displayed in the pathway. Choose to color entities by their probe names or condition in the selected interpretation or select from entity list associated data like p-values or Fold Change if present in the selected entity list. See To overlay data on a pathway on page 921 for detailed instructions on how to overlay this type of data on a pathway.

31.3.5.6 To filter pathway lists by p-value cut-off

The pathway list panel on the left side of The Pathway View displays all the pathways in the open pathway list. In pathway lists that resulted from an SEA or MOA workflow, this list contains all the pathways available for the chosen pathway sources and pathway organism. The first time such a pathway list is opened, the default p-value cut off and the minimum number of matches (see To filter pathway lists by minimum number of matches on page 915) are set to '1.0' and '1', respectively. The pathways listed in the pathway list panel fulfill both the p-value cut off and minimum number of matches criteria.

To restrict the spreadsheet to pathways that satisfy a particular p-value cut-off, use the available option provided in the Filter panel of the Pathway View for each input experiment.

- Enter the desired value in the **p-value cutoff** field.

The list of pathways now only contains pathways with p-values below the selected value. The p-value cut off most commonly chosen for representing statistical significance is '0.05'. For more information on p-values see p-value Computation on page 1013.

When you close the Pathway View, the current p-value setting is saved and will be displayed the next time you open this pathway list.

NOTE The p-value filter option is only available for pathway lists that were created as a result of an SEA or MOA with Genomics, Transcriptomics, and Proteomics experiments.

In pathway lists that were created as a result of an MOA workflow, you can choose different values for the p-value cut off criteria for each input entity list and choose whether the list of pathways should be restricted to the pathways that fulfill both conditions, or at least one of them. You can also choose to apply only the filter criteria for one or none of the experiments.

- Select **Show pathways that pass either filter**.
  The pathway list panel displays only pathways with p-values below the determined cut-off in either experiment.
- Select **Show pathways that pass both filters**.
  The pathway list panel displays only pathways with p-values below the determined cut-off in both experiments.
- Unselect the filter panel section associated with the experiment that you want to disregard for filtering the pathway list.
  The respective filter panel is greyed out and the pathway list panel only displays pathways that pass the filter of the remaining active filter panel. If you unselect both experiments, the table contains all the matched and unmatched pathways for the chosen pathway organism from the sources that were selected in Step 1 of the MOA wizard.

31.3.5.7 To filter pathway lists by minimum number of matches

The pathway list panel on the left side of The Pathway View displays all the pathways in the open pathway list. In pathway lists that resulted from an SEA or MOA workflow, this list contains all the pathways available for the chosen pathway sources and pathway organism. The first time such a pathway list is opened, the default p-value cut off (see To filter pathway lists by p-value cut-off on page 914) and the minimum number of matches in the Filter panel are set to '1.0' and '1', respectively. The pathways listed in the pathway list panel fulfill both the p-value cut off and minimum number of matches criteria.

To restrict the spreadsheet to pathways with a particular minimum number of matches, use the input field provided in the Pathway View for each input experiment.

- Enter the desired value in the **Min # of matches** field.

The list of pathways now only contains pathways which have at least the determined minimum number of matches.

When you close the Pathway View, the current minimum number of matches setting is saved and will be displayed the next time you open this pathway list.

NOTE The **Min # of matches** filter option is only available for pathway lists that were created as a result of an SEA or MOA workflow.

In pathway lists that were created as a result of an MOA workflow, you can choose different values for the minimum number of matches criteria for each input entity list and choose whether the list of pathways should be restricted to the pathways that fulfill both conditions, or at least one of them. You can also choose to apply only the filter criteria for one of the experiments.

- Select **Show pathways that pass either filter**.
  The pathway list panel displays only pathways with at least the minimum number of matches in either experiment.
- Select **Show pathways that pass both filters**.
  The pathway list panel displays only pathways with at least the determined minimum number of matches in both experiments.
- Unselect the filter panel section associated with the experiment that you want to disregard for filtering the pathway list.
  The respective filter panel is greyed out and the pathway list panel only displays pathways that pass the filter of the remaining active filter panel. If you unselect both experiments, the table contains all the matched and unmatched pathways for the chosen pathway organism from the sources that were selected in Step 1 of the MOA wizard.

31.3.5.8 To zoom in or out on a pathway

You can zoom in and out of the pathway as a whole, or selected regions of the pathway.

- Scroll to zoom in or out on the entire pathway, or use the **Zoom in** and **Zoom out** icons from the Pathway View Toolbar.
- Right-click on the pathway viewer and drag a rectangle over the area you want to zoom in to.

31.3.5.9 To select entities in a pathway

There are multiple ways to select entities in a pathway:

- Click the chosen entity in the pathway or in the Heatmap Pane spreadsheet. Press **Ctrl** and click the chosen entities to select multiple entities.
- Click into an empty area in the pathway viewer and drag a rectangle over the area of interest in the pathway. All entities within the selected area are selected.
- Click the **Select all entities** , **Select matched entities** , or **Select highlighted entities** icons from the Pathway View Toolbar.
- Right-click on the pathway viewer and select **Select all**, **Select all entities**, **Select matched entities**, or **Select highlighted entities**.

Selected entities are marked in the pathway according to the default color settings illustrated in the Legend. Selecting entities in the pathway viewer restricts the rows displayed in the Heatmap Pane to those entities.

Selected entities can be moved or their size changed for better visualization.

31.3.5.10 To create a new entity list from selected entities in a pathway

This functionality varies depending on whether the pathway list was created from an SEA or MOA workflow. The behavior for pathway lists created from a Pathway Experiment or the To create a new pathway operation is the same as that for the SEA workflow.

NOTE Only entities that are also present in the **All Entities** entity list can be saved as new entity lists using the **Create Entity List** operation. If the selection contains entities from both All Entities entity lists of the input experiments in an MOA experiment two separate entity lists are created, one for each set of entities also present in either one of the input experiments.

For pathway lists created from an SEA workflow:

1. Select the desired entities as described in To select entities in a pathway on page 916.
2. Click the **Create entity list from selection** icon from the Pathway View Toolbar.
   An Inspector opens.
3. Rename the new entity list and edit notes if desired. Click **OK**.

The new entity list is saved as a child list of the **All Entities** entity list in the same experiment as the open pathway list.

For pathway lists created from an MOA workflow:

1. Select the desired entities as described in To select entities in a pathway on page 916.
2. Click the Create entity list from selection icon from the Pathway View Toolbar.
   If the selection contains entities from both input entity lists a Multiple Objects Inspector opens.
   If the selection contains entities from only one of the input entity lists an Inspector opens.
3. Rename the new entity lists or list and edit notes if desired. Access the second entity list in a Multiple Objects Inspector by clicking on the entity list name listed in the left panel. When you are satisfied with the details click **OK**.

The new entity lists are saved as child lists of the respective input entity lists saved in the MOA experiment.

31.3.5.11 To visualize selected entities in a profile plot

1. Select entities as described in To select entities in a pathway on page 916.
2. Click the **Launch profile plot on selected entities** icon from the Pathway View Toolbar.

A Profile Plot dialog opens. Only entities that are present in an input experiment are shown in the profile plot. If the pathway list was a result of an MOA workflow and the selected entities are present in both input experiments, the dialog contains two tabs: One each for the entities in the two input experiments. In these profile plots each entity is identified by a different color.

31.3.5.12 To highlight matching entities from other entity lists in the Experiment Navigator

In **GeneSpring**, you can visualize matching entities between the displayed pathway and the active entity list in any open experiment in the **Experiment Navigator**. **GeneSpring** uses the available annotations listed in Table 31.1 to identify matching entities between the selected entity list and the pathway. This functionality is available for all types of pathway that can be imported into, created/edited in **GeneSpring**.

- Click on an entity list in an open experiment in the **Experiment Navigator**. This entity list becomes the active entity list (indicated by the bold script).

Matching entities between this active entity list and the displayed pathway are indicated by a blue border.

31.3.5.13 To choose a different entity list

You can choose a different entity list to be overlaid on the displayed pathway, in order to see which entities are common between the pathway and the new entity list.

1. Click the **Choose another Entity List** icon from the Pathway View Toolbar.

   The Entity List Chooser dialog opens.
2. Click **Choose...** next to the respective experiment name for which you want to change the entity list in the Entity List Chooser dialog.

   The Choose Entity List dialog opens.
3. Select the desired entity list and click **OK**.

   The dialog closes and the name of the selected entity list appears in the corresponding field of the Entity List Chooser dialog.
4. Repeat steps 2-3 to choose a different entity list for the second experiment (this only applies to pathways saved as a result of an MOA workflow). Click **OK**.

Only the coloring of the pathway entities is updated to reflect the matches with the different entity list or lists. The other elements of The Pathway View continue to display the data associated with the original input entity list.

NOTE The Entity List that was chosen using the toolbar option described here is not saved when the pathway list is closed. The p-value columns in the pathway list panel are not updated with new p-values corresponding to the different entity list.

NOTE If the technology of an Entity List differs from the technology of the Entity List on which the **Pathway Analysis** was performed, it is disabled and not available for selection in **Choose a different entity list** dialog.

31.3.5.14 To choose a different interpretation

You can choose a different interpretation according to which data is overlaid on the displayed pathway.

1. Click the **Choose another Interpretation** icon from the Pathway View Toolbar.
   The Interpretation Chooser dialog opens.
2. Click **Choose...** next to the respective experiment name for which you want to change the interpretation in the **Interpretation Chooser** dialog.
   The Choose Interpretation dialog opens.
3. Select the desired interpretation and click **OK**.
   The dialog closes and the name of the selected interpretation appears in the corresponding field of the **Interpretation Chooser** dialog. Only averaged interpretations and un-averaged interpretations with one sample per condition can be selected in this dialog.
4. Repeat steps 2-3 to choose a different interpretation for the second experiment (this only applies to pathway lists saved as a result of an MOA workflow). Click **OK**.

The data in the Heatmap Pane and the Heatstrips or Quilt Plots in the Pathway Viewer are updated to reflect the different interpretation.

NOTE The interpretation that was chosen using the toolbar option described here is not saved when the pathway list is closed.

31.3.5.15 To display, remove, and customize quilt plots and heatstrips in the Pathway Viewer

Heatstrips and Quilt Plots represent experimental data associated with matching entities directly in the Pathway Viewer. You can choose whether quilt plots or

heatstrips are displayed and whether they are displayed for all experiments (default setting), for only one of two experiments (in MOA only), or not displayed at all.

- By default, **GeneSpring** displays quilt plots in the Pathway Viewer. If you want to display experimental data as heatstrips instead of quilt plots, select **Heatstrips** from the **Choose Quilt Plots/Heatstrips to display** ( ) drop-down options.
- Select the desired option from the **Select experiment(s) to display Quilt Plots/Heatstrips** ( ) drop-down options available from the Pathway View Toolbar to display heatstrips or quilt plots for all experiments (default setting), for only one of two experiments (in MOA only), or not at all.
- Determine which parameters are plotted against the x-axis and y-axis in a quilt plot. Click the **Choose parameters along axes** icon to open the **Choose Quilt Plot Parameters** dialog and assign the available parameters to the desired axis fields.

You can also customize the color scheme for the displayed heatstrips, determine a color for missing values in quilt plots and remove/add the experiment indicators from/to quilt plots and heatstrips in MOA experiments.

1. Click the **View and Modify Display Settings** icon from the Pathway View Toolbar, or select Display Settings on right-clicking over the Pathway Viewer.

   The Pathway Property Dialog opens. It contains five tabs: **Data**, **Entity**, **Heatstrip**, **Quilt Plot**, and **Select/Highlight**.
2. Click the **Heatstrip** tab.
3. Select the condition for which you want to change the color.
4. Click **Set Color** to define the new color.
5. Repeat steps 2 and 3 until you have defined the heatstrip color for each condition.
6. In a MOA experiment, click the tab with the name of the second experiment to reveal the conditions of the selected interpretation from that experiment. Then repeat steps 2 and 3 until you have defined the heatstrip color for each condition.
7. To switch off the display of the experiment indicator next to heatstrips in MOA experiments, deselect the **Display Experiment Indicator** option.
8. Click the **Quilt Plot** tab.
9. Click the **Assign color** field to type in a new set of RGB values that define the desired color, or click the drop-down arrow to select a color from the displayed options. To access the **Choose a Color** dialog click **More Colors**. In the **Choose a Color** dialog click **OK** to apply any changes, click **Cancel** to close the dialog without applying changes, or click **Reset** to return to the original color setting.
10. Click **OK**.

Any changes are applied to the pathway displayed in the pathway viewer.

31.3.5.16 To overlay data on a pathway

New in version 13.1, you can now overlay numerical as well as categorical data associated with the entities in the selected entity list onto the corresponding entities displayed in the pathway. You can choose to color entities by their probe names or condition in the selected interpretation or select from entity list associated data like p-values or Fold Change if present in the selected entity list. Using the options available in the Pathway Property Dialog you can customize the default color settings and what should be displayed if an entity matches both experiments in a MOA experiment.

1. Click the **View and Modify Display Settings** icon from the Pathway View Toolbar, or select Display Settings on right-clicking over the Pathway Viewer.

   The Pathway Property Dialog opens. It contains four tabs: **Data**, **Entity**, **Heatstrip**, and **Select/Highlight**.

2. Click the **Entity** tab. Use the **Color By** options to overlay the desired data on the pathway.

   - By default, **GeneSpring** colors entities by **Matched Experiment** if the pathway list results from Pathway Analysis. Click **Customize...** under **Color Settings** to define a custom fill color for the entities matching the experiment in a SEA pathway list. In a MOA experiment, choose the entity fill colors for entities matching either input experiment.
   - Choose **Data (Signal/Abundance Values)** to color matching entities by their ProbeName or condition in the input interpretation. Select the required value type from the drop-down list. Click **Customize...** to change the default color setting if desired. When multiple values present in the experiment correspond to a single pathway entity, the experiment values are summarized in the same way as in the Heatmap Pane and the entity is rendered in the corresponding color.
   - Choose **Entity List Associated Values** to color matching entities by associated values present in the input entity list, for example p-value or Fold Change. Select the required value type from the drop-down list. Click **Customize...** to change the default color setting if desired:
     - **Categorical Values** - Each value has a distinct color assigned to it. In the **Color Selector** dialog under the **Color By** section, choose the desired colors for each of the data values. In some cases multiple values from the experiment will be present for the same pathway entity. You can determine how **GeneSpring** displays these entities by selecting the desired option in the **In case of multiple values** section of the dialog. Choose **Assign Distinct Color** and define a color that is distinct from the colors chosen for the individual data values to highlight such entities using the drop-down list. Or you can select the desired option from the **Color by** drop-down list. The available options are **Minimum Frequency**, **Maximum Frequency**, and **First Value**.
     - **Numerical Values** - Each value is represented as the shade of color corresponding to the value on a color range. In the **Color Selector** dialog under the **Color Range Selector** section, use the sliders or manually input the desired Minimum color, Middle color, and Maximum color to define the color range. Also define a **Missing value color** from the drop-down list. In some cases multiple values

from the experiment will be present for the same pathway entity. You can determine how **GeneSpring** displays these entities by selecting the desired option from the **If multiple values are available, use** drop-down list. The available options are **Minimum**, **Maximum**, **Sum**, **Average**, **Median**, and **First Value**.
By default, if any of the columns related to Fold Change (e.g. FC (abs), FC, Log FC) are selected from the **Color by** drop-down list, **GeneSpring** selects and displays the **Maximum** value. If a column containing p-values (p-value, p-value (corrected)) is selected from the **Color by** drop-down list, **GeneSpring** selects and displays the **Minimum** value, by default. For all other columns, the default selection and display is **Minimum**.

- In a MOA experiment, select the desired option under **Settings for pathway entities matched in both experiments**.

In KEGG pathways, one pathway entity can represent many entities for which different experiment data might exist. By default, **GeneSpring** displays the appropriate color for the first entity matched in the experiment under the displayed pathway entity. The tool tip reveals the entities with a background color corresponding to the individual values for each of these entities. The same color is also displayed next to each of these entities in the KEGG Entity Properties dialog (see Entity Properties - KEGG pathways).

3 Click **OK**.

Any changes are applied to the pathway displayed in the pathway viewer.

NOTE

Any changes made to the **Matched Experiment** option of the **Color By** section are applied to all the other pathways present in the same experiment as the currently open pathway.
Any changes made to the **Data (Signal/Abundance Values)** and **Entity List Associated Values** options of the **Color By** section are only applied to the open pathway list.

NOTE

When the **Data (Signal/Abundance Values)** option is selected and you choose one of the conditions in the selected interpretation to color matching pathway entities, the fill colors in the matching pathway entities and the cell color in the pathway view heatmap pane for the corresponding entity will not always be the same, although the actual signal/abundance value is the same in both instances. This is due to the fact that the color range applied for pathway entities data overlay (as shown in the **Color Settings** tab in the Legend) is with reference to the values of the **selected condition across all entities** present in the technology, while the color range in the heatmap pane (as shown in the **Heatmap** tab in the Legend) is with reference to the values of **all conditions** in the interpretation across **all pathway entities** present in the technology.

31.3.5.17 To view and modify display settings

By default, normalized signal values are displayed in the Heatmap Pane and in the Heatstrips or Quilt Plots. The color settings for heatstrips and quilt plots in the Pathway Viewer, the Heatmap Pane, and pathway entities are provided in the Legend. You can change the displayed data to raw signal values and the color settings for the heatstrips and pathway entities, for example for better contrast or visualization of experimental data in a particularly large pathway.

1 Click the **View and Modify Display Settings** icon from the Pathway View Toolbar, or select Display Settings on right-clicking over the Pathway Viewer.

   The Pathway Property Dialog opens. It contains five tabs: **Data**, **Entity**, **Heatstrip**, **Quilt Plot Layout**, and **Select/Highlight**.

2 Make the desired changes via these five tabs. Click **OK**.

The changes are applied to the pathway viewer, the Heatmap Pane, and the Legend and persist across all pathway lists in the current experiment, with the exception of the **Data (Signal/Abundance Values)** and **Entity List Associated Values** options from the **Color By** section of the **Entity** tab. These settings only persist across the currently open pathway list.

NOTE The **Data Settings** tab allows you to choose between displaying raw and normalized signal values. For raw and normalized signal values, the Heatmap Pane contains experimental data for matching entities of the **All Entities** entity list of the input experiment. Raw values cannot be displayed for any Two Color experiment type.

31.3.5.18 To view the pathway legend in Full Screen mode

When working with a pathway list in Full Screen Mode, you can access the Legend explaining the sample/condition layout of quilt plots as well as the color coding for quilt plots and heatstrips and matching pathway entities as follows:

- Click the **Legend Info** tab displayed at the top right corner of the pathway viewer.

Once expanded it contains the same information as the Legend pane that is part of the GeneSpring desktop visible in the ordinary Pathway View.

31.3.5.19 To change pathway layout

GeneSpring offers a number of different pathway layouts (see Pathway Layouts on page 959) to apply to pathways while viewing them in the pathway viewer.

- Select the desired option from the **Layout graph** drop-down ( ) available from the Pathway View Toolbar.

Pathway layouts are available only for pathways resulting from the NLP workflows and BioPAX Level 2 and Level 3 formats. For any non-NLP pathways the new pathway layout is not saved when the Pathway View is closed.

31.3.5.20 To inspect pathway entities and relations

You can access more information about the entities and relations in a pathway (valid for all non-GPML format pathways, except KEGG pathways).

- Double-click the desired entity or relation node.
- Right-click on the desired entity and select properties.

In a non-GPML pathway, an **Entity Properties** dialog opens. It contains two tabs: **Properties** and **Style**. The **Properties** tab contains all available annotation information on the entity or relation. The **Style** tab allows you to modify the display settings for the entity.

Double-clicking on an entity in a GPML pathway opens an **Entity Properties** dialog that contains basic annotation information for the given entity.

31.3.5.21 To inspect KEGG pathway entities

You can access more information about the entities and enzyme information in a KEGG pathway.

- Double-click the desired entity.
- Right-click on the desired entity and select properties.

In a KEGG pathway, an **Entity Properties - KEGG pathways** dialog opens. It contains two tabs: **Entities** and **Enzyme** (available only for metabolic pathways). The **Entities** tab contains all available annotation information on the entity. The **Enzyme** tab contains the enzyme information and notes on the entity.

31.3.5.22 To copy selected entities and relations

You can copy selected entities and relations of a NLP-derived pathway in the pathway viewer to create a new pathway around the selection. You can also copy the entire pathway if you want to edit it and then save edited version as a new pathway (see Creating and Editing Pathways on page 878).

1. **Select entities** and/or relations in the pathway viewer. Click the **Copy Selection** icon in the Pathway View Toolbar.
2. Click the **Create new pathway** icon in the main toolbar.
   An empty pathway viewer opens in the main window.
3. Paste the selected entities and/or relations by clicking the **Paste Selection** icon in the Pathway View Toolbar.

The selected entities and/or relations are displayed in the pathway viewer. See Creating and Editing Pathways on page 878 for the additional pathway editing options.

31.3.5.23 To create a Venn diagram on pathway lists

In **GeneSpring**, you can create Venn diagrams to compare entity lists (see Entity List) and pathway lists. You can identify common pathways between up to four pathway lists in any of the open experiments in your current project by displaying them in a Venn diagram.

NOTE The comparison of pathways is on their Name and Organism and not on the entities in the pathway. We recommend to always name pathways and interaction networks appropriately to avoid confusion later.

---

1. Click the drop-down arrow of the **Venn Diagram** icon and select **Pathway List**.

   The **Venn Diagram** dialog opens.
2. Select up to four pathway lists. Click **OK**.

The **Venn Diagram** view of the selected pathway lists is displayed in the main window. For more information on Venn Diagrams in **GeneSpring** go to The Venn Diagram on page 173.

It is important to note, that **GeneSpring** only creates the Venn diagram from the pathways of a selected pathway list that satisfy the p-value cut-off (see To filter pathway lists by p-value cut-off) and minimum number of matches (see To filter pathway lists by minimum number of matches) filter specified the last time you worked in that pathway list. For example, if the pathway list contains 30 pathways and only 17 passed the p-value and minimum number of matches filter the last time you closed the selected pathway list, only those 17 pathways are used for creating the Venn diagram.

31.3.5.24 To find pathways containing any/all selected entities

In a pathway list that contains a Heatmap Pane you can search for additional pathways that contain any or all of a selection of entities listed in the Heatmap Pane.

1. Select the desired entities (rows) in the heatmap pane spreadsheet. To select multiple rows, hold **Ctrl** and select all the desired rows.

   The selected rows are highlighted.
2. Right-click on one of the selected rows and select **Find pathways containing any selected entities** or**Find pathways containing all selected entities** as desired.

   The Organism Chooser dialog opens
3. Choose a pathway organism from the drop-down and select the desired pathway sources from **Curated pathways** and/or **Literature Derived Networks** for the search. Click **OK**.

   GeneSpring starts searching the selected pathway sources for matching pathways. If you selected any of the WikiPathways options and you have

not previously downloaded WikiPathways for the chosen organism, you have the option to download them now. These pathways are then automatically included in the search. A **Pathwaylist Inspector** listing all the identified pathways and details about how the list was created opens.

4 Rename the pathway list and edit the **Notes** if desired. Click **OK**.

The new pathway list is saved in the **Imported pathway lists** folder inside the Analysis folder of the experiment hierarchy tree. If this is the first time you imported a new pathway list the folder is newly created for saving the new pathway list.

31.3.5.25 To select the appropriate isomer for heatstrip and quilt plot data overlay in a pathway when multiple isomers in the experiment are mapped to only one entity in the pathway

Metabolites in Mass Profiler Professional (MPP) experiments are characterized by their mass, retention time and annotation id. Multiple metabolites that share the same mass but have varying retention times typically indicate that they are isomers of the same compound. Following identification via IDBrowser, metabolites with the same mass but varying retention times are assigned the same identifier. Since experimental data can vary significantly between different isomers, it is important to be able to distinguish between existing isomers when overlaying experimental data on a pathway.

You can make this distinction and determine for which isomer (identified by its retention time) you want to display Heatstrips or Quilt Plots in the Pathway Viewer in the Heatmap Pane of the The Pathway View. Each entity with the same identifier, but different retention times is grouped under a common compound name in the Heatmap Pane spreadsheet. The presence of isomers is indicated through an arrow before the compound name in the **Compound Name** column and the presence of an additional **Review Status** column, which contains a tick box. In addition, the number of compounds with isomers present in the spreadsheet is indicated by the phrase **Conflict: N** (where N is the number of compounds with isomers for which a selection has not yet been made) added after the experiment name in the title tab of the corresponding spreadsheet.

By default, **GeneSpring** displays heatstrips or quilt plots for the isomer with the shortest retention time. To see the different isomers and select the one for which you would like to display heatstrips or quilt plots in the pathway follow these steps:

1 Click the row containing a compound with isomers.

The available isomers appear in ascending order from shortest to longest retention time below the common compound name. The difference in retention time compared to the isomer with the shortest retention time is indicated after the compound name in each row.

2 Double-click the desired isomer.

The corresponding experimental data is displayed as a heatstrip or quilt plot in the pathway. The tick box in the **Review Status** column is ticked and N in the title tab for the spreadsheet reduces by '1'.

3 To change the selection to another isomer, double-click any one of the other isomers.

Experimental data for the corresponding isomer is displayed as a heatstrip or quilt plot in the pathway.

4 To undo the selection, double-click the **Review Status** tick box.

The tick mark is removed and N in the title tab for the spreadsheet increases by '1'. Experimental data for the isomer with the shortest retention time is once again displayed in the pathway.

31.3.5.26 To export Gene Annotations of entities

You can export the entity names and Entrez Gene Ids of either all the entities or selected entities of the pathway displayed in the pathway viewer to a .txt file.

1 Select the desired entities in the pathway viewer. (If you want to export all pathway entities, no selection is required.)

The Heatmap Pane spreadsheet is restricted to the rows corresponding to the selected entities.

2 Right-click anywhere in the pathway viewer and select **Export Gene Annotations**.

The **Export Gene Annotations** dialog opens.

3 Select whether you want to export **Selected Entities** or **All Entities** under **Select Entities to Export**.

4 Click **Browse** to change the default name ($Exported_Entities.txt$) and location where the file will be saved. If you are renaming the file, add ".txt" at the end of the new name in the **Choose a file** dialog. Click **OK**.

The new name and location are displayed under **Choose Output File**. Click **OK**

The file is saved in the specified location. If you did not add ".txt" at the end of the file when changing the file name, right-click the file and choose **Open with**. Select "WordPad" as the program with which to open the file.

31.3.5.27 To export pathway entities to eArray

You can export all the entities or selected entities of the pathway displayed in the pathway viewer to Agilent's eArray portal to create custom arrays for these entities.

NOTE Export to eArray/SureDesign exports entities with their RefSeq identifiers either directly or by translation using BridgeDb (make sure the BridgeDb database for the desired organism has previously been downloaded into **GeneSpring**). Only entities with RefSeq identifiers, or for which BridgeDb can translate an available identifier into a RefSeq identifier are exported. **GeneSpring** can export a maximum of 10,000 entities with this feature.

---

1 Select the desired entities in the pathway viewer. (If you want to export all pathway entities, no selection is required.)

The Heatmap Pane spreadsheet is restricted to the rows corresponding to the selected entities.

2 Right-click anywhere in the pathway viewer and select **Export to eArray**. The **Export to eArray** dialog opens.

3 Select whether you want to export **Selected Entities** or **All Entities** under **Select Entities to Export**.

4 Select for which **Application** you are exporting the entities: **Expression** or **RNA Enrichment**. Click **OK**.

Your default browser opens Agilent's eArray portal website. You have to log in to the portal with your details to continue. Depending on your choice of **Application** in the dialog, the **Expression** or **RNA Enrichment** page of the eArray portal will appear.

31.3.5.28 To copy entities to the Clipboard

You can copy either Entrez Gene IDs or SwissProt/UniProt IDs of selected entities of the pathway displayed in the pathway viewer to the Clipboard. For entities with differing annotations from the selected annotation, **GeneSpring** converts those annotations to either Entrez Gene IDs or SwissProt/UniProt IDS using BridgeDb - ID Mapping.

You can also copy all of the available annotations for selected pathway entities displayed in the pathway viewer to the clipboard as tab separated value text (.tsv).

1 Select the desired entities in the pathway viewer.
The Heatmap Pane spreadsheet is restricted to the rows corresponding to the selected entities.

2 Right-click anywhere in the pathway viewer and select the desired annotation from the submenu of **Copy to Clipboard as**.
The **Copied to Clipboard** information dialog displays the number of identifiers of the chosen annotation that were copied to the Clipboard for the selected entities. Click **OK**.
The copied identifiers can now be retrieved in any file that allows you to paste information from the Clipboard.

The copied identifiers can now be retrieved in any file that allows you to paste information from the Clipboard.

31.3.5.29 To add pathway entities to the Pathway List Clipboard

The **GeneSpring** Pathway List Clipboard is a repository of selected pathway entities of interest and allows you to compile custom entity lists comprising entities from multiple pathways independent of their format for export to file or directly to eArray/SureDesign. There are two ways to add pathway entities to the Pathway List Clipboard, (1) only selected entities in a pathway and (2) all entities in selected pathways:

- Select the entities of interest in a pathway in the pathway viewer. Right-click over the pathway viewer and select **Add to Pathway List Clipboard**.
  The selected entities and their associated properties are added to the Pathway List Clipboard.
- Select the pathways of interest in the pathway list panel. Right-click over the pathway list panel and select **Add to Pathway List Clipboard**.
  All the entities in the selected pathways are added to the Pathway List Clipboard.

31.3.5.30 To inspect the Pathway List Clipboard

You can inspect the contents of the Pathway List Clipboard from an open pathway list.

1. Click the **Pathway List Clipboard Inspector** icon in the **Pathway View Toolbar**.
   The Pathway List Clipboard Inspector opens.
2. Review the entities saved in the Pathway List Clipboard using the available options from the right-click menu (see Common Operations on Table Views on page 122). You can export selected entities to file or eArray/SureDesign (see "To export entities from the Pathway List Clipboard to eArray/SureDesign" on page 930) or delete selected entities (see "To delete entities from the Pathway List Clipboard" on page 930).
3. Click **Close** to return to the Pathway View.

The content of the Pathway List Clipboard for a given pathway list is saved across multiple sessions. However, the content is deleted when the project containing the pathway list is exported.

31.3.5.31 To export entities from the Pathway List Clipboard to file

You can export selected properties (columns) for either all or the selected pathway entities from the Pathway List Clipboard in a text (.txt) file format.

1. Click the **Pathway List Clipboard Inspector** icon in the **Pathway View Toolbar**.
   The **Pathway List Clipboard Inspector** opens.
2. Select the entities that you want to export to file. If you want to export all the entities in the Pathway List Clipboard, you do not have to make any selection in the Inspector.
3. Click the **Export to file** icon.
   The **Export to File** dialog opens.
4. Choose whether to export **All Entities** or **Selected Entities**.
5. Select the columns to export by moving the desired properties from the **Available Columns** to the **Selected Columns** are using the provided buttons.
6. If desired, change the order of the **Selected Columns** using the provided buttons.
7. Click **Browse** to determine where **GeneSpring** saves the exported entities and specify the desired file name.
   The chosen location appears in the **Export to** field.
8. Click **Export**.

The pathway entities are exported as a *.txt file to the specified location.

31.3.5.32 To export entities from the Pathway List Clipboard to eArray/SureDesign

You can export either all or the selected pathway entities from the Pathway List Clipboard directly to eArray or SureDesign to create custom assays for these entities.

NOTE

Export to eArray/SureDesign exports entities with their RefSeq identifiers either directly or by translation using BridgeDb (make sure the BridgeDb database for the desired organism has previously been downloaded into **GeneSpring**). Only entities with RefSeq identifiers, or for which BridgeDb can translate an available identifier into a RefSeq identifier are exported. **GeneSpring** can export a maximum of 10,000 entities with this feature.

---

1. Click the **Pathway List Clipboard Inspector** icon in the **Pathway View Toolbar**.
   The **Pathway List Clipboard Inspector** opens.
2. Select the entities that you want to export to eArray/SureDesign. If you want to export all the entities in the Pathway List Clipboard, you do not have to make any selection.
3. Click the **Export to eArray/SureDesign** icon.
   The **Export to eArray/SureDesign** dialog opens.
4. Choose whether to export **All Entities** or **Selected Entities**.
5. Choose the **Application**, either **Expression** or **RNA Enrichment**.
6. Select the organism from the **Organism** drop-down menu. This menu contains all the organisms associated with pathways that were added to the Pathway List Clipboard in their entirety as well as entities belonging to a particular organism.
7. Click OK.

Your default browser opens Agilent's eArray/SureDesign portal website. You have to log in to the portal with your details to continue. Depending on your choice of **Application** in the dialog, the **Expression** or **RNA Enrichment** page of the eArray portal will appear.

31.3.5.33 To delete entities from the Pathway List Clipboard

You can delete pathway entities from the Pathway List Clipboard.

1. Click the **Pathway List Clipboard Inspector** icon in the **Pathway View Toolbar**.
   The **Pathway List Clipboard Inspector** opens.
2. Select the entities that you want to delete from the Pathway List Clipboard. To delete all entities, use Ctrl-A to select all entities.

3 Click the **Delete selected entities** icon.
The selected entities are permanently removed from the Pathway List Clipboard.

4 Click **Close** to return to the Pathway View.

31.3.6 Working in an MOA Experiment

This section describes the analysis options that are available in an MOA experiment, compared to pathway lists created from an SEA workflow, which are described in the Working with Pathway Lists section.

Unlike the result of an SEA workflow - a list of pathways saved under the input entity list of the chosen experiment - the result of an MOA workflow is an MOA experiment that contains the list of entity lists and interpretations from the input experiments at the time of creating the MOA experiment. The MOA experiment is saved as a new experiment and added to the Experiment Navigator below all the other experiments in the open project.

Due to its status as an independent experiment with entity lists from two different experiment types and a pathway list if Multi-Omic Pathway Analysis was performed, or a Correlation Heatmap if Multi-Omic Correlation Analysis was performed, a limited number of analysis options for these items are available for the MOA experiment in the Workflow Browser and the main toolbar. See "Working in a Multi-Omic Analysis Experiment" on page 609 for details.

31.4 Multi-Omic Analysis Workflow Reference

This section describes the user interface elements of the wizard-driven MOA workflow, toolbar icons, dialog boxes, and right-click options.

31.4.1 The Wizard-driven Multi-Omic Analysis (MOA) Workflow

This section describes the window components of the four steps in the wizard-driven MOA workflow that allow you to choose experiments, pathway organism, pathway sources, interpretations, entity lists, and annotation in Steps 1 and 2. The section also covers the options you have to review the workflow results and save them in Steps 3 and 4.

31.4.2 Multi-Omic Analysis (Step 1 of 4)

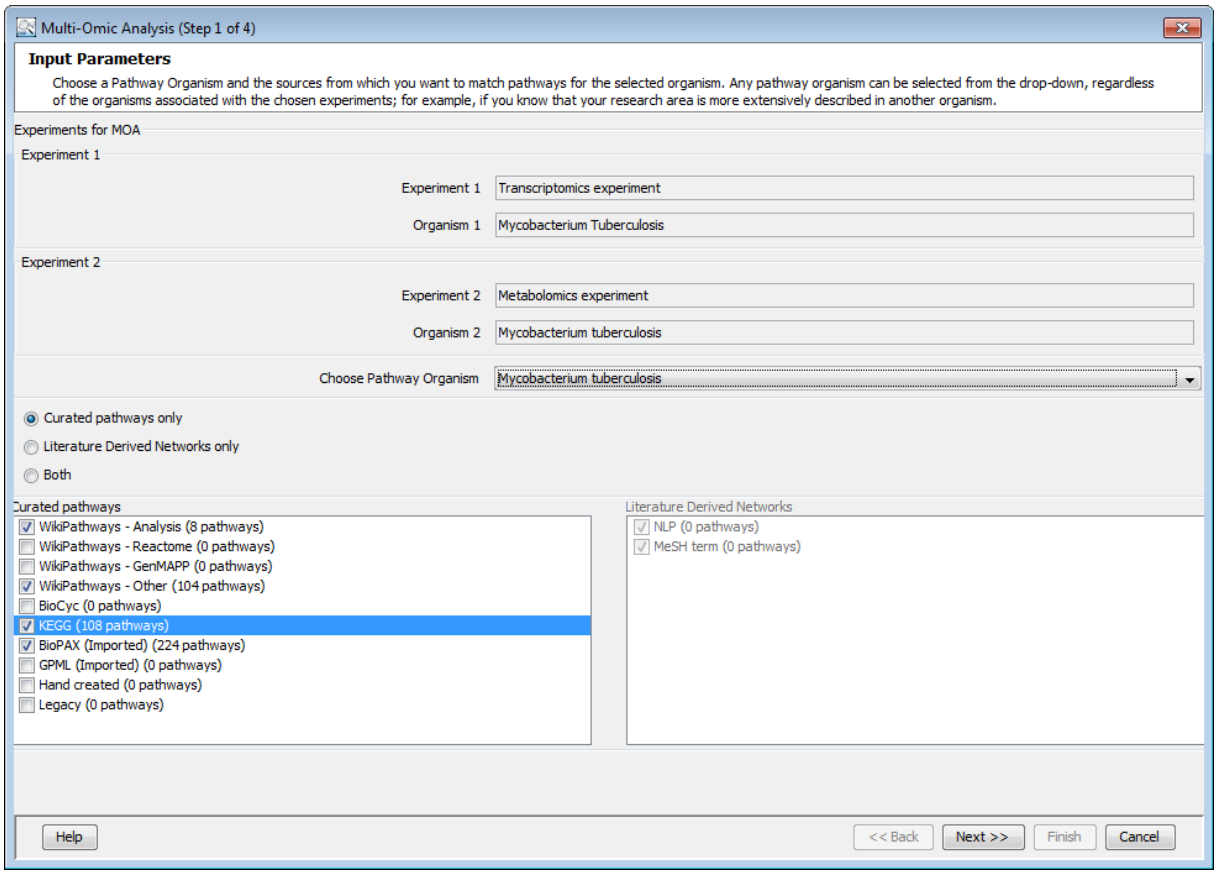


Figure 31.28: Multi-Omic Analysis workflow - Step 1 of 4

In Step 1 of 4 of the wizard-driven MOA workflow you can choose the pathway organism, and the pathway sources for finding matched pathways.

Choose Pathway Organism Use this drop-down menu to select the pathway organism. See To choose a pathway organism on page 901.

Pathway Sources The number of pathways available in **GeneSpring** for the chosen experiment organism is displayed in brackets next to each source. Select the preferred category of pathway sources, from these options:

- Curated pathways
- Literature Derived Networks
- Both

See To select pathway sources on page 903.

Next ≫ Move on to the next step in the workflow.

Cancel Aborts the workflow.

Help Opens the **GeneSpring** manual in a new browser window.

31.4.3 Multi-Omic Analysis (Step 2 of 4)

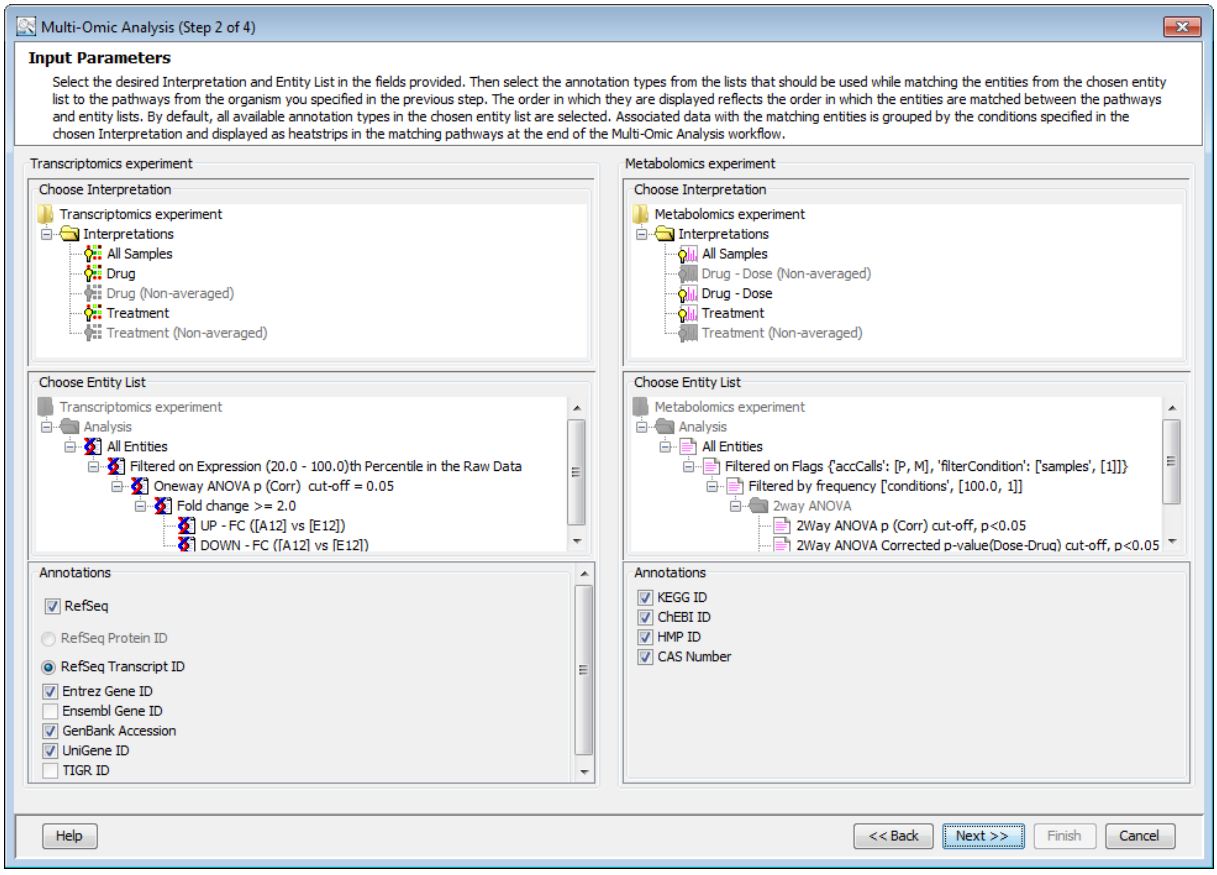


Figure 31.29: Multi-Omic Analysis workflow - Step 2 of 4

In Step 2 of 4 of the wizard-driven MOA workflow you can choose the entity lists and interpretations from each input experiment (selected in Step 1). In the SEA workflow you can only choose one entity list and one interpretation. In this step you can also select the desired annotations from the available annotations in the experiment.

Choose Interpretation	Select desired interpretation.
Annotations	Select desired annotations for matching pathway entities with entity lists.
Choose Entity List	Select desired entity list.
≪ **Back**	Return to the previous step in the workflow.
Next ≫	Move on to the next step in the workflow.
Cancel	Aborts the workflow.
Help	Opens the **GeneSpring** manual in a new browser window.

31.4.4 Multi-Omic Analysis (Step 3 of 4)

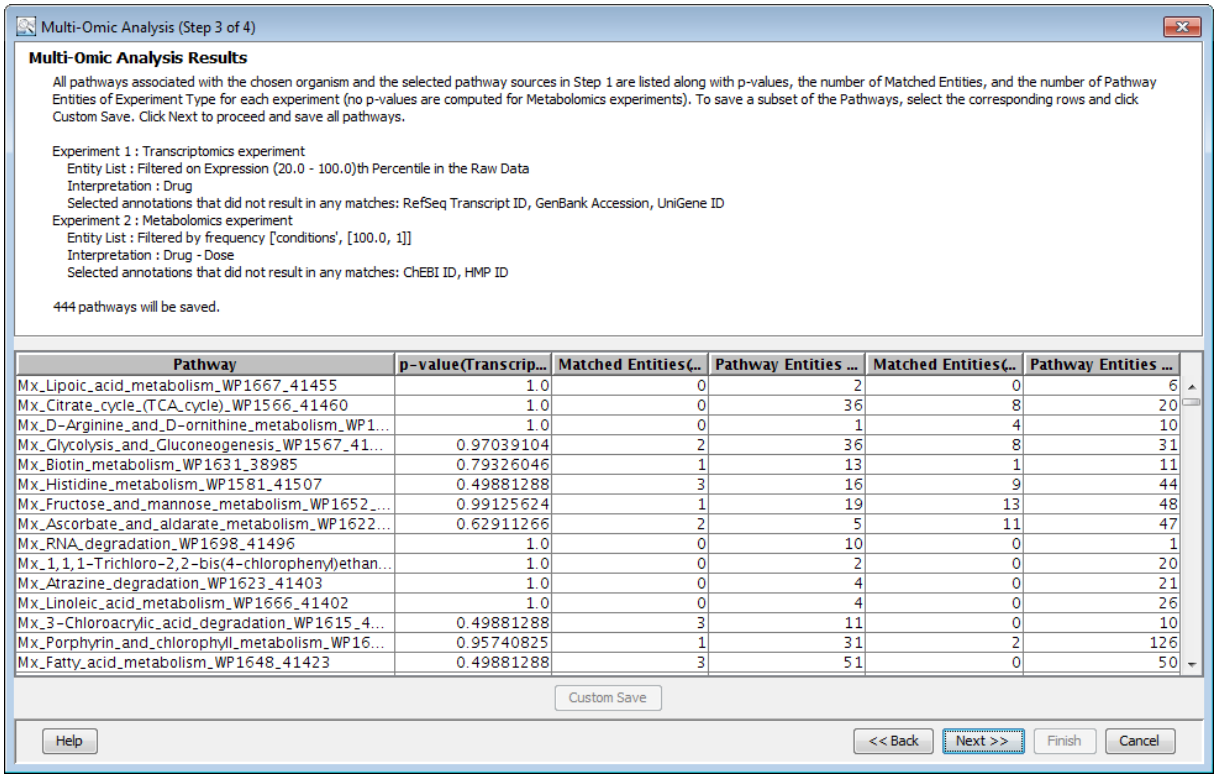


Figure 31.30: Multi-Omic Analysis workflow - Step 3 of 4

In Step 3 of 4 of the wizard-driven MOA workflow all the pathways that are associated with the organism and pathway sources that were chosen in Step 1 are listed as a table. For each matched pathway the additional columns contain p-values that were computed for each input experiment. For unmatched pathways, the default p-value is '1.0' The table also displays the number of matching entities in the input entity list and the total number of pathway entities of the same experiment type as the input experiment. You can custom save a list of selected pathways. Proceed with the workflow to save all the listed pathways as a pathway list.

Results table	Lists all pathways that are associated with the organism and pathway sources that were chosen in Step 1.
Custom Save	Save a selection of the listed pathways as an MOA experiment (MOA workflow only). The selection is saved as a pathway list in an SEA workflow.
≪ **Back**	Return to the previous step in the workflow.
Next ≫	Move on to the next step in the workflow.
Cancel	Aborts the workflow.
Help	Opens the **GeneSpring** manual in a new browser window.

31.4.5 Multi-Omic Analysis (Step 4 of 4)

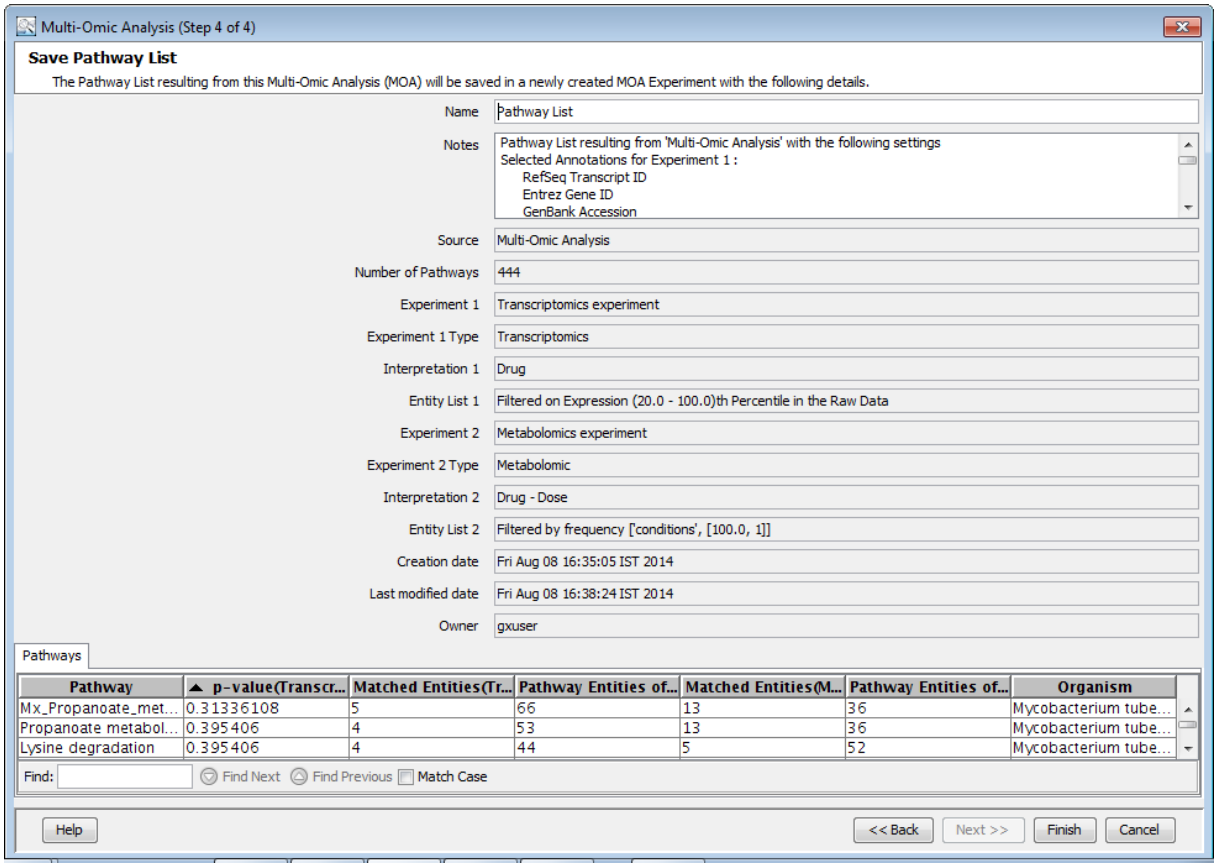


Figure 31.31: Multi-Omic Analysis workflow - Step 4 of 4

Step 4 of 4 of the wizard-driven MOA workflow summarizes the workflow parameters and allows you to name the pathway list and edit **Notes** if required. You can also review the list of pathways that will be saved as a pathway list in the new MOA experiment on clicking **Finish**.

Name	Allows you to change the default name.
Notes	Allows you to review and edit notes.
Pathway table	Lists all pathways that are associated with the organism and pathway sources that were chosen in Step 1.
≪ **Back**	Returns to the previous step in the workflow.
Finish	Completes the workflow and saves the result pathway list in the source experiment for SEA and in the Pathway Analysis folder of the MOA experiment for MOA.
Cancel	Aborts the workflow.
Help	Opens the **GeneSpring** manual in a new browser window.

31.4.6 The Pathway View

The following sections describe all the display components and functional elements present in the Pathway View, including the toolbar icons, different view components, right-click menus, and dialog boxes. In the Pathway View you can analyze, manipulate, and interpret the saved pathway list resulting from MOA and SEA.

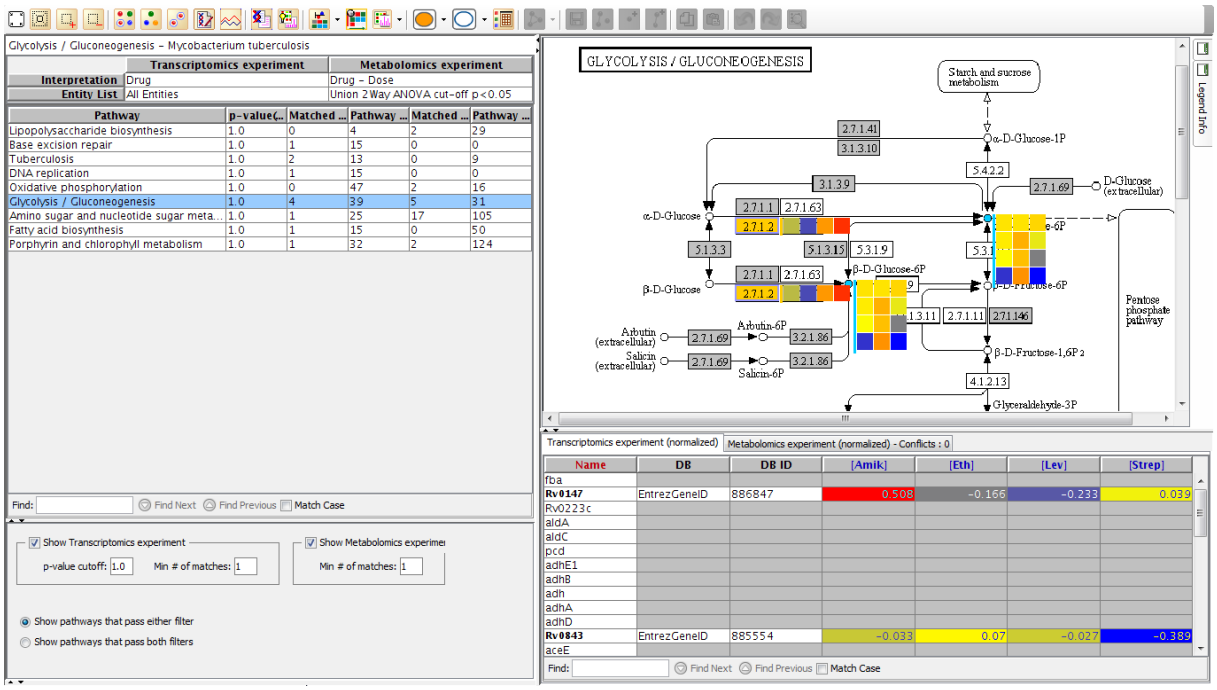


Figure 31.32: Pathway View of a pathway list resulting from MOA

31.4.6.1 Pathway View Toolbar

The Pathway View Toolbar provides access to the most important and common visualization and analysis functions in the Pathway View, including zooming in and out of the pathway, selecting entities, displaying quilt plots or heatstrips in the Pathway Viewer, or changing the display settings of the pathway, quilt plots, heatstrips, and the Heatmap Pane. The availability of each function represented by an icon in the toolbar depends on the pathway format (non-GPML (excluding BioPAX Level 3 and KEGG) or GPML, BioPAX Level 3, and KEGG pathways) and whether you are looking at an existing pathway or creating a new one. Table 31.2 details the icons, their utility, and availability depending on the pathway format. Table 31.3 details the icons, and availability depending on whether you are looking at an existing non-GPML (excluding BioPAX Level 3 and KEGG) pathway or creating a new one.

Table 31.2: Toolbar Icons available in the Pathway View depending on pathway type/format

Toolbar Icon	**Function**	**Non-GPML pathways (excluding BioPAX Level 3 and KEGG)**	**GPML, BioPAX Level 3, and KEGG pathways**
	Fullscreen view	✓	✓
	Zoom to fit view to visible area	✓	✓
	Zoom in	✓	✓
	Zoom out	✓	✓
	Select all entities	✓	✓
	Select matched entities	✓	✓
	Select highlighted entities	✓	✓
	Create Entity List from selection	✓	✓
	Launch Profile Plot on selected entities	✓	✓
	Choose another Entity List	✓	✓
	Choose another Interpretation	✓	✓
	Choose Quilt Plots/Heatstrips to display	✓	✓
	Choose parameters along axes	✓	✓
	Select experiment(s) to display Quilt Plots/Heatstrips	✓	✓
	Select Entity Background Color to display	✓	✓
	Select Highlight display	✓	✓
	View and modify display settings	✓	✓
	Layout graph drop down	✓	*X*
	Save Pathway	*X*	*X*
	Remove unlinked entities from the view	*X*	*X*
	Add entity	*X*	*X*
	Add interaction	*X*	*X*
	Copy selection	✓	*X*
	Paste selection	*X*	*X*
	Undo	*X*	*X*
	Redo	*X*	*X*
	Pathway List Clipboard Inspector (if it contains entities)	✓	✓

Table 31.3: Toolbar Icons available in the Pathway View for existing non-GPML pathways *vs* creating new pathways

Toolbar Icon	**Function**	**Existing non-GPML pathways (excluding BioPAX Level 3 and KEGG)**	**Creating new pathways**
	Fullscreen view	✓	✓
	Zoom to fit view to visible area	✓	✓
	Zoom in	✓	✓
	Zoom out	✓	✓
	Select all entities	✓	✓
	Select matched entities	✓	*X*
	Select highlighted entities	✓	*X*
	Create Entity List from selection	✓	*X*
	Launch Profile Plot on selected entities	✓	*X*
	Choose another Entity List	✓	*X*
	Choose another Interpretation	✓	*X*
	Choose Quilt Plots/Heatstrips to display	✓	*X*
	Choose parameters along axes	✓	*X*
	Select experiment(s) to display Quilt Plots/Heatstrips	✓	*X*
	Select Entity Background Color to display	✓	*X*
	Select Highlight display	✓	*X*
	View and modify display settings	✓	*X*
	Layout graph	✓	✓
	Save Pathway	*X*	✓
	Remove unlinked entities from the view	*X*	✓
	Add entity	*X*	✓
	Add interaction	*X*	✓
	Copy selection	✓	✓
	Paste selection	*X*	✓
	Undo	*X*	✓
	Redo	*X*	✓
	Pathway List Clipboard Inspector (if it contains entities)	✓	*X*

31.4.6.2 Pathway Viewer

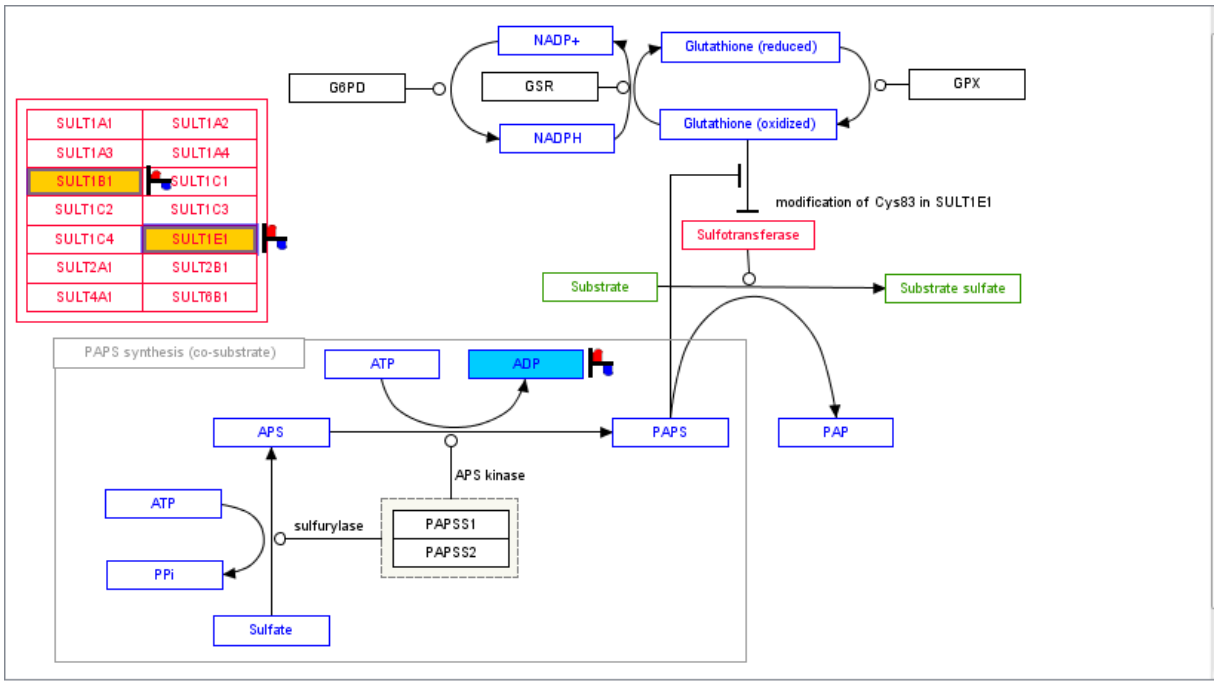


Figure 31.33: Pathway Viewer pane of the Pathway View

The pathway viewer pane displays the pathway that was selected from the pathway list in the Pathway list panel. Details of the pathway (pathway name, pathway organism, associated experiments, entity lists, and interpretations) are displayed above the pathway list panel.

NOTE By default, when the pathway viewer is first launched after SEA or MOA, the first pathway in the pathway list panel is displayed. In all subsequent launches, the last opened pathway is displayed.

Legend Tab in Pathway Viewer

To view the **Legend** explaining the sample/condition layout of quilt plots as well as the color coding for quilt plots and heatstrips and matching pathway entities while interrogating a pathway list in Full Screen mode, click the Legend tab at the top right edge of a pathway in the pathway viewer. Once expanded it contains the same information as the Legend pane that is part of the GeneSpring desktop visible in the ordinary Pathway View.

Interactive Pathway Behavior (All pathways excluding KEGG pathways)

For interactive pathway behavior, which indicates matched entities, highlighted entities, selected entities (see Figure 31.34), and experimental data in the form of Heatstrips and Quilt Plots, the color coding and quilt plot layout is explained in the Legend pane of the application window.

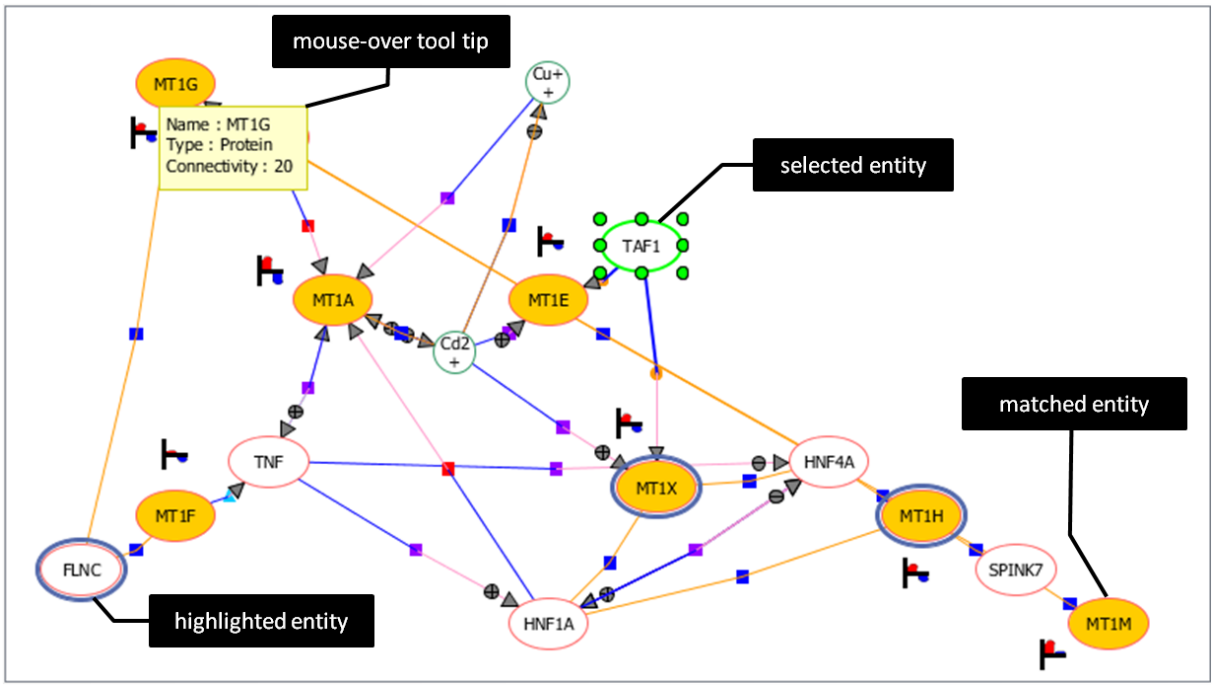


Figure 31.34: Interactive pathway behavior and display

In a pathway viewer, the entities with a blue border (highlighted entities) match the entities in the active entity list in the Experiment Navigator. **GeneSpring** matches highlighted entities for both GPML and non-GPML pathways in the same way as in the MOA workflow to find matched entities, including direct matches and indirect matches via BridgeDb. The only difference between the algorithms used in these two scenarios is that you cannot select annotations to consider for matching of highlighted entities and matching is always done using all available annotations. To switch off the default highlighting of matching entities with the active entity list in the Experiment Navigator use the **Select Highlight display** icon drop-down options.

In the absence of BridgeDb databases and HomoloGene Groups for all participating organisms, this mapping process is based on direct matches only.

You can change the default colors for matching experiments in the Entity tab of the Pathway Properties Dialog, or change the entity fill color settings to another data type (Color by Signal/Abundance Values, or Entity List Associated Values). See To overlay data on a pathway. To switch off entity fill coloring, or toggle between experiments, use the **Select Entity Background Color to display** icon drop-down options.

Interactive Pathway Behavior for KEGG pathways

A KEGG pathway map is a molecular interaction/reaction network diagram

represented in terms of the KEGG Orthology (KO) groups, so that experimental evidence in specific organisms can be generalized to other organisms through genomic information.

Basic graphics objects in the reference KEGG pathway maps are:

- boxes - ortholog (KO) groups identified by K numbers and, in metabolic maps, reactions identified by R numbers as well
- circles - other molecules, usually chemical compounds identified by C numbers, but including glycans identified by G numbers
- lines - reactions identified by R numbers in metabolic maps; ortholog (KO) groups identified by K numbers in global metabolism maps

and in organism specific pathway maps that are computationally generated:

- boxes - genes or gene products identified by the combination of the KEGG organism code and gene identifiers

For each KEGG pathway, only the first reference pathway map is manually drawn (see Figure 31.35) and organism specific maps are computationally generated (see Figure 31.36). For metabolic pathways, each box (or line) in the reference map is linked to the K number (KO identifier), the EC number, and the R number (reaction identifier). For all metabolic and non-metabolic maps, K numbers are converted to gene identifiers in each organism to generate organism-specific pathways (For more information, see: http://www.genome.jp/kegg/kegg3a.html).

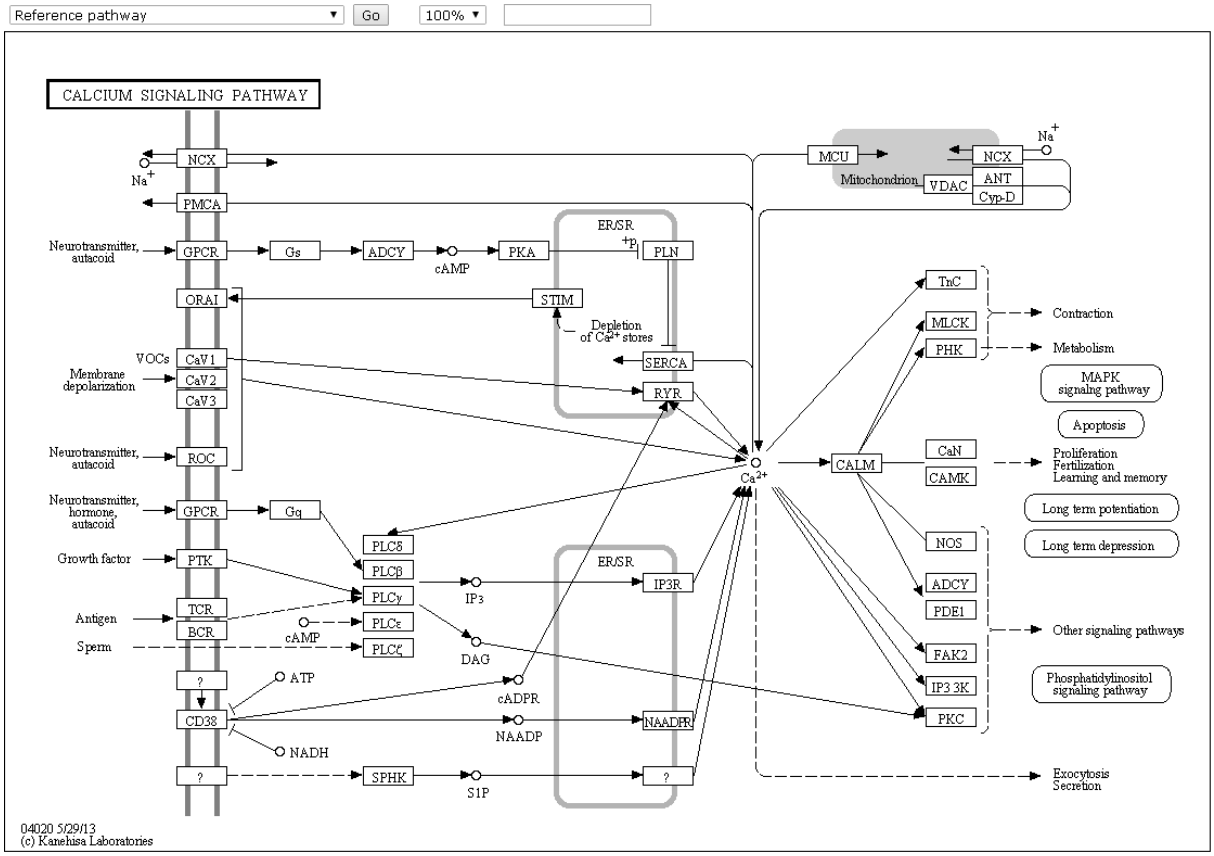


Figure 31.35: Reference KEGG pathway (from http://www.genome.jp/kegg/kegg3a.html)

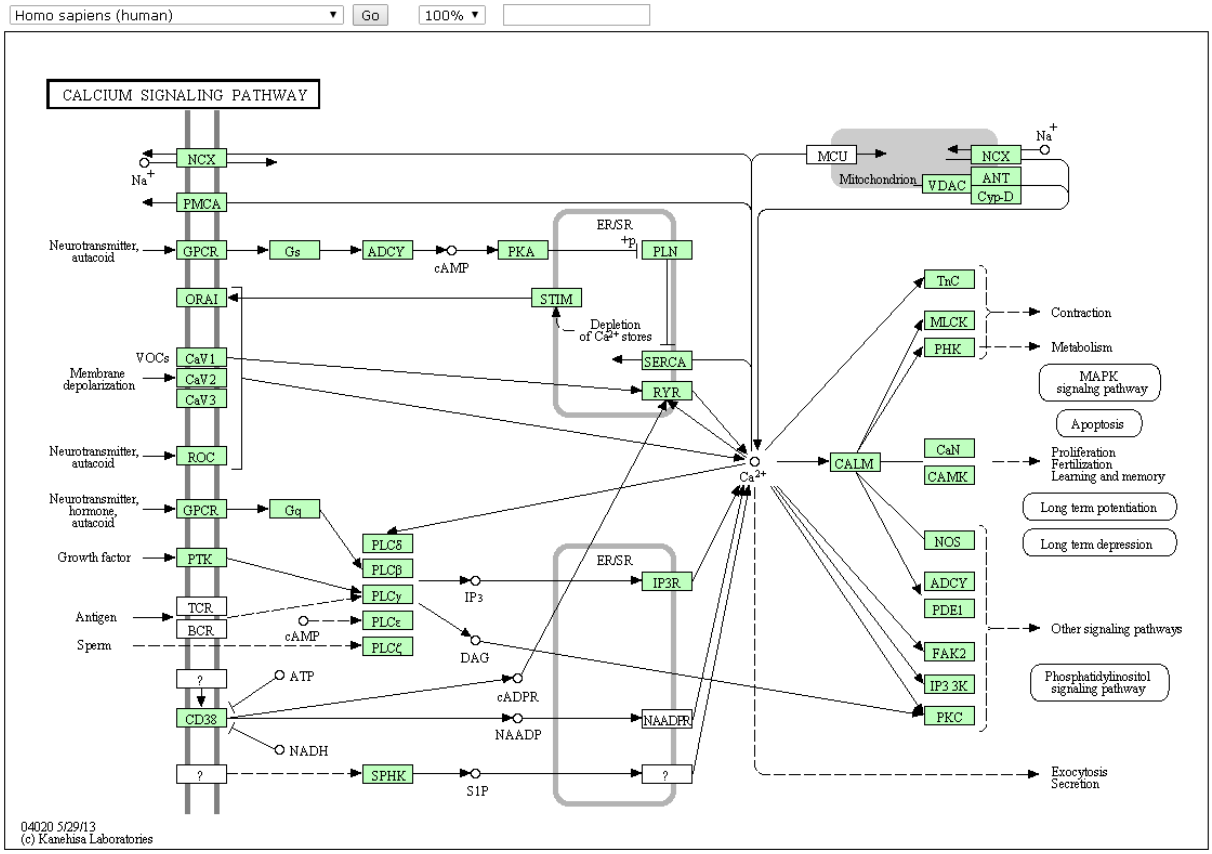


Figure 31.36: Organism specific KEGG pathway derived from reference pathway (from http://www.genome.jp/kegg/kegg3a.html)

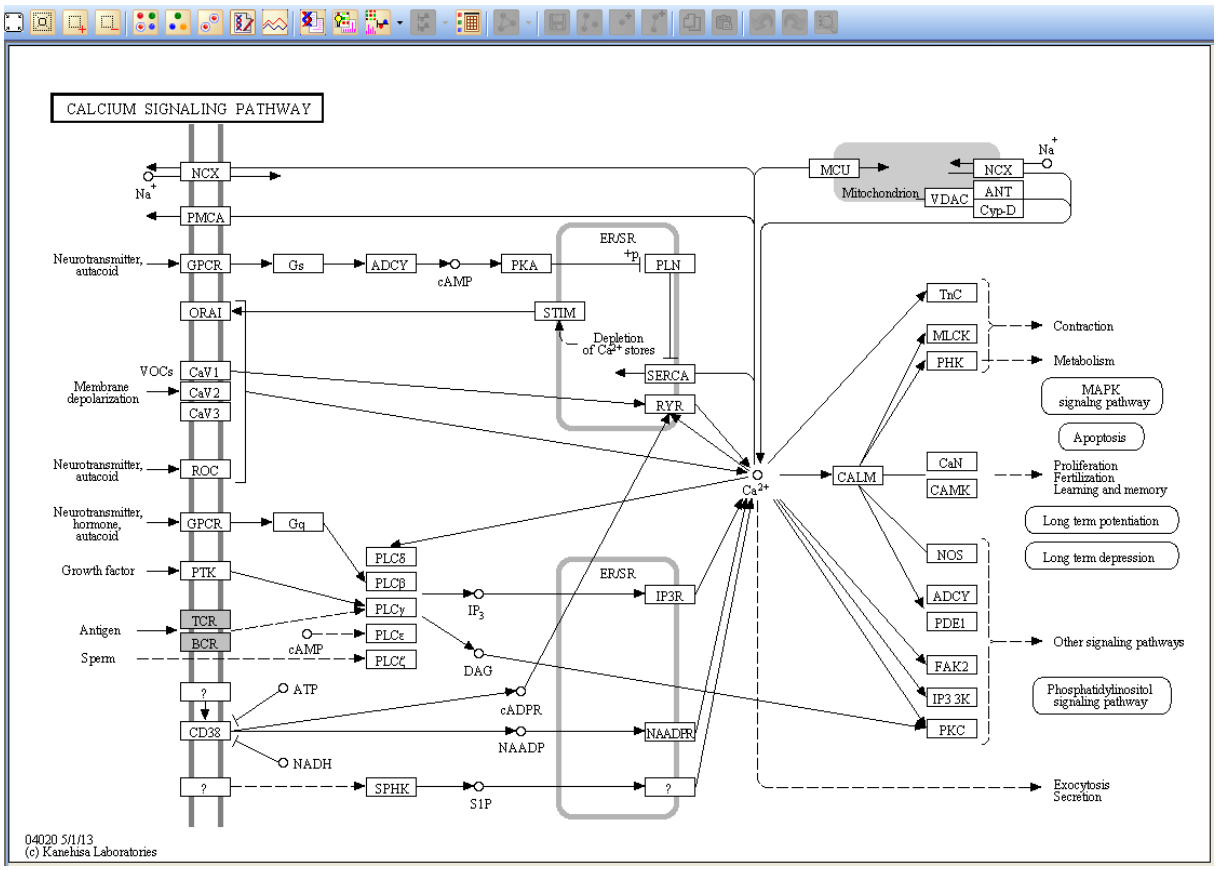


Figure 31.37: Organism specific KEGG pathway in **GeneSpring** (without experiment matching colors)

KEGG pathways in GeneSpring

Only organism specific pathways can be imported and viewed in **GeneSpring** (see Figure 31.37). If a gene identifier corresponding to the K number in the reference map is not available in the pathway organism, it is displayed with a grey background color in **GeneSpring**. Thus, KEGG pathway nodes in **GeneSpring** with grey background color indicate that they are absent or not yet discovered in the pathway organism. For interactive KEGG pathway behavior, which indicates matched entities, highlighted entities, selected entities, absent entities (see Figure 31.38), and experimental data in the form of Heatstrips or Quilt Plots, the color coding is explained in the Legend pane of the application window.

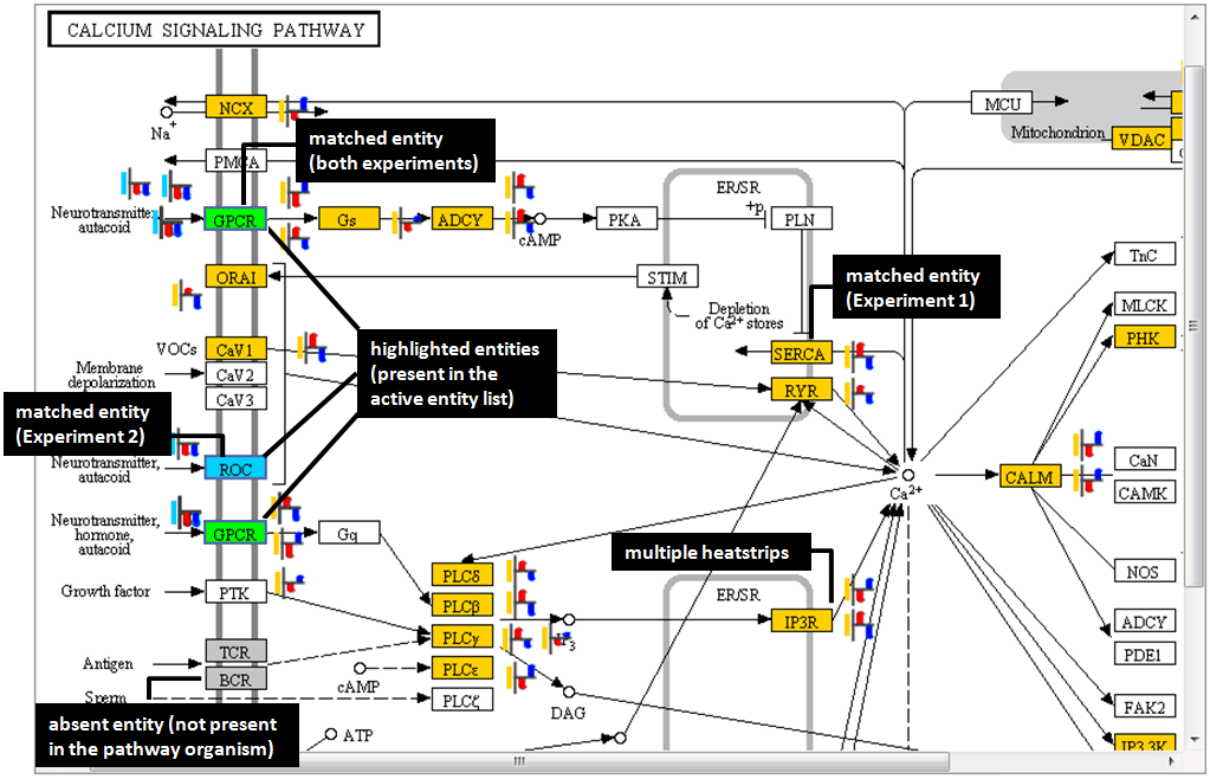


Figure 31.38: Interactive pathway behavior and display of KEGG pathways in **GeneSpring**

In a pathway viewer for KEGG pathways, the entities with a blue border (highlighted entities) match the entities in the active entity list in the Experiment Navigator. You can switch off highlighting using the the **Select Highlight display** icon drop-down options. Selected entities are displayed with a green border, and matched entities for each individual experiment (in MOA) are in Yellow (matching in Experiment 1) and Light Blue (matching in Experiment 2) colors. Entities matching to both the experiment types (in MOA) are shown in green color. These colors are also displayed in the tool tip information and the properties dialog (see Entity Properties - KEGG pathways) of the respective entities. You can change the default colors for matching experiments in the **Entity** tab of the Pathway Properties Dialog, or change the entity fill color settings to another data type (Color by Signal/Abundance Values, or Entity List Associated Values). To switch off entity fill coloring, or toggle between experiments, use the **Select Entity Background Color to display** icon drop-down options.

Some entities in KEGG pathways can have multiple heatstrips or quilt plots for the same experiment as a KEGG pathway entity can represent multiple entities in experiment(s). You can view the details of these multiple entities from the Entity Properties dialog box. Where multiple entities are present, those matching the input experiment are displayed in bold in the tool tip and the properties dialog.

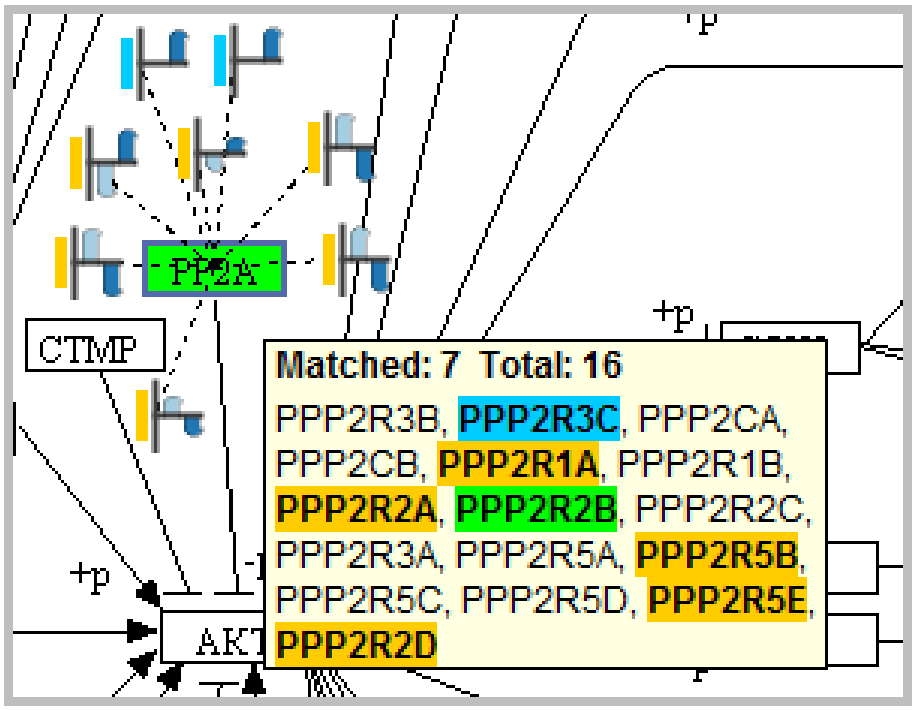


Figure 31.39: Tool tip information for a KEGG pathway entity representing multiple entities that match with both experiments as well as just one of the two experiments in an MOA experiment

31.4.6.3 Pathway components

For GPML pathways, legends or additional notes that explain the pathway components are imported (if available) with the pathway file.

For non-GPML pathways (BioPAX Level 2 and 3, Text format, NLP/hand created and imported from Pathway Architect), the respective legends identifying the pathway components (entities, relations, and edges) are provided in separate tabs on the right edge of the Pathway Viewer pane.

For KEGG pathways, a legend explaining the pathway components (notation) is provided in a tab on the right edge of the Pathway Viewer pane.

Pathways created in **GeneSpring**, or imported in the BioPax Level 2 format, Text format or from Pathway Architect appear as network (graphical) representations of the biological pathways based on the entities and relations present in the Interaction Databases (see Interaction Databases). Similarly, BioPAX Level 3 pathways are also represented as networks, **but do not depend on or are connected with the Interaction Databases**. The entities and relations in each of these networks are represented by the following graphical elements.

Entities

BioPAX Level 2, Text format, NLP/hand created pathways

The following categories (corresponding to the entity types available in the Interaction Database) are used when displaying entities in Text or BioPAX Level 2 format pathways and interaction networks created using NLP:

- **PROTEINS/ GENES**

- **SMALL MOLECULES**
- **ENZYMES**
- **PROCESSES**
- **FUNCTIONS**
- **COMPLEX**
- **FAMILY**

The **Entities** legend (see Figure 32.1) displays the default graphical representation for each type of entity.

You can review information for each entity by double-clicking the entity, which opens its property table (see Figure 31.40). For details on the properties of each entity, see Database Entities on page 1041.

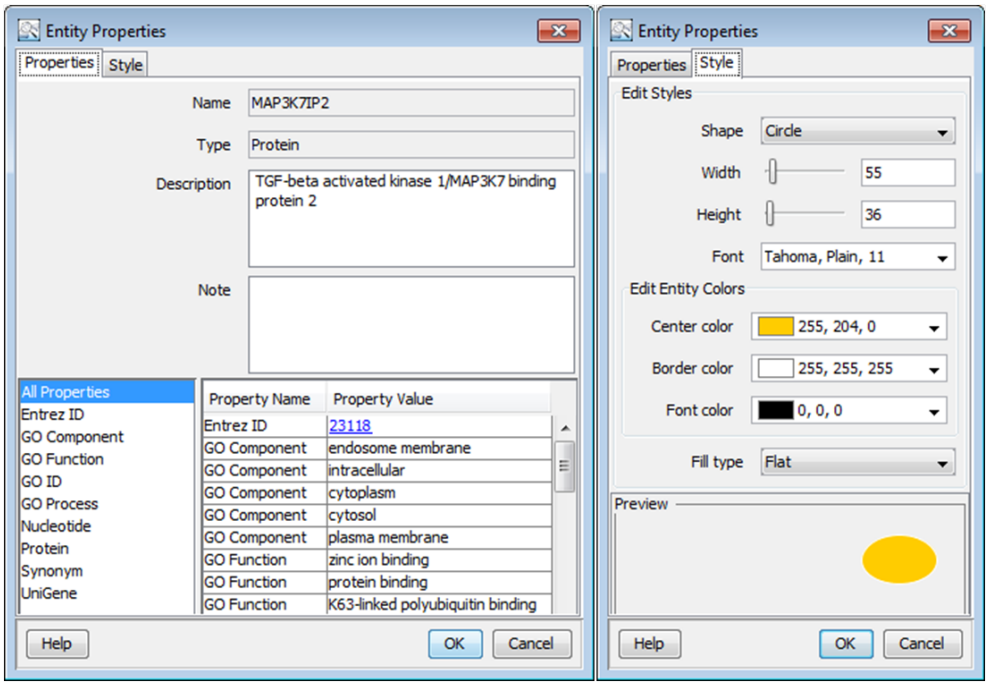


Figure 31.40: Entity Properties

BioPAX Level 3

The following types of entities are present in BioPAX Level 3 pathway networks:

- **PROTEIN**
- **SMALL MOLECULE**
- **COMPLEX**
- **DNA**
- **DNA REGIONS**
- **RNA**
- **RNA REGIONS**
- **GENE**

- **PHYSICAL ENTITY**

They are visualized as text only and can be distinguished by their color, as indicated in the Legend (see Figure 31.41).

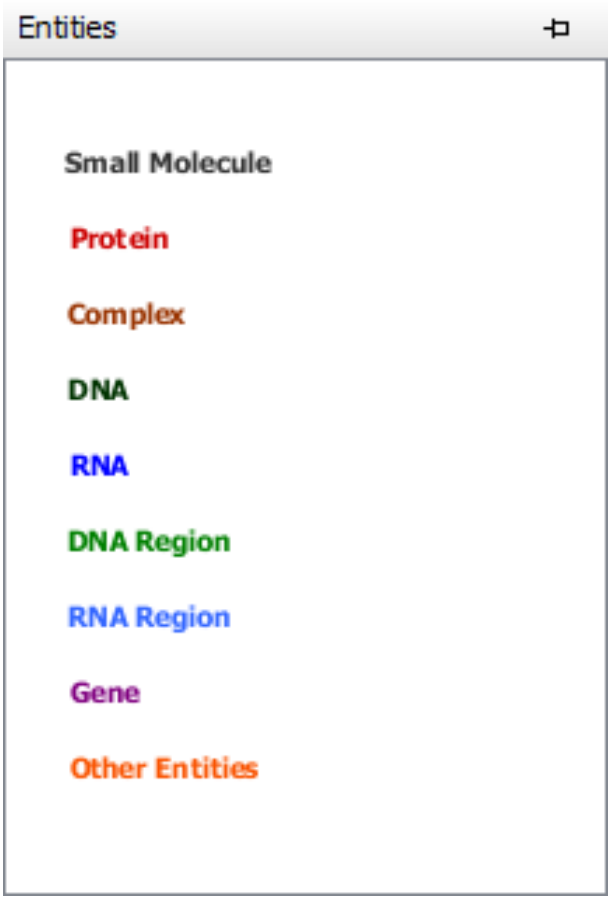


Figure 31.41: BioPAX Level 3 Entities Legend

Double-click or right-click and select **Properties** an entity to inspect its properties or edit the default appearance with the options available from the Style tab of the **Entity Properties** dialog box. For details on the entities in a BioPAX Level 3 pathway, you can consult the documentation available from BioPAX (`http://www.biopax.org/release/biopax-level3-documentation.pdf`).

Relations

BioPAX Level 2, Text format, NLP/hand created pathways

The following molecular relations are distinguished in Hand created networks and pathways in the BioPAX (Level 2) and Text format and pathways imported from Pathway Architect in **GeneSpring**: BINDING, REGULATION, PROTEIN MODIFICATION, EXPRESSION, PROMOTER BINDING, METABOLISM, TRANSPORT, MEMBER (see Relations on page 1045). They are represented as illustrated in Figure 31.42.

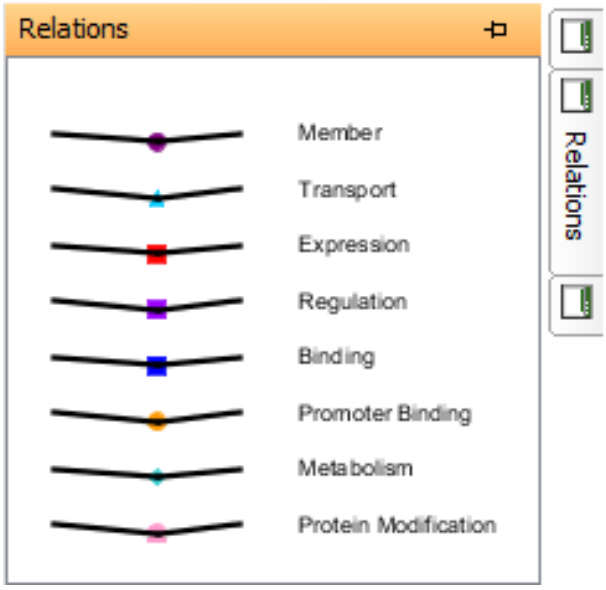


Figure 31.42: Relations Legend

The relations between various molecules and/or processes are represented by the edges of the network as illustrated in Figure 31.43.

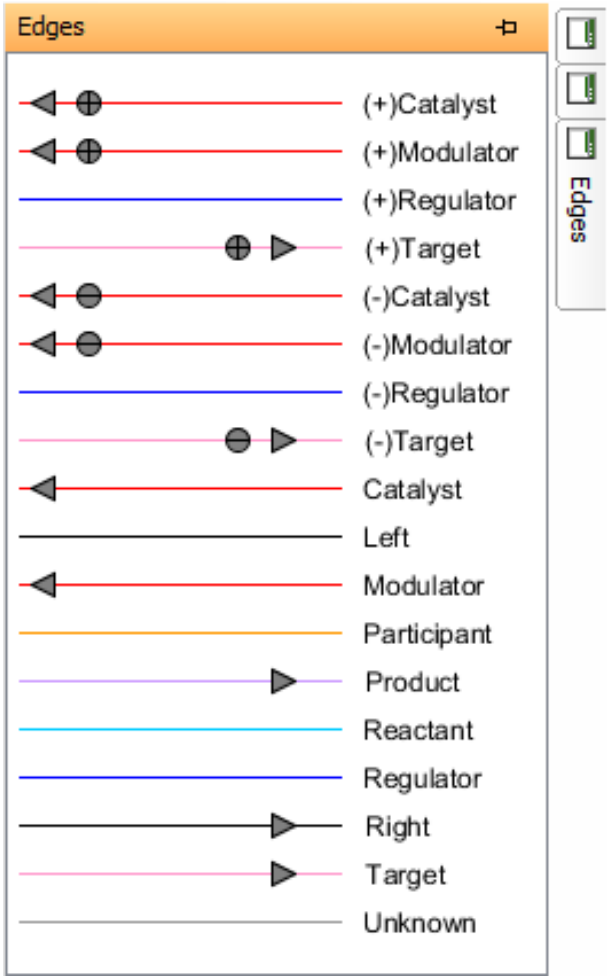


Figure 31.43: Edges Legend

The properties of each relation can be viewed in a table by double-clicking on the relation.

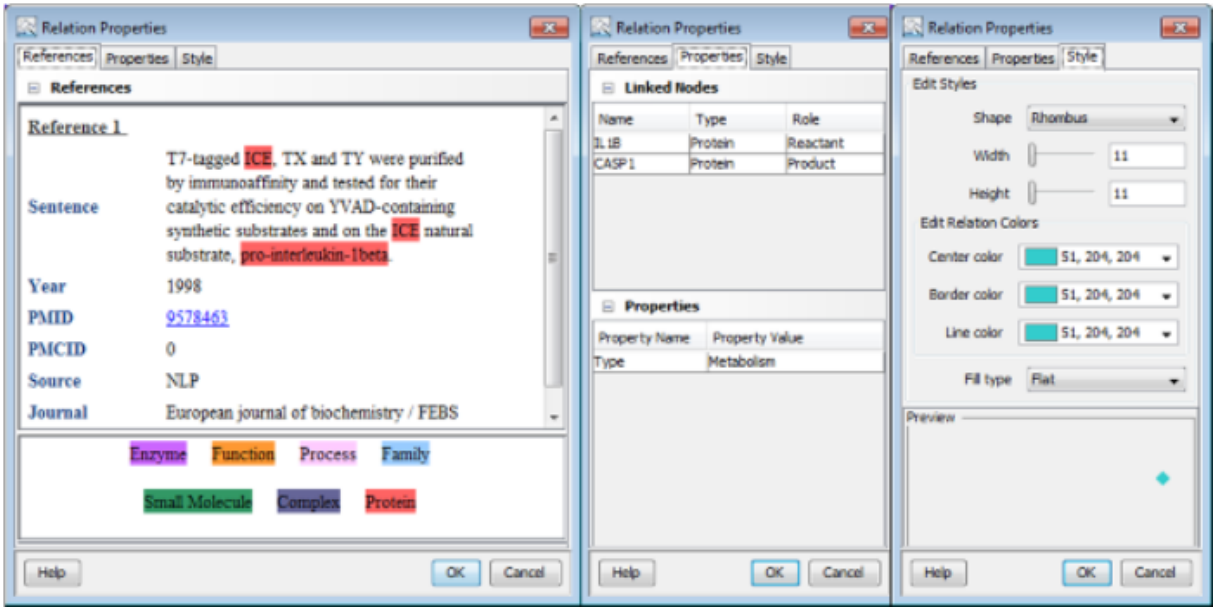


Figure 31.44: Relation Properties

Each relation is characterized by the following criteria:

1. Participating entities and their roles (see Table 32.2 for more details). The edges of the relation specify the roles of the participating entities.
2. Mechanisms, for example, phosphorylation, ubiquitination, etc. The full list of mechanisms available in **GeneSpring** is provided in Table 32.3
3. Score that determines the quality of the relation. The number of references, source of the data and a reference score for each associated reference contribute to a relation score.
4. Each reference that contributes to the relation is associated with the Journal, Year of publication and a reference score. The exact sentence that was interpreted by the NLP can also be viewed.

Further details on relation properties are described in Relations on page 1045.

BioPAX Level 3

The following molecular relations are distinguished in BioPAX (Level 3) pathways in **GeneSpring**: COMPLEX ASSEMBLY, TRANSPORT, GENETIC INTERACTION, BIOCHEMICAL REACTION, TEMPLATE REACTION, DEGRADATION, TRANSPORT WITH BIOCHEMICAL REACTION, MOLECULAR INTERACTION.

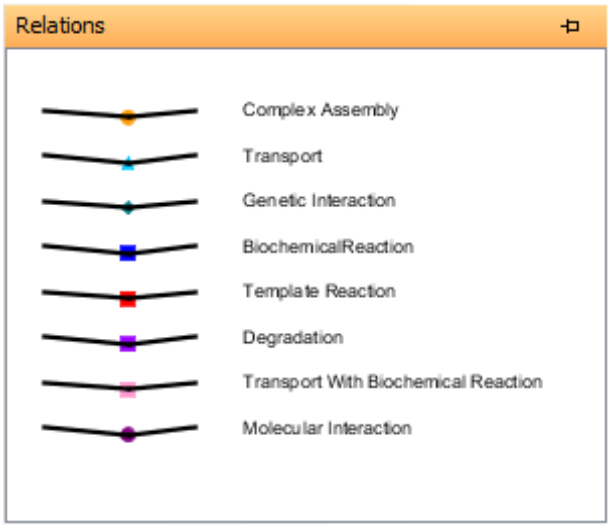


Figure 31.45: Relations Legend for BioPAX Level 3 Pathways

The edges of a BioPAX Level 3 pathway are the same as above (see Figure 31.43), and Properties can also be inspected on right-click. For details on the relations in a BioPAX Level 3 pathway, you can consult the documentation available from BioPAX (`http://www.biopax.org/release/biopax-level3-documentation.pdf`).

Notation

KEGG pathways

KEGG pathways display a multitude of graphical representations for the type of entities (gene product, proteins, RNA, chemical compounds, glycan, drug, and small molecules) in the pathway and the relations (interactions) between them. Individual KEGG pathways also display a link (not active in **GeneSpring**) to other KEGG pathways wherever applicable.

The **Notation** legend (see Figure 31.46) displays the default graphical representation for the entities and the relations in KEGG pathways.

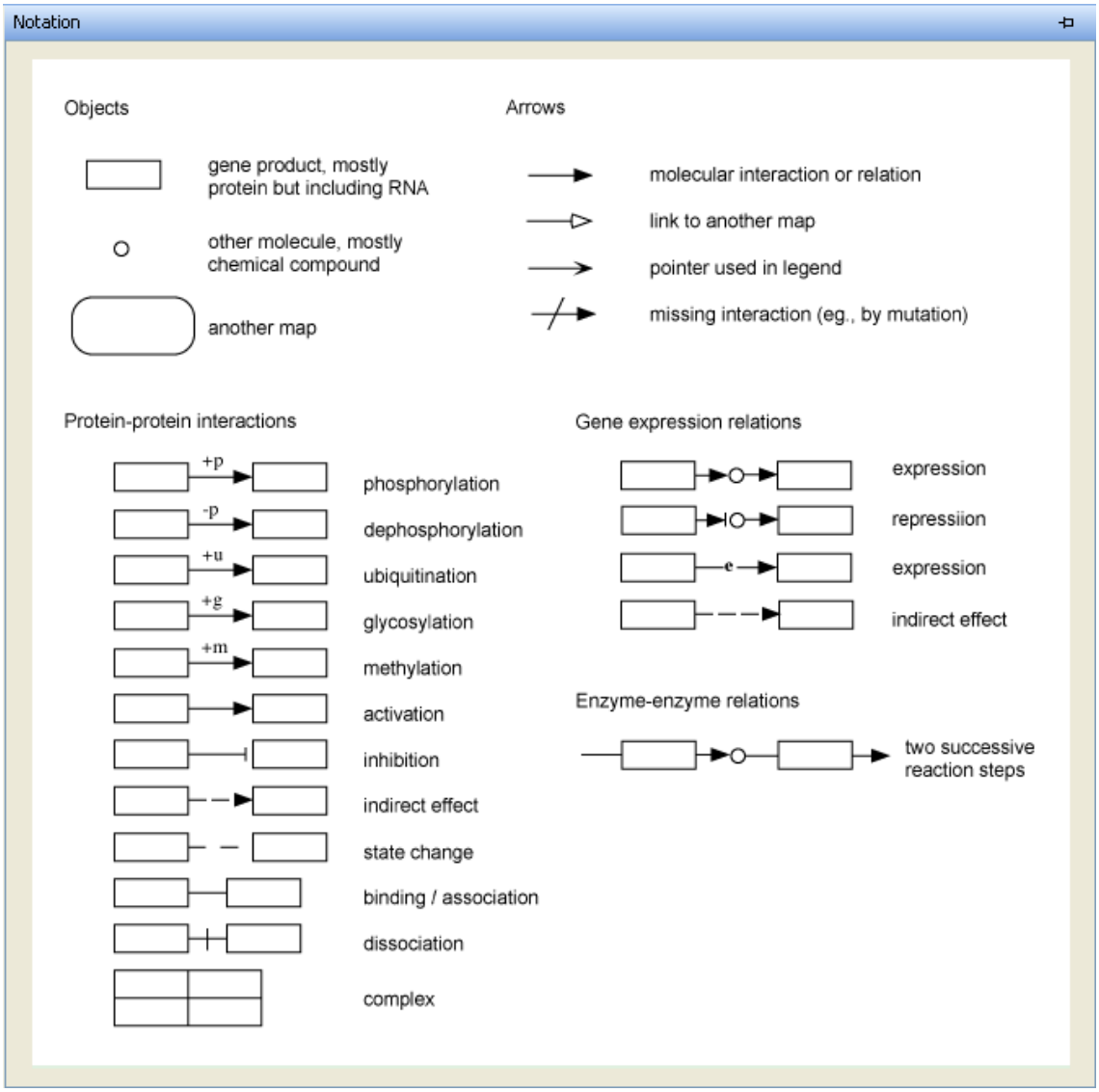


Figure 31.46: Notation Legend for KEGG pathways

You can review information for each entity by double-clicking the entity, which opens its property table (see Figure 31.47). An Enzyme tab is present for entities of KEGG Metabolic pathways to review the enzyme information of the selected entity (see Figure 31.48).

For details on the entities in a KEGG pathway, you can consult the documentation available on the KEGG website (`http://www.genome.jp/kegg/`).

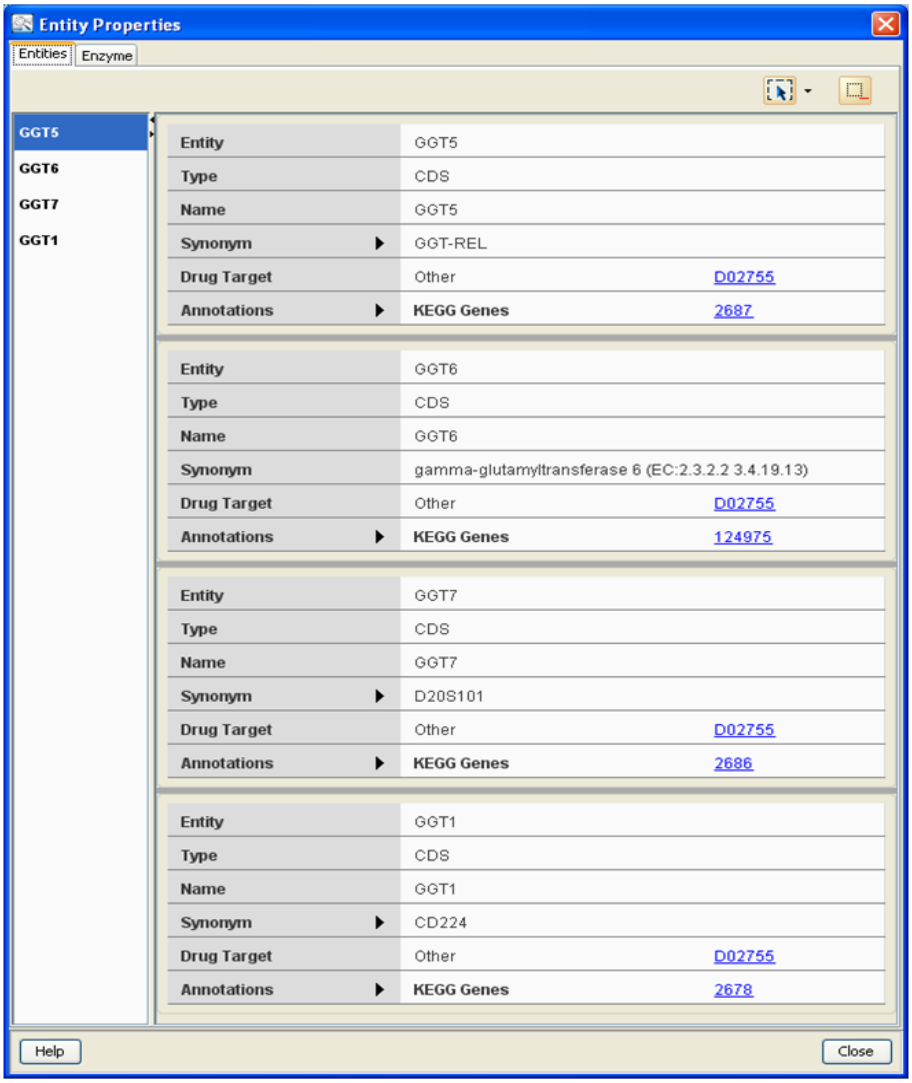


Figure 31.47: KEGG Entity Properties

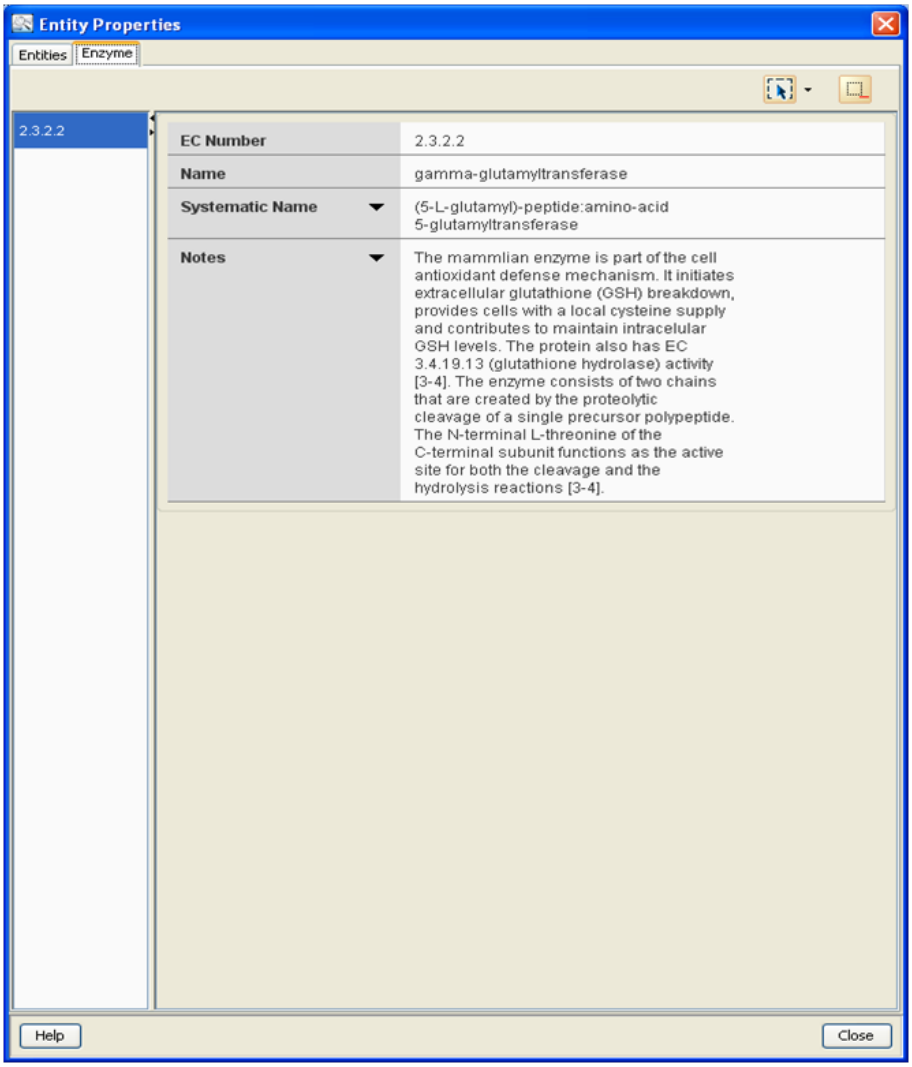


Figure 31.48: KEGG Entity Properties with Enzyme information

Pathway Themes

You can customize the pathway image. The shapes, colors, fonts, and formats of entities, relations, and edges of pathways in the BioPAX and Text format and pathways imported from Pathway Architect, can all be altered using the **Edit Pathway Theme** dialog (see Figure 31.49), by selecting **Tools** $\longrightarrow$ **Edit Pathway Theme** from the main menu. Use the options in the **NLP Networks Theme** tab to change the appearance of pathways in the BioPAX Level 2 and Text format and pathways imported from Pathway Architect. Use the options in the **BioPAX Level 3 Theme** tab to change the appearance of BioPAX Level 3 pathways.

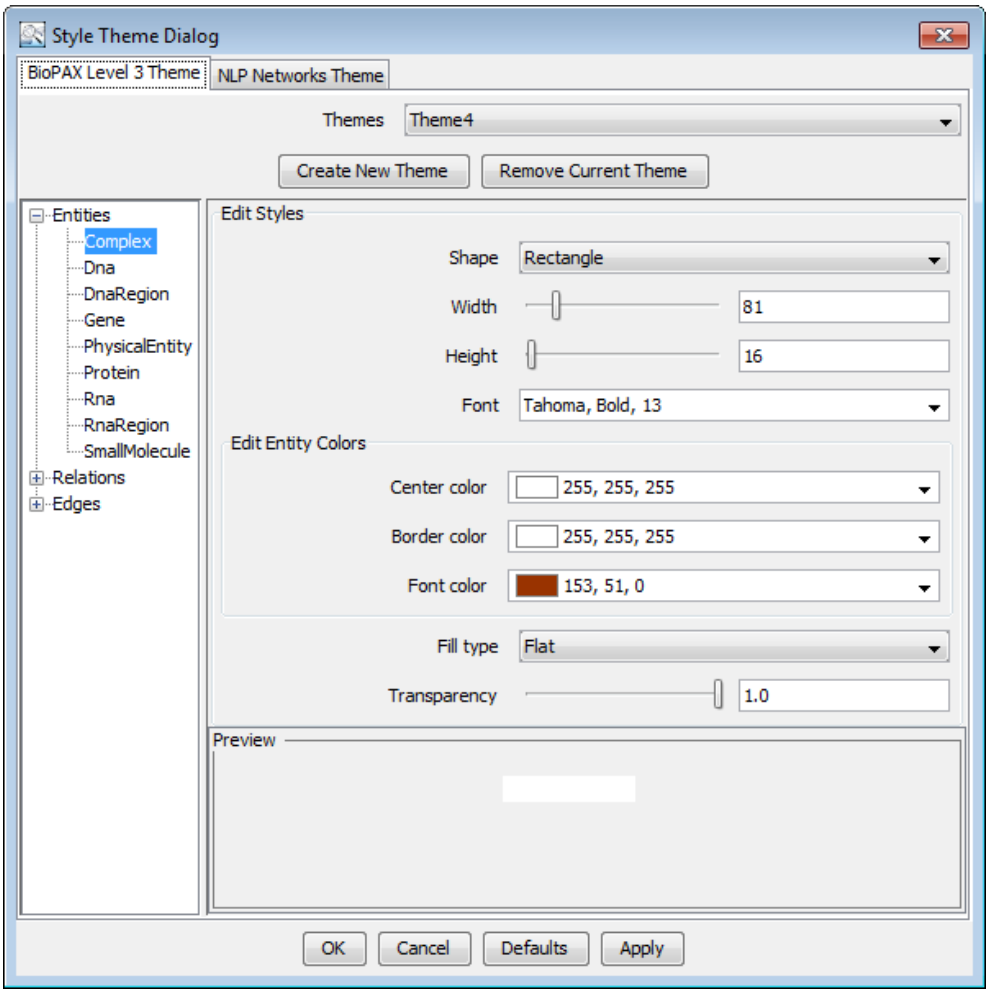


Figure 31.49: Style Theme Dialog

Click **Apply** to observe changes in a displayed pathway or click **OK** to apply the changes and close the dialog box and return to the pathway view.

NOTE After editing a Pathway Theme, the indicator for matching entities (by default this is a yellow background) is lost. Refresh the pathway by double-clicking its name in the pathway list panel to retrieve this indicator.

The appearance of individual entities and relations can be altered using the options available in the **Style** tab of the **Entity properties** dialog (see Figure 31.40) and **Relation Properties** (see Figure 31.44) dialog, respectively.

Pathway Layouts

Any entity or relation can be dragged and moved around the view as desired. To

move multiple entities and relations, select them and drag the selection to the desired position in the pathway viewer.

In addition, **GeneSpring** offers a choice of six different pathway layouts. The Layout graph icon in the Pathway View Toolbar has several drop-down options for layouts. These are:

- QuadTree FDP
- Neato
- FDP
- Dot
- Twopi
- Cellular (disabled for BioPAX Level 3 pathways)
- Simple (available for BioPAX Level 3 pathways only)

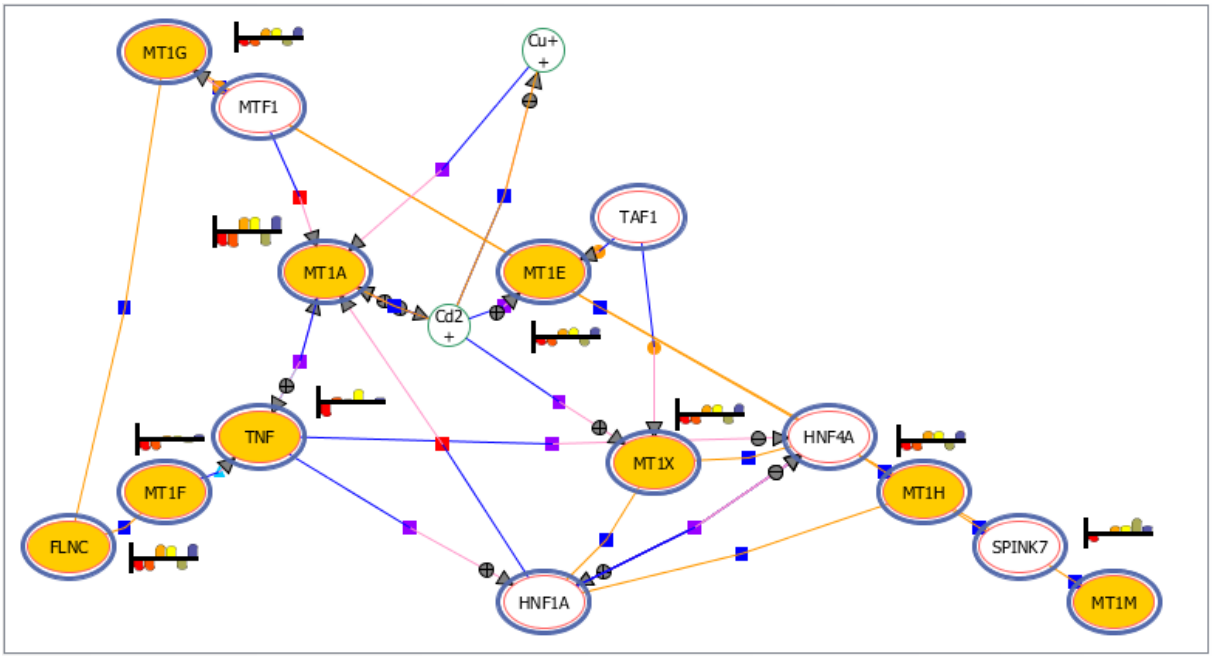


Figure 31.50: Twopi layout

The cellular layout includes information of the proteins sub-cellular localization within a cell. This information is derived from the Cellular component of Gene Ontology. Cellular localization is described in Database Entities on page 1041.

31.4.6.4 Heatstrips and Quilt Plots in the Pathway Viewer

GeneSpring offers two options for overlay of experiment data of matching entities (pathway entities that are also present in the input entity list) on pathways: Quilt Plots (default display from version 14.0) and heatstrips. You can switch between the two options using the **Choose Quilt Plots/Heatstrips to display** toolbar icon.

Quilt Plots

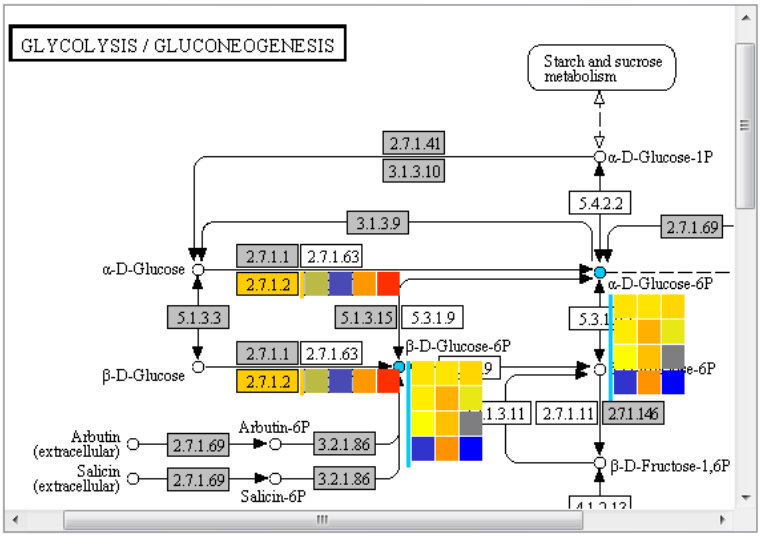


Figure 31.51: Quilt plots as displayed in the pathway viewer of a KEGG pathway in a MOA experiment

A quilt plot in the pathway viewer displays experiment data for an entity from the selected experiment or experiments for all conditions in the selected interpretation. By default, normalized data for the entity is displayed in a quilt plot. The **Quilt Plot Layout** tab in the **Legend** illustrates which parameter is drawn on which axis. You can change which parameters in the interpretation are drawn across the x-axis and y-axis according to your own requirement or preference using the **Choose parameters along axes** toolbar icon.

The quilt plot contains as many cells as conditions in the selected interpretation. Each cell of the quilt plot corresponds to a condition and is colored according to the same color range that is used for coloring the cells in the Heatmap Pane of the Pathway View based on the abundance value for the shown condition.

If both the input entity lists in an MOA experiment contain data for a given entity, two quilt plots (one for each experiment) are displayed in the pathway viewer for the same entity. The corresponding experiment is indicated by a colored bar to the left of the quilt plot. You can hide this indicator using the Display Settings options (see To view and modify display settings on page 923).

Heatstrips

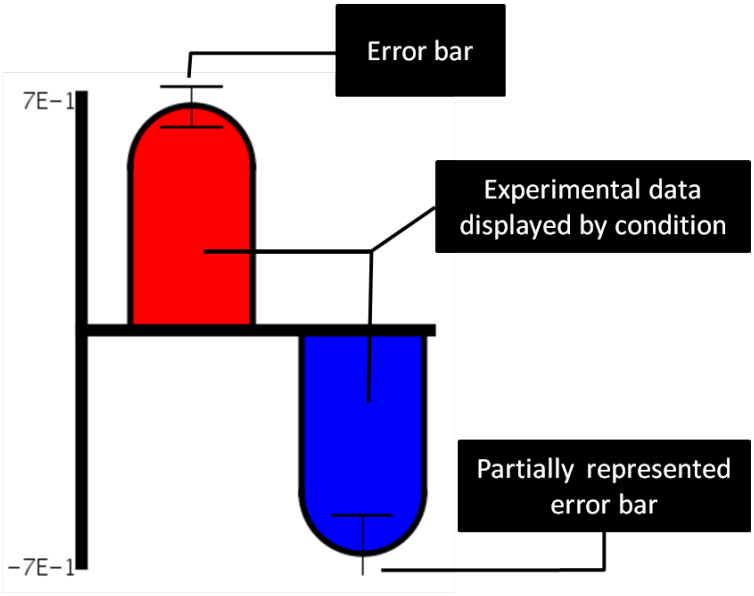


Figure 31.52: A sample heatstrip as displayed in the pathway viewer

A heatstrip displays data for an entity from the selected experiment or experiments across all conditions in the selected interpretation along with their respective standard errors in the form of error bars. By default, normalized data for the entity is displayed in a heatstrip. The error bar on a heatstrip indicates the standard error of the mean for samples within a condition.

See Heatstrip Error Bar Computation on page 1014 for more information on how the standard errors are calculated.

Each colored bar on the heatstrip corresponds to a condition in the selected interpretation and the heatstrip contains as many bars and colors as conditions. You can change the colors of the heatstrips (see To view and modify display settings on page 923).

An individual bar on a heat strip when selected highlights the respective interpretation's bar in all the displayed heatstrips. The height and direction of each bar in a heatstrip reflects the data values of the entity in the corresponding condition. The scale is adjusted for each heatstrip individually, based on values for that entity across all conditions in the interpretation. Error bars that exceed the highest value of the displayed experimental data are only represented partially (see Figure 31.52).

If both the input entity lists in an MOA experiment contain data for a given entity, two heatstrips (one for each experiment) are displayed in the pathway viewer for the same entity. The corresponding experiment is indicated by a colored bar to the left of the heatstrip (see Figure 31.38). You can hide this indicator using the Display Settings options (see To view and modify display settings on page 923).

31.4.6.5 Heatmap Pane

The Heatmap Pane is located below the pathway viewer and is a spreadsheet that lists all the pathway entities of the experiment type in the **Pathway Entity** column. Pathway entities that matched with the input entity list are displayed in bold in the Heatmap Pane spreadsheet.

The other columns and their contents vary depending on the technology of the input experiment, the interpretation that was selected during the MOA or SEA workflow, and whether HomoloGene Translation for Gene/Protein Identifiers was required during the workflow because multiple organisms were selected.

The annotation and identifier displayed in the **DB** and **DB ID** columns depend on whether HomoloGene translation was required to map the pathway entities with the experiment entities or not:

- If HomoloGene translation took place, these columns contain Entrez Gene IDs for the pathway organism for all matched entities.
- If no HomoloGene translation was required, these columns contain the annotation and identifier of the input experiment that led to a match with the pathway. See To select annotations for more details on the mapping process.

For all technology types, experimental data is grouped by the conditions of the chosen interpretation. The cells that contain experimental data are colored according to a default Heatmap Pane view, which is explained in the corresponding **Heatmap** tab of the Legend.

The way the experimental data is displayed in a Heatmap Pane depends on the technology type of the experiment:

Metabolomics: The Heatmap Pane groups isomers and corresponding data in the experiment under a single 'Parent'. See To select the appropriate isomer for heatstrip and quilt plot data overlay in a pathway when multiple isomers in the experiment are mapped to only one entity in the pathway on page 926 for more information on this feature and on how to select the desired isomer for data overlay in the pathway.

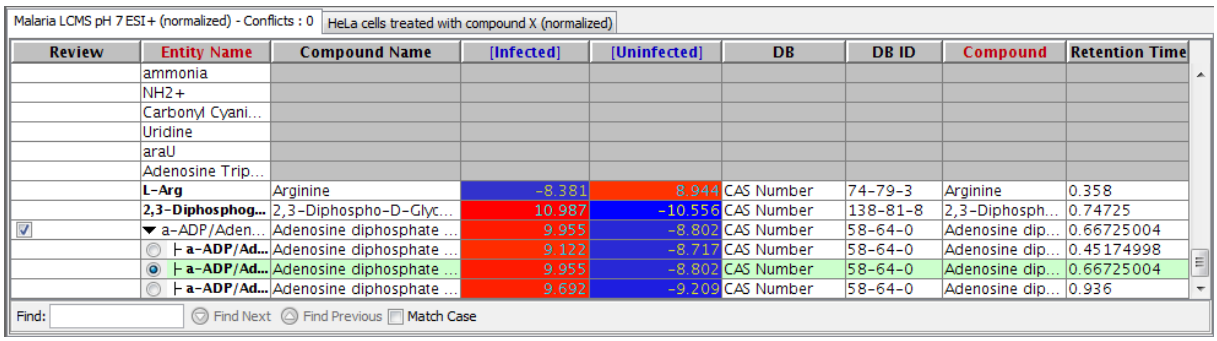


Figure 31.53: A Heatmap Pane containing data from a Metabolomics experiment

Transcriptomics and Proteomics: If multiple entities (with identical identifiers in the experiment (probes/compounds) map to the same pathway entity, the data associated with these entities is summarized and the arithmetic mean of individual probes is displayed in the Heatmap Pane if normalized values are selected in the data settings. For raw values, the geometric mean is displayed in the Heatmap Pane spreadsheet.

The default setting for displaying experimental data in the Heatmap Pane is up to three decimal places. Any values that start with '0.000...' or '-0.000...' appear as '0' or '-0', respectively, regardless of the number at the fourth decimal place. You can change this default display setting by selecting **Properties** on right-clicking

over the Heatmap Pane. In the **Properties** dialog, choose the desired option from the **Display precision** drop-down in the **Visualization** tab.

NOTE Pathway entities with missing data source or identifiers are excluded from the analysis and are not displayed in the Heatmap Pane. In addition, **GeneSpring** only recognizes certain data source labels depending on the pathway format (see Supported Data Sources).

Furthermore, if multiple entities in transcriptomics or proteomics experiments map to the same pathway entity, the calculation of the displayed value varies between matched entities and unmatched entities:

- For **matched** entities (displayed in bold) only the signal values available for the probes/compounds present in the **input entity list** are summarized.
- For **unmatched** entities the signal values of the probes/compounds present in the **experiment** are summarized.

CAUTION Using the **Create Entity List** function to create an entity list from **Matched Entities**, **Selected Entities**, or **Highlighted Entities** will always create an entity list with all the probes in the experiment, not just those present in the input entity list. To discard any entities that had been filtered out by previous analysis steps leading up to the input entity list, you can use the Venn Diagram functionality.

31.4.6.6 Pathway list panel

The pathway list panel on the left side of The Pathway View displays the list of pathways contained in the open pathway list in a spreadsheet format.

If the pathway list was created using the SEA workflow there are four columns in the spreadsheet containing the following information:

- Pathway Name
- p-value (name of Experiment 1)
- Matched entities (name of Experiment 1)
- Pathway entities of Experiment Type (name of Experiment 1)

In a pathway list from an MOA experiment the spreadsheet contains an extra three columns, which contain the data for the second experiment.

- p-value (name of Experiment 2)
- Matched entities (name of Experiment 2)

- Pathway entities of Experiment Type (name of Experiment 2)

The **Matched entities** columns contain the number of entities in the pathway that matched with the input entity list using the selected annotations during the SEA or MOA workflow.

The **Pathway entities of Experiment Type** columns contain the number of entities in the pathway that are of the experiment type: the number of genes and proteins in the pathway if the experiment type is Genomics/Proteomics/Transcriptomics and the number of metabolites if the experiment type is Metabolomics, regardless of whether the entities are present in the technology or not.

Pathway	p-value(HeLa cells treated...	Matched Entities(HeLa cell...	Pathway Entities of Experi...	Matched Entities(Malaria L...	Pathway Entities of Experi...
Rn_Renin_-_Angiotensin_System_WP376_41750	0.022946538	5	56	0	3
Rn_Catecholamine_synthesis_WP513_41663	0.1292645	1	5	0	12
Rn_TNF-alpha_and_mucus_production_in_lun...	0.5490754	1	26	0	0
Rn_SIDS_Susceptibility_Pathways_WP1304_41...	0.043908074	5	60	0	4
Rn_The_effect_of_Glucocorticoids_on_target_g...	0.043346554	2	14	0	0
Rn_Nuclear_receptors_in_lipid_metabolism_an...	0.6208322	1	32	0	5
Rn_p38_MAPK_Signaling_Pathway_(BioCarta)_...	0.6462351	1	34	0	0
Rn_p53_pathway_WP655_41760	0.1706993	3	46	0	0
Rn_Matrix_Metalloproteinases_WP278_41770	0.010900854	4	29	0	0
Rn_Complement_and_Coagulation_Cascades_...	0.08949964	4	62	0	1
Rn_Striated_Muscle_Contraction_WP316_41640	0.09382169	3	36	0	0
Rn_Adipogenesis_WP155_41714	0.019233914	9	130	0	0
Rn_p53_signal_pathway_WP656_41652	0.5644363	1	30	0	3
Rn_Endochondral_Ossification_WP1308_41790	0.116580546	4	61	0	2
Rn_Monoamine_GPCRs_WP276_41763	0.017966514	4	33	0	4
Rn_Toll-like_receptor_signaling_pathway_WP1...	0.2968436	4	91	0	0
Rn_Tryptophan_metabolism_WP270_41758	0.44408855	2	45	0	0
Rn_Cytokines_and_Inflammatory_Response_(B...	0.009473142	4	26	0	0
Rn_Mitochondrial_LC-Fatty_Acid_Beta-Oxidati...	0.10121488	2	17	0	1

Find: Find Next Find Previous Match Case

Figure 31.54: A pathway list panel from the Pathway View of a pathway list resulting from Multi-Omic Pathway Analysis

31.4.6.7 Filter panel

With the options provided in this panel you can filter pathway lists by p-value cut-off or the minimum number of matches for each input entity list in order to identify significant pathways. This panel is only available for pathway lists created from an SEA or MOA workflow. The functionality of this panel is described in To filter pathway lists by p-value cut-off on page 914 and To filter pathway lists by minimum number of matches on page 915.

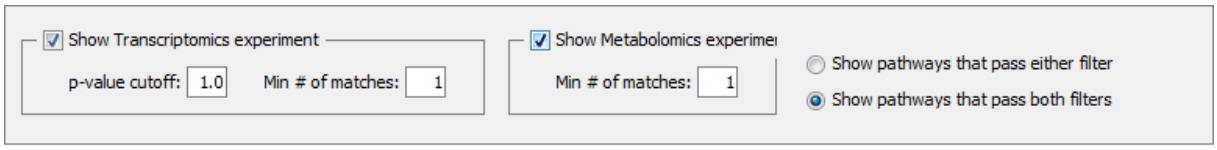


Figure 31.55: A sample filter panel from the Pathway View of a pathway list resulting from Multi-Omic Pathway Analysis

31.4.6.8 Workflow Browser

The Workflow Browser in an MOA experiment contains the options to perform either multi-omic **Pathway Analysis** or multi-omic **Correlation Analysis** on any of the entity lists present in the MOA experiment:

31.4.6.9 Legend

The legend at the bottom right corner of the main **GeneSpring** window contains three tabs which hold the color-coding and layout information for **Heatstrips** or **Quilt Plots** in the Pathway Viewer, the **Heatmap Pane**, as well as **Color Settings** for matched, highlighted, and selected entities as displayed in the Pathway Viewer. For more information on how to change these display settings see To view and modify display settings on page 923.

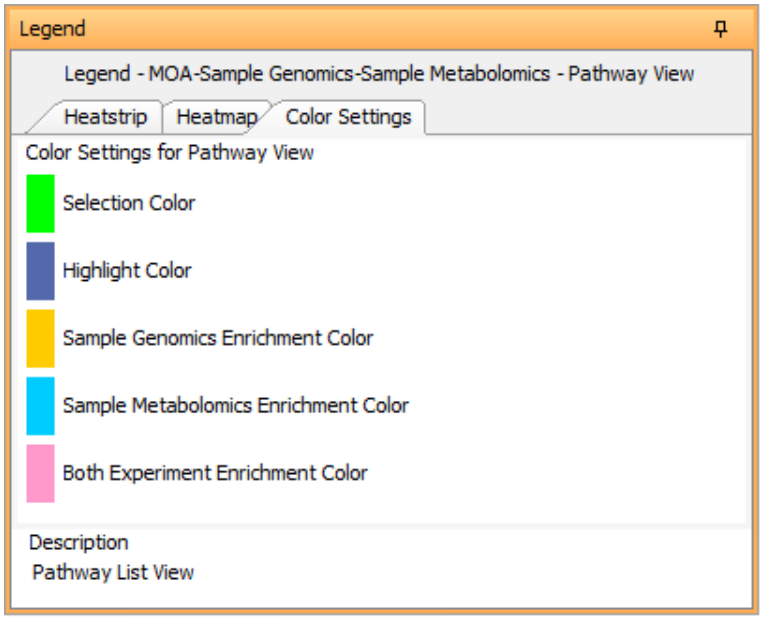


Figure 31.56: Legend displaying the Color Settings for the Pathway View of an MOA experiment

31.4.7 Dialog Boxes

This section describes the dialog boxes that can appear when you analyze and interpret data in an MOA experiment or pathway lists.

31.4.7.1 Choose Entity List

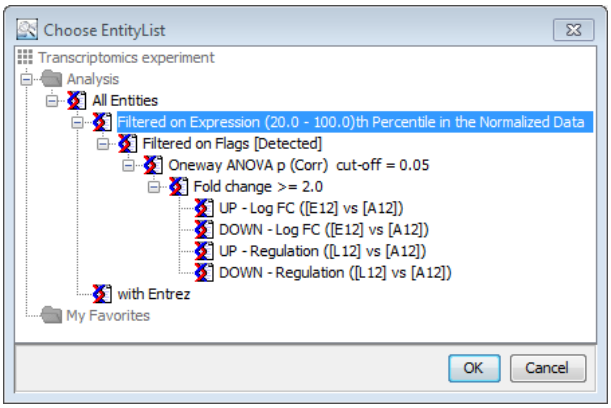


Figure 31.57: Chose Entity List dialog box

Purpose: Select an entity list from the available entity lists to replace the entity list with which the displayed pathway is currently associated.

To open: Click **Choose...** in the Entity List Chooser dialog.

OK Confirms the selected entity list, closes the dialog and returns to the Entity List Chooser dialog.

Cancel Aborts the selection process and returns to the Entity List Chooser dialog.

31.4.7.2 Choose Interpretation

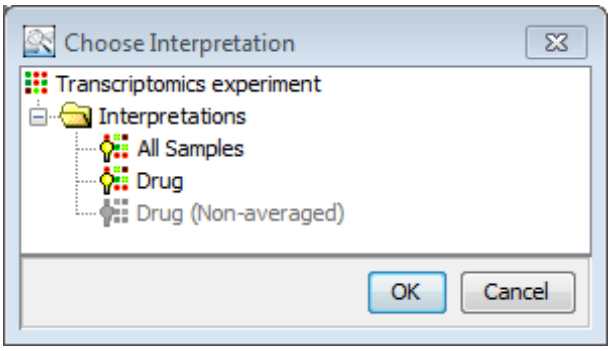


Figure 31.58: Choose Interpretation dialog box

Purpose: Select an interpretation from the available interpretations to replace the interpretation with which the displayed pathway is currently associated. See To choose a different interpretation on page 919 for more information about this feature.

To open: Click **Choose...** in the Interpretation Chooser dialog.

OK Confirms the selected interpretation, closes the dialog and returns to the Interpretation Chooser dialog.

Cancel Aborts the selection process and returns to the Interpretation Chooser dialog.

31.4.7.3 Choose Quilt Plot Parameters

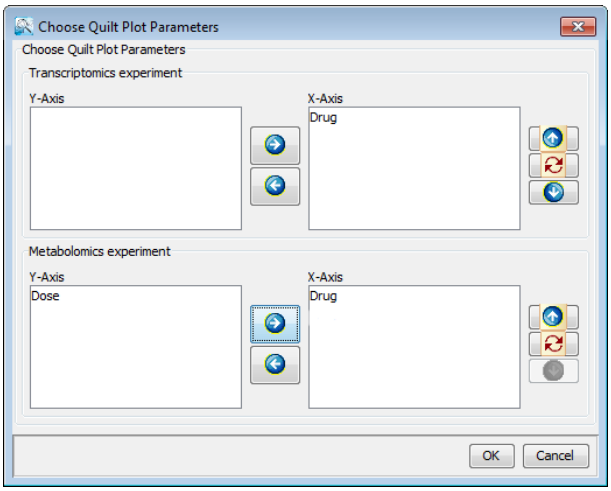


Figure 31.59: Choose Quilt Plot Parameters dialog box

Purpose: Determine which parameters in the selected interpretation(s) are plotted against the x- and y-axes in a quilt plot. See To display, remove, and customize quilt plots and heatstrips in the Pathway Viewer on page 919 for more information about this feature.

To open: Click the **Choose parameters along axes** icon in the **Pathway View Toolbar**.

OK Confirms any changes and returns to the **Pathway View**.

Cancel Closes the dialog box and returns to the **Pathway View** without applying any changes.

31.4.7.4 Confirm BioCyc Pathways Download

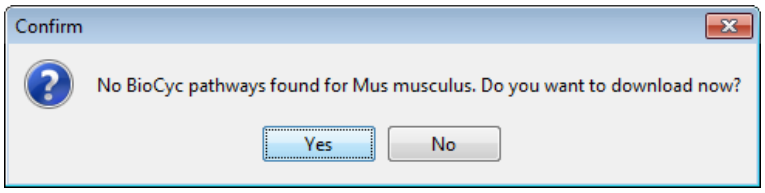


Figure 31.60: Confirm BioCyc Pathways Download dialog box

Purpose: Confirm whether you want to download BioCyc pathways for the chosen pathway organism now.

To open: In **Step 1** of the wizard-driven MOA workflow when a pathway organism is chosen, for which no BioCyc pathways have previously been downloaded during MOA.

Yes Downloads all available pathways for the chosen organism from the Agilent Server.

No If no other pathways from other selected sources are available for the chosen organism, an error message appears, as the workflow cannot be continued with the current selection of pathway sources.

31.4.7.5 Confirm KEGG Pathways Download

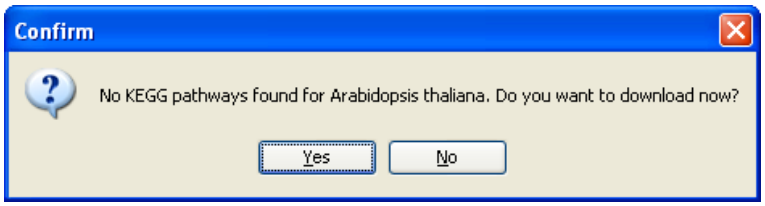


Figure 31.61: Confirm KEGG Pathways Download dialog box

Purpose: Confirm whether you want to download KEGG pathways for the chosen pathway organism now.

To open: In **Step 1** of the wizard-driven MOA workflow when a pathway organism is chosen, for which no KEGG pathways have previously been downloaded during MOA.

Yes Downloads all available pathways for the chosen organism from the Agilent Server.

No If no other pathways from other selected sources are available for the chosen organism, an error message appears, as the workflow cannot be continued with the current selection of pathway sources.

31.4.7.6 Confirm WikiPathways Download

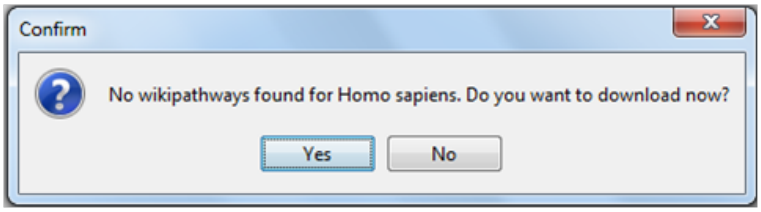


Figure 31.62: Confirm WikiPathways Download dialog box

Purpose: Confirm whether you want to download WikiPathways for the chosen pathway organism now.

To open: In **Step 1** of the wizard-driven MOA workflow when a pathway organism is chosen, for which no WikiPathways have previously been downloaded during MOA.

Yes Downloads all available pathways for the chosen organism from the WikiPathways protal (http://www.wikipathways.org).

No If no other pathways are available for the chosen organism, an error message appears, as the workflow cannot be continued with the current selection of pathway sources.

31.4.7.7 Entity List Chooser

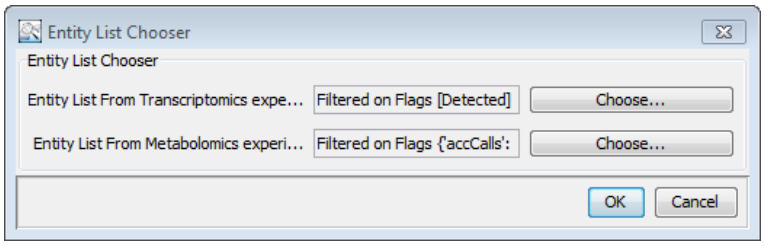


Figure 31.63: Entity List Chooser dialog box

Purpose: Choose a different entity list from only one or both source experiments of the **MOA** workflow to reflect in the displayed pathway. If the pathway was saved as a result of a **Single Experiment Analysis** workflow, you can only choose a different entity list from one experiment.

To open: Click the **Choose another Entity List** icon from the Pathway View Toolbar in an open pathway list.

Choose... Opens the Choose Entity List dialog.

OK Confirms the chosen entity list and closes the dialog. **GeneSpring** updates the displayed pathway in The Pathway View.

Cancel Aborts the selection process and returns to the Pathway View.

31.4.7.8 Entity Properties - GPML pathways

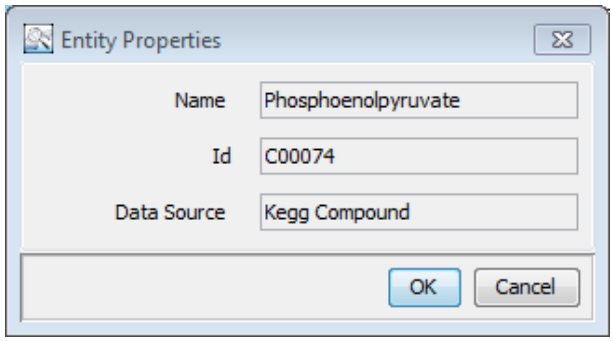


Figure 31.64: Entity Properties dialog box - GPML pathways

Purpose: Inspect properties of a pathway entity (see To inspect pathway entities and relations).

To open: Double-click an entity in a GPML pathway or right-click and select **Properties**.

OK Closes the dialog and returns to the Pathway View.

Cancel Closes the dialog and returns to the Pathway View.

31.4.7.9 Entity Properties - non-GPML pathways

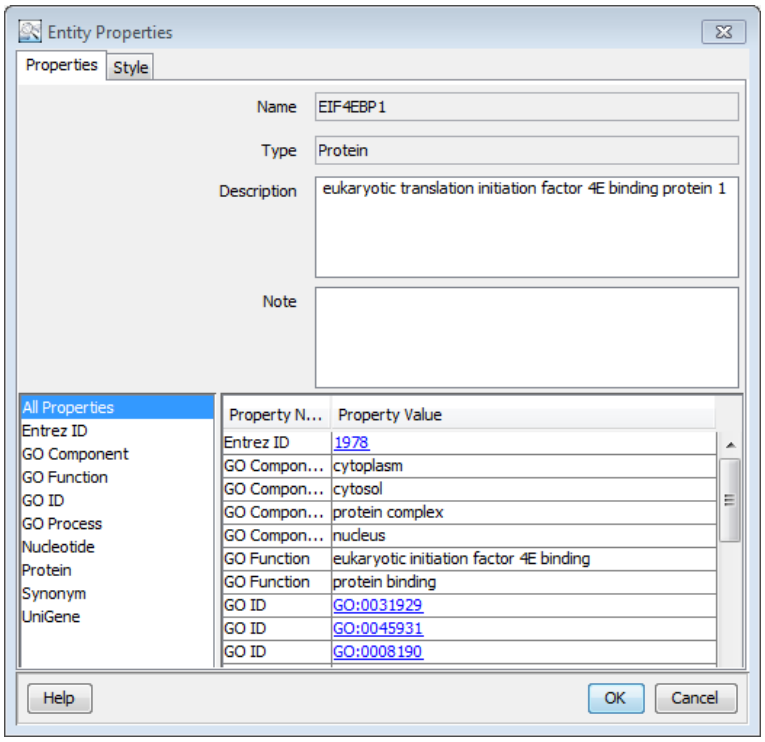


Figure 31.65: Entity Properties dialog box - non-GPML pathways

Purpose: Inspect properties of a pathway entity and edit the entity's appearance in the pathway (see To inspect pathway entities and relations).

To open: Double-click an entity in a non-GPML pathway or right-click and select **Properties**.

For BioPAX Level 2, Text format, and NLP/hand created pathways, the Properties tab represents all the available information about the entity, while the Style tab allows you to customize the appearance of that particular entity.

For BioPAX Level 3 pathways, a Reference tab provides references for the entity and its placement in the pathway in addition to the Properties and Style tabs.

Help Opens the **GeneSpring** manual in a new browser window.

OK Closes the dialog and applies any changes made in the Style tab to the entity in the pathway.

Cancel Closes the dialog and returns to the Pathway View without applying any changes.

31.4.7.10 Entity Properties - KEGG pathways

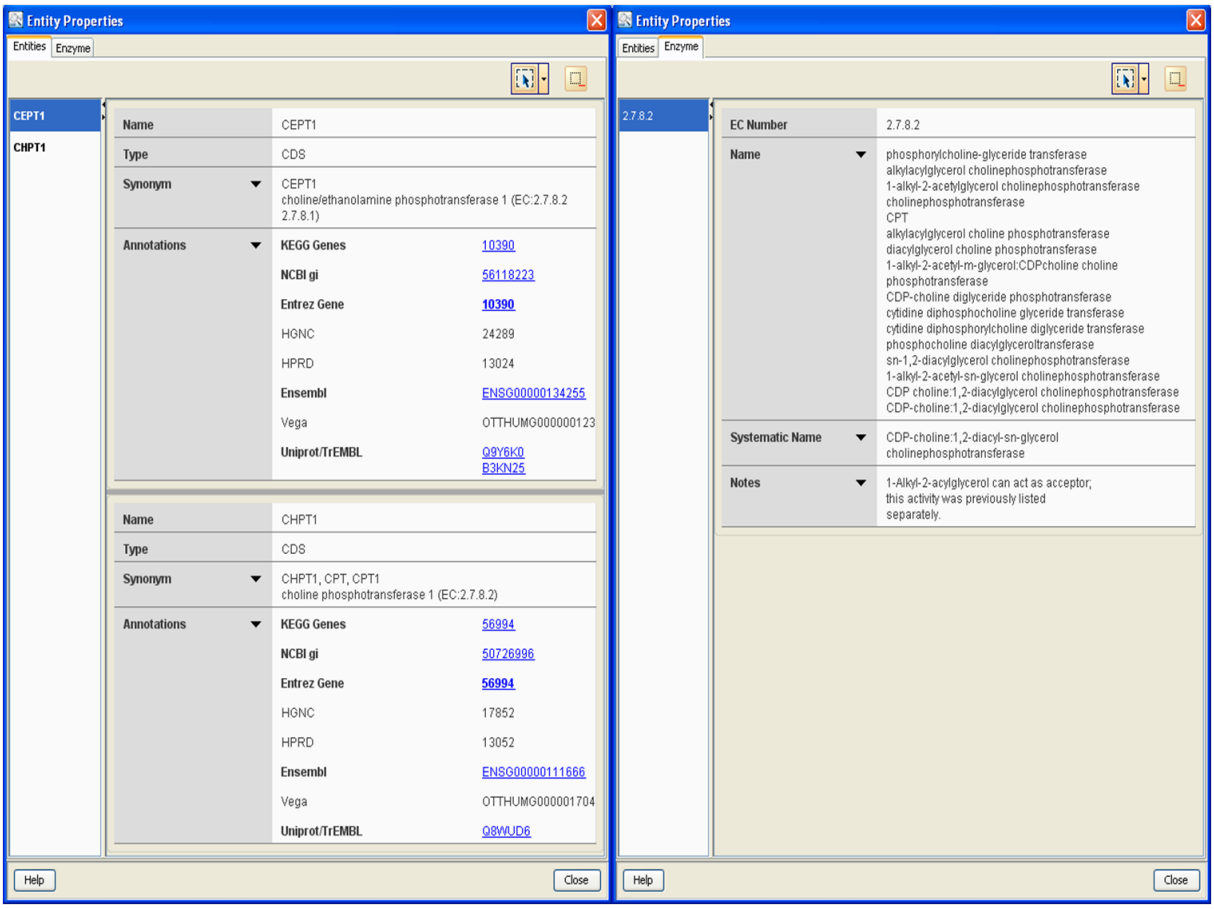


Figure 31.66: Entity and Enzyme Properties dialog box - KEGG pathways

Purpose: Inspect properties of a pathway entity and enzyme information (see To inspect KEGG pathway entities).

To open: Double-click an entity in a KEGG pathway or right-click and select **Properties**.

For KEGG pathways, the Entities tab represents all the available information about the entity (including Synonyms, Drug/ Disease Target information, and annotations), while the Enzyme tab (where Enzyme information is available) contains the enzyme level information of the entity (including EC number, Name, Systematic Name, and Notes). Selecting any hyper-linked information in the Entities tab will open the relevant webpage in a new browser window.

In the Entity properties dialog of a KEGG pathway entity that represents multiple entities, those entities displayed in bold in the left pane denote that they are present/matched (either directly or via BridgeDb matching) in the input Entity List(s) of the analysis while non-bold entities are absent in the input entity list (See Figure 31.67). Under the Annotations property section, annotations that are supported for matching in **GeneSpring** are highlighted in bold. The annotation identifier that actually led to the entity being matched is also highlighted in bold.

Use the **Property Filter** drop-down and **Collapse All/Expand All** / icons at the top of the dialog box to display only selected properties for each entity and collapse or expand the details provided under each property.

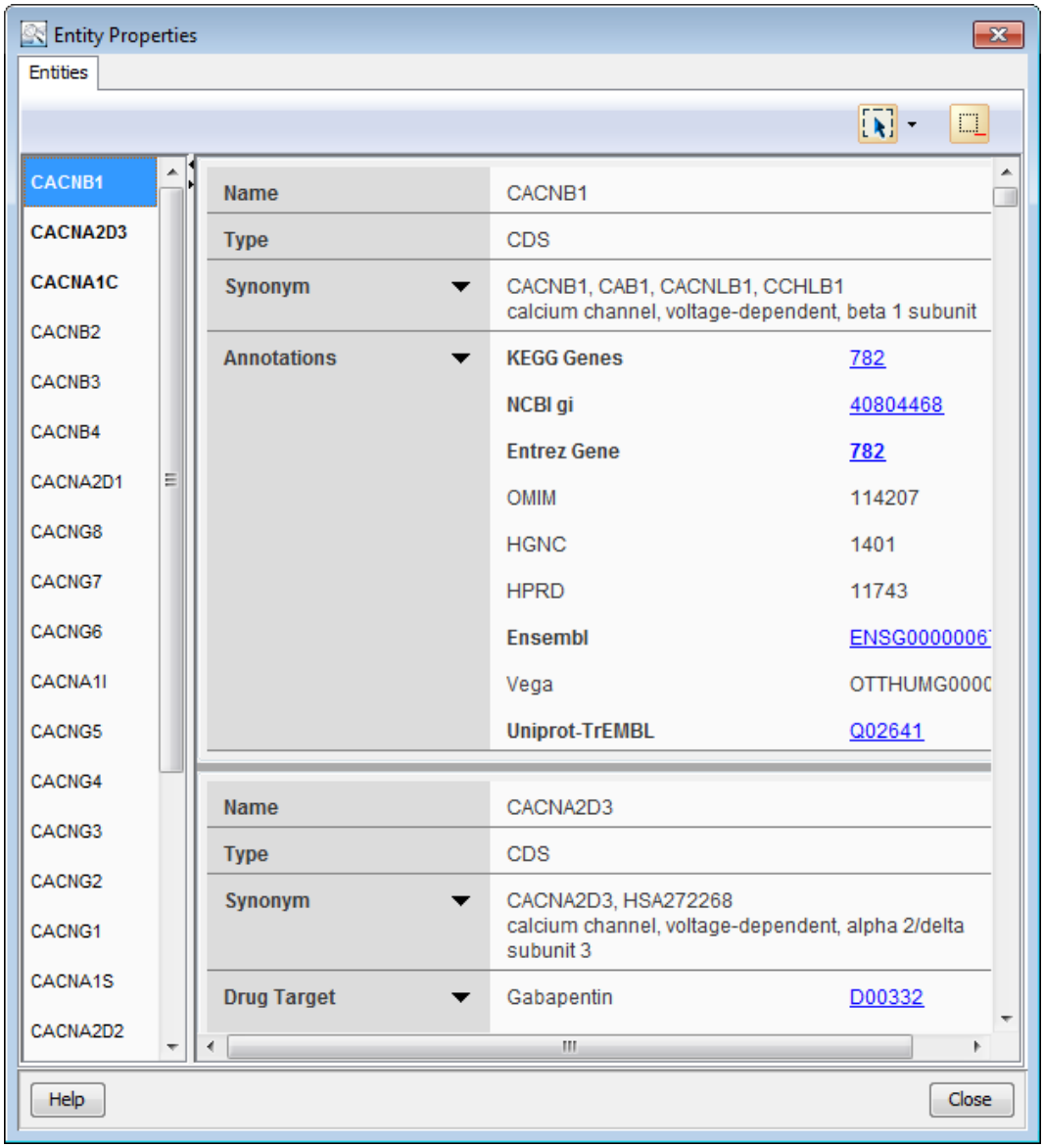


Figure 31.67: Entity and Enzyme Properties dialog box - KEGG pathways

Help Opens the **GeneSpring** manual in a new browser window.

Cancel Closes the dialog and returns to the Pathway View.

31.4.7.11 Export pathway entities to eArray

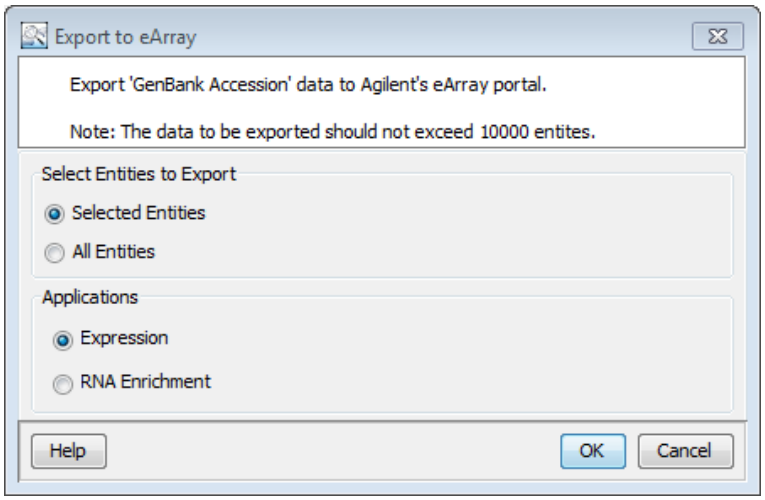


Figure 31.68: Export to eArray dialog box

Purpose: Export all entities of the displayed pathway, or a selection of those entities to Agilent's eArray portal (see To export pathway entities to eArray).

To open: Right-click on the Pathway Viewer and select **Export entities to eArray**.

Help Opens the **GeneSpring** manual in a new browser window.

OK Closes the dialog and launches the eArray portal's website in your default browser.

Cancel Closes the dialog and returns to The Pathway View.

31.4.7.12 Export pathway entities in the Pathway List Clipboard to eArray/SureDesign

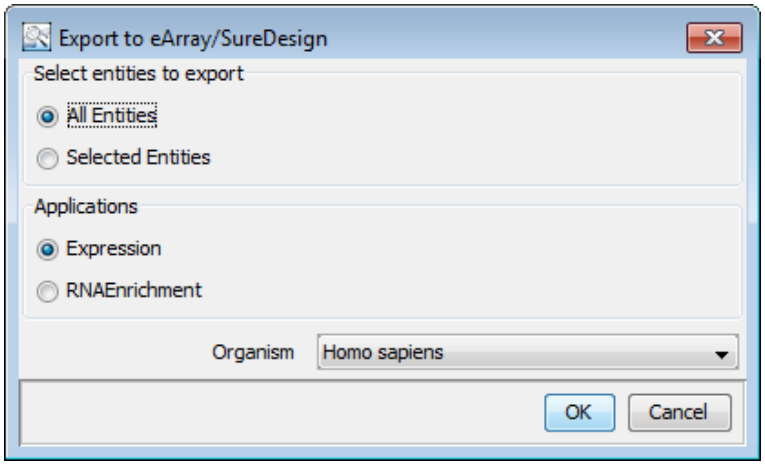


Figure 31.69: Export to eArray/SureDesign dialog box

Purpose: Export all entities or a selection of the entities in the Pathway List Clipboard to eArray/SureDesign (see To export entities from the Pathway List Clipboard to eArray/SureDesign).

To open: Click the **Export to eArray/SureDesign** icon.

OK Closes the dialog and launches the eArray/SureDesign portal's website in your default browser.

Cancel Closes the dialog and returns to the Pathway List Clipboard Inspector.

31.4.7.13 Export pathway entities in the Pathway List Clipboard to File

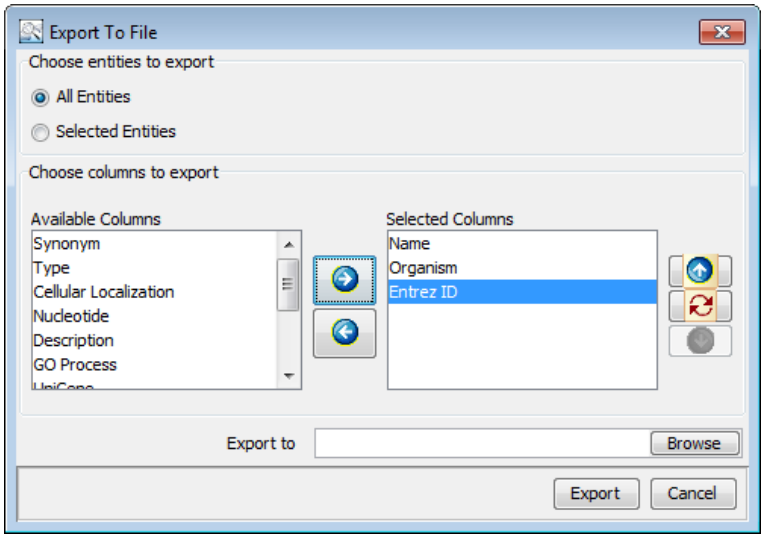


Figure 31.70: Export to File dialog box

Purpose: Export all entities or a selection of the entities in the Pathway List Clipboard to file (see To export entities from the Pathway List Clipboard to file).

To open: Click the **Export to file** icon.

Export Closes the dialog and exports all or the selected entities to file.

Cancel Closes the dialog and returns to the Pathway List Clipboard Inspector.

31.4.7.14 Input Node Properties

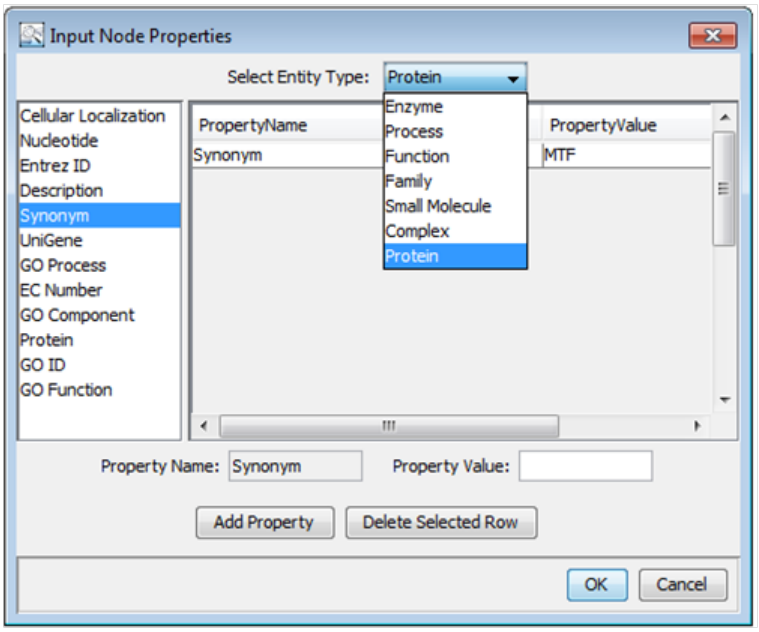


Figure 31.71: Input Node Properties dialog box

Purpose: Select the desired entity type, property name (depending on the entity type, and provide a property value to add the corresponding entity to a new pathway. You can find more information about the available entity types in Database Entities on page 1041.

To open: Click the **Add Entity** icon from the Pathway View Toolbar in the Pathway View when Creating and Editing Pathways (see To add a new entity to the pathway via the toolbar on page 881).

Add Property	Adds the specified entity as a new row to the spreadsheet view in the center of the dialog.
Delete Selected Row	Deletes a selected row from the spreadsheet view.
OK	Closes the dialog and adds the entities listed in the spreadsheet view to the new pathway.
Cancel	Aborts the selection process and returns to the Pathway View.

31.4.7.15 Input Relation Properties

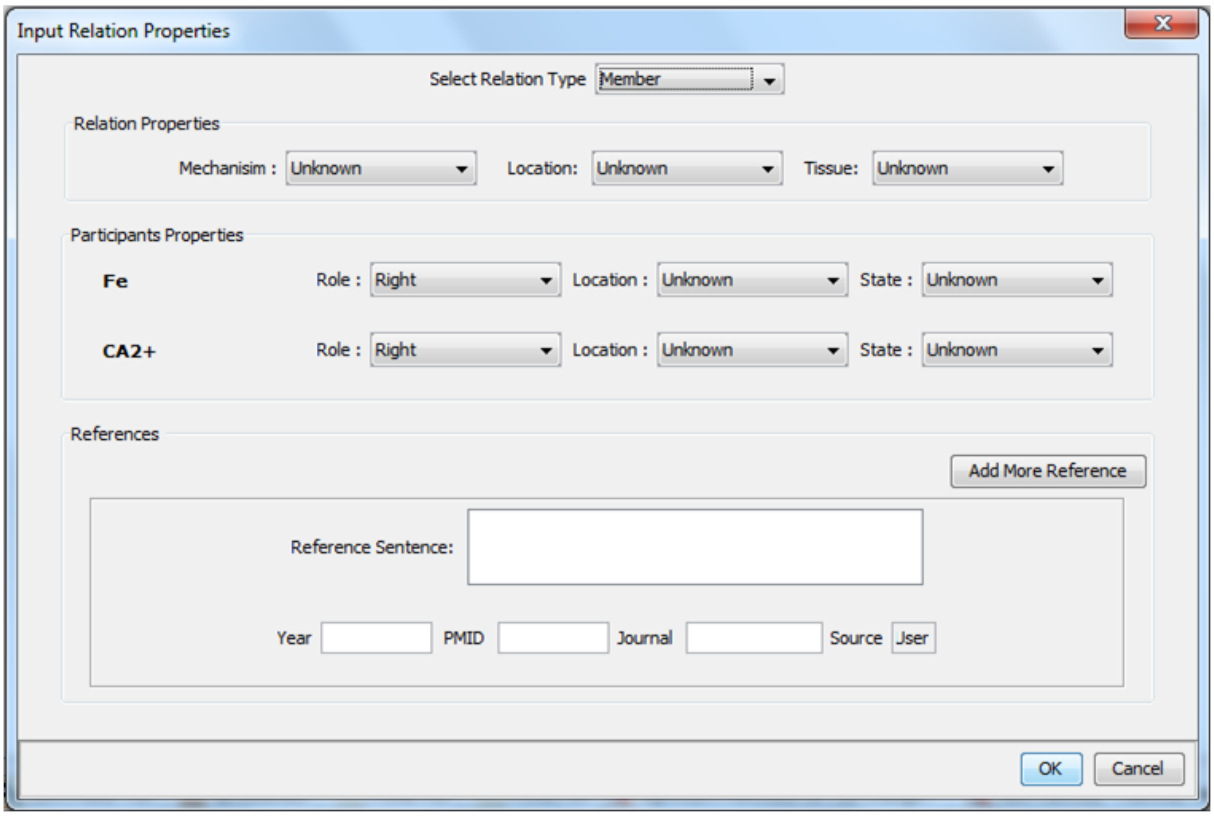


Figure 31.72: Input Relation Properties dialog box

Purpose: Select the desired **Relation Type**, **Relation Properties** (depending on the relation type), the **Participants Properties**, and provide **References** for the new relation from the respective drop-down menus. You can find more information about the available relation types in Relations on page 1045.

To open: Click the **Add Interaction** icon from the Pathway View Toolbar in the Pathway View when Creating and Editing Pathways (see To add an interaction between two entities in the pathway on page 883).

Add More Reference	Adds further editable fields to the References area to allow you to create additional references for the new relation.
OK	Closes the dialog and adds the new relation to the pathway.
Cancel	Aborts the selection process and returns to the Pathway View.

31.4.7.16 Interpretation Chooser

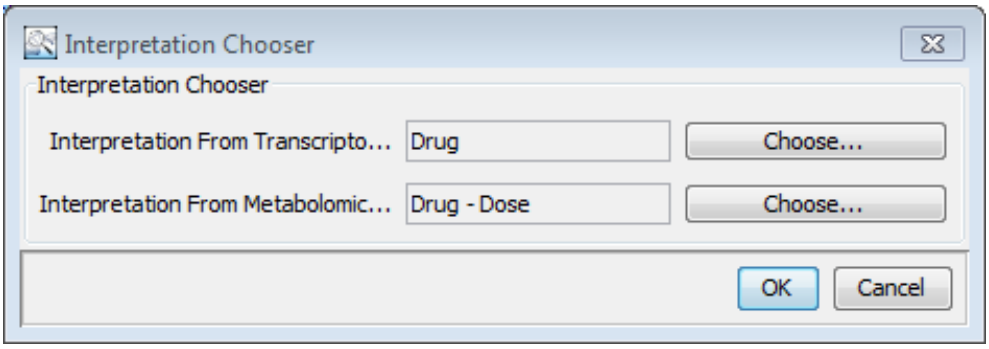


Figure 31.73: Interpretation Chooser dialog box

Purpose: Choose a different interpretation from only one or both source experiments of the **MOA** workflow. If the pathway was saved as a result of a **Single Experiment Analysis** workflow, you can only choose a different interpretation from one experiment.

To open: Click the **Choose another Interpretation** icon from the Pathway View Toolbar in an open pathway list.

Choose... Opens the Choose Interpretation dialog.

OK Confirms the interpretation and closes the dialog. **GeneSpring** updates the displayed pathway and Heatmap Pane in The Pathway View.

Cancel Aborts the selection process and returns to the Pathway View.

31.4.7.17 Multiple Objects Inspector

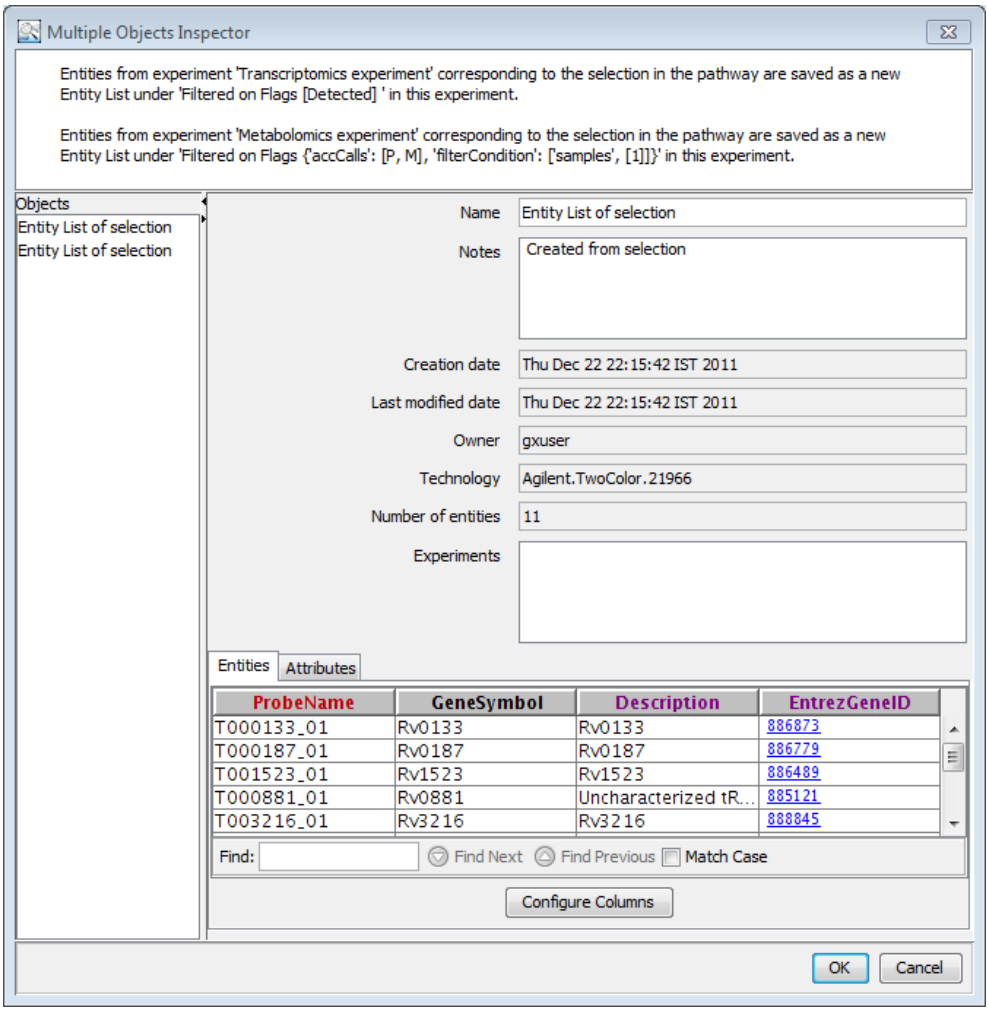


Figure 31.74: Multiple Objects Inspector dialog box

Purpose: Inspect properties of multiple objects, for example pathways (when inspected from the Search Wizard) or entity lists. You can rename the individual objects and edit their **Notes**.

To open: Click the **Inspect selected pathways** icon in Step 3 of the Search Wizard to inspect selected pathways from a search result. Click the **Create Entity List from selection** icon in the Pathway View of a pathway list created from an MOA workflow. A Multiple Objects Inspector opens if the selection contains entities from both input experiments.

Objects Lists the objects being inspected. Click an object name to display the details of that object in the right panel, for example the **Entities** in an Entity List or the **Entities** and **Relations** in a pathway.

OK Closes the dialog and accepts any changes.

Cancel Closes the dialog box. No changes are applied.

31.4.7.18 New Multi-Omic Analysis Experiment

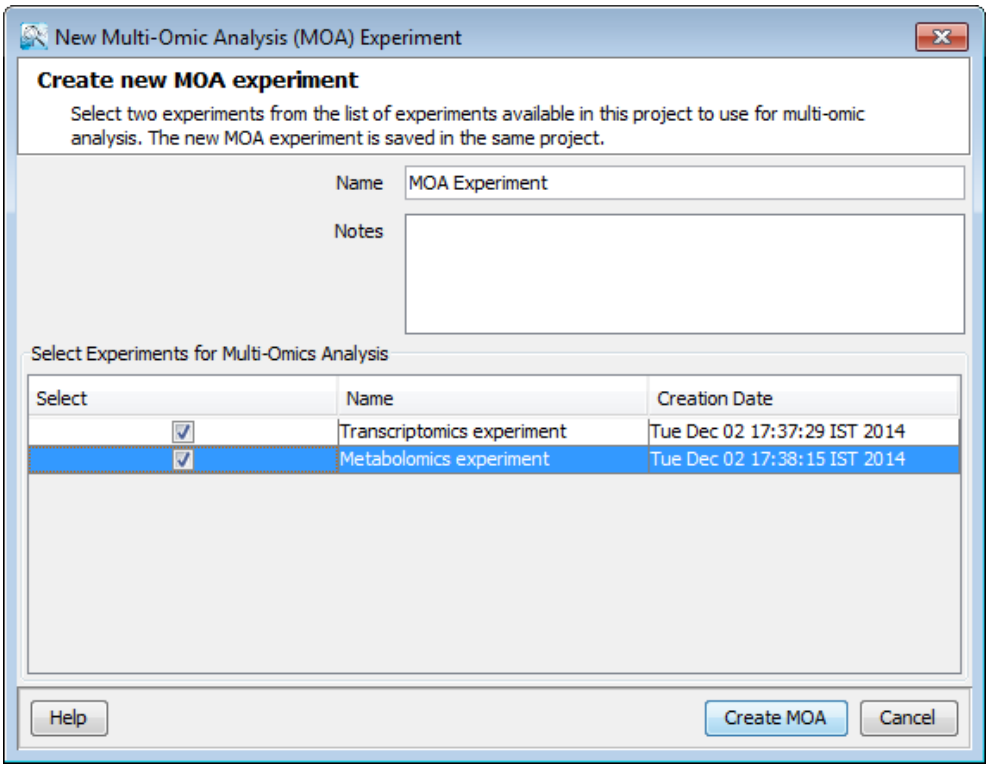


Figure 31.75: New Multi-Omic Analysis Experiment dialog box

Purpose: Enter details and select experiments required for the creation of a Multi-Omic Analysis (MOA) Experiment.

To open: Select **Multi-Omic Analysis** from the **Pathway Analysis** section of the Workflow Browser.

Name	Field to provide a name for the new MOA experiment.
Description	Field to provide a description of the new MOA experiment.
Select Experiments for Multi-Omic Analysis	A table that contains all the experiments in the open project that are available for the creation of an MOA experiment. Select two experiments from this table.
Help	To find out more about MOA experiment creation, click Help. The user manual opens in the default browser of your computer.
Create MOA	Prompts **GeneSpring** to create the new MOA experiment from the selected two experiments. The experiment is saved in the **Project Navigator** below any other existing experiments.
Cancel	Closes the dialog box. The MOA experiment is not created and the MOA workflow does not start.

31.4.7.19 Organism Chooser

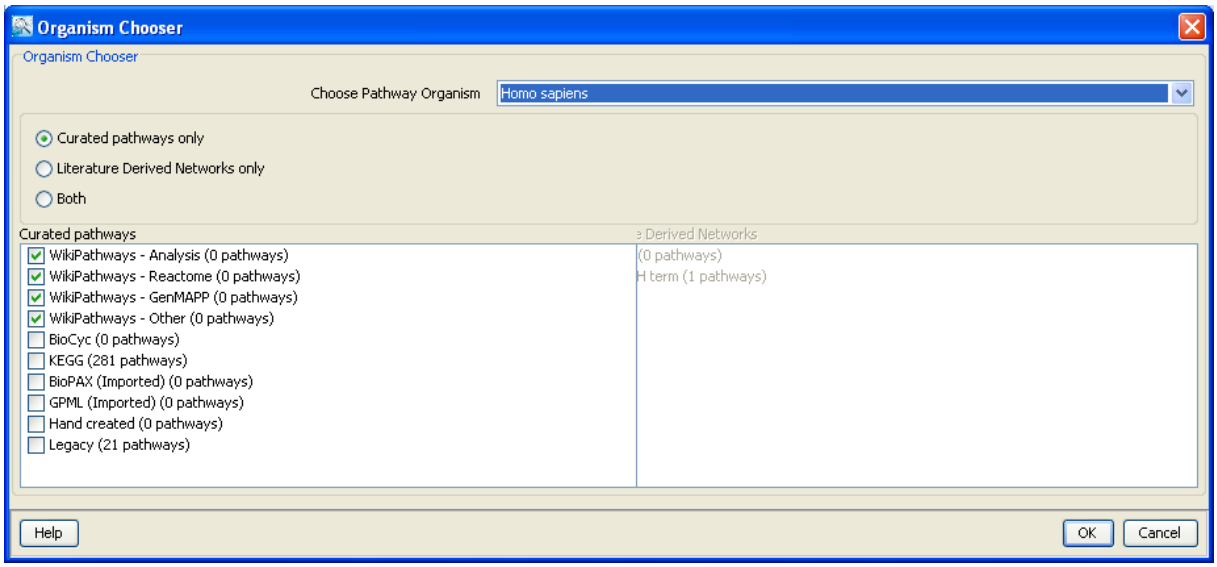


Figure 31.76: Organism Chooser dialog box

Purpose: Specify the pathway organism from the **Choose Pathway Organism** drop-down and the pathway sources to be searched by selecting the desired options from **Curated pathways** and/or **Literature Derived Networks**.

To open: Select **Find pathways containing all selected entities** or **Find pathways containing any selected entities** from the right-click options on the Heatmap Pane of The Pathway View.

OK Confirms the chosen organism and pathway sources. If any of the WikiPathways options in **Curated pathways** are selected and no WikiPathways have previously been downloaded for the chosen organism, **GeneSpring** gives you the option to download them in this step first and **GeneSpring** automatically includes them in the search. Otherwise, **GeneSpring** starts searching all the selected pathway sources when you click **OK**.

Cancel Aborts the selection process and returns to the Pathway View.

31.4.7.20 Pathway Inspector

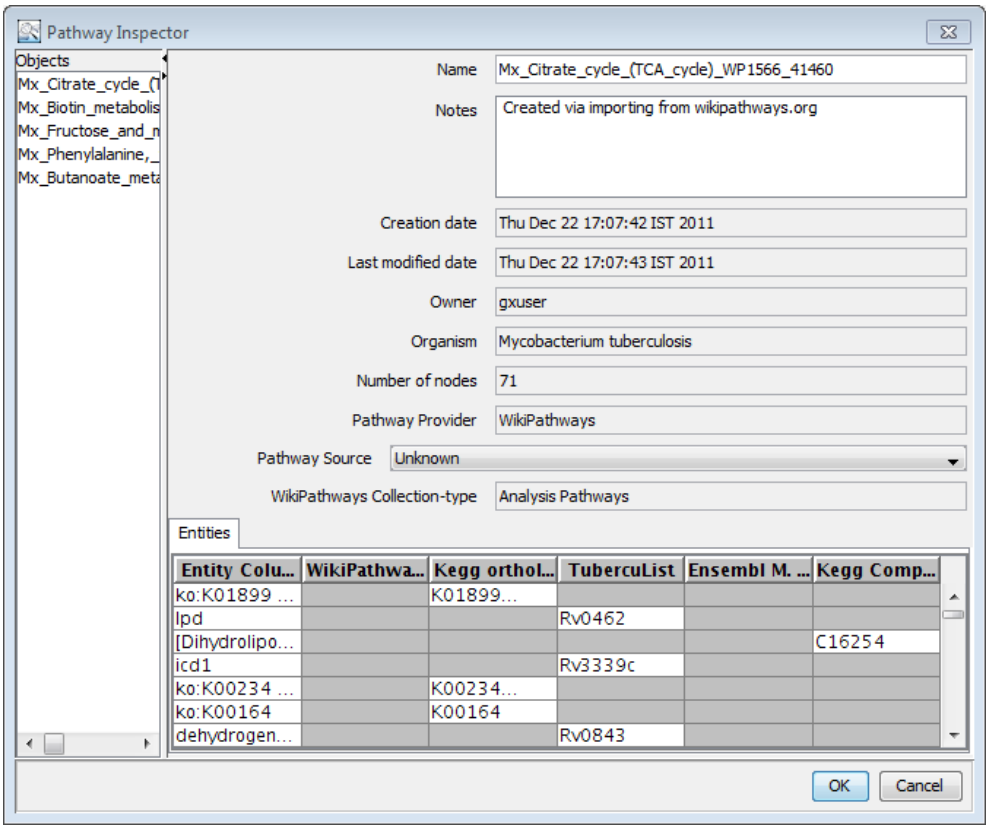


Figure 31.77: Pathway Inspector dialog box

Purpose: Inspect properties of one or multiple pathways listed in a Pathway List Inspector. You can rename the pathways, edit their **Notes** and change the **Pathway Source** via a drop-down menu.

To open: Double-click one pathway or a selection of pathways in a Pathway List Inspector.

Objects Lists the pathways being inspected. Click a pathway name to display the **Entities** and **Relations** in that pathway in the right panel. For GPML pathways only **Entities** are displayed.

OK Closes the dialog. Any changes made to the **Name** or **Notes** are applied.

Cancel Closes the dialog box. No changes are applied.

31.4.7.21 Pathway List Clipboard Inspector

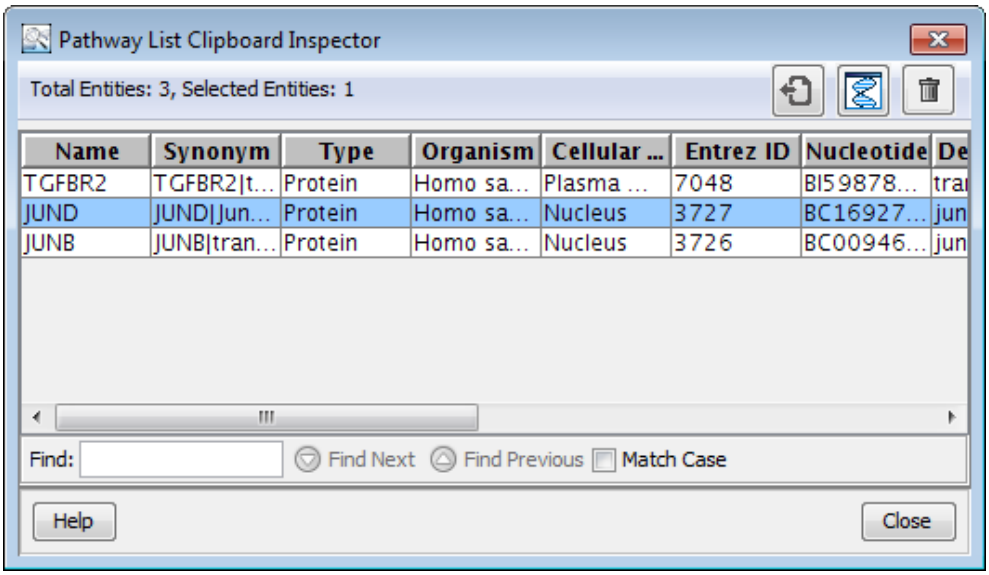


Figure 31.78: Pathway List Clipboard Inspector dialog box

Purpose: Inspect the pathway entities that have been added to the Pathway List Clipboard. Lists all the pathway entities and their properties (for example, Name, Synonym, Entrez Gene ID, etc.).

To open:

Click the **Pathway List Clipboard Inspector** icon in the **Pathway View Toolbar**. This icon is only active if you have previously added pathway entities to the Pathway List Clipboard (see To add pathway entities to the Pathway List Clipboard on page 928).

Find Allows you to search the table for a specific term.

Help Opens the **GeneSpring** manual in a new browser window.

Close Closes the dialog.

31.4.7.22 Pathway List Inspector

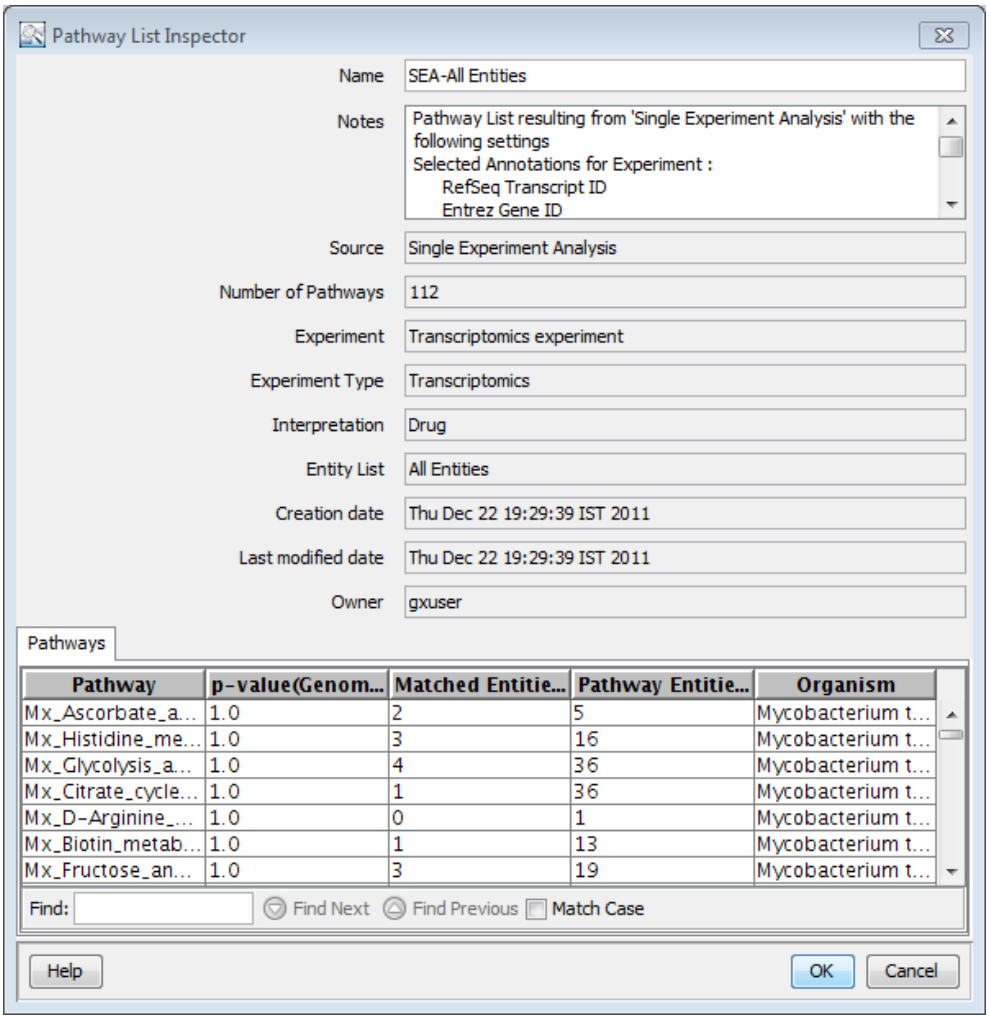


Figure 31.79: Pathway List Inspector dialog box

Purpose: Inspect the details of a pathway list. Lists all the pathways present in the pathway list and their associated data (for example, p-value, Matched Entities, and Pathway Entities of Experiment Type, Organism for a pathway list resulting from MOA or SEA). You can rename the pathway list and edit Notes.

To open: Right-click on the pathway list in the Experiment Navigator and select **Inspect Pathway List** (see To inspect a pathway list on page 913).

Help Opens the **GeneSpring** manual in a new browser window.

OK Closes the dialog and accepts any changes.

Cancel Closes the dialog box and returns to the previous state of the application. No changes are applied.

31.4.7.23 Pathway Property Dialog

Purpose: Inspect the data and display settings of the current pathway view. You can choose the data you would like to be displayed by clicking **Data** and selecting the desired data type from the **Data Chooser** drop-down. You can change the default color settings for matched entities (select **Entity**), heatstrips (select **Heatstrip**), quilt plots (select **Quilt Plot Layout**), and selection and highlight indicators (select **Select/Highlight**). You can also overlay data associated with matched entities, like the condition in the selected interpretation, p-values, or Fold Change on the pathway. If the pathway list is a result from an MOA workflow, you can choose different data settings, entity fill colors, and heatstrip coloring for each participating experiment.

To open: Click the **View and Modify Display Settings** icon or right-click on the pathway viewer and select **Display Settings**.

Help Opens the **GeneSpring** manual in a new browser window.

OK Closes the dialog and accepts any changes.

Cancel Closes the dialog box and returns to the Pathway View. No changes are applied.

Data tab

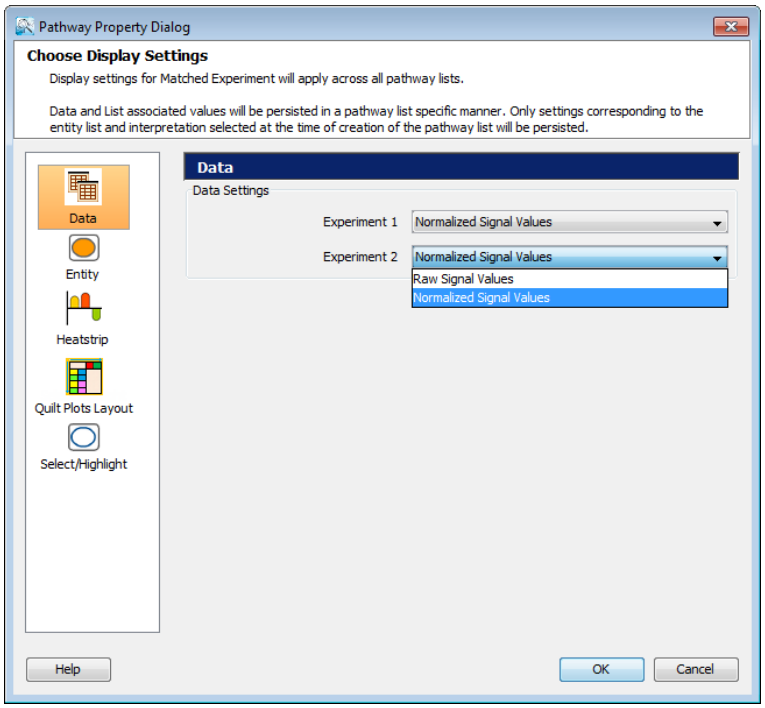


Figure 31.80: Pathway Properties Dialog dialog box - Data tab

Choose whether to display **Normalized Signal Values**, or **Raw Signal Values** in the heatstrips or quilt plots in the Pathway Viewer and Heatmap Pane of the Pathway View.

Data Settings Select the desired data type (Normalized Signal Values or Raw Signal Values if available) from the drop-down list for each experiment.

Entity tab

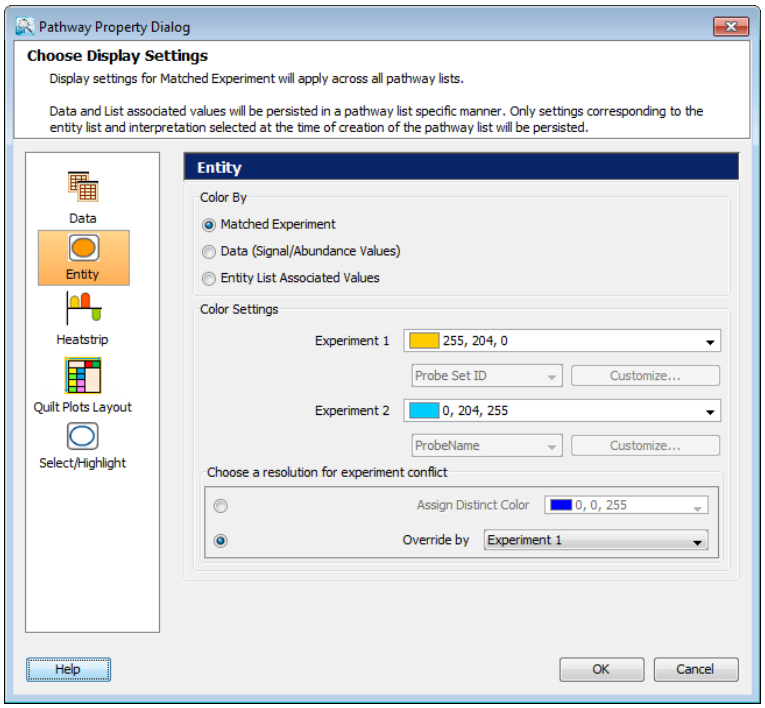


Figure 31.81: Pathway Properties Dialog dialog box - Entity tab

Define the fill color for matched entities based on experiment-level (Matched Experiment) or entity-level (probe name, condition in the selected interpretation, p-value, Fold Change, etc.) information. Use these options for data overlay on any pathway resulting from single experiment or multi-omic Pathway Analysis.

Color By Provides the following three options for data overlay on matching entities between the input experiment and the pathway.

Matched Experiment to color pathway entities by the matching experiment.

Data (Signal/Abundance Values) to color entities by probe name or a condition in the selected interpretation.

Entity List Associated Values to color entities by probe name or an entity list associated value if present in the selected entity list, for example p-value or Fold Change.

Color Settings Customize entity fill color for matching entities. Select the desired data type (probe name, condition in selected interpretation, entity-list associated values present in the selected entity list) from the drop-down list below the selected color. Click **Customize...** to modify the default entity fill color.

Settings for pathway entities matched in both experiments Select from the following two options what you want **GeneSpring** to do if an entity matches both experiments in a MOA experiment.

Assign Distinct Color to color entities in a third distinct color different from the assigned colors for Experiment 1 and Experiment 2.

Override by to color entities either in the chosen color for Experiment 1 or Experiment 2. Select the desired option from the drop-down list.

Heatstrip tab

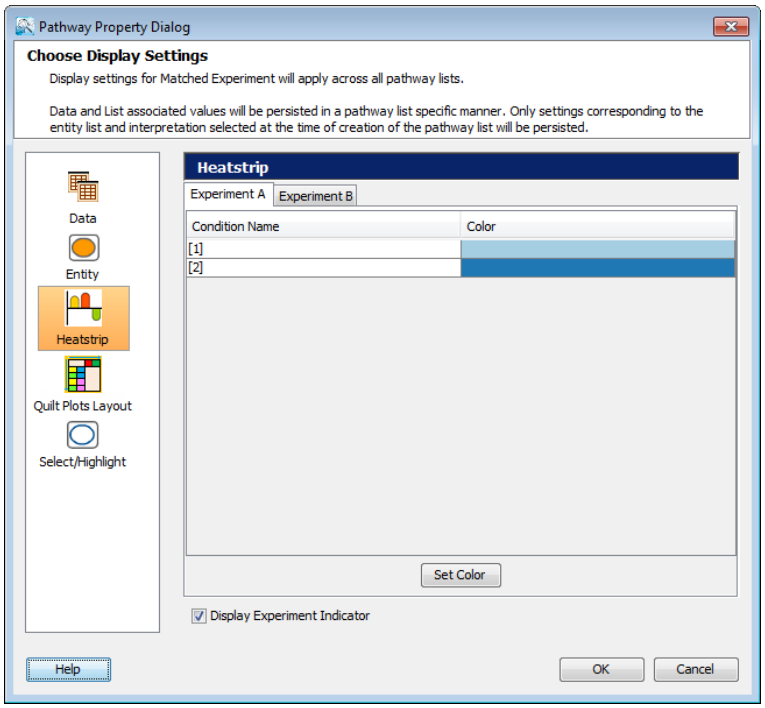


Figure 31.82: Pathway Properties Dialog dialog box - Heatstrip tab

All the conditions in the selected interpretation and the default colors assigned to each condition are displayed as a table in the Heatstrip panel. In a MOA experiment, there are two separate tabs for each experiment. Customize the heatstrip colors and choose whether to display the experiment indicator or not in a MOA experiment.

Set Color To change the color of a heatstrip, select the corresponding row in the table and click **Set Color**.

Display Experiment Indicator Deselect this option to remove the colored indicator bar that identifies the experiment with which a heatstrip is associated in a MOA experiment. By default the colors used for these indicator bars correspond to the chosen colors for the **Matched Experiment** option in the Entity tab of the Pathway Properties Dialog.

Quilt Plot Layout tab

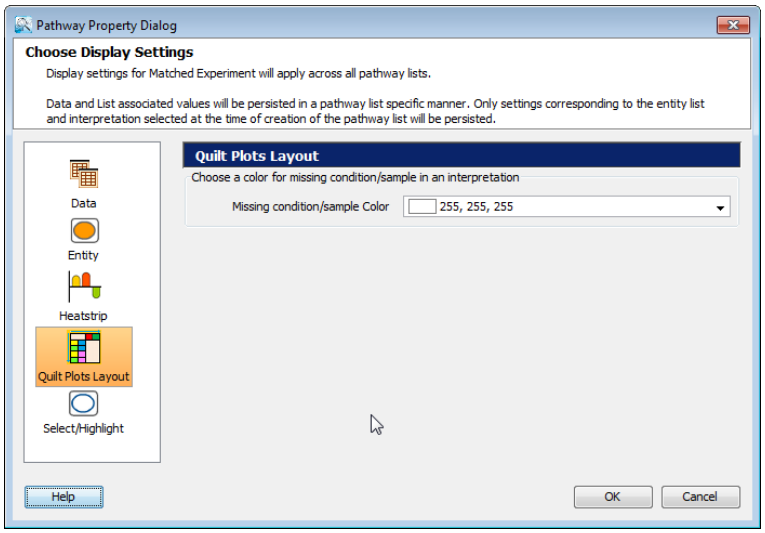


Figure 31.83: Pathway Properties Dialog dialog box - Quilt Plot Layout tab

Determine the color that indicates missing samples/conditions in a quilt plot displayed in the pathway viewer.

Missing condition/sample Color To change the color used to indicate missing samples/conditions in the quilt plot click the **Missing condition/sample Color** drop-down and choose from the available options or click **More Colors** to open the **Change a Color** dialog to define a new color.

Select/Highlight tab

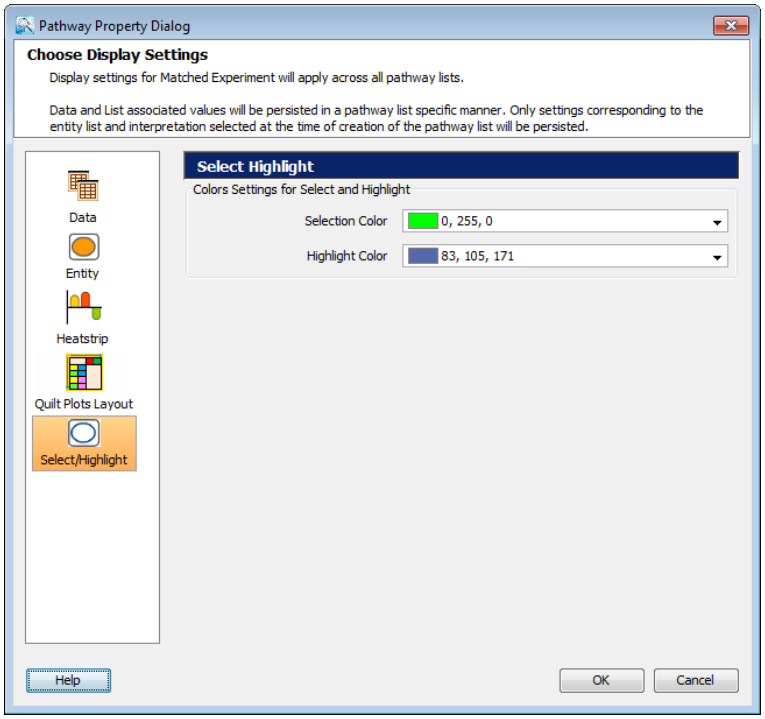


Figure 31.84: Pathway Properties Dialog dialog box - Select Highlight tab

Customize the colors used for indicating selected and highlighted entities in a pathway.

Color Settings for Select and Highlight Use the drop-down list to change the default for the **Selection Color** and **Highlight Color**.

31.4.7.24 Profile Plot

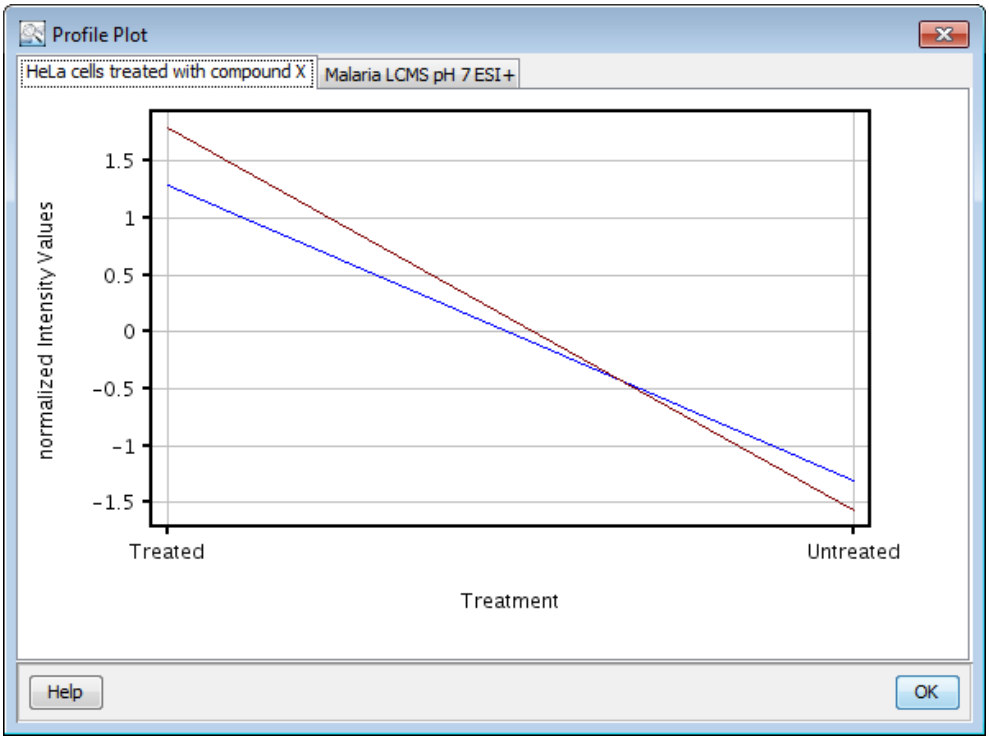


Figure 31.85: Profile Plot dialog box

Purpose: View a profile plot of selected entities in the displayed pathway in The Pathway View. If the entities stem from two different entity lists from an MOA experiment, two separate profile plots are created and can be accessed by clicking on the respective tabs.

To open: Click the **Launch profile plot on selected entities** icon from the Pathway View Toolbar in an open pathway list.

Help Opens the **GeneSpring** manual in a new browser window.

OK Closes the dialog.

31.4.7.25 Relation Properties

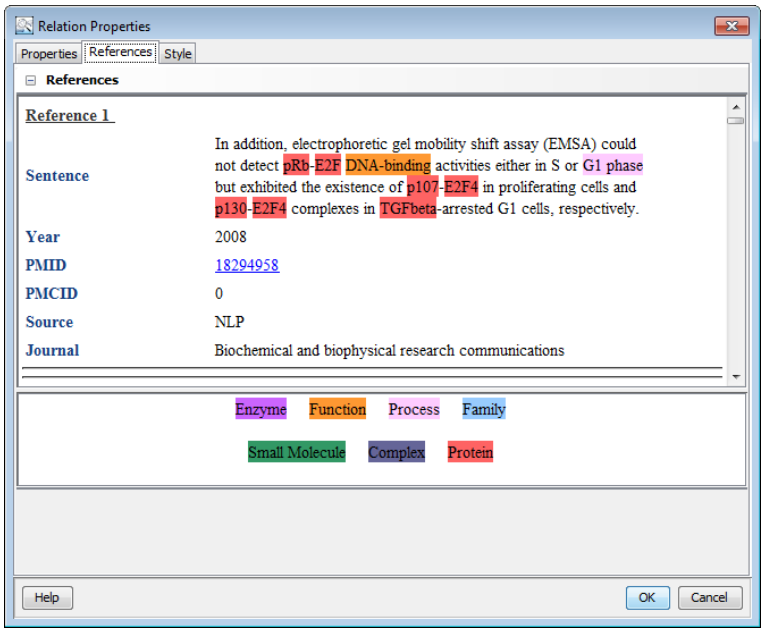


Figure 31.86: Relation Properties dialog box

Purpose: Inspect a relation's references, properties and edit the relation's appearance in the pathway (see To inspect pathway entities and relations).

To open: Double-click a relation in a non-GPML pathway or right-click and select **Properties**.

Help Opens the **GeneSpring** manual in a new browser window.

OK Closes the dialog and applies any changes made in the Style tab to the relation in the pathway.

Cancel Closes the dialog and returns to the Pathway View.

31.4.7.26 Export Gene Annotations

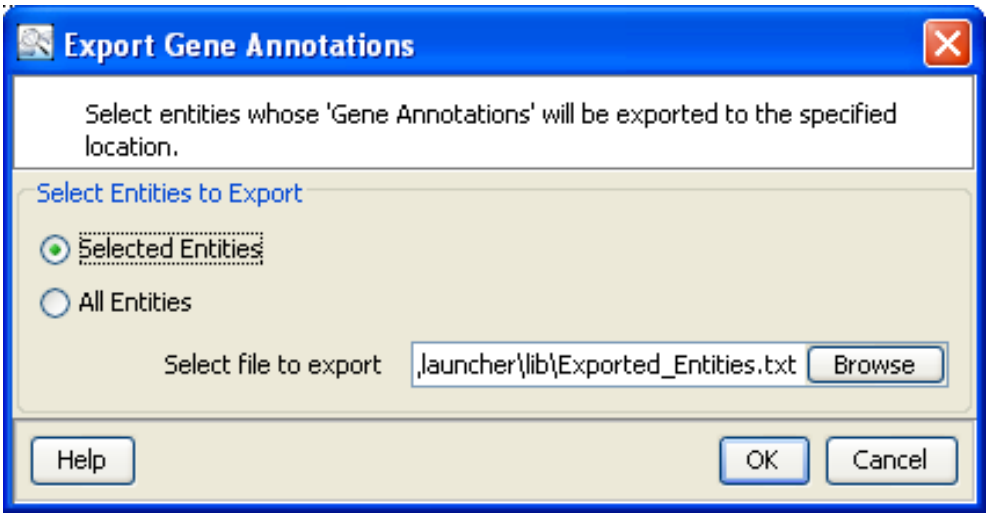


Figure 31.87: Select Entities to Export dialog box

Purpose: Export Entrez Gene IDs or SwissProt/UniProt IDs of a selection of the entities of the pathway displayed in the pathway viewer (see To export Gene Annotations of entities).

To open: Select entities in the displayed pathway, then right-click on the Pathway Viewer and select **Export Gene Annotations**.

Browse Opens the **Choose a file** dialog, which allows you to specify a new file name and location.

Help Opens the **GeneSpring** manual in a new browser window.

OK Closes the dialog and saves the file in the specified location.

Cancel Closes the dialog and returns to The Pathway View.

31.4.7.27 Venn Diagram (pathway lists)

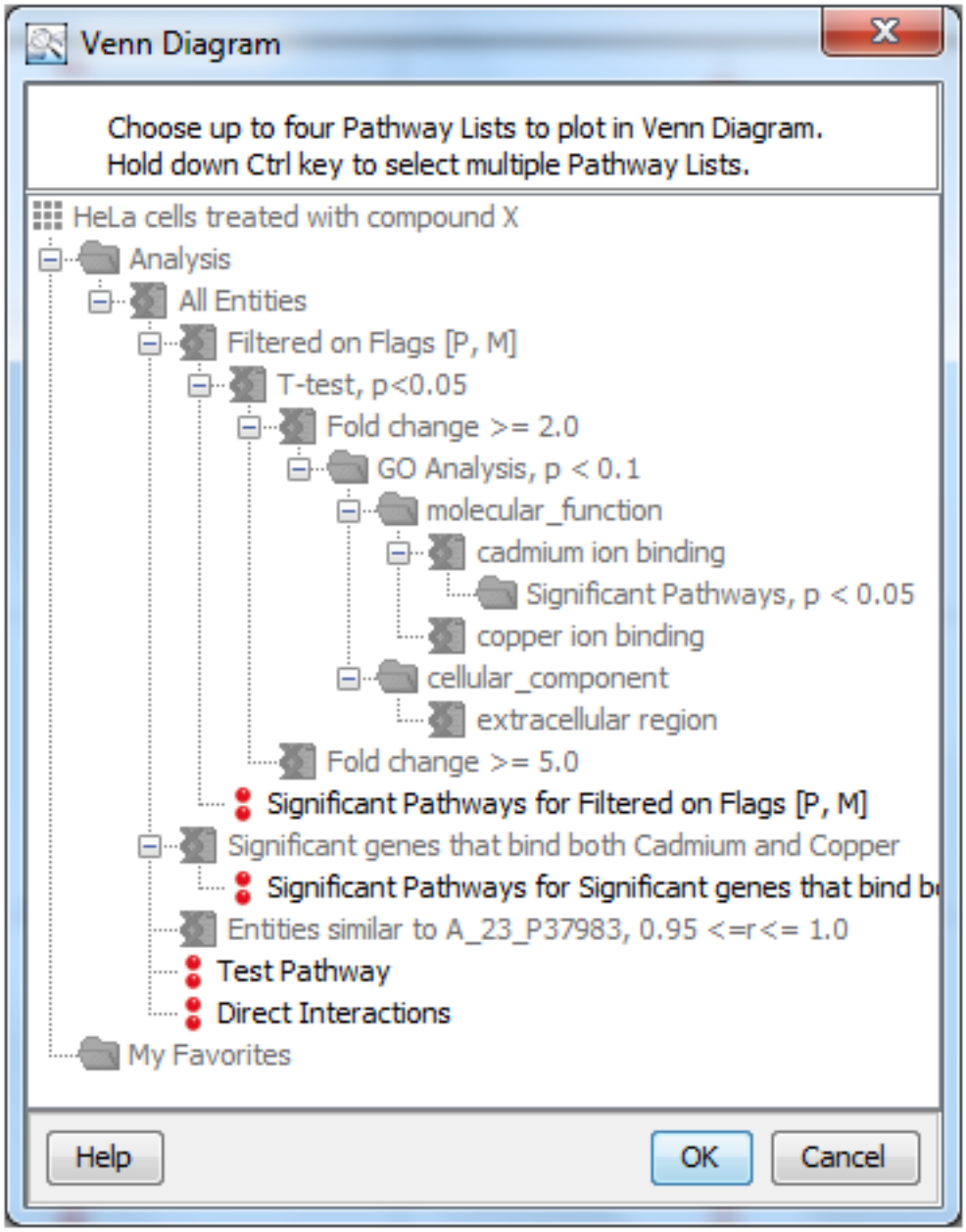


Figure 31.88: Venn Diagram dialog box

Purpose: Select up to four pathway lists from open experiments to identify common pathways between them.

To open: Select Pathway List from the **Venn Diagram** drop-down icon in the main toolbar.

Help Opens the **GeneSpring** manual in a new browser window.

OK Accepts the selection and opens the Venn Diagram view.

Cancel Aborts the selection process and returns to the previous view.

Table 31.4: Right-click options on the Heatmap Pane

Option	Description
Find pathways containing all selected entities	Finds pathways that contain all the selected entities in the Heatmap Pane.
Find pathways containing any selected entities	Finds pathways that contain any one of the selected entities in the Heatmap Pane.
Collapse all rows	Collapses all rows. Particularly useful when isomers are present.
Expand all rows	Expands all rows.
Display Heatmap only	Displays experimental data for the selected interpretation in a colored mode.
Display Data only	Displays experimental data for the selected interpretation in a numerical mode.
Display Heatmap and Data	Displays experimental data for the selected interpretation in both the colored and numerical mode.
Select all rows	Selects all the visible rows in the pathway list.
Invert row selection	Selects all the visible rows that were previously not selected.
Clear row selection	Unselects all previously selected rows.
Limit to row selection	Limits display to the selected rows only.
Copy to Clipboard as • Entrez Gene IDs • Swiss-Prot/UniProt IDs	Converts the identifiers of the selected entities to the selected annotation using BridgeDb and copies the corresponding identifiers to the Clipboard.
Freeze columns before	Freezes columns to the left and enables horizontal scrolling of the columns to the right.
Unfreeze Columns (Only available if columns are frozen.)	Unfreezes previously frozen columns.
Copy	Copies the selected pathway and related metrics.
Copy view	Copies all the pathways in the current view with their metrics.
Print	Opens an html browser with the current view for printing.
Publish	n/a
Export as • Image • HTML • Text	Exports the heatmap table in the chosen format. The 'Image' option only exports the area of the Heatmap Pane visible in the current view. The 'Text' option exports selected rows or the whole table if no rows are selected.
Properties	Opens a properties dialog box to alter rendering, visualization, columns to be included and view description of pathway lists.

31.4.8 Right-click Options

This section lists the various options available through right-clicking on the different elements associated with pathways.

31.4.8.1 Heatmap Pane

The right-click options for the Heatmap Pane displayed below the Pathway Viewer in The Pathway View are listed in Table 31.4.

Table 31.5: Right-click options on pathway list name in the Experiment Navigator

Option	Description
Open Pathway List	Opens the pathway list in the Pathway View.
Inspect Pathway List	Opens the Pathway List Inspector.
Remove Pathway List	Removes the pathway list only from the experiment.
Delete Pathway List	Permanently deletes the pathway list from **GeneSpring**.
Share Pathway List	Enabled only in **WorkGroupPlus** mode.
Change owner	Enabled only in **WorkGroupPlus** mode.

Table 31.6: Right-click options on the pathway list panel

Option	Description
Select all rows	Selects all the visible rows in the pathway list.
Invert row selection	Selects all the visible rows that were previously not selected.
Clear row selection	Unselects all previously selected rows.
Limit to row selection	Limits display to the selected rows only.
Reset Filters	Clears all filters applied to any of the columns.
Freeze columns before	Freezes columns to the left and enables horizontal scrolling of the columns to the right.
Unfreeze Columns (Only available if columns are frozen.)	Unfreezes previously frozen columns.
Copy	Copies the selected columns and rows to the clipboard.
Add to Pathway List Clipboard	Adds selected entities in the displayed pathway to the Pathway List Clipboard.
Copy View	Copies the pathway list panel to the clipboard as an image.
Print	Opens an html browser with the current view for printing.
Export as • Image • HTML • Text	Exports pathway list spreadsheet in the chosen format.
Properties	Opens a properties dialog box to alter rendering, visualization, columns to be included and view description of pathway lists.

31.4.8.2 Pathway List Name in the Experiment Navigator

The right-click options for a pathway list in the Experiment Navigator are summarized in Table 31.5.

31.4.8.3 Pathway List Panel

The available right-click options for the Pathway list panel in The Pathway View are listed in Table 31.6.

31.4.8.4 Pathway Viewer

The available options on right-clicking over the pathway displayed in the Pathway Viewer of The Pathway View depend on whether a pathway is in the GPML or BioPAX Level 3 format or not. They are listed in Table 31.7. Right-click options for KEGG pathways in the Pathway Viewer of The Pathway View are listed in Table 31.8.

Table 31.7: Right-click options for Pathway Viewer depending on pathway source

Option	Description	Non-GPML pathways (excl. BioPAX Level 3)	GPML and BioPAX Level 3 pathways
Select all	Selects all entities and relations in the pathway.	✓	✓
Select all entities	Selects all entities in the pathway.	✓	✓
Select matched entities	Selects all the entities in the pathway that match entities from the source entity list/lists.	✓	✓
Select highlighted entities	Selects all the entities that are highlighted (with a blue halo) in the pathway, indicating matching entities with the active entity list in any of the open experiments in the Experiment Navigator.	✓	✓
Clear selection	Clears all selections and highlights in the displayed pathway.	✓	✓
Export Gene Annotations	Export Entrez Gene IDs of the selected or all entities from the pathway viewer as a Text file.	✓	✓
Export to eArray	Export selected or all entities from the pathway viewer to eArray.	✓	✓
Add to Pathway List Clipboard	Adds selected entities in the displayed pathway to the Pathway List Clipboard.	✓	✓
Copy to Clipboard as Tab Separated text (.tsv)	Copies the available annotation information for the selected entities as displayed in the Properties dialog to the Clipboard in the tab separated text format.	✓	✓
NLP Network Discovery	Creates a new pathway from the relations between any selected entities.	✓	*X*
Refresh connectivity	Refreshes the connectivity between the entities in the displayed pathway.	✓	*X*
Show report for: • Entities Table • Selected Entities Table • Relations Table • Selected Relations Table	Opens a new spreadsheet view displaying the corresponding data for the selection.	✓	*X*
Filter by type	Filters the pathway by relation and entity types.	*X*	*X*
Reset filter	Resets any filters set using **Filter by type**.	*X*	*X*
Search by property	Searches entities in the displayed pathway based on their properties.	✓	*X*
Delete selection	Deletes selected entities from the displayed pathway.	*X*	*X*
Keep selection in foreground	Brings all selected entities to the foreground and grays out other entities.	✓	*X*
Bring all to foreground	All entities are brought to the foreground.	✓	*X*
Expand selection		*X*	*X*
Export as • Report • Image • Navigable HTML	Exports the pathway in the selected format. Selecting **Image** opens a dialog that allows you to specify file type, size, and resolution. Selecting Navigable HTML creates an html file where each entity and relation is hyper-linked to its properties. To transfer this saved file, you have to transfer it along with the properties folder.	✓	✓
Display settings	Opens the Pathway Property Dialog, which allows you to alter **Data settings**, **Entity Coloring**, and **HeatStrip Coloring**.	✓	✓
Properties	Displays the pathway properties.	✓	*X*

Table 31.8: Right-click options for KEGG pathways in the Pathway Viewer

Option	Description
Select all entities	Selects all entities in the pathway.
Select matched entities	Selects all the entities in the pathway that match entities from the source entity list/lists.
Select highlighted entities	Selects all the entities that are highlighted (with a blue halo) in the pathway, indicating matching entities with the active entity list in any of the open experiments in the Experiment Navigator.
Clear selection	Clears all selections and highlights in the displayed pathway.
Export Gene Annotations	Export Entrez Gene IDs of the selected or all entities from the pathway viewer as a Text file.
Export to eArray	Export selected or all entities from the pathway viewer to eArray.
Add to Pathway List Clipboard	Adds selected entities in the displayed pathway to the Pathway List Clipboard.
Copy to Clipboard as • Entrez Gene IDs • Swiss-Prot/UniProt IDs • Tab Separated text (.tsv)	Converts the identifiers of the selected entities to the selected annotation using BridgeDb and copies the corresponding identifiers to the Clipboard. Copies the available annotation information for the selected entities as displayed in the Properties dialog to the Clipboard in the tab separated text format.
Export as • Report • Image • Navigable HTML	Exports the pathway in the selected format. Selecting **Image** opens a dialog that allows you to specify file type, size, and resolution. Selecting Navigable HTML creates an html file where each entity and relation is hyper-linked to its properties. To transfer this saved file, you have to transfer it along with the properties folder.
Display settings	Opens the Pathway Property Dialog, which allows you to alter **Data settings**, **Entity Coloring**, and **HeatStrip Coloring**.

31.5 Mapping and Statistical Algorithms

This section explains the mapping and statistical algorithms used when performing MOA:

- Homologene translation (for genes) in case the target organism of the pathways is different from the organism selected in the input experiment.
- ID mapping using BridgeDb if the entities in the entity list refer to the same bio-entity as the one in the pathway but both use identifiers from different biological databases to refer to them.
- Supported data sources for mapping entities
- p-value computation for genes, metabolites, and proteins.
- Error bar computation for heatstrips.

31.5.1 HomoloGene Translation for Gene/Protein Identifiers

GeneSpring uses HomoloGene annotations, based on the HomoloGene tool available on the National Center for Biotechnology Information (NCBI) website (`http://www.ncbi.nlm.nih.gov/homologene`), to identify homologs of a gene in diverse organisms. This process is necessary for identifying matched entities for an input entity list when the chosen pathway organism differs from the organism of the input experiments. **GeneSpring** automatically downloads the HomoloGene Groups for the selected organisms in an SEA or MOA workflow when they are required. You can inspect for which organism HomoloGene Groups were downloaded, or download HomoloGene groups for additional organisms using the Annotations Manager (see To download HomoloGene groups for gene / protein identifiers on page 888).

For HomoloGene translation the experimental organism is designated as the source organism and the pathway organism becomes the target organism. HomoloGene translation uses Entrez Gene IDs to identify the homologs in different organisms and the genes from each organism are assigned to a homologroup. Genes from different organisms that map to the same homologroup are considered homologs of each other.

For this mapping process, **GeneSpring** attempts to translate annotations for the entities in both the experiment and the pathway to Entrez Gene IDs and identifies any homologs as matches.

In the SEA or MOA workflow, if Entrez Gene IDs are available and selected, it is given the first priority. For an entity that does not have an Entrez Gene ID, the remaining annotations are used in the order of priority as listed in Step 2 of the SEA or MOA workflow to retrieve the corresponding Entrez Gene ID via BridgeDb (see BridgeDb - ID Mapping on page 1012) based on the selections made in Step 2 of the SEA or MOA workflows (see To select annotations on page 908). Therefore it is recommended that you do not unselect Entrez Gene ID from the Annotations field when working with multiple organisms. Unselecting Entrez Gene ID leads to entities being disregarded for the mapping process if they only have Entrez Gene IDs, which can significantly lower the number of matching entities in the result.

31.5.2 BridgeDb - ID Mapping

To determine if a pathway is over-represented for a given experiment, **GeneSpring** matches entities in the experiment and pathway using the annotation-identifier combination for an entity in each. However, the experiments and the pathways often refer to the same bio-entity with identifiers from different annotations. For example, if the annotations for small molecules present in a Metabolomics experiment are CAS Numbers, a pathway can have ChEBI IDs or PubChem identifiers or a combination of both for the same entities.

BridgeDb (`http://www.bridgedb.org`) and custom ID Mapper databases resolve this mapping problem by mapping an entity's identifier in one annotation database to the corresponding identifier in another annotation database. Of the data sources available in BridgeDb, **GeneSpring** currently supports a restricted number of data sources for Metabolites, Genes, and Proteins (see To select annotations on page 908) for a number of organisms. Annotation databases across which mappings are supported are determined by the BridgeDb and ID Mapper databases specific to each organism. The organisms for which BridgeDb and ID Mapper databases are available are listed on the update server. You will be notified whenever updates to the BridgeDb or ID Mapper databases or new custom ID Mapper databases become available to download from the Agilent Server (see To download an organism's BridgeDb or ID Mapper database on page 894).

31.5.3 Supported Data Sources

The data sources that **GeneSpring** recognizes for identifying the entity type (Gene, Protein, Metabolite) of a pathway entity varies and matching that entity to the experiment depending on the pathway format.

GPML and BioPAX Level 3 Pathways:

- **Genes**: MGI, Ensembl B. subtilis, UniGene, Rfam, Ensembl Xenopus, CTD Gene, TAIR Gene, Ensembl Cow, ZFIN, GeneDB, Aclame, Gramene Arabidopsis, WormBase, SGD [50], Gramene Genes DB, TAIR, HOVERGEN, EMBL, Gramene Maize, Wheat gene catalog, Ensembl Pig, Other, Ensembl, RefSeq, Ensembl Chicken, Ensembl Yeast, PlantGDB, KEGG Genes, Ensembl Dog, Gramene Rice, FlyBase, Ensembl M. tuberculosis, TubercuList, HGNC, WikiGenes, EC Number, Ensembl Zebrafish, Oryzabase, TIGR, HGNC Accession number, Ensembl C. elegans, Ensembl Mosquito, OMIM, Ensembl Fruitfly, UCSC Genome Browser, Ensembl Human, Wheat gene names, Entrez Gene, IRGSP Gene, dbSNP, MaizeGDB, Gene Wiki, Ensembl E. coli, RGD, NASC Gene, miRBase Sequence, BioCyc, Ensembl Horse, Ensembl Rat, Ensembl Chimp, Rice Ensembl Gene, HsGene, Ecoli, Ensembl Mouse, GenBank, uniprot, swissprot
- **Proteins**: Database of Interacting Proteins, Wormpep, PhosphoSite Residue, Ligand Expo, KEGG Orthology, PRIDE, PDB, Rhea, MEROPS, TAIR Protein, SABIO-RK Reaction, UniParc, ChEMBL target, SPRINT, Pfam, ProDom, TTD Target, InterPro, MACiE, PeptideAtlas, Enzyme Nomenclature, SMART, Signaling Gateway, Uniprot/TrEMBL, RESID, Transport Classification Database, SABIO-RK EC Record, Conserved Domain Database, KEGG Reaction, Ligand-Gated Ion Channel database, PANTHER, PROSITE, CluSTr, PhosphoSite Protein, PIRSF, YeastCyc

- **Metabolites**: TTD Drug, PharmGKB Drug, CTD Chemical, PubChem-substance, 3DMET, KEGG Glycan, Kegg Compound, ChEBI, ChemIDplus, NuGO wiki, T3DB, DrugBank, PubChem-bioassay, LipidBank, LIPID MAPS, Wikipedia, KEGG Drug, ChEMBL compound, Anatomical Therapeutic Chemical, HMDB, CAS, Chemspider, PubChem, GlycomeDB

Pathways imported from Pathway Architect or in the Biopax Level 2 or Text format:

- **Proteins**: Entrez ID, Unigene, Protein. **GeneSpring** maps the Protein field to the data source Swiss-Prot (UniProt/TrEMBL) for matching. Of these annotations, only the annotations present in the annotations list specific to the experiment type will be considered for matching. Thus, values in the Entrez ID and Unigene fields will be considered for matching in Genomics and Transcriptomics experiments and the values in the Protein field will be considered for matching in Proteomics experiments.
- **Metabolites**: CAS

If multiple values are present in the pathway for the same annotation, **GeneSpring** only considers the first value that matches with the experiment.

31.5.4 p-value Computation

GeneSpring uses the Hypergeometric method to compute p-values. The process of computing these p-values, including the exact formula applied in **GeneSpring** is described under Find Similar Entity Lists on page 287.

No p-values are computed for entities from Metabolomics experiments during an MOA or SEA workflow to avoid a misrepresentation of the significance of matching pathways caused by the fact that **GeneSpring** uses the technology (All Entities list) as a reference for p-value computation. The technology of a metabolomics experiment is limited to only the measured metabolites with an observable abundance in the experiment. Pathways on the other hand are likely to contain many other metabolites that may not be present in the technology. This results in a pathway p-value computed with the technology as reference to be higher than a more realistic p-value computed with a comprehensive reference set of global entities. To avoid this apparent increase in significance, **GeneSpring** only reports the number of matched entities for a Metabolomics experiment.

In the absence of the applicable **BridgeDb** database for the experiment organism, the p-values calculated during an SEA as part of the **Analysis: Biological Significance** workflow type available during experiment creation and from the **Utilities** section of the Workflow Browser will differ from the p-values calculated when running an SEA directly from the Workflow Browser, which can only be completed if the applicable BridgeDb database has been downloaded. **GeneSpring** is pre-packaged with metabolite BridgeDb and human BridgeDb, and so, there is no difference in p-values computed for experimental data from humans. You can download BridgeDb for all other organisms following the steps described under To download an organism's BridgeDb databse on page 894.

31.5.5 Heatstrip Error Bar Computation

The error bar for a condition in a Heatstrip indicates the standard error (SE) of the mean for the condition in the selected interpretation.

It is defined as: $SE_{\bar{x}} = \frac{s}{\sqrt{n}}$

where, 's' is the condition standard deviation and 'n' is the size (number of samples) of the condition.

If multiple probes/ compounds in the input entity list map to the same pathway entity, the error is computed as the average of the squares of errors for that condition in each probe.

31.6 Analyzing Pathways Using External Tools Accessible from GeneSpring

In addition to the pathway analysis features available in **GeneSpring** (see Pathway Analysis in **GeneSpring** on page 896) you can also connect to other well-known, subscribed and open source tools directly from within **GeneSpring** and seamlessly move your data from one tool to the other. **GeneSpring** supports this functionality for Ingenuity Pathways Analysis (IPA), MetaCore™, and Cytoscape. The respective workflows for each of these tools are described in the following sections.

31.6.1 Ingenuity Pathways Analysis (IPA) Connector

The **GeneSpring**-IPA Connector enables users of **GeneSpring** and Ingenuity Pathways Analysis (IPA) to exchange information seamlessly between the two applications. Genes of interest can be identified using powerful statistical and analytical tools in **GeneSpring**. The biological context and significance of these findings can then be assessed in IPA using its various analysis tools. With the **GeneSpring**-IPA Connector, users will be able to send Gene Lists and associated expression data directly to IPA. Once in IPA, users can perform network analyses, build pathways, view relevant canonical pathways, and obtain proprietary information on protein interactions and pathways. Users can then send a list of genes back to **GeneSpring** (only available for some experiment types), allowing further iterative analysis of those genes in **GeneSpring**. Both options can be accessed from the **Pathway Analysis** section of the Workflow Browser.

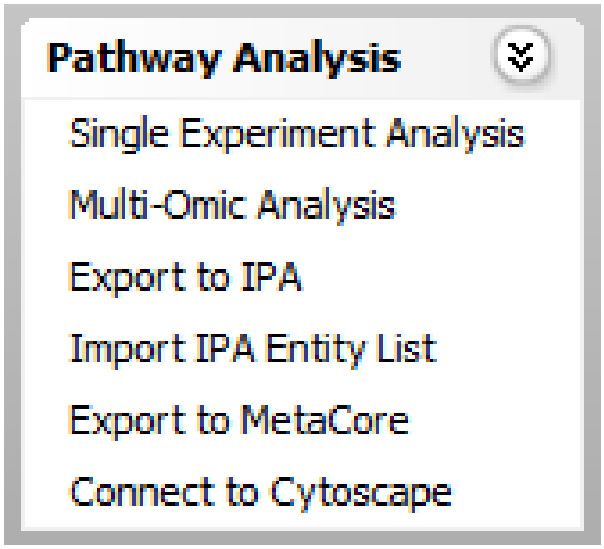


Figure 31.89: IPA options in the workflow browser

Select **Export to IPA** from the workflow to access the three IPA options supported in **GeneSpring**:

- Create Pathway in IPA
- Perform Data Analysis on Experiment
- Perform Data Analysis on Entity List

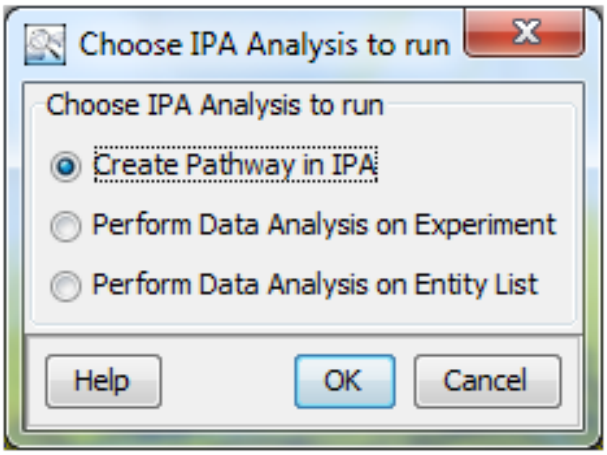


Figure 31.90: The Export to IPA dialog

31.6.1.1 Create Pathway in IPA

Select **Create Pathway in IPA** from the Export to IPA dialog which allows you to send an Entity List from **GeneSpring** to IPA and use those genes to create a pathway in IPA. This pathway can then be subjected to further manipulation and analysis in IPA by growing a node, removing nodes and interactions, and interrogating a node or an interaction.

The **Create New Pathway** dialog (Figure 31.91) opens with the following options:

- **Entity List**: By default, the active entity list is selected. Click **Choose...** to select a different entity list.
- **IPA Server address**:Name of the server running IPA. By default this will point to the main server of Ingenuity "`analysis.ingenuity.com`". To choose a different server, enter the appropriate server address. The IPA Server address can also be permanently configured from **Tools** $\longrightarrow$ **Options** $\longrightarrow$ **Miscellaneous** $\longrightarrow$ **Pathway Analysis (IPA)**.
- **Pathway Name**: The name for the new pathway to be created. By default, the name of the entity list that was originally selected is used. If you selected a different entity list in this dialog the name for the pathway is not updated to reflect the new selection.
- **Project Folder**: The name of the project in IPA under which this pathway should be stored. By default, **GeneSpring** uses the name of the **GeneSpring** project.
- **Gene Identifier Column**: This indicates the type of gene identifier that will be used to map genes in the entity list to genes in the Ingenuity Pathways Knowledge Base (IPKB). The type of identifier selected determines from which annotation column in the **GeneSpring** technology the identifiers are retrieved. These identifiers are used to map genes in the list to genes in IPKB. Only identifiers that can be matched to genes in IPKB are used to build a new pathway in IPA. The supported identifiers are:

  - Entrez Gene ID
  - Locus Link ID
  - Affymetrix Probeset ID
  - UniGene
  - GenBank Accession
  - Swiss-Prot
  - Agilent Probe ID
  - RefSeq Protein ID
- **Save Pathway**: Indicates whether the pathway generated will be saved in IPA. If selected, the new pathway is saved in IPA to the specified Project Folder, within My Pathways, under the specified Pathway Name.

Figure 31.91: Create New Pathway

Click **OK** to create the pathway in IPA. Your default browser will start up and connect to the IPA Server.

The connection to IPA is performed through the IPA HTTP API and on some systems this can result in the session being blocked due to the security settings as shown below. If this happens, simply click on the yellow bar and select **Allow Blocked Content**. There is no danger in allowing this access.

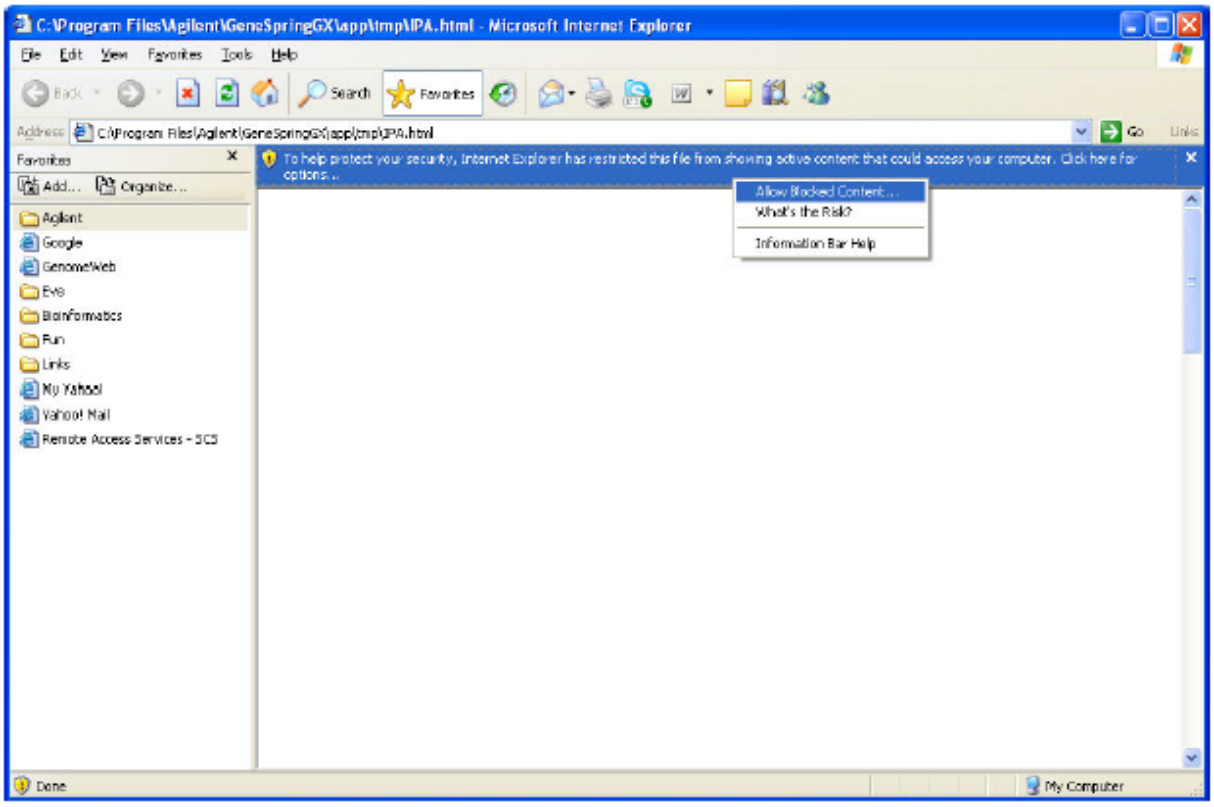


Figure 31.92: IPA Pathway Creation - security warning

IPA uses Java Webstart and a dialog appears indicating JAVA is starting up.

Figure 31.93: Java Startup

The login dialog will appear. Enter your IPA login credentials here.

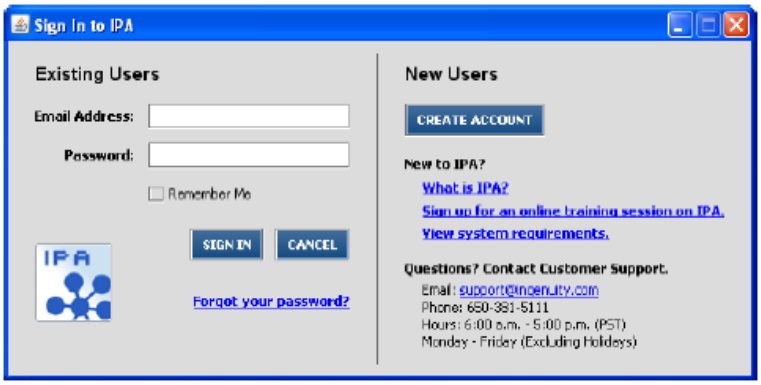


Figure 31.94: IPA Login Dialog

IPA will start up and show the pathway as illustrated in Figure 31.95. At this point you can continue your analysis in IPA.

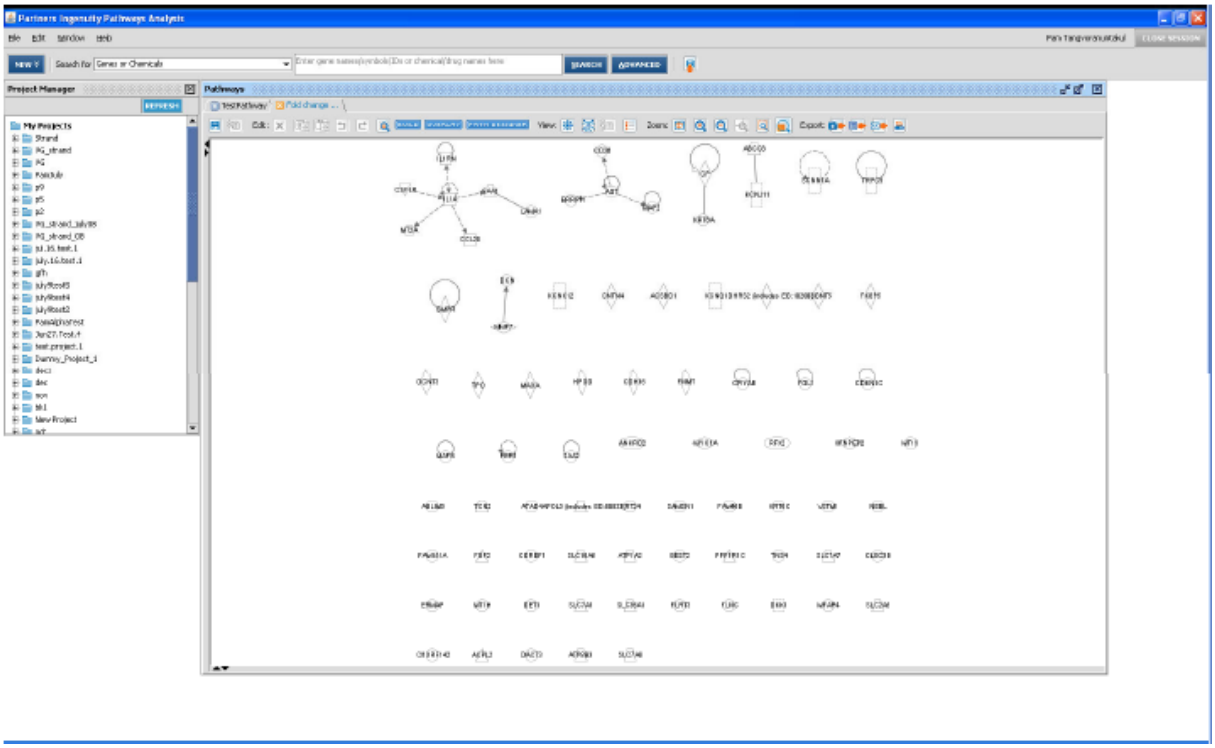


Figure 31.95: Pathway Analysis in IPA

31.6.1.2 Import IPA Entity List

To send a list of entities from IPA back to **GeneSpring** for further analysis, create a list in IPA and use the **Import IPA Entity List** option in the **Pathway Analysis** section of the Workflow Browser. To create a list of pathway genes in IPA:

1. Select the molecules of interest by dragging a rectangle around the molecules of interest in the pathway viewer. These molecules will then be highlighted. Click **Add Selected Molecules to List** and select **New List** to create a new list or **Saved List** to add the selected genes to an existing gene list. See Figure 31.96.

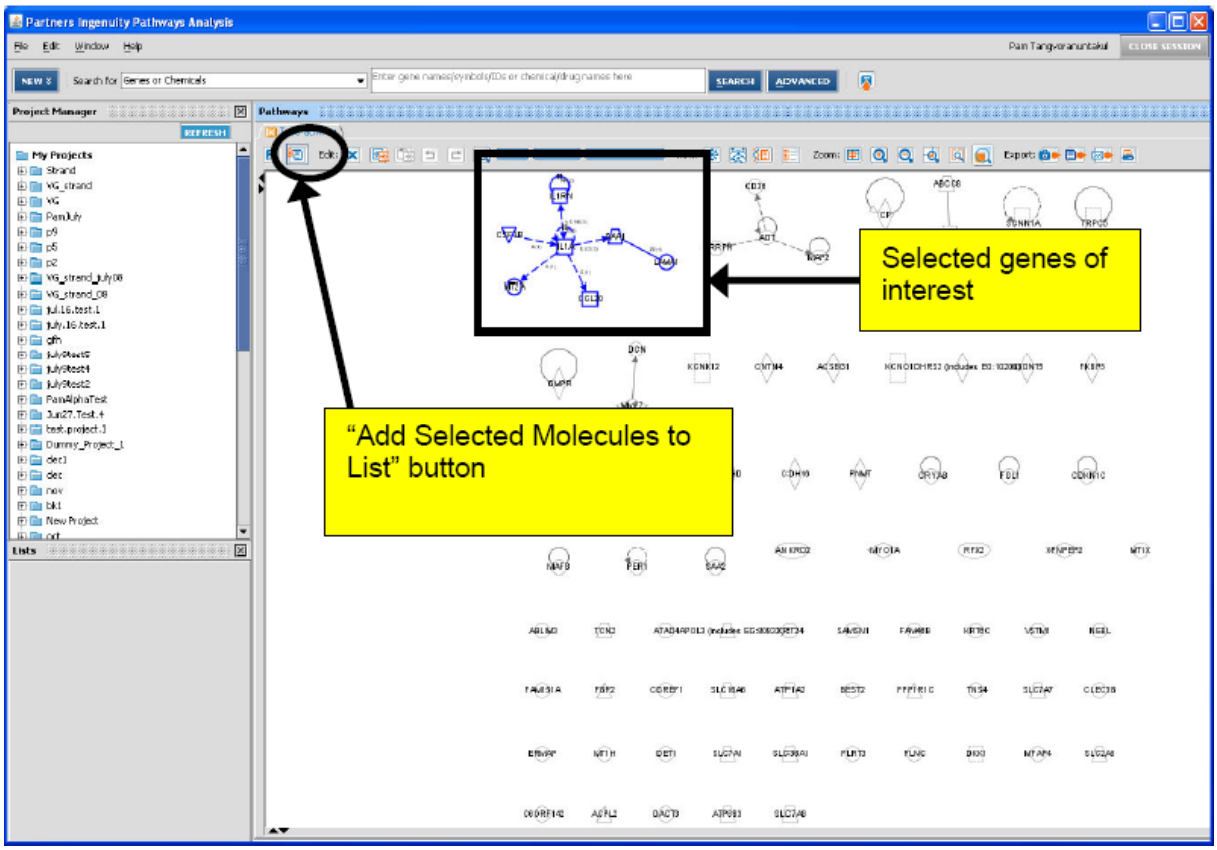


Figure 31.96: Creation of Entity List

Alternatively, select the genes for export to **GeneSpring** in the Network tab of the analysis results and create a list from them. See Figure 31.97.

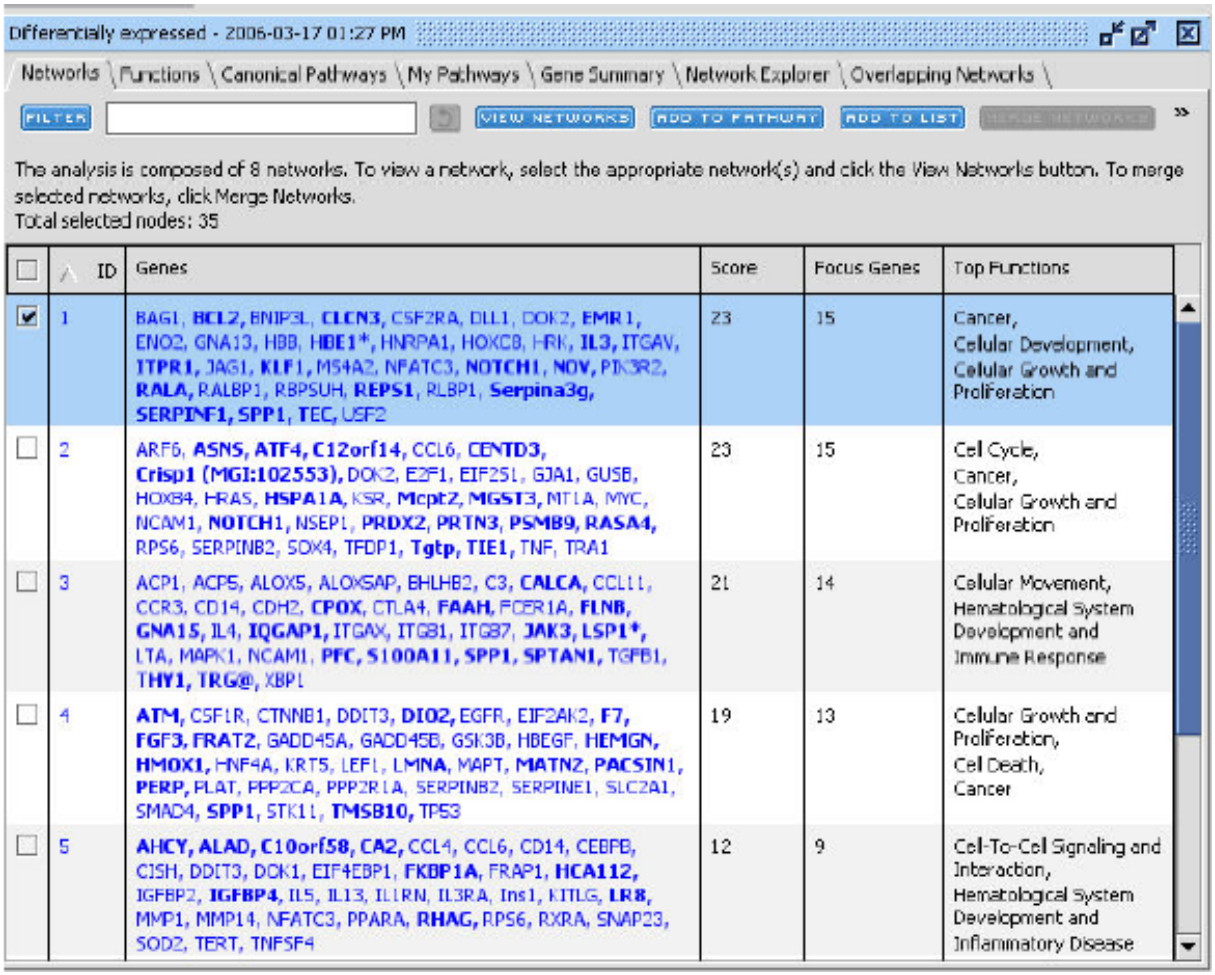


Figure 31.97: Creation of Entity List

2. Indicate List Name and Project to save the gene list in the specified Project folder under My Lists. See Figure 31.98.

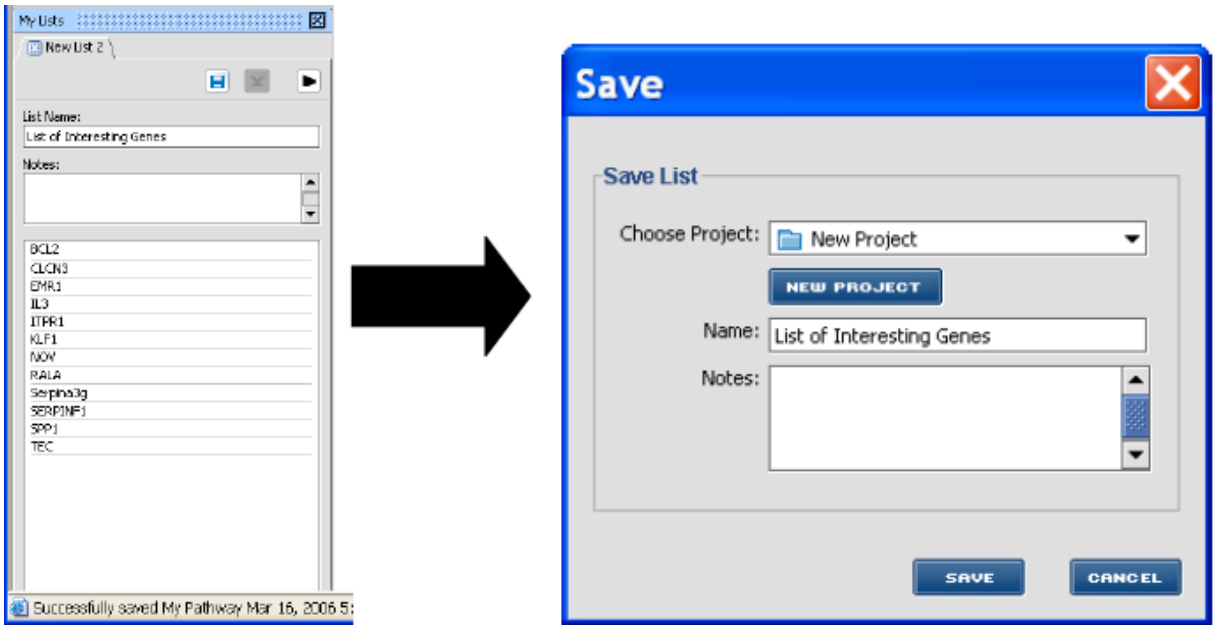


Figure 31.98: Save List

3. Click on the black arrow. Select **Send to** and select **GeneSpring**. Specify the settings such as the type of identifier to use to map genes in the gene list to genes in the **GeneSpring** experiment. Click **OK**. See Figure 31.99.

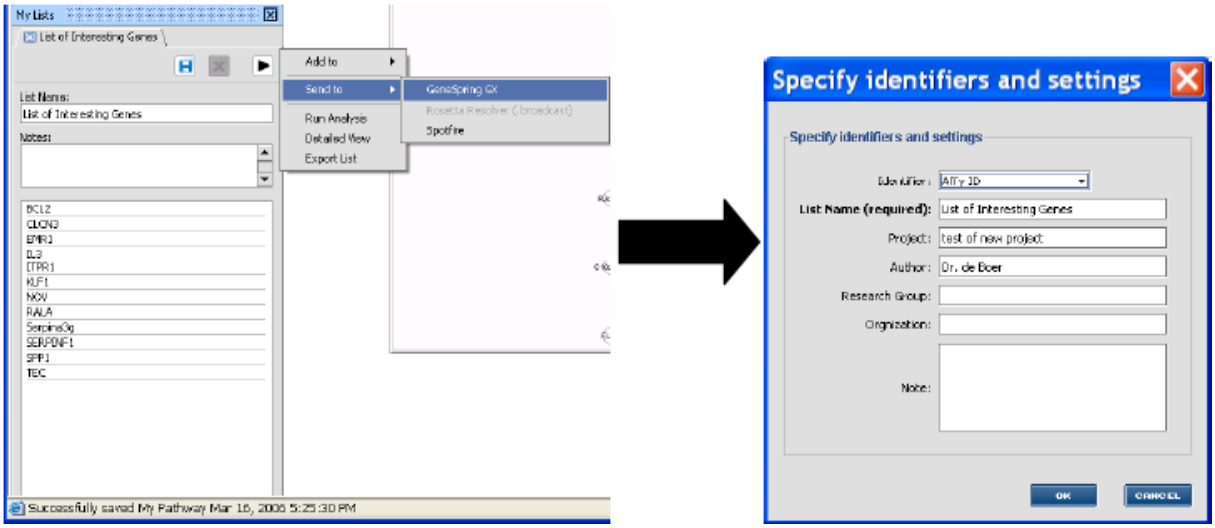


Figure 31.99: **GeneSpring** suitable list creation

4. Indicate the location to where the file will be saved. The gene list will be saved as a **GeneSpring**.zip file. See Figure 31.100.

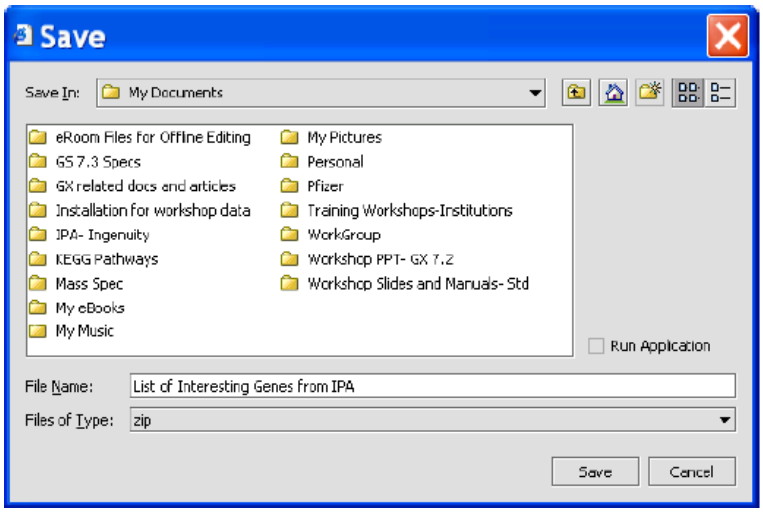


Figure 31.100: Saved List Location

To import the gene list into **GeneSpring**, use the option **Import List from IPA** from the Pathway Analysis workflow.

This prompts you with a dialog to select the .zip file created from IPA in the previous step. Select the *.zip* file and click **Open**.

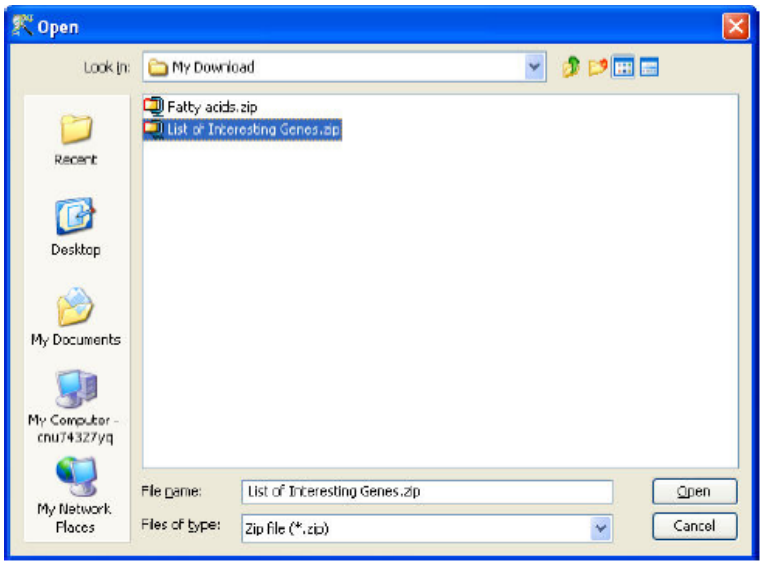


Figure 31.101: Selection of Folder

The option will import the gene list stored in this .zip file to a new Entity List with the same name you provided in the previous step in IPA. The new Entity List is saved in a folder called Imported Lists, which is created automatically. See Figure 31.102.

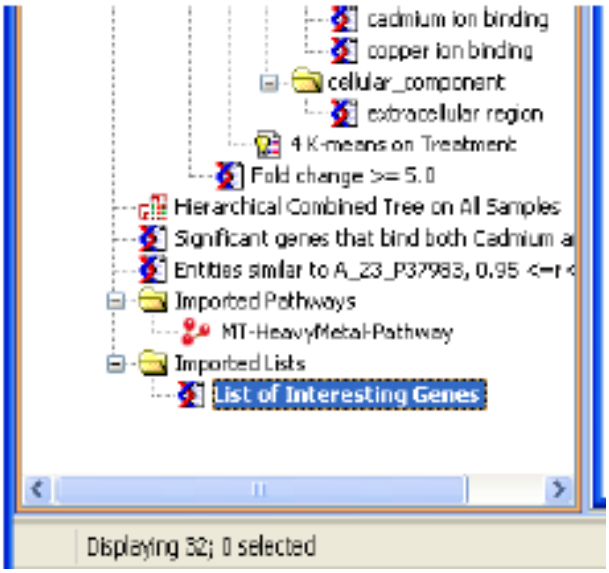


Figure 31.102: Entity List Creation

NOTE Only identifiers listed in the previous section are supported in matching the imported genes to entities in the current technology. If no entities can be matched, then the option shows this error.

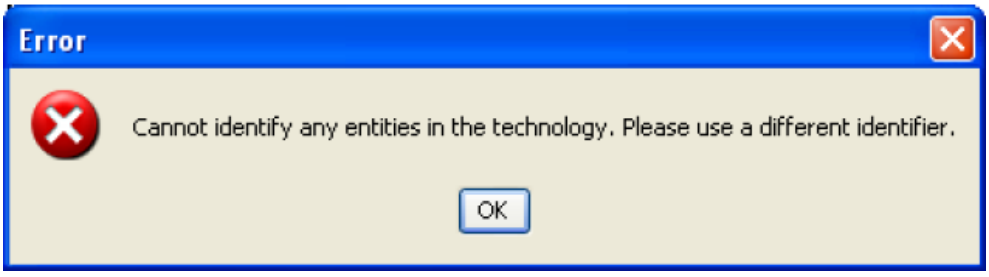


Figure 31.103: Import Entity List from IPA - Error Message

As indicated in the error, you can select a different identifier in IPA to represent the chosen genes. It is recommended to choose Entrez Gene IDs since many of the technologies in **GeneSpring** are likely to have this type of identifier.

31.6.1.3 Perform Data Analysis on Experiment

This option allows you to send an entity list and the associated gene expression data from **GeneSpring** to IPA to perform data analysis in IPA. Genes in the entity list that are also found in IPKB are used as Focus Genes to build networks. The networks can be subjected to further manipulation and analysis in IPA by

growing a node, removing nodes and interactions, interrogating a node or an interaction, and perform Function, Canonical Pathways, My Pathways, Gene Summary, and Overlapping Networks analyses. You are able to create gene lists from the generated networks and send the gene lists back to **GeneSpring** (see Import IPA Entity List on page 1020).

Select **Export to IPA** from the **Pathway Analysis** section in the Workflow browser. The IPA dialog appears with the three activities that are supported in **GeneSpring**.

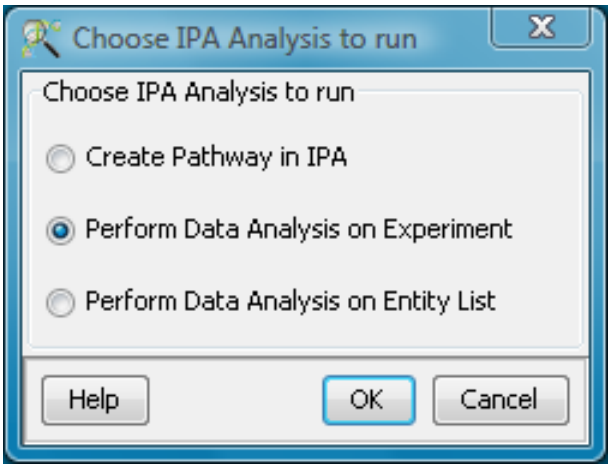


Figure 31.104: Perform Data Analysis on Experiment

Select **Perform Data Analysis on Experiment** and the corresponding dialog opens with the following options:

- **Entity List**: The active entity list is selected. To use a different entity list, cancel the option, select a different entity list in the Experiment Navigator and select the option again.
- **Experiment Interpretation**: The active interpretation is selected. To change the interpretation, click **Choose...**. The experiment interpretation defines the conditions for which the log_2 values for each gene are sent to IPA. Log_2 values for the conditions in the selected experiment interpretation are sent to IPA for analysis. The name of the data set used in IPA is named after the source experiment in **GeneSpring**.

NOTE IPA only allows unique names for datasets per project. To analyze the same experiment more than once, either change the name of the experiment or change the project name.

- **IPA Server address**: Name of the server running IPA. By default this points to the main server of Ingenuity "analysis.ingenuity.com". To choose a different server, enter the server address of the desired server. The IPA server address can also be permanently configured from **Tools** $\longrightarrow$ **Options** $\longrightarrow$ **Miscellaneous** $\longrightarrow$ **Pathway Analysis(IPA)**.

- **Project Folder**: The name of the project in IPA under which you want to store this pathway. By default, IPA will use the name of the **GeneSpring** project.
- **Use both Direct and Indirect relationships**: If you select **Yes**, IPA builds networks using both direct and indirect molecular interactions between genes. If you select **No**, IPA builds networks using only direct interactions between genes.
- **Knowledge Base content**: This indicates which database is searched for information to build the network. An empty string indicates all available Knowledge Bases that are searched and information from all sources is used to perform analysis.
- **Include My Pathways in Enrichment Score**: If you select **Yes**, all pathways saved under My Pathways are included in the scoring process.
- **Review Settings and ID Mapping before Running Analysis**: If you select **No**, data analysis is automatically performed using the settings defined in this option. If you select **Yes**, the **Create Analysis** dialog opens for you to review and modify settings before running analysis.
- **Gene Identifier Column**: This indicates the type of gene identifier that is used to map genes in the entity list to genes in the Ingenuity Pathways Knowledge Base (IPKB). The type of identifier selected, determines from which annotation column in the **GeneSpring** technology the identifiers are retrieved. These identifiers are used to map genes in the list to genes in IPKB. Only identifiers that can be matched to genes in IPKB are used to build a new pathway in IPA. The list of supported identifiers is:
  - Entrez Gene ID
  - Locus Link ID
  - Affymetrix Probeset ID
  - UniGene
  - GenBank Accession
  - Swiss-Prot
  - Agilent Probe ID
  - RefSeq Protein ID

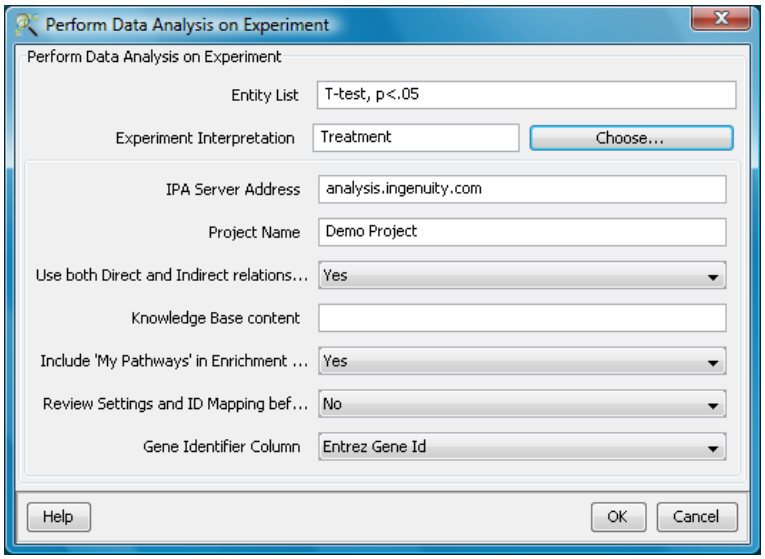


Figure 31.105: Perform Data Analysis on Experiment

Click **OK** to start the analysis in IPA. Your default browser will start up and connect to the IPA Server.

IPA analysis

The connection to IPA is performed through the IPA HTTP API and on some systems this may result in the session being blocked due to the security settings as shown below. If this happens, simply click on the yellow bar and select 'Allow Blocked Content'. There is no danger in allowing this access. See Figure 31.92.

IPA is using Java Webstart and a dialog appears indicating JAVA is starting up. See Figure 31.93. The login dialog appears. Enter your login credentials here. See Figure 31.94.

If **Review settings and ID mapping before running an analysis** was selected, the following screen 31.106 in IPA appears, if not, the main window opens immediately.

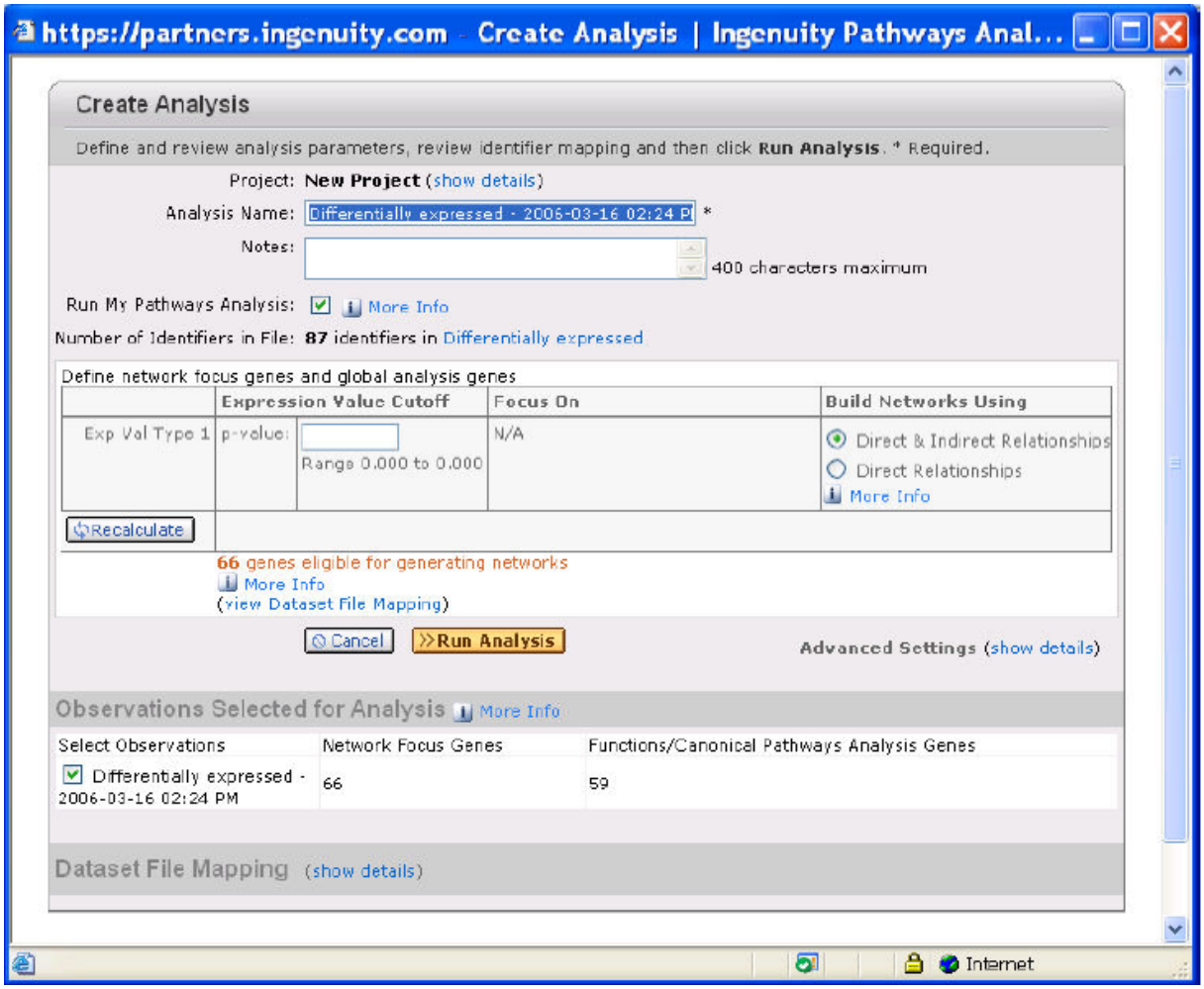


Figure 31.106: Create Analysis

Review the analysis settings and click **Run Analysis** to proceed. The analysis is performed and a little clock appears while it is running. When the analysis is finished, the clock disappears and the analysis is ready for review.

NOTE IPA only allows unique names for datasets per project. To analyze the same experiment more than once, either change the name of the experiment or change the project name.

31.6.1.4 Perform Data Analysis on Entity List

This option allows you to send an entity list, with or without list-associated values, from **GeneSpring** to IPA to perform data analysis in IPA. Genes in the entity list that are also found in IPKB are used as Focus Genes to build networks. The networks can be subjected to further manipulation and analysis in IPA by growing a node, removing nodes and interactions, interrogating a node or an interaction, and perform Function, Canonical Pathways, My Pathways, Gene Summary, and Overlapping Networks analyses. You are able to create gene lists from the generated networks and send the gene lists back to **GeneSpring** (see Import IPA Entity List on page 1020).

Select **Export to IPA** from the **Pathway Analysis** section in the Workflow Browser. The IPA dialog appears with the three activities that are supported in **GeneSpring**.

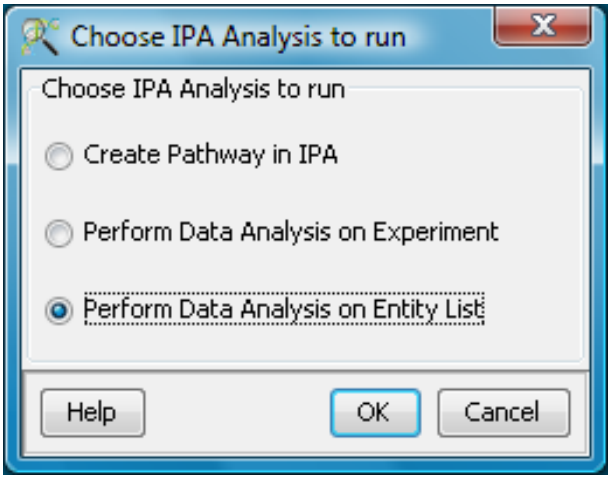


Figure 31.107: Data Analysis on Entity List

Select **Perform Data Analysis on Entity List** and the corresponding dialog appears with these options:

- **Entity List**: The active entity list in the open **GeneSpring** experiment is pre-selected by default. To use a different entity list, cancel the option, select a different entity list in **GeneSpring** and select the option again.

NOTE IPA will use the name of the Entity List to name the data set in IPA. IPA will only allow unique names for the datasets per project. To perform a different analysis on the same entity list, perform the analysis from within IPA or change the project name in this dialog.

- **IPA Server address**: Name of the server running IPA. By default, this will point to the main server of Ingenuity "analysis.ingenuity.com". To choose a different server, enter the appropriate server address. The IPA Server address can also be permanently configured from **Tools** ⟶ **Options** ⟶ **Miscellaneous** ⟶ **Pathway Analysis(IPA)**
- **Project Folder**: The name of the project in IPA under which this pathway should be stored. By default, the name of the **GeneSpring** project is used.
- **Use both Direct and Indirect relationships**: If you select **Yes**, IPA builds networks using both direct and indirect molecular interactions between genes. If you select **No**, IPA builds networks using only direct interactions between genes.
- **Knowledge Base content**: This indicates what database is searched for information to build the network. An empty string indicates all available Knowledge Bases are searched and information from all sources are used to perform analysis.
- **Include My Pathways in Enrichment Score**: If you select **Yes**, all pathways saved under My Pathways are included in the scoring process.
- **Review Settings and ID Mapping before Running Analysis**: If you select **No**, data analysis is automatically performed using the settings defined in this option. If you select **Yes**, the **Create Analysis** window is displayed for you to review and modify settings before running analysis.
- **Gene Identifier Column**: This indicates the type of gene identifier that is used to map genes in the entity list to genes in the Ingenuity Pathways Knowledge Base (IPKB). The type of identifier selected, determines from which annotation column in the **GeneSpring** technology the identifiers are retrieved. These identifiers are used to map genes in the list to genes in IPKB. Only identifiers that can be matched to genes in IPKB are used to build a new pathway in IPA.

  The list of supported identifiers is:
  - Entrez Gene ID
  - Locus Link ID
  - Affymetrix Probeset ID
  - UniGene
  - GenBank Accession
  - Swiss-Prot
  - Agilent Probe ID
  - RefSeq Protein ID
- **Associated Value Column**: If the selected entity list has one or more associated value columns, this drop-down contains the names of the columns and one column can be selected.
- **Associated Value Type**: This sets the type of value that is represented by the associated value.

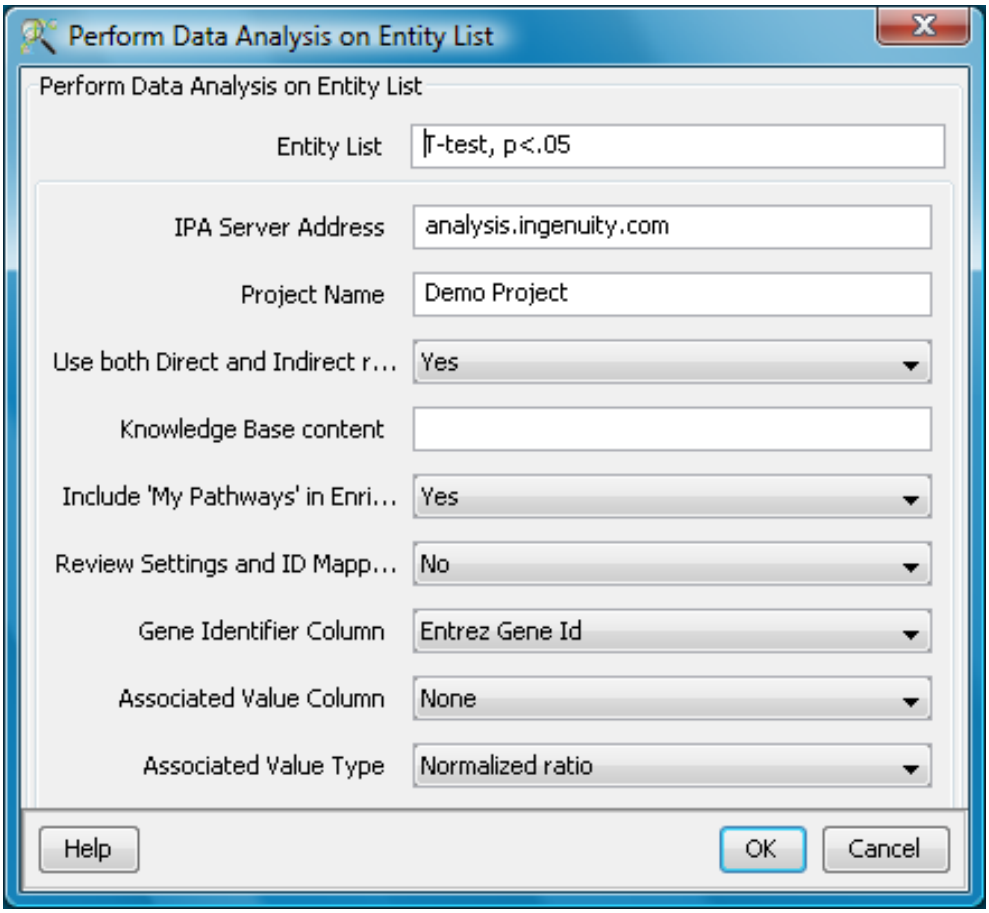


Figure 31.108: Perform Data Analysis on Entity List

Click **OK** to start the analysis in IPA. Your default browser will start up and connect to the IPA Server.

For the fold change selection, **GeneSpring** converts the **GeneSpring** notation (2, up and 4, down, etc.) to the IPA notation (2 and -4). The gsgs notation is stored in two columns but is converted to one number in accordance with the IPA notation. All other settings do not result in any conversion and it is up to you to choose the correct Value Type.

IPA analysis

The connection to IPA is performed through the IPA HTTP API and on some systems this may result in the session being blocked due to the security settings as shown below. If this happens, simply click on the yellow bar and select 'Allow Blocked Content'. There is no danger in allowing this access. See Figure 31.92.

IPA uses Java Webstart and a dialog appears indicating JAVA is starting up. See Figure 31.93. The login dialog appears. Enter your login credentials here. See Figure 31.94.

If **Review settings** and **ID mapping before running an analysis** was selected, the screen (See Figure 31.106) in IPA appears, if not, the main window opens immediately.

Review the analysis settings and click **Run Analysis** to proceed. The analysis is performed and a little clock appears while it is running. When the analysis is

finished, the clock disappears and the analysis is ready for review.

31.6.2 MetaCore™ Connector

The MetaCore™ Connector enables you to seamlessly export **GeneSpring** entity lists and their associated expression data to GeneGo MetaCore. Genes of interest can be identified using the powerful statistical and analytical tools in **GeneSpring**. The biological context and significance of these findings can then be assessed in MetaCore.

31.6.2.1 Using the MetaCore Connector

The MetaCore feature can be accessed from the **Export to MetaCore** link from the **Pathway Analysis** section of the Workflow browser. A dialog opens with the following options:

- **Entity List**: By default, the active entity list is selected. Click **Choose...** to select a different entity list. You can select the **All Entities** entity list to send all the data in the experiment to MetaCore or you can select any other sub-entity list of your choice, which only sends the corresponding subset of data to MetaCore.
- **Interpretation**: By default, the active interpretation is selected. Click **Choose...** to select a different interpretation. You can only send one column of data at a time to MetaCore. The choice of interpretation allows you to control which type of data is sent to MetaCore (sample-wise or condition-wise). If an averaged interpretation is chosen, then one particular condition's data is sent to MetaCore. In this case, the intensity values have already been averaged across samples in that condition. If a non-averaged interpretation is chosen, then you can send data one sample at a time.
- **MetaCore Server Address**: Points to `http://portal.genego.com` by default. It can be changed to point to your organization's installation. You must have a valid account on the server in order to be able to login to the portal at the end of this process.
- **Experiment Prefix**: The exported data is contained within an experiment in MetaCore. This option sets a prefix string to the name of this experiment. The default is a time-stamped string.
- **Gene Identifier Column**: This option sets the identifier of the data column that is exported to MetaCore. Currently, there is only one option: Entrez Gene ID. The rest of this section assumes that the Entrez Gene ID is selected as the identifier. If the Entrez Gene ID annotation is not present for the technology of the chosen entity list, then you will have to update the technology with Entrez Gene ID annotations before proceeding (see Update Technology Annotations on page 83). Only those entities with an associated Entrez Gene ID annotation are exported to MetaCore. The exported data in MetaCore consists of Entrez Gene IDs and an expression value (normalized intensity value in log_2 scale) for each Entrez Gene ID.

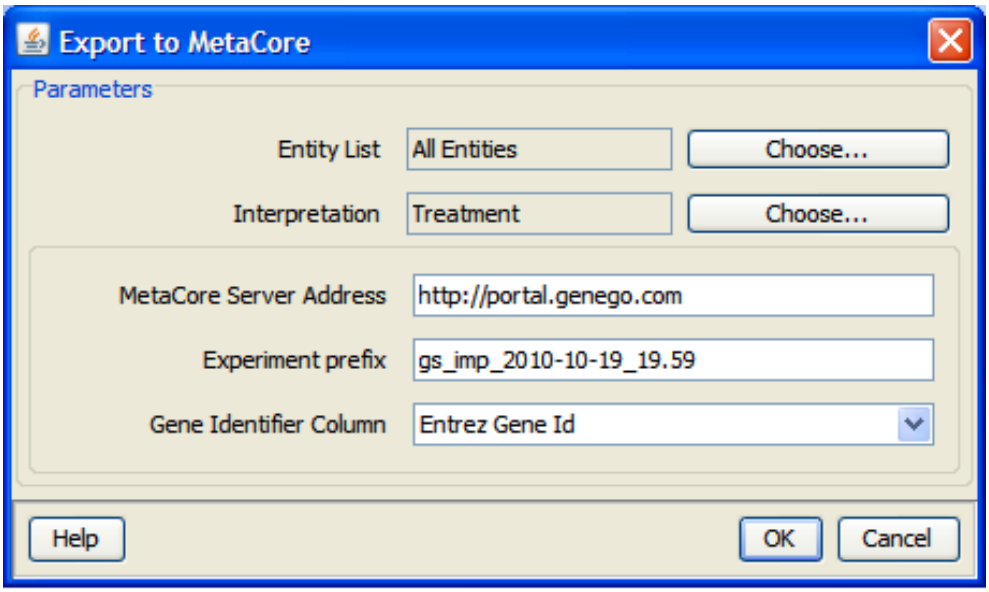


Figure 31.109: Export to MetaCore - Choose Parameters

The following rules determine the translation between the entity list identifier column and the Entrez Gene ID annotation. Let i be an entity in the chosen entity list and let E be the corresponding Entrez Gene ID annotation for that entity, then the following scenarios are possible:

1. **Null entry:** There is no Entrez Gene ID annotation for i, therefore E is null or empty: In this case, i is ignored and there is no corresponding row in the exported MetaCore data.
2. **One-to-one mapping:** i has just one Entrez Gene ID E: This corresponds to one row in the exported MetaCore data, with the expression value for i associated with E.
3. **One-Entity-to-many-Entrez-Gene-IDs mapping:** i has multiple Entrez Gene IDs, say E_1, E_2, etc.: This corresponds to multiple rows in the exported MetaCore data (as many rows as Entrez Gene IDs). Each Entrez Gene ID E_i will have the same expression value i.
4. **Many-Entities-to-one-Entrez-ID mapping:** Multiple entities, say i_1, i_2 etc. have the same Entrez Gene ID E: This corresponds to one row in the exported MetaCore data. The expression value for E is the geometric mean of the intensity values corresponding to the entities i_1, i_2, etc.

On clicking **OK**, another dialog opens where you can choose:

- **Data Column**: You can choose a single column of associated expression values to send to MetaCore i.e. only one expression value per Entrez Gene ID can be sent. Further, that expression value is always the normalized expression value. That expression value can be the normalized expression value per sample (in case the chosen interpretation is an un-averaged interpretation) or the normalized expression value per condition (if the chosen interpretation is an averaged interpretation).
- **Experiment Suffix**: This option sets the suffix string to the name of the created experiment in MetaCore. If left blank, the suffix will be the name of the chosen data column.

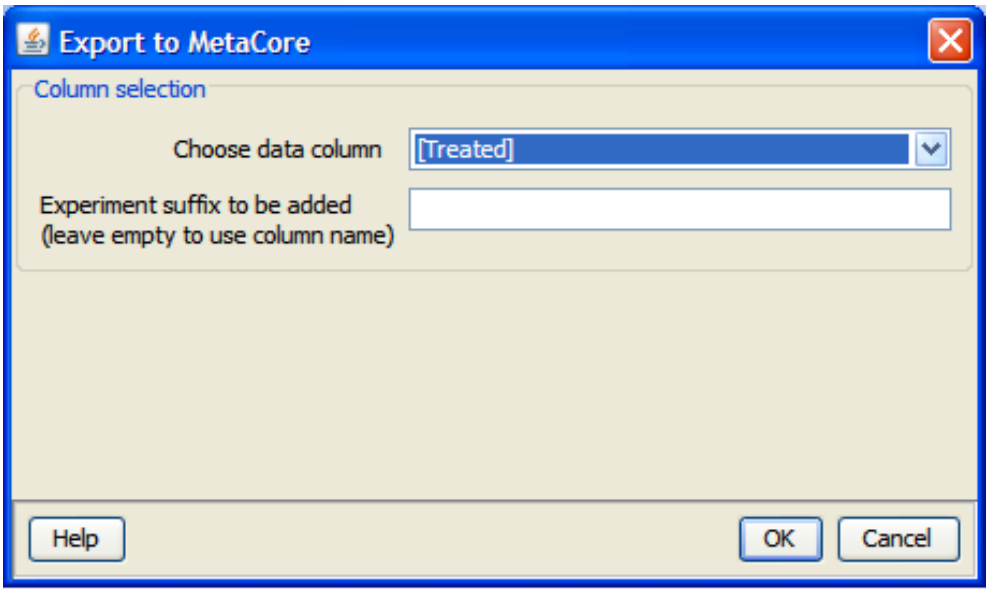


Figure 31.110: Export to MetaCore - Choose Data Column

Upon clicking **OK**, the MetaCore page corresponding to the exported entity list will open in a new browser window. Enter your login details when prompted.

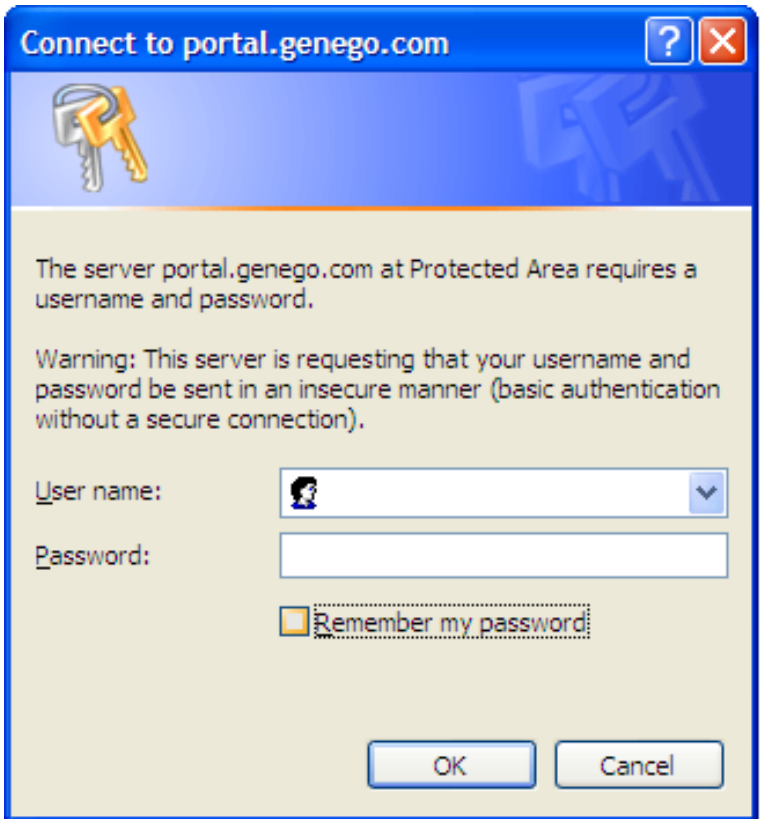


Figure 31.111: Export to MetaCore - Login Details

The newly opened MetaCore browser window displays the **Background Processing Status** page where you can see the exported data attached to the named experiment. After the status of the experiment shows **FINISHED**, you can click the experiment name to continue working on the experiment.

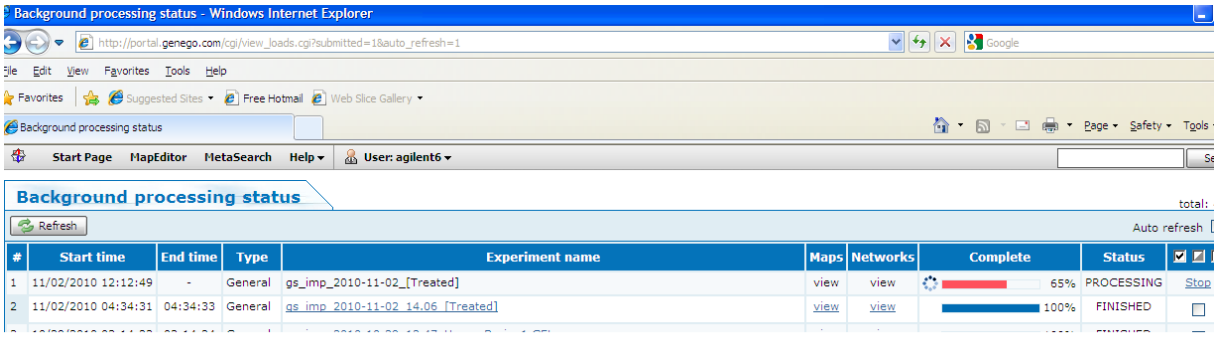


Figure 31.112: Export to MetaCore - Processing

Your experiments can also be seen in the **EXPERIMENTS** folder under **My Data**, as illustrated in Figure 31.113. The export process is now complete.

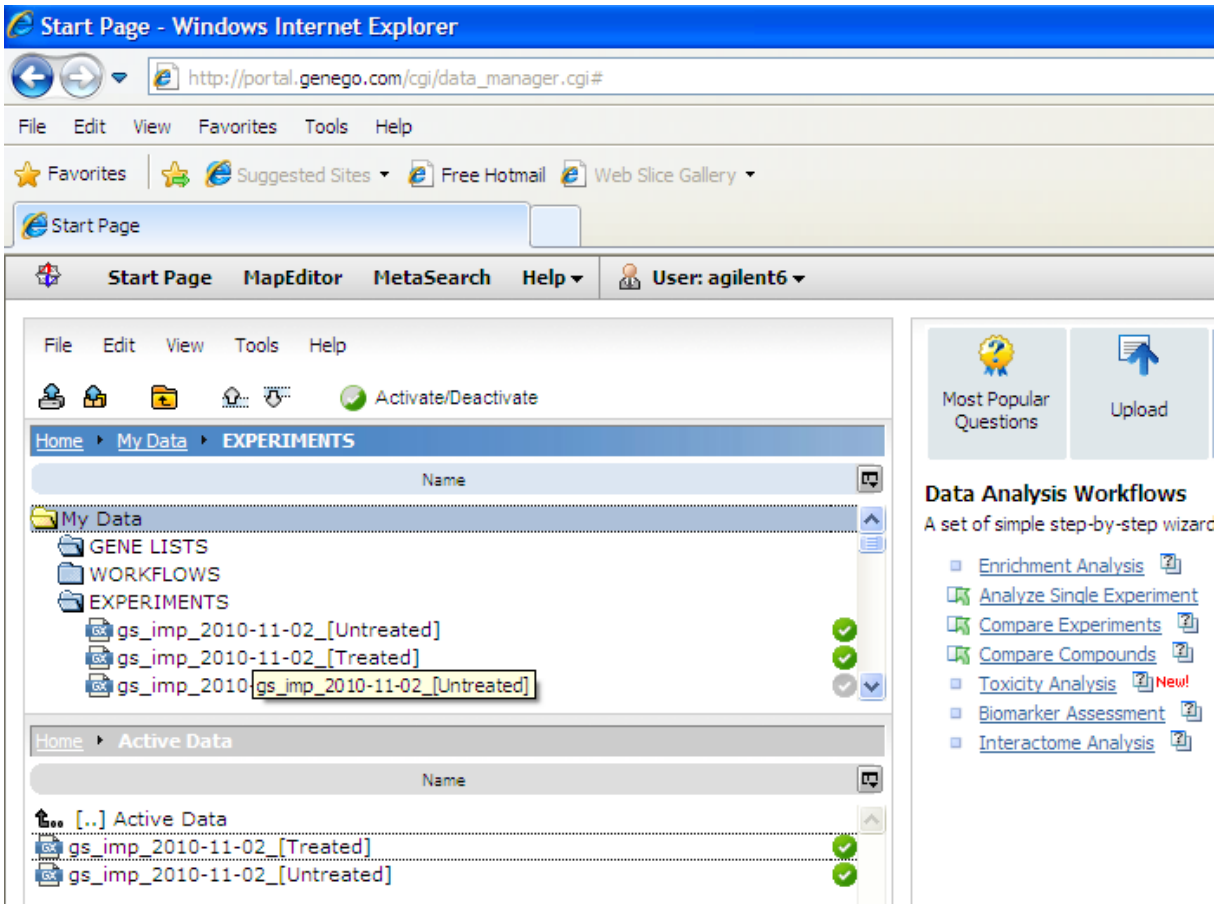


Figure 31.113: Export to MetaCore - Experiments View

31.6.3 Cytoscape Connector

The Cytoscape Connector in **GeneSpring** enables you to seamlessly export **GeneSpring** entity lists and their associated expression data to your installation of Cytoscape.

This option is only available for Windows operating systems. In addition, Cytoscape has to be installed in a folder for which you have full write permissions if you are working in Windows 7.

To connect to Cytoscape:

- Navigate to **GeneSpring** `installation_directory\app\cystoscape\lib` and copy all the nine *.jar* files to `Cytoscape_installation_directory\plugins`
- Configure the installation path for Cytoscape via **Tools** $\longrightarrow$ **Options** $\longrightarrow$ **Miscellaneous** $\longrightarrow$ **Cytoscape Installation Directory Path**.

Then choose **Connect to Cytoscape** in the **Pathway Analysis** section of the Workflow Browser to start exporting entity lists from **GeneSpring** to Cytoscape.

NOTE

If the **Connect to Cytoscape** feature fails to launch the Cytoscape Connector, check the Java heap size memory setting for Cytoscape which determines the amount of memory for launching the Cytoscape. Configure the **Memory for Cytoscape** via **Tools** $\longrightarrow$ **Options** $\longrightarrow$ **Miscellaneous** $\longrightarrow$ **Cytoscape Installation Directory Path**. If you get an out of memory error or **GeneSpring** shuts-down, increase the memory value (default is **1586 MB**) in steps of 100 and try connecting to Cytoscape. If Cytoscape fails to launch or the exception says, '*Could not reserve enough space for object heap*', decrease the memory value in steps of 100 and try connecting to Cytoscape.

Chapter 32

NLP Networks

This section provides detailed information on the basic principles of the proprietary Natural Language Processing (NLP) algorithm that is used to create the Interaction Database for all the supported organisms and describes the content of the Interaction Database. It also describes the various options available in **GeneSpring** to create pathways from PubMed abstracts, the MeSH database, selected entities, or personal data sources using NLP.

These features are part of the Pathway Analysis module of GeneSpring, which require a separate license (see Licensing on page 862) to be enabled.

32.1 Natural Language Processing (NLP) in GeneSpring

A majority of relations in the **GeneSpring** Interaction Databases are derived using a Natural Language Processing (NLP) algorithm that runs on published Medline abstracts and are updated quarterly. NLP is based on a ‘‘deep parsing” method and is driven by an elaborate sentence grammar that maximizes accuracy and has control over different aspects of a sentence without compromising recall. The NLP system operates on a sentence-by-sentence manner and extracts only those relations that are completely within a sentence. The four main phases - entity recognition, syntax analysis, semantic analysis, and semantic inference - are described in greater detail in this section.

32.1.1 Entity Recognition

Entities occurring in a sentence are tagged by consulting entity dictionaries. Since there are variations in how terms appear in literature (hyphens, Greek letter), **GeneSpring** uses a tokenization procedure, which breaks every term into tokens, identifies matching tokens and strings back collection of matching tokens into terms. Some of the cases handled by entity recognition are:

- alternate ways of writing the same name: NFkappa-B, NF Kappa B.
- local abbreviation expansion: abbreviations defined early in the abstract are expanded in the rest of the abstract before entity recognition. For example, caspase 1, 3 is recognized as caspase 1, caspase 3.

Only sentences in which at least two entities have been recognized are passed on to the next stage.

32.1.2 Syntax Analysis

Using a set of rules, the syntactic tree structure of the sentence is derived using context free grammar rules for English. The syntactic structure breaks up the sentence into its underlying linguistic constituents like noun, verb, and adverbial/propositional clauses and phrases.

GeneSpring also captures the functional roles of different parts of the sentence like the subject, object, subject modifier, object modifier, and predicate modifier. Before sentences in which more than two entities were recognized are parsed for their syntax tree, a lexical analysis step refers each word to a dictionary of English words derived from the UMLS Lexicon. This allows **GeneSpring** to handle variations caused by tense, number, person, etc.

Several different syntactical structures of sentences can carry the same semantic context:

- A regulates B
- B is regulated by A
- A plays a role in the recognition of B
- A belongs to a family of B regulators
- B activity was found to be regulated by the addition of A

The next two stages: Semantic analysis and inference recognize sentences with different structures that carry the same meaning.

32.1.3 Semantic Analysis

This phase converts all syntax trees into semantic trees using a semantic dictionary that maps all words of interest to semantic concepts. For example, the word ‘‘modulate” maps to ‘‘regulate”. The semantic tree identifies which entity regulates another entity using the sentence structure imposed by the syntax tree. The relationships captured by the semantic tree are only direct relationships, which are only a fraction of the total number of relations. Indirect relations can be captured by making inferences across multiple relationships present.

32.1.4 Semantic Inference

An example of a sentence that requires **GeneSpring** to infer a relationship is as follows: ‘‘Treatment by A causes an increase in the amount of B.” The semantic tree yields the following concepts: cause, treatment, increase, amount. To conclude that A regulates B, **GeneSpring** has to make inferences across these four semantic concepts, using agents in one relationship to fill the missing holes in other relationships. The process is driven by domain specific inference rules which make several passes through the semantic tree trying to unify the various concepts present in the semantic tree and inferring new concepts. **GeneSpring** can augment the semantic network by adding new semantic nodes that would represent the inferred concepts. **GeneSpring** extracts relations by searching the resulting semantic network for relation nodes that contain all the required arguments. A reference consisting of the following fields is created for each parsed relation:

- PMID (PubMed ID)
- actual sentence
- year of publication
- journal name
- reference score
- source

A signature is created for each relation depending upon the participants, their roles, and their mechanisms. References to the relation are added based upon the signature derived from the reference sentence.

32.2 Interaction Databases

The Pathway module of **GeneSpring** is seamlessly integrated with a database of relations between various biological molecules and processes. The molecules and processes are depicted as **Entities** and their biological interactions as Relations. In a Pathway View entities form the nodes of the graph and the edges depict the relations.

You can download and update Interaction Databases from the Agilent Server or with a **GeneSpring** update file for a number of organisms as described in To download an organism's Interaction Database on page 890.

If you are working with an organism that is not currently available in **GeneSpring** you can create a new organism as described in Supported organisms and annotations on page 894.

32.2.1 Interaction Database Organization Overview

Each organism database comprises the following entities: proteins, small molecules, processes, functions, enzymes, complexes, and families. Of these, only proteins are organism specific while the others are largely organism independent. With this in mind and for reasons of efficiency, the Interaction Database is organized in a hierarchical fashion with two levels. The top level is termed *Generic* and contains information that is common across organisms. The second level comprises the various organism specific entities (predominantly proteins) and relations specific to the organism.

This organization is also reflected in the manner in which NLP is run on Medline abstracts. From the overall MeSH hierarchy, about 3000 species terms were identified. These terms were used to divide the millions of Medline abstracts available at the the time of creating the Interaction Databases into two groups: those which have a species MeSH term and those which do not. The latter set of Medline abstracts are marked as *Generic* and the non-protein relations arising from this set of abstracts are shared by all organisms.

Organism specific relations arise from:

- abstracts that have MeSH terms relevant to the organism
- relations from "Generic" abstracts that have protein participants that can be mapped to protein entities of the organism.

Each organism specific database itself has two sections - a systems section and a user section. All prepackaged entities and relationships reside in the systems section while any new entities and relationships added by you (for example by Importing Pathways into **GeneSpring**) are added to the user section. Any pathway construction or entity matching query accesses both the systems and the user section, as well as the *Generic* database to provide a complete answer to the query. **GeneSpring** is also capable of recognizing elements in pathways imported or created by you that are *Generic* and treats them as such in future queries.

32.2.2 Database Entities

The Pathway database contains the following basic types of entities:

- **Complexes**: This class of molecules consists of more than one protein which physically bind each other and are biologically active and stable in their combined form.
- **Enzymes**: Proteins acting as biocatalysts in a metabolic reaction. These entities are particularly important in depicting a biochemical network.
- **Family**: A group of proteins related by structure, function, or another biological parameter.
- **Function**: A class of molecular functions, for example protein-kinase activity.
- **Process**: A class of known biological processes, for example apoptosis.
- **Protein**: Represents genes, their transcripts and protein products for a given organism.
- **Small Molecules**: Represents classes of molecules like drugs, metabolites, metals, etc.

The default representation of these entities is depicted in Figure 32.1. You can change the default visual representation of the entities and relations in a pathway view by selecting **Tools** $\longrightarrow$ **Edit Pathway Theme** from the main menu.

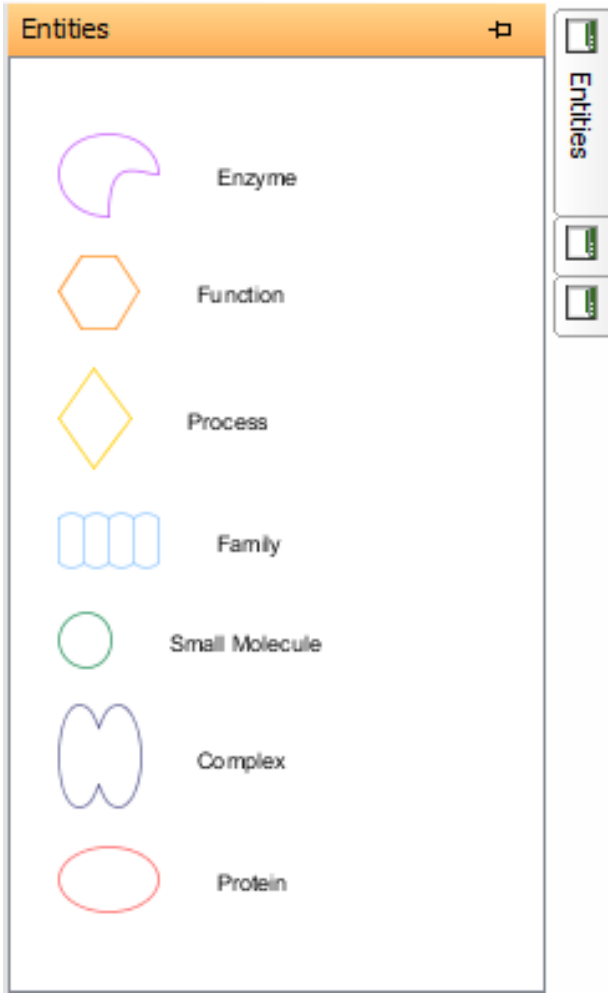


Figure 32.1: Entities Legend

The databases for each of these entities and information on their properties was extracted from various public databases as described here:

- **Proteins**: Protein entities were obtained by parsing Entrez Gene ID data available from NCBI for each organism. Protein entities differ between organisms. All the entities from different species of *Drosophila*, *E. coli*, etc. are present in a single database. The available properties for proteins are;

  Synonym, Description, Cellular Localization, Entrez Gene ID, GO ID, GO Process/Function/Component, Note, Nucleotide, Protein (accessions), and UniGene ID.

- **Cellular Localization**: The following eight GO Cellular Component terms form the ‘‘Cellular Localization” lexicon:
  - Plasma membrane
  - Cytoplasm
  - Nucleus
  - Extracellular matrix
  - Mitochondrion

- Chloroplast (plants only)
- Golgi apparatus
- Endoplasmic reticulum

Distance of these terms to each of their descendant terms in the GO hierarchy is computed. Proteins which have GO Cellular Component annotations are assigned a single cellular localization term based on this distance computation.

- **Small Molecules**

The list of small molecules was created using four primary sources:

- Medical Subject Headings (MeSH) 2008 descriptor records
- MeSH 2008 supplementary concept records
- Chemical Entities of Biological Interest (ChEBI)
- PubChem

Details on how the sources were used are given below:

MeSH substance descriptor records: Each MeSH record is associated with a set of concepts and semantic types. The records were limited to the following semantic types:

- Antibiotic
- Biologically Active Substance
- Body Substance
- Carbohydrate
- Chemical Viewed Functionally
- Chemical Viewed Structurally
- Chemical
- Eicosanoid
- Element, Ion, or Isotope
- Hazardous or Poisonous Substance
- Hormone
- Immunologic Factor
- Indicator, Reagent, or Diagnostic Aid
- Inorganic Chemical
- Lipid
- Neuroreactive Substance or Biogenic Amine
- Organic Chemical
- Organophosphorus Compound
- Pharmacologic Substance
- Steroid
- Substance
- Vitamin

Some filtering was carried out to remove records likely to be only proteins. For each record the name was set as the MeSH descriptor name, while all the terms listed in the concepts section were stored as aliases. In addition, Registry numbers (if available) were stored as CAS property values. MeSH supplementary concept records were parsed in a similar manner.

PubChem: PubChem records having CAS Numbers that were mentioned in Medline annotations were retrieved.

ChEBI: Data from ChEBI was incorporated.

To limit the size of the final database, we consider only those records that contain terms participating in NLP derived relations. Because of the overlap between the data sources, it is possible that the same term is present in multiple records. Filtering of these records providing "useful" terms was carried out as follows:

- *Generic* records which provide terms that are also provided by a larger set of entities from a different source were removed.
- Sets of "equivalent" records (records providing the same set of "useful" terms) were identified.
  One representative record from each equivalence set was retained (with preference given to ChEBI records).
- Records were clustered based on shared terms. Large clusters were broken up by discarding records that were too generic.
- In each cluster, records that were redundant were removed.

All sources are reviewed periodically and any changes to the existing database are updated on the basis of the review.

Properties: All small molecule records have Synonym and CAS Number properties. Additionally, some records from ChEBI and PubChem also have Molecular WT and Formula properties.

- **Enzyme**: All enzyme data were obtained from Expasy and MeSH.
- **Family**: The family names were computed by scanning all Medline abstracts and extracting all sequences of words which satisfy the following three criteria:
  - each word has occurred in the name/alias/description of some protein
  - the first word has occurred as the first token of some protein
  - the last word is either family/subfamily or the plural form of the last word in some protein

  Approximately 110,000 unique potential family names were identified in this manner with some degree of manual curation. The *E. coli* database also includes approximately 299 operons as families. Family entities have no properties other than Synonym.
- **Process, Function and Complex**: The Process and Function entities in the **GeneSpring** databases correspond exactly to the Process and Function sections of the Gene Ontology (GO) hierarchy, respectively. GO is a hierarchical organization of process, function, and cellular component terms that can assist in the annotation of proteins.

  The Complex entities in the **GeneSpring** database correspond to the GO IDs that are in the section of the GO cellular component hierarchy rooted at GO:0043234 ("protein complex"). The database does not contain complex constituent information at present.

  All properties of the entities are directly taken from the gene-ontology.obo file available at GO as of March 2012 and are updated every six months.

  Properties: The properties for the entity types **Process**, **Function**, and **Complex** properties have been deciphered from Gene Ontology and are listed in Table 32.1.

Table 32.1: Process, Function, and Complex Properties

Process	Function	Complex
Alias	Alias	Alias
Description	Description	Description
GO ID	GO ID	EC Number
Name	Name	GO Component
Note	Note	GO ID
		Name
		Note

All database entities are also tagged with **Connectivity** information. Connectivity of an entity represents how well the entity is connected to other entities. In the context of a pathway view (network) the **Local connectivity** of an entity is defined as the number of other entities in the view that are connected to it. **Global connectivity** of an entity is independent of the view and is defined as the total number of relations in which the entity participates.

For example, while expanding a network, you have the option to limit the number of new entities to be included in the new pathway. You can either choose to expand based upon **Local connectivity** or the **Local to global connectivity ratio** (see **Expand Interactions** on page 1056).

32.2.3 Relations

To extract relation information from literature, the **GeneSpring** Natural Language Processing (NLP) algorithm uses a dictionary consisting of the entities described in Database Entities. A relation represents molecular interactions between these entities and is characterized by a set of participating entities. Entities can play several roles within each relation. The relations in the **GeneSpring** Interaction Database are of the following types:

- Binding
- Expression

- Metabolism
- Promoter Binding
- Protein Modification
- Regulation
- Transport
- Member (these relation are not included in the **GeneSpring** database, but can be provided upon request)

Each relation is characterized by two or more entities playing the following primary roles. In the plain text format, a basic relation with two participants is represented as A⟶ B, where A is the "regulator" node and B the "target" node. In addition, several entities may influence the relation of (A⟶ B). These participants are referred to as controllers. In the plain text format, they are represented as C⟶ (A ⟶ B), where C is the controller. Some of these roles are defined with more specific names in the context of specific types of relations. Example: Controllers of a metabolism reaction are called Catalysts.

The "effect" attribute of a relation, reflects how the change in A, which can be qualitative or quantitative, influences B. If an increase/decrease in A is followed by an increase/decrease in B (same direction), then the effect of A on B is considered to be positive. If an increase in A is followed by a decrease in B or vice versa, then the effect of A on B is considered to be negative. The various combinations and their text format representations are shown below:

Change in A, Change in B, Effect, and Representation

Regulation with unknown effect in text format is represented as A⟶ B . Positive and negative effects are shown as `[A --+> B]`, `[A --| B ]`, respectively.

Table 32.2 summarizes the various roles played by the participants of each type of relation, including their effects.

Table 32.2: Participant Roles

Relation Type	**Everyone**	**Left**	**Right**	**Controllers**
Binding	Participant			Modulator (+) Modulator (-) Modulator
Member		Left	Right	Modulator (+) Modulator (-) Modulator
Metabolism		Reactant	Product	Catalyst (+) Catalyst (-) Catalyst
Regulation, Protein Modification, Promoter Binding, Expression, Transport		Regulator (+) Regulator (-) Regulator	Target (+) Target (-) Target	Modulator (+) Modulator (-) Modulator

The different types of relations are explained in the following sections:

- **Regulation**

  Regulation is the most basic relation type. All relation types in **GeneSpring** (except Member) can be considered as more detailed versions of the regulation relations. In **GeneSpring** an entity A ‘‘regulates” another entity B, if A has some influence on B. The participant entities in the Regulation relation are ‘‘regulator”, ‘‘target”, and ‘‘modulator”. Each participant either has a positive, negative, or unknown effect.

  The ‘‘mechanism’’ attribute indicates the way in which B was influenced by A. At present this attribute is always set to ‘‘None’’ for Regulation relations.

- **Binding**

  Two entities, A and B, participate in a ‘‘Binding” relation, if they physically interact with each other. Since both A, B take part equally in the interaction, this is an undirected relation and there is no ‘‘regulator” or ‘‘target” node. Both entities are referred to as ‘‘participants”. The plain text representation takes the form A --- B. If multiple entities, say A, B, and C bind together to form a complex, the line representation is discarded and the relation is depicted as A, B, C. There are no effects ascribed to the participants. However, ‘‘modulators” which can exert either a ‘‘positive”, a ‘‘negative”, or an ‘‘unknown” effect can control binding relations.

  Data from two-hybrid screens, immunoprecipitation, and other

laboratory interaction detection methods can best be represented as binding relations.

- **Promoter Binding**

  If protein A binds to the 5' upstream (or promoter) region of gene B, the relation is represented as a Promoter Binding relation, A $\longrightarrow$ B. Note that this is a directed relation like **Regulation**, in which the "target" node is the gene entity. This relation type is characterized by "regulator", "target", and "modulator" participants exerting either positive, negative, or unknown effects on the relation.

  Transcription factor binding information can best be represented in the form of Promoter Binding relations.

- **Expression**

  If A causes a quantitative change in the amount of protein/mRNA B, the relation is represented as an Expression relation. The participant roles are exactly the same as defined in **Regulation**.

- **Transport**

  Relations involving the change of an entity because of it being transported between cellular compartments are represented by Transport relations. At present, we do not capture the "mechanism" attribute of transport relations. The participant roles are exactly the same as defined in **Regulation**. The "target" entity is the molecule that is transported in the cell and the "regulator" entity regulates this transport.

- **Protein Modification**

  Relations in which an entity A causes the post translational modification of a protein B are represented by Protein Modification relations of the form A $\longrightarrow$ B. The participant roles are exactly the same as defined in **Regulation**. The "target" entity is usually a protein. (However, **GeneSpring** allows the "target" entity to be enzymes and family terms also because of the significant overlap among these entity types). This type of relation has the additional attribute "mechanism" that explains the type of protein modification undergone by the target. The **Mechanisms** included in **GeneSpring** are listed in Table 32.3.

Table 32.3: Protein Modification Mechanisms

Glutamylation	ISGylation	Hydroxylation
Citrullination	De-Ubiquitination	Isoprenylation
Sulfation	Ubiquitination	Acetylation
Unknown	Ribosylation	De-Formylation
De-Neddylation	ADP-ribosylation	De-Phosphorylation
De-Geranylation	Amidation	De-Glycosylation
De-Imination	Myristoylation	De-Amidation
Palmitoylation	De-Methylation	Sumoylation
Glycosylation	De-Palmitoylation	De-Sumoylation
Ethylation	De-Myristoylation	Lipoylation
Biotinylation	De-Acetylation	Methylation
Neddylation	De-Hydroxylation	Glycylation
Farnesylation	Geranylgeranylation	Formylation
Phosphopantetheinylation	Alkylation	
Phosphorylation	De-Farnesylation	

At present this relation type is the only one which uses the "mechanism" attribute in a non-trivial manner. All other relation types have the mechanism attribute value set to "Unknown".

- **Member**

  This is used to store relations of the form "A protein is a member of family B", "A protein is a part of complex B". The text format is A $\longrightarrow$ B. These relations are not parsed from NLP and hence not included in the **GeneSpring** database. They can be provided on request.

- **Metabolism**

  Relations involving the quantitative change of small molecule entities are represented as **Metabolism** relations. For example: "Enzyme E catalyzed the conversion of small molecule A to B". In this directed relation, the left participant A is the "reactant", and the right participant B is the "product". Reactants and products do not possess any effect. The controller of this reaction (relation) is enzyme E, which is referred to as the catalyst for the relation.

Depending on the source of the relation information (see Database Statistics on page 1050), each relation is assigned a **Relation score**. This property indicates a confidence matrix on the quality of relations in the Interaction Database in **GeneSpring**. All relations derived from curated databases like IntAct [32] and all user-created relations are given a score of 10 (highest score). NLP-derived relations are graded on a scale of 1-9, the best being 9 and the weakest being 1. The score properties are internally calculated based on the number of references and the syntax of the sentences.

Each reference associated with a relation is graded on a scale of 1-9, based on the semantic rules of the NLP algorithm, the weakest being 1 and the strongest being 9. This property is called the **RefScore** of the reference. Any relation supported by at least one reference of Refscore 9 or having 3 or more references supporting it, is graded as 9.

32.2.4 Database Statistics

The entities and relations in the **GeneSpring** Interaction Databases are derived from PubMed abstracts using a proprietary Natural Language Processing (NLP) algorithm (see NLP in **GeneSpring** on page 1038) as well as interactions from IntAct (`http://www.ebi.ac.uk/intact/`), and the Saccharomyces Genome Database (SGD [50], `http://yeastmine.yeastgenome.org/yeastmine/aspect.do?name=Interactions`) for yeast. The **GeneSpring** Interaction Databases are updated every six months and you are notified when an update can be downloaded from the Agilent server.

You can look up how many entities and relations (split by their types) are in each downloaded organism's Interaction Database by selecting **Annotations** $\longrightarrow$ **Pathway Database Statistics** from the main menu. Each organism database is characterized by its unique protein entities. Proteins and genes are not distinguished in the **GeneSpring** database.

The number of entries available for the different entity and relation types varies significantly, for example the number of entries for the Promoter Binding relation is much lower compared to other relations. This difference reflects the fact that this area of research is not as well covered in publications as for example research on metabolites. As a consequence you are less likely to get Promoter Binding related pathways when looking for significant pathways. However, this does not mean that they are any less significant in terms of interpreting your experiment results.

32.3 NLP Network Discovery

You can perform NLP Network Discovery on entity lists and selected entities in a pathway viewer as described in this section. To perform NLP Network Discovery on custom lists of entities you can create a Pathway experiment as described in Creating a Pathway Experiment on page 1072.

Organism specific Interaction Databases are available as updates to **GeneSpring**. The relations in the database are mainly derived from published literature abstracts using a proprietary Natural Language Processing (NLP) algorithm (see NLP in **GeneSpring** on page 1038). Additional interactions from experimental data, available in public repositories like IntAct are also included in the **GeneSpring** databases. You can find more information on the sources for these databases under Database Statistics on page 1050.

32.3.1 The NLP Network Discovery Workflow

You can query an Interaction Database by selecting entities from a pathway or choosing an entity list and create a new pathway with the selected entities or entities in the chosen entity list and other related entities in the database. A variety of algorithms are available to do this. In each case, the database that is queried corresponds to the organism of the technology of the current experiment. **GeneSpring** uses Entrez Gene ID, Swiss-Prot, and Gene Symbol from the technology for this query to map to available Entrez Gene IDs, and available entries in the Protein field, and the Symbol field of the Interaction Database, respectively. Hence it is important that both the technology and the Interaction Database contain at least one of these annotations.

You have two different options for NLP Network Discovery:

Simple or Guided Analysis: This option allows you to explore the most common functionalities of network discovery in **GeneSpring**. The default settings for guiding you through the simple network discovery workflow include:

- matching the selected entities to entities in the database,
- retrieving relevant relations between the set of matched entities,
- displaying the results in a graphical view in the form of a network.

Advanced Analysis: This option allows you to explore all the functionalities of network discovery in **GeneSpring** in detail and enables you to change the settings at every step of the process as desired.

The steps of the wizard-driven workflows for **Simple/Guided Analysis** and **Advanced Analysis** are as follows:

Step 1 of 5: The first step allows you to select an entity list of interest, or when launching via right-click from the pathway viewer a number of selected entities in the displayed pathway and enter the parameters for analysis. In **Step 1 - Input Parameters** choose the analysis type and algorithm for building the pathway. For guided network discovery, select **Simple** as the type of analysis.

If you would like to make changes to the default parameters for pathway creation, select **Advanced Analysis**. Then choose the network algorithm that is used to query the relations database and create a pathway network from the selected entities or entity list.

The available options for the **Simple Analysis** workflow are:

Direct Interactions: Finds relations that connect the selected entities.

Network Targets and Regulators: Finds entities that are upstream and downstream of two or more entities from the original list.

Network Targets: Finds downstream entity targets that connect two or more entities from the original list of selected entities.

Network Regulators: Finds upstream entity regulators that connect to two or more entities from the original list of selected entities.

Network Binders: Finds entities that "bind" (are connected by binding interactions) to two or more entities from the original list of selected entities.

Network Modifiers: Finds protein entities that are either regulators or targets of biochemical protein modifications (for example, phosphorylation, ubiquitination, etc.) of two or more proteins from the original list of selected entities.

Transcription Regulators: Finds protein entities regulating mRNA expression of, or whose expression is regulated by two or more entities from the original list of selected entities.

Transport Regulators: Finds all molecules that are regulating the transport of other molecules.

Metabolism Regulators: Finds molecules that are regulating the metabolism of bio-molecules.

Small Molecules: Finds all small molecules (drugs) regulators and targets of two or more entities from the original list of selected entities.

Biological Processes: Finds all biological process entities connected to two or more entities from the original list of selected entities.

Shortest Connect: Finds the smallest set of relations that connects all entities in a given list into a single network.

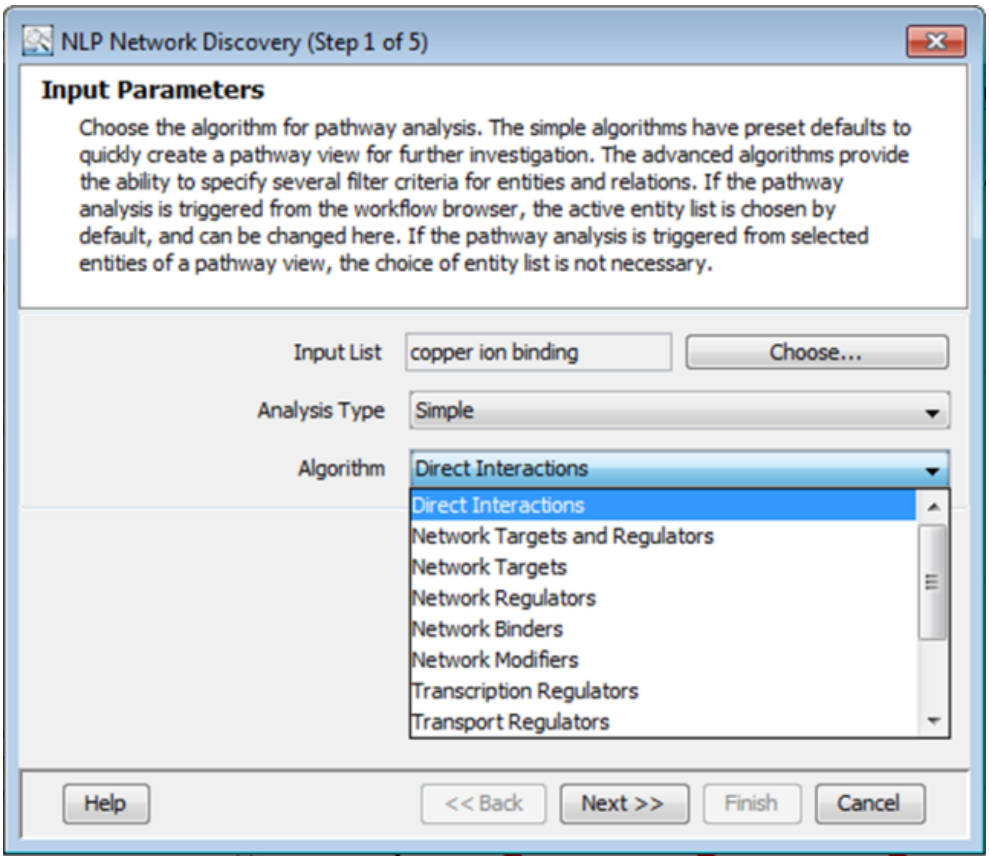


Figure 32.2: Algorithm options for the Simple Analysis workflow

Choosing **Simple Analysis** launches only Steps 1,4, and 5 of this wizard-driven workflow. After the algorithm is specified, the filters for **Simple Analysis** are set to these default settings:

Algorithms type: local/ global.

Connectivity : Connectivity relevance: 50 Connectivity: ≤ 2

Entity filter: All entities are selected.

Quality filter: ≥ 9

Relation filter: All relation types are selected.

You can find more details on each of these settings in the corresponding sections as described for the **Advanced Analysis** workflow.

For the **Advanced Analysis** workflow, you can choose from the following three algorithms:

Direct Interactions: Finds relations that connect the selected entities.

Expand Interactions: Expands the existing network to include the first-degree neighbors of the selected entities.

Shortest Connect: Finds the smallest set of relations that connects the set of selected entities into a single network. Note that some intermediate entities may be introduced in this process.

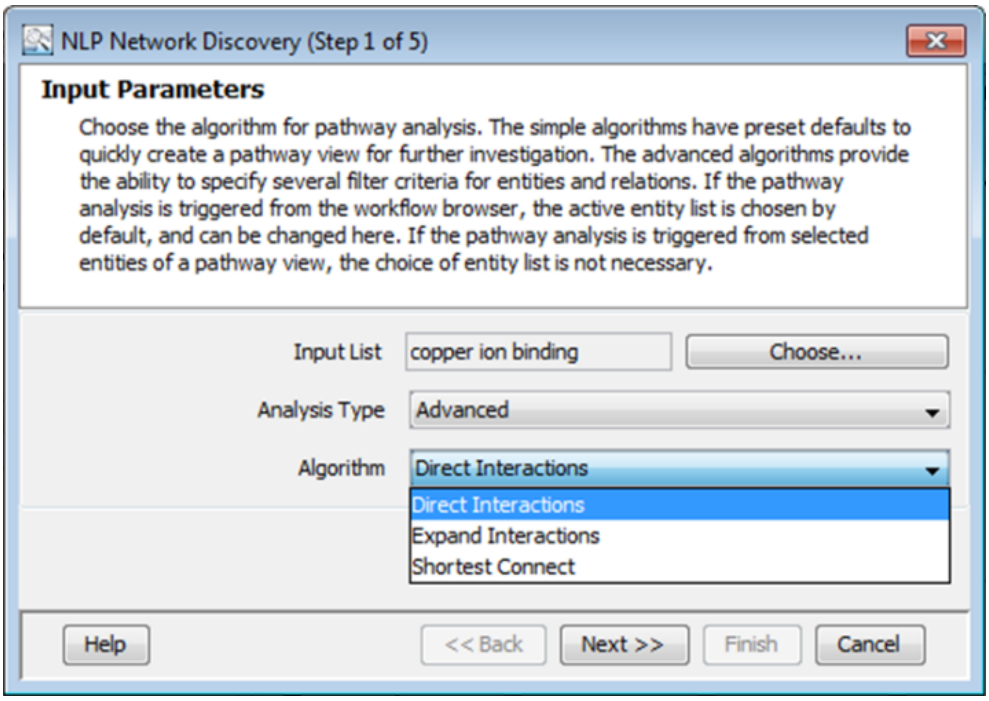


Figure 32.3: Algorithm options for the Advanced Analysis workflow

If the selected algorithm does not find any entities that honor the algorithm criteria, **GeneSpring** prompts you to select a different algorithm or other entities for analysis.

Step 2 of 5: This step only appears in the **Advanced Analysis** workflow and displays the **Matching Statistics** after matching the selected entities to entities in the database. The **Match Result** column indicates whether a match was found or not. For matched rows, additional details like **Probe Set Id**, **Name** (Gene Symbol), **Type** (of matched entity), and **Global Connectivity** (total number of relations in which the matched entity participates) are also displayed. Here you can select a subset of the matched rows before moving on to the next step.

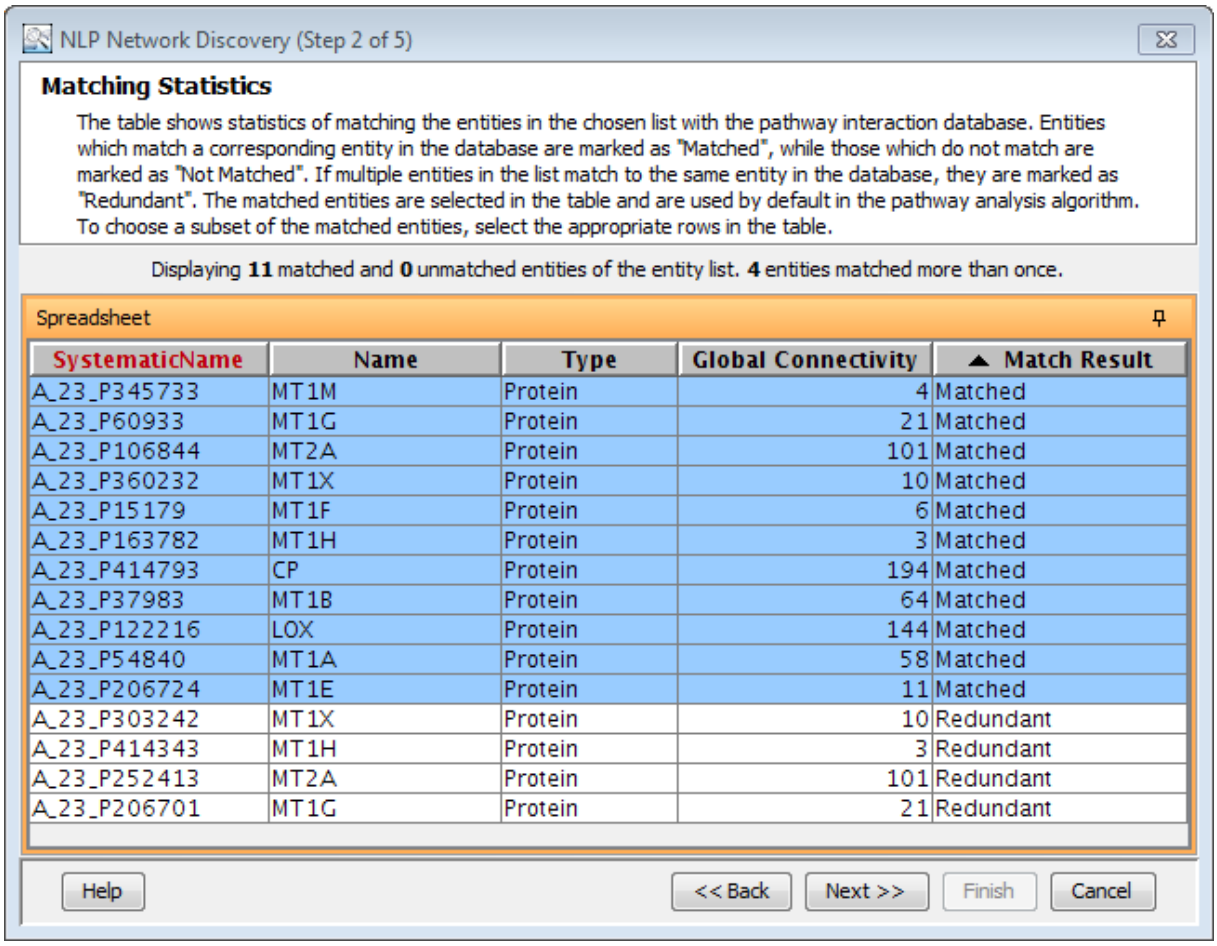


Figure 32.4: Matching Statistics in Step 2 of the Advanced Analysis workflow

Step 3 of 5: This step is only available in the **Advanced Analysis** workflow. It allows you to set filters for the available analysis parameters depending on the algorithm that was selected in Step 1.

For the **Direct Interactions** algorithm the parameters are:

Relation Filter

Relation score: You can build or expand a network by including only relations with a particular quality score. The concept of assigning a quality score to every relation is explained in Relations on page 1045. The default score defining relation quality is set to ≥ 9.

Relation types: You can select one or more types of relations (see Relations for more details on each available type) to generate the network from the selected entities, depending upon the biology of the problem. Check boxes are provided for multiple selections. For example, you can choose to select only the 'expression' and 'promoter binding' types of relations to create a network of transcriptional analysis. The default settings use all types of relations to create the network.

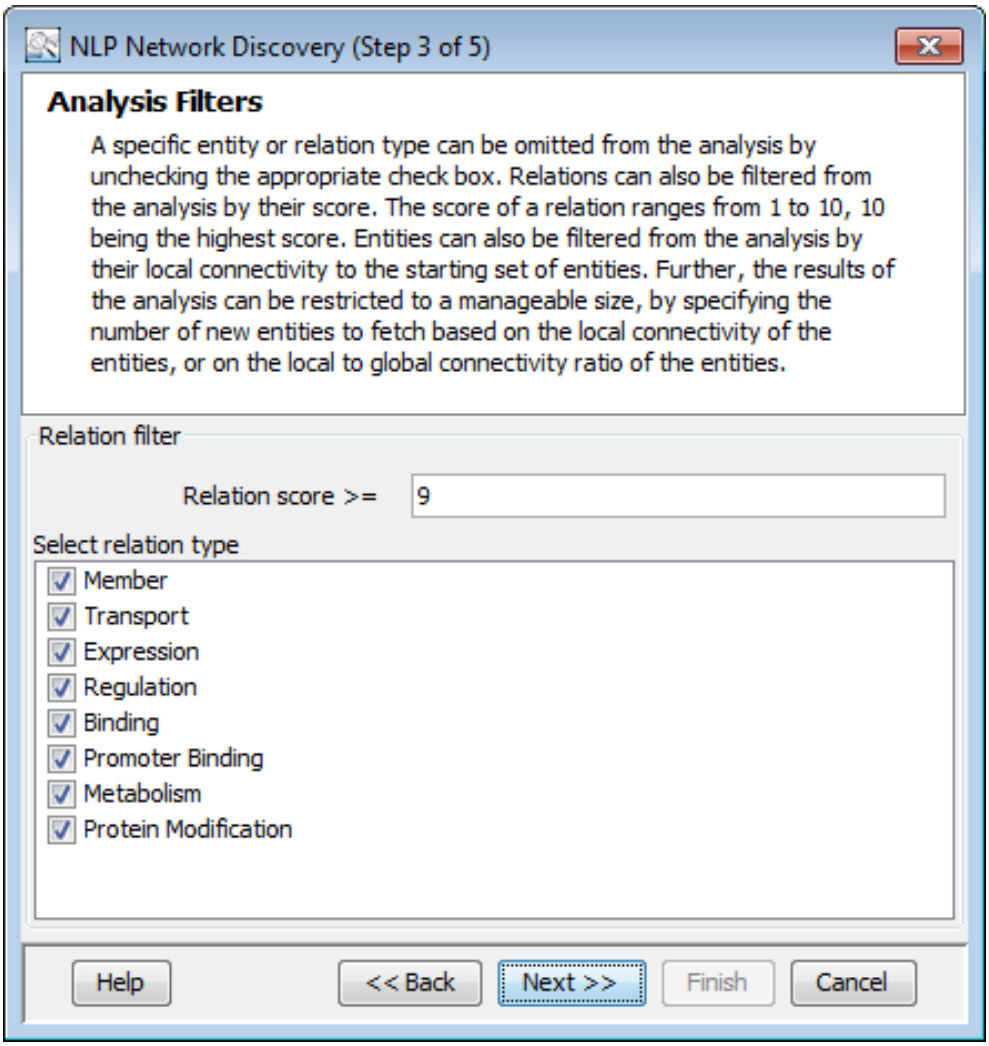


Figure 32.5: Analysis Filters for Direct Interactions in the Advanced Analysis workflow

For the **Expand Interactions** algorithm, the following filter settings are allowed to create a network:

Relation Filter: This filter option is the same as the one for the **Direct Interactions** algorithm described on page 1055.

Entity Filter:

- *Entity local connectivity*: This filter allows you to add new entities to a given network by ranking the new entities based upon their local connectivity. The default setting is > 2, to consider only those entities which are connected to two or more entities from the starting list of selected entities.
- *Entity type*: You can select the types of entities to add to the expanded network, depending on the biological question. Check boxes are provided for multiple selections. For example, you can opt to study all biological processes and functions regulating a list of selected genes. In this case, **GeneSpring** expands the list of entities for the network by limiting the newly added entities to **Processes and Functions**. The default setting uses all types of entities to create the network.

Limit analysis results based on:

- *Local connectivity*: Allows you to add a certain number of entities to the given network based on their rank with regards to local connectivity. New entities are ranked with decreasing priority, based on how many entities they are connected with within a given list of entities.
- *Local to global connectivity ratio*: A local/global connectivity ratio is computed for each new entity. Local connectivity is based on the number of entities to which it connects within a given list and global connectivity is the number of relations that it participates in within the entire database. New entities are ranked with decreasing priority based on this local/global connectivity ratio.

The default setting is **Local to global connectivity ratio**.

Maximum number of new entities: This field allows you to limit the maximum number of entities to be added to a network. The default is set to 50.

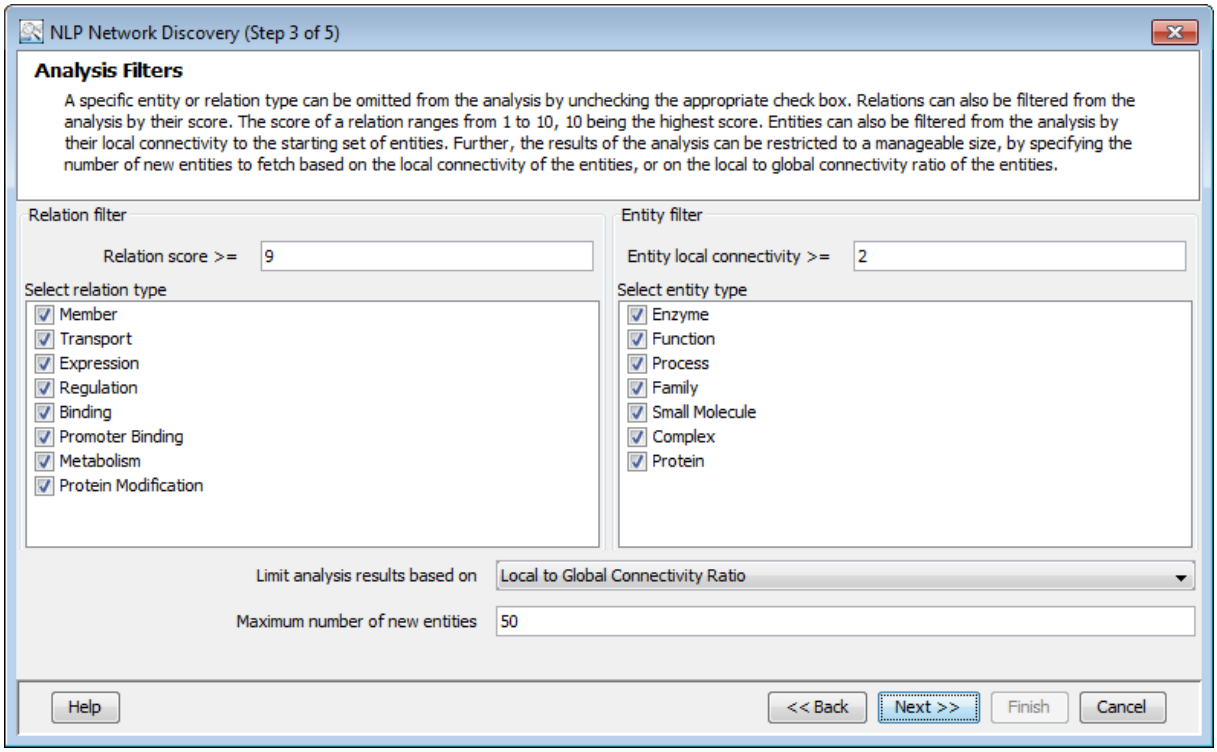


Figure 32.6: Analysis Filters for Expand Interactions in the Advanced Analysis workflow

For the **Shortest Connect** algorithm in the **Advanced Analysis** workflow, the following filter settings are allowed to create a network:

Relation Filter: This filter option is the same as the one for the **Direct Interactions** algorithm described on page 1055.

Entity Filter:

Entity global connectivity -This filter allows you to add new entities to connect two disconnected network clusters by ranking the new entities based on their global connectivity. The default setting is 100 and above.

Entity Filter: This filter option is the same as the one for the **Expand Interactions** algorithm described on page 1056.

NOTE

Internally the algorithm tries to create a single connected component by carrying out a series of expansions. If the number of expansion steps is more than 10 or if the number of entities from which expansion needs to be carried out next exceeds 10,000, the algorithm will abort. Some small molecules, such as calcium, have a very high number of connections (greater than 5000). If these molecules are encountered in the initial stages, the algorithm is likely to abort because of the large size of the expansion set. Two suggestions to make the algorithm find the shortest path are:

- Use the filter to turn off the Small Molecules entity type. This prevents Calcium and Glucose from jamming the algorithm before it can determine the shortest path.
- Increase the global connectivity of your protein entities. The default is set to 100. For example, you can increase this number to 1000 as well as unselect Small Molecules.

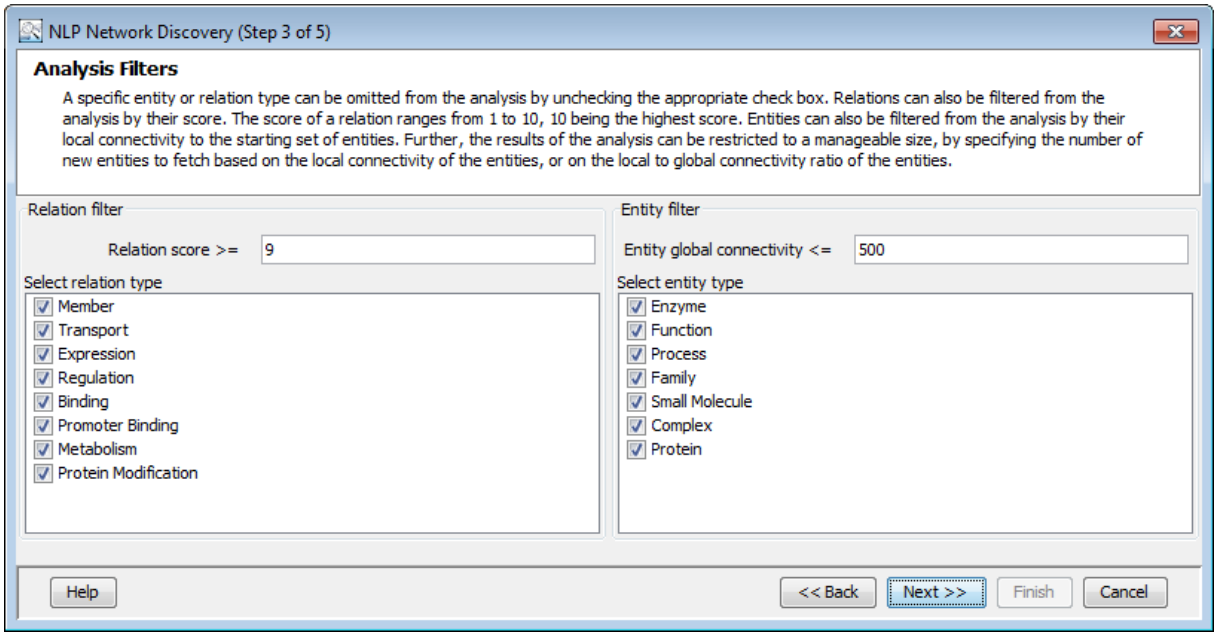


Figure 32.7: Analysis Filters for Shortest Connect in the Advanced Analysis workflow

Step 4 of 5: This step is called **Analysis Result** and displays the created pathway. The initial number of entities, the number of new relations and the number of new entities are also displayed.

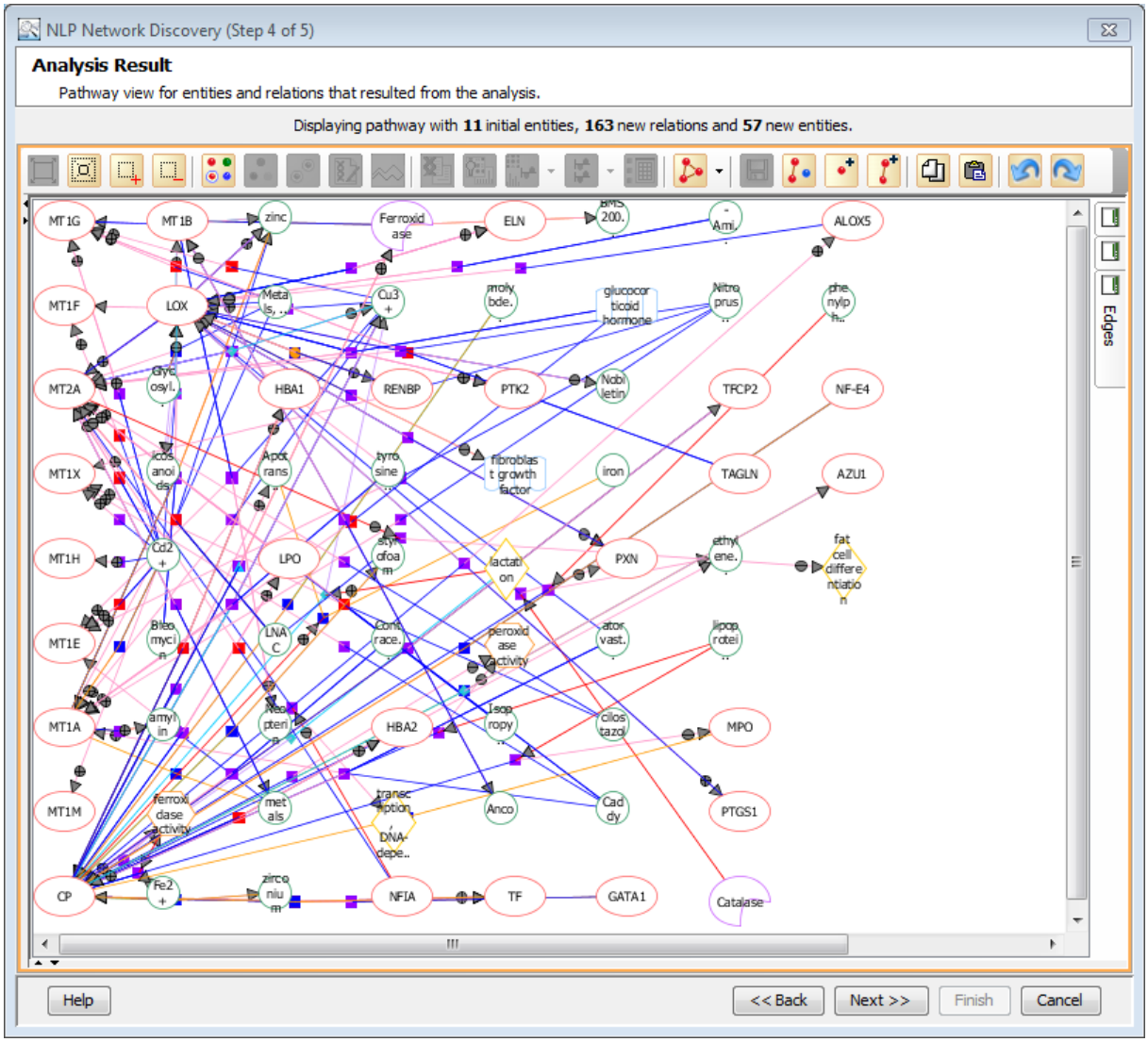


Figure 32.8: Analysis Result of the Simple/Guided and Advanced Analysis workflows

Step 5 of 5: Before saving the new pathway list, details of the newly created network/pathway appear in the form of an inspector as Step 5 of both the **Simple/Guided** and **Advanced Analysis** workflows. Here you can specify a **Name** and add to the **Notes** that are automatically generated. Clicking **Finish** creates a new pathway list that is saved under the input entity list that was used to create the pathway from which the original entities were selected. The new pathway opens in The Pathway View.

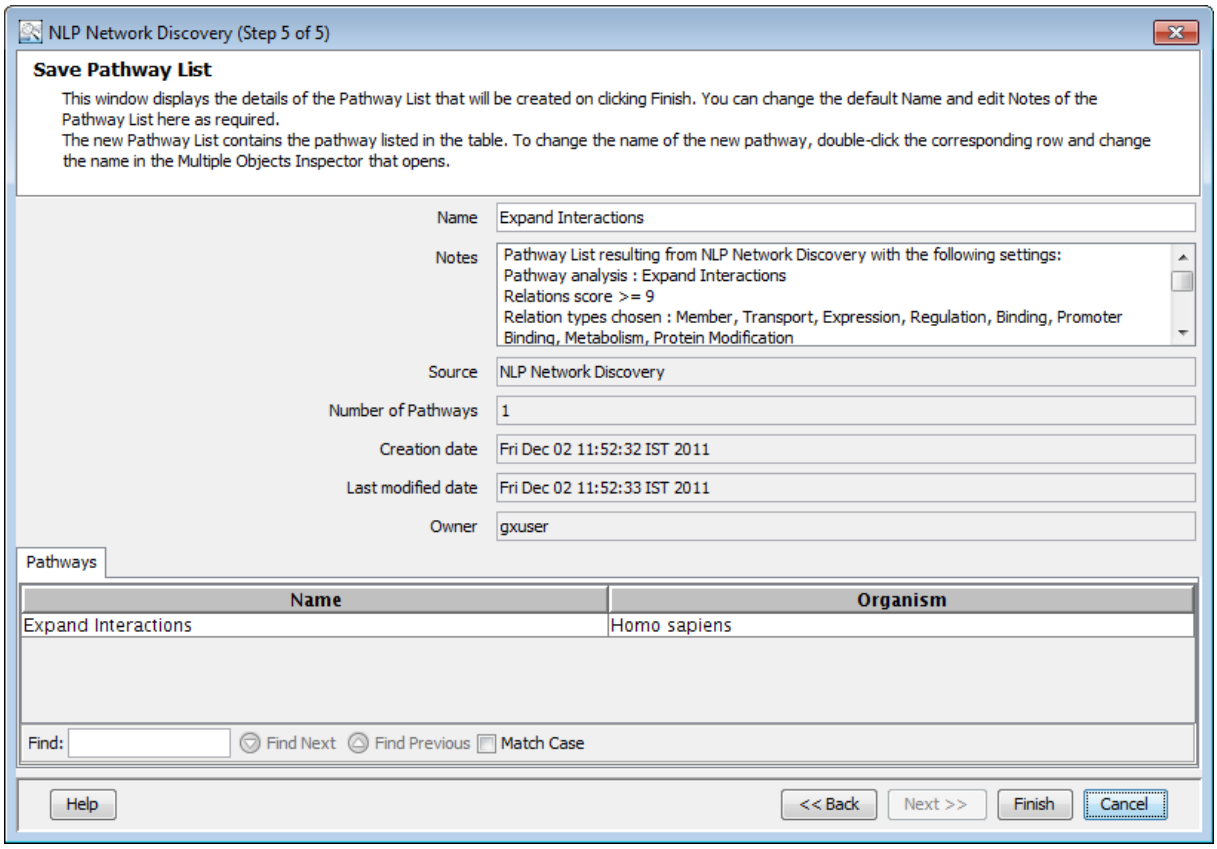


Figure 32.9: Save Pathway List

32.4 Extract Relations via NLP

Running Natural Language Processing (NLP) is another mechanism to create new pathways. You can run the in-built NLP algorithm (see Natural Language Processing (NLP) in **GeneSpring** on page 1038) directly on PubMed abstracts or on local pdf/doc/html files. NLP first recognizes entities in sentences and then performs information extraction to identify relationships between these entities. The entities that NLP can recognize are restricted to those packaged in the generic Interaction Database and the Interaction Databases for the organism of the currently active experiment (see Interaction Databases on page 1040). Additional entities imported into these databases via BioPAX (Level 2 only) import, for example, do not participate in NLP.

Launch this NLP functionality by selecting **NLP Networks** ⟶ **Extract Relations via NLP** from the Workflow Browser.

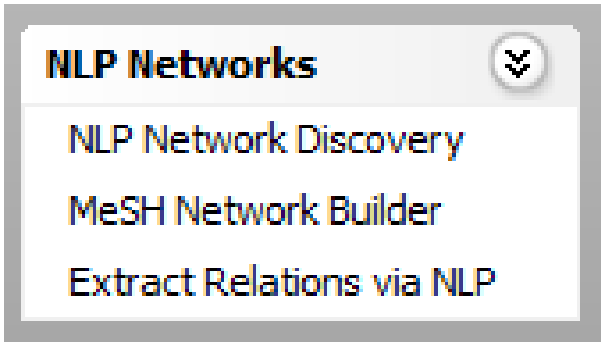


Figure 32.10: Extract Relations via NLP

A wizard-driven workflow opens with the following steps to complete:

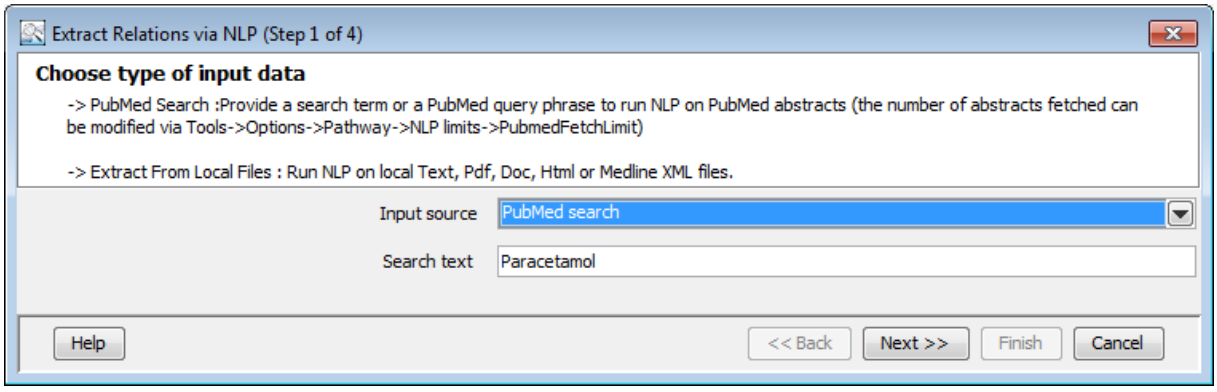


Figure 32.11: Input Data

Step 1 of 4: Choose type of input data If the chosen **Input source** is "PubMed search", you can specify a search query that is submitted to PubMed. The documents returned by PubMed form the

basis for the next step. You also have the choice of running NLP on local files. Non-text format files, like .doc and .pdf files are first converted into text using publicly available converters (antiword and pdftotext, respectively) and the text content is then passed on to the next stage.

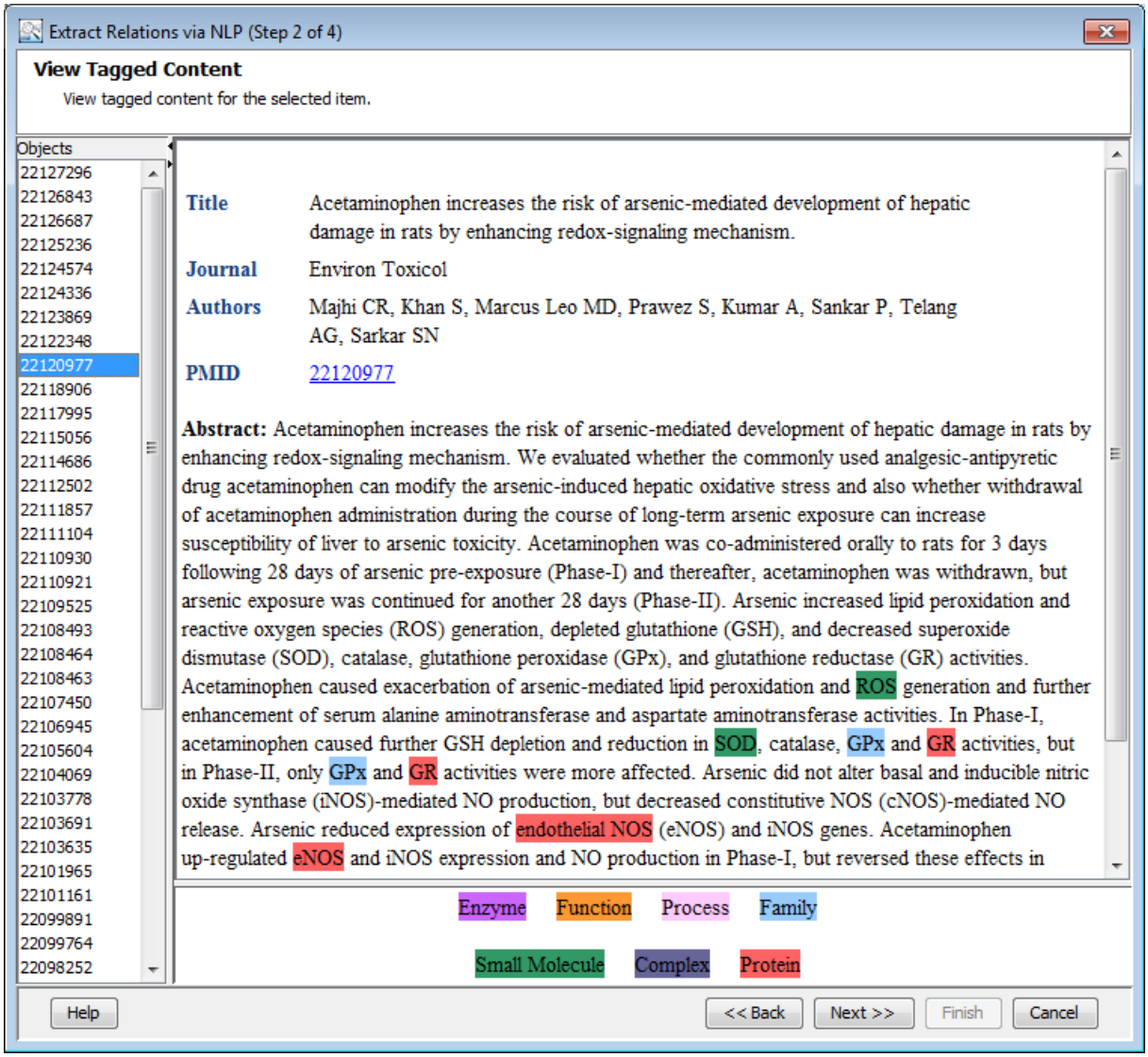


Figure 32.12: View Tagged Content

Step 2 of 4: View tagged content In this step the target documents containing the search terms are identified and tagging is performed using the entity dictionary. All molecular and biological process/function entities present in the **GeneSpring** Interaction Databases are tagged in these searched documents. Matching entities are highlighted according to default color settings, with the corresponding legend displayed below the searched content. In the case of PubMed articles (or Medline XML files), the PubMed ID is shown in the left hand column. In all other cases, the name of the

file is displayed. Each tagged document can be viewed separately in the right-hand column by selecting the corresponding document name or PubMed ID.

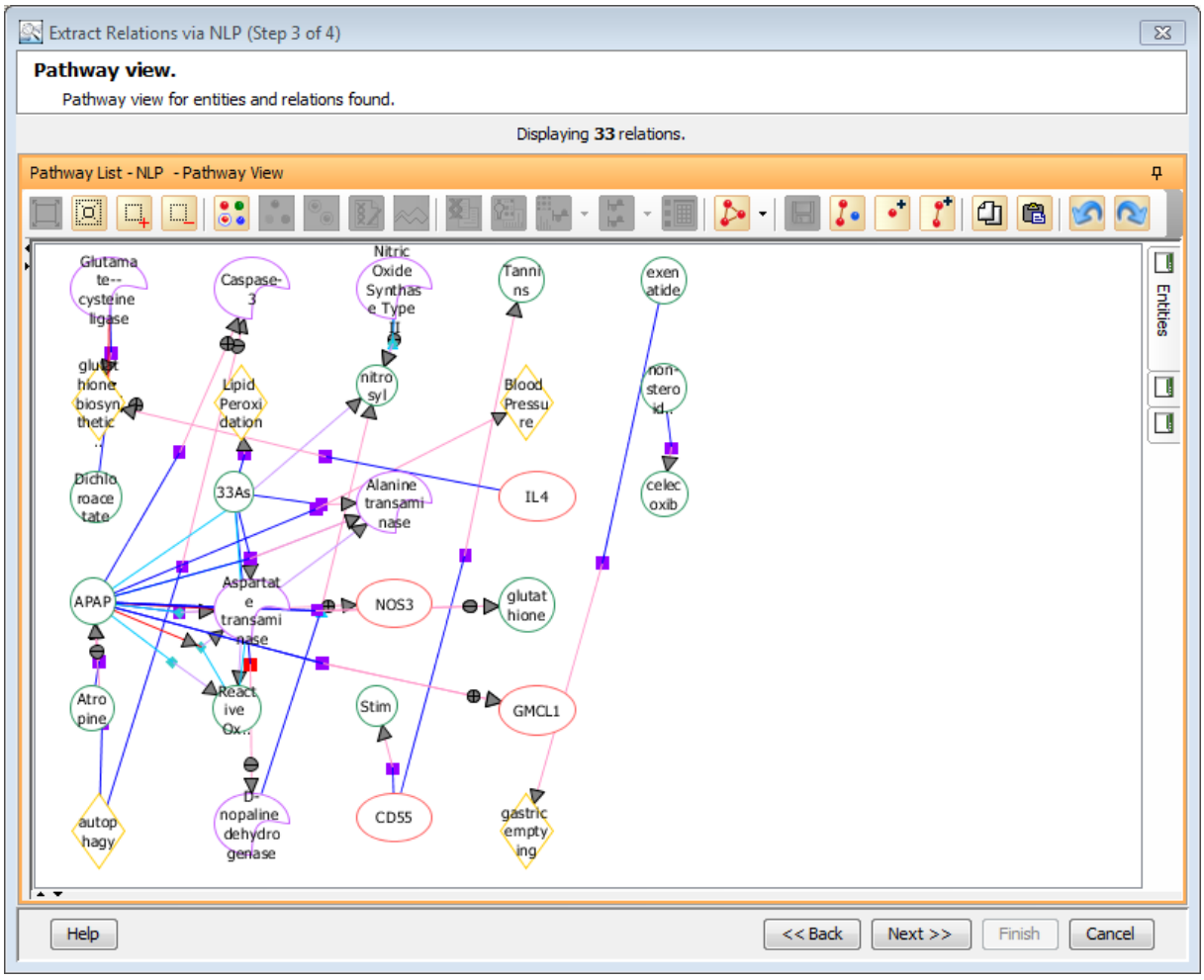


Figure 32.13: Pathway View

Step 3 of 4: Pathway view All sentences which have at least two tagged entities are analyzed by the NLP algorithm. A **Pathway View** is launched with the relations extracted from these sentences. The total number of extracted relations and new entities is also shown. The sentences from which the relations were derived can be examined by double-clicking on the relations.

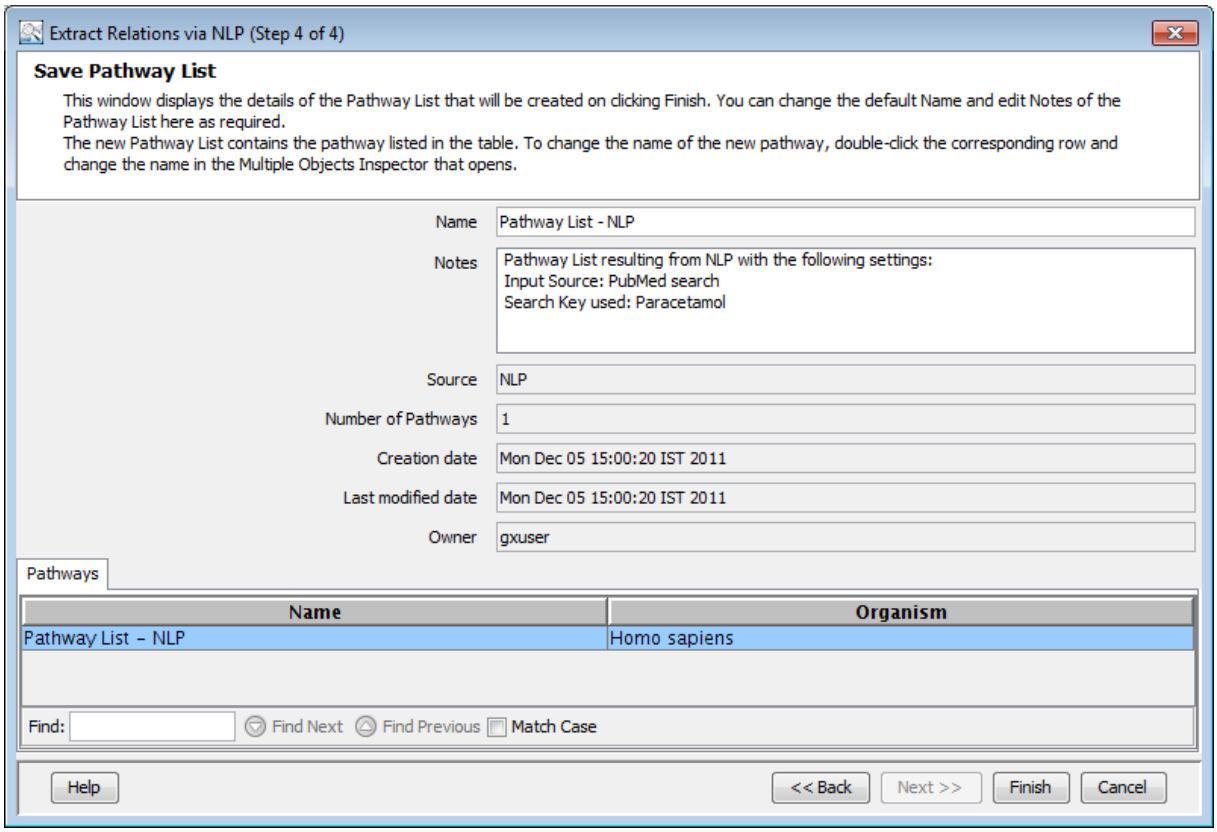


Figure 32.14: Extract Relations via NLP - Save Pathway List

Step 4 of 4: Save Pathway List In this step, statistics about the Pathway List that is about to be created are displayed. Double-click the row in the **Pathways** table at the bottom of this window to find more details on the **Entities** and **Relations** in the new pathway in a Pathway Inspector. The names of the participating entities, the entity type, the Entrez Gene ID of the proteins, and the CAS Number of metabolites, where applicable, are displayed.

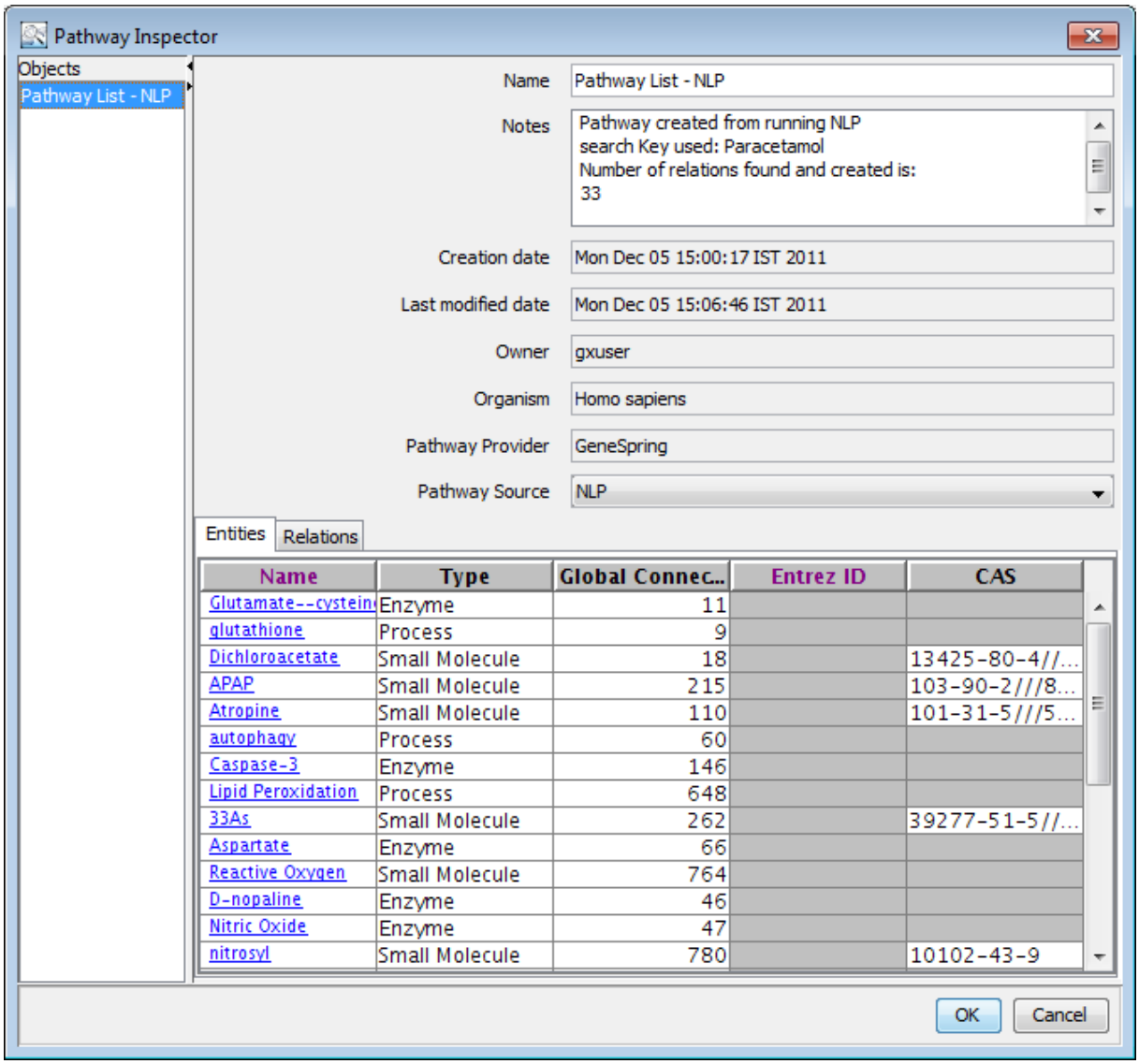


Figure 32.15: Extract Relations via NLP - Pathway Inspector

Click **Finish** to create the pathway. This new pathway is saved in the **user space** of **GeneSpring** 's Interaction Database, containing all the pathways that were prepackaged, imported, or newly created. It is now available for future attempts to identify matched pathways (see Pathway Analysis in **GeneSpring** on page 896) for selected entity lists and can be searched and added to the current experiment using the Search Pathways option.

32.4.1 NLP Settings

Some of the phases in the NLP process are time and memory intensive. In particular, due to the inherent ambiguities in the English language,

and because of inconsistent use of grammar, it is possible for a sentence to have multiple syntax and semantic parse trees. To prevent **GeneSpring** from crashing while running NLP, the algorithm is restricted to a set of defaults that limit the power of the NLP engine when run in this workflow. Some of the differences compared to the algorithm used to create the Interaction Databases (see Natural Language Processing (NLP) in **GeneSpring** on page 1038) are:

- smaller lexicon
- limit on the sentence length
- limit on the number of alternate parses that are processed
- limit on semantic inference

At present these settings cannot be altered.

Another difference compared to the NLP algorithm used to create the databases is that MeSH terms are not used to filter out irrelevant abstracts. If you wish to do so, please consult PubMed help for details on how to construct the corresponding PubMed queries appropriately:

http://www.ncbi.nlm.nih.gov/books/NBK3827/#pubmedhelp.Search_Field_Descrip.

32.5 MeSH Network Builder

The **MeSH Network Builder** in **GeneSpring** is a novel way of creating pathways around a *Term* or a *Concept*.

- This functionality works independently of experiments and entities.
- It only requires a *Term/Concept* as input.
- **GeneSpring** pulls out the MeSH terms relevant to this input term/concept from the MeSH database (see `http://www.nlm.nih.gov/mesh/meshhome.html`), by exact/substring/synonym matching.
- It obtains all the interactions containing these MeSH terms from the Interaction Database packaged with **GeneSpring**. A list of the relevant MeSH headings along with the number of relations (exact or all relevant interactions) is shown. You can then choose one or more of these terms to create a pathway with all identified interactions.
- You can also filter the interactions based on the frequency of occurrence and the type of relation (direct or indirect) to the 'term'.
- Once the pathway is opened in The Pathway View, all the regular operations available for non-GPML (excl. BioPAX Level 3) pathways can be carried out. See Working with Pathway Lists on page 912 for details on the possible operations in the Pathway View.

32.5.1 Launching MeSH Pathway Builder

MeSH Network Builder can be launched by selecting **NLP Network Discovery** $\longrightarrow$ **MeSH Network Builder** from the experiment workflow. A four-step wizard-driven workflow opens in a separate window.

Step 1: Input Page: Type the *Concept* or the **MeSH Term** in the provided field. Note that the term does not have to be technical. A simple phrase or phenomenon of interest is sufficient, for example "Mouth ulcer".

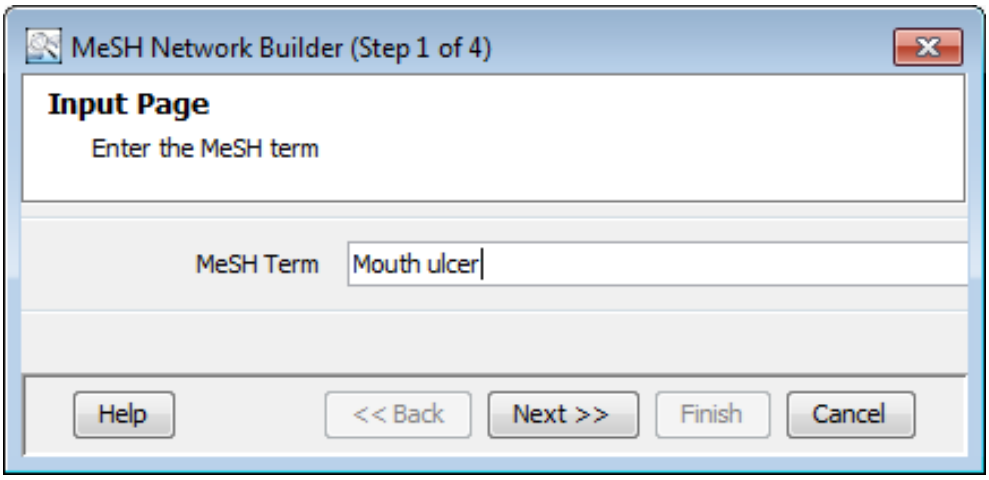


Figure 32.16: Step 1: Input Page

Step 2: Select relevant MeSH Terms: A table is shown that lists all the MeSH headings containing the input *Term/Concept*. Select one or more MeSH headings and define the filtering options explained in Table 32.4.

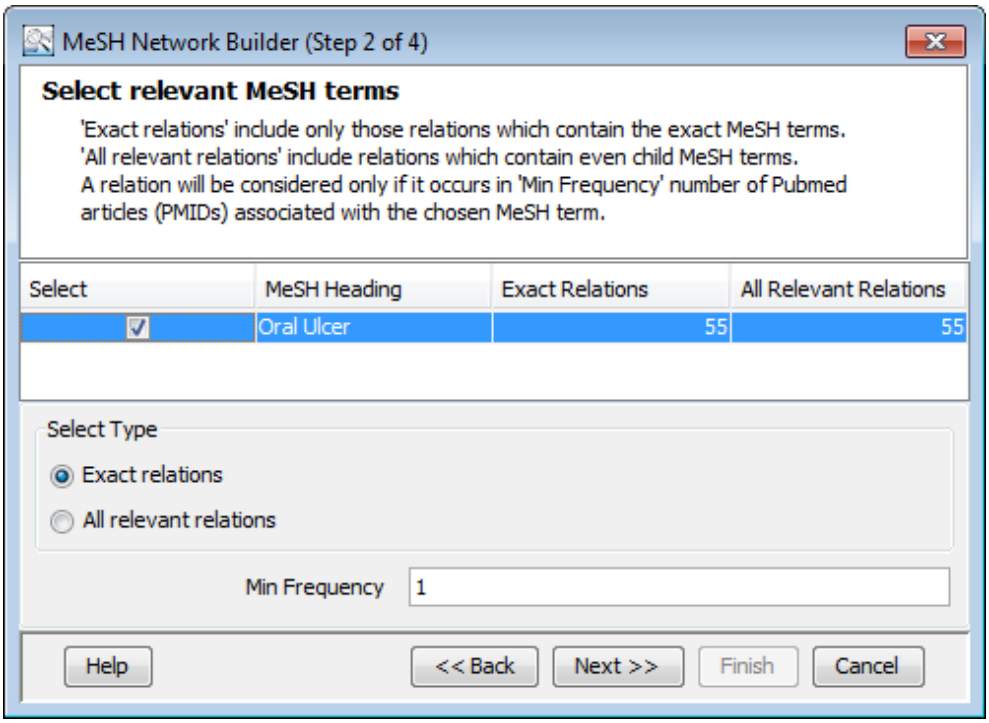


Figure 32.17: Step 2: Select Relevant MeSH Terms

Table 32.4: Type of relationship and Frequency

Exact Relations	Includes only those interactions that contain the exact MeSH headings that were selected.
All Relevant Relations	Includes all interactions that contain either the exact MeSH heading or the child MeSH terms.
Min Frequency	Defines the minimum number of PubMed articles (PMIDs) associated with the MeSH term that an interaction should have. For example, if the **Min Frequency** is set to ‘‘5”, the pathway will include only those interactions which have at least 5 PMIDs that contain the relevant MeSH term supporting. The default setting for this parameter is ‘‘1”.

Step 3: MeSH Pathway: The pathway built out of all the interactions associated with the chosen MeSH term or terms is shown here.

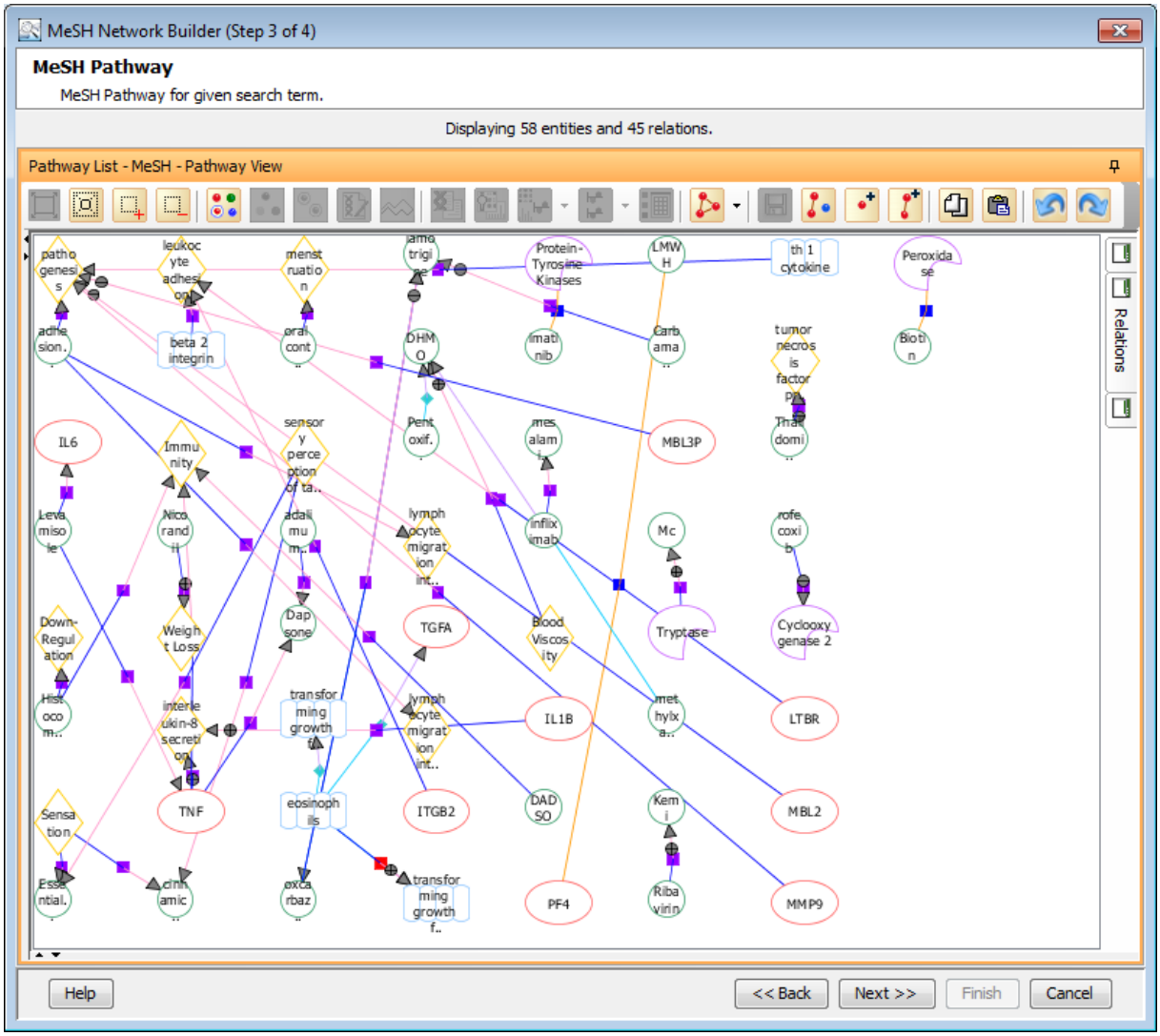


Figure 32.18: Step 3: MeSH Pathway

Step 4: Pathway created from MeSH: In this step all the details about the creation of the new network is summarized. Double-click the row in the **Pathway** table at the bottom of the window. Information on both entities and relations contained in the new pathway is displayed in the Pathway Inspector.

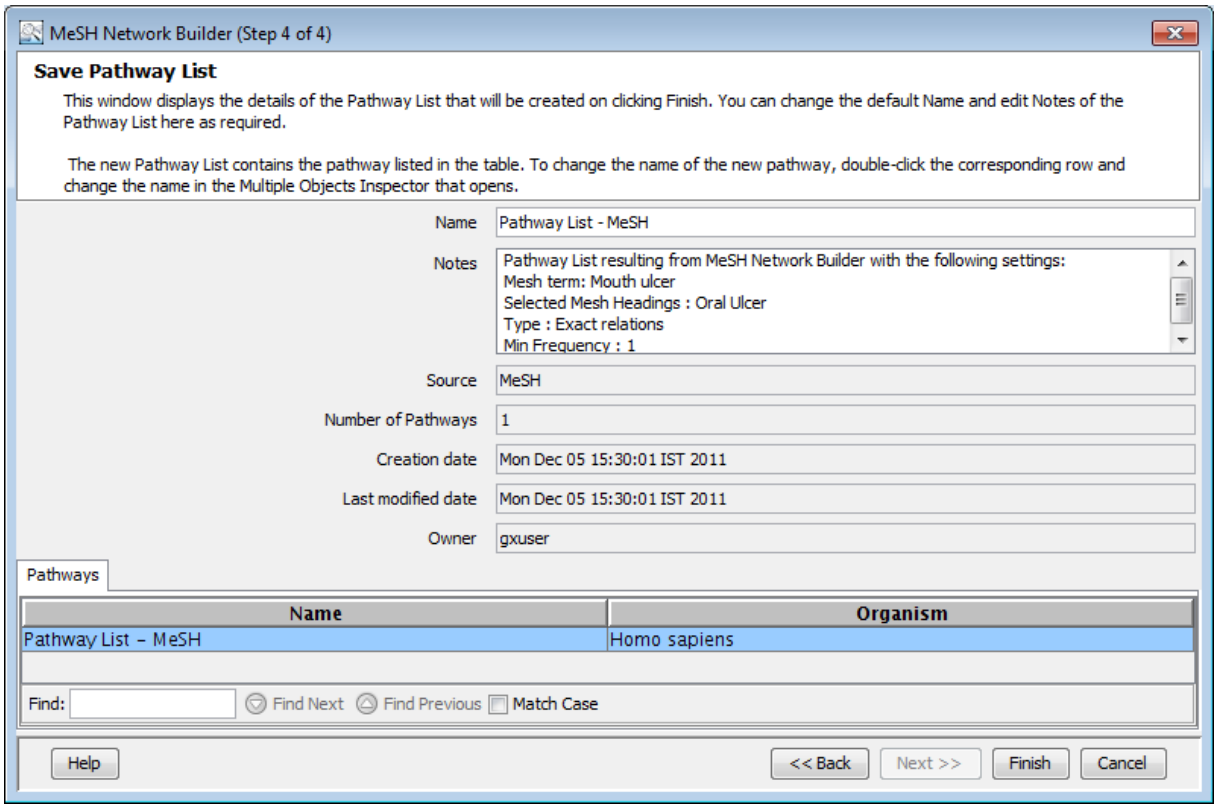


Figure 32.19: MeSH Network Builder Step 4: Pathway created from MeSH

32.6 Creating a Pathway Experiment

GeneSpring offers a separate pathway analysis experiment, independent of the microarray analysis workflow. This feature aims to provide a wider use of pathway functionalities to cover different types of data. The input to a pathway experiment is a tab-delimited text file. The file should contain a column with standard public identifiers for the entities (Entrez Gene ID, UniGene ID, or Gene Symbol for proteins), associated data columns for all samples, for example signal intensity values for every sample and list-associated values such as fold change or p-values from standard statistical tests. Only the ID column is mandatory in the file, the rest of the columns are useful to visualize the pathway in the context of the data using data overlay. Different types of data can be imported into **GeneSpring** which may include:

1. Genomic Data
2. Proteomic Data
3. Metabolomic Data
4. Combinations of proteins and small molecules
5. Genes corresponding to interesting biological processes and functions defined by gene ontology.

32.6.1 Launching a Pathway Experiment

When creating new experiments in **GeneSpring** one of the experiment types on offer is a **Pathway** Experiment. In the **New Experiment** dialog, select Pathway from the **Analysis type** drop-down.

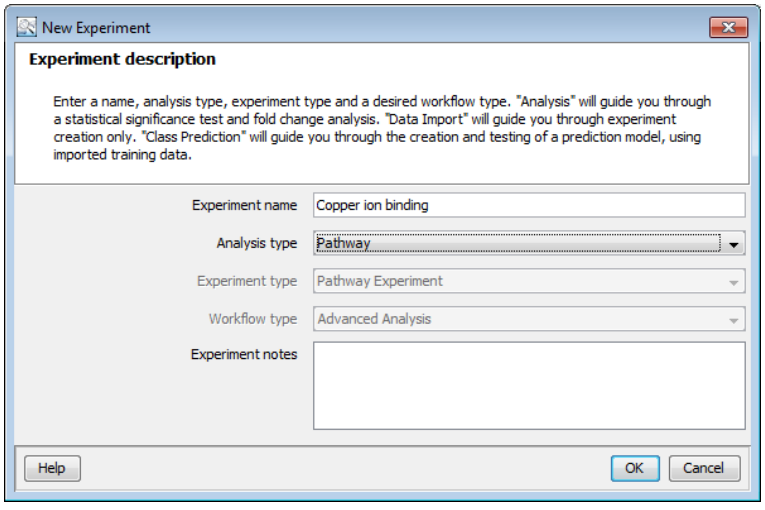


Figure 32.20: Pathway Experiment Creation

Data Import

Step 1 of 4: As in all other **GeneSpring** experiments provide the Input parameters which include a name to the technology associated with the experiment and specify an organism from the drop-down. You also have to specify the data file for upload that contains a list of the entities for network discovery.

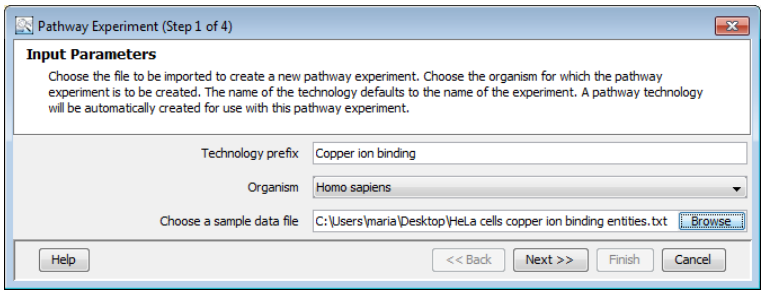


Figure 32.21: Input Parameters

Step 2 of 4: Choose the column of the input entity list that contains the entities' identifiers. As default, the first column is selected. You can select another column from the **Identifier column in the file** drop-down. **GeneSpring** performs an automated ID match. The available IDs are present under **Available Items** and the matched ID is displayed under Selected items. Select the Identifier type in your entity list by moving it from **Available Items** to **Selected items**. A mix of IDs for different entity types is permitted. **GeneSpring** recognizes all standard public IDs for protein entities (Entrez Gene ID, Nucleotide, UniGene ID and Gene Symbol). All associated numerical data columns are imported with the entity list.

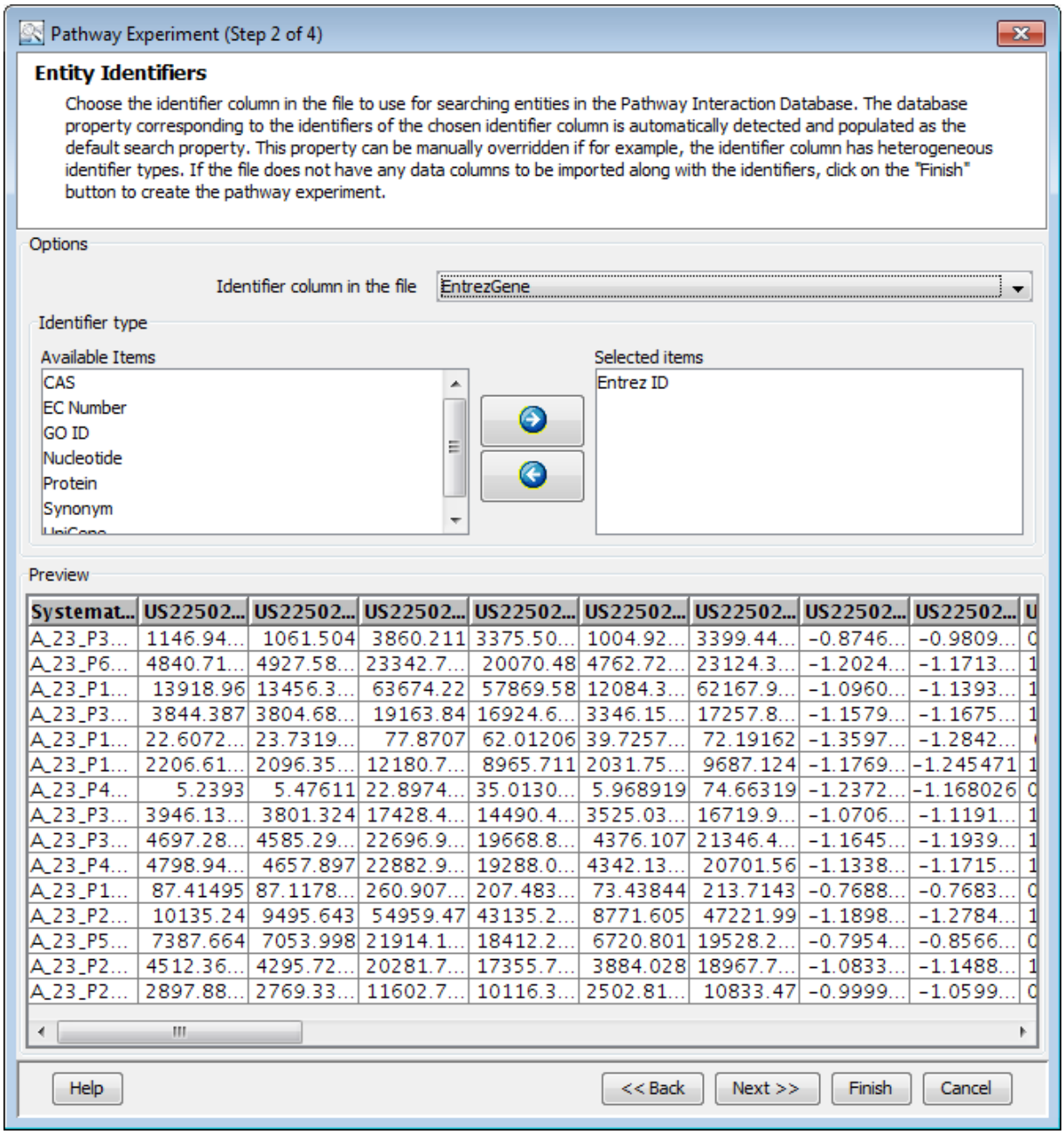


Figure 32.22: Import List from File

Step 3 of 4: This step, called **Input parameters - Choose signal column** allows you to select any columns that contain real numbers as sample data columns. This step is optional. If your entity list does not have any data columns, click finish, which takes you directly to the Entity List Inspector. If multiple rows in the input file map to the same entity in the database, the average of the data rows is computed and displayed.

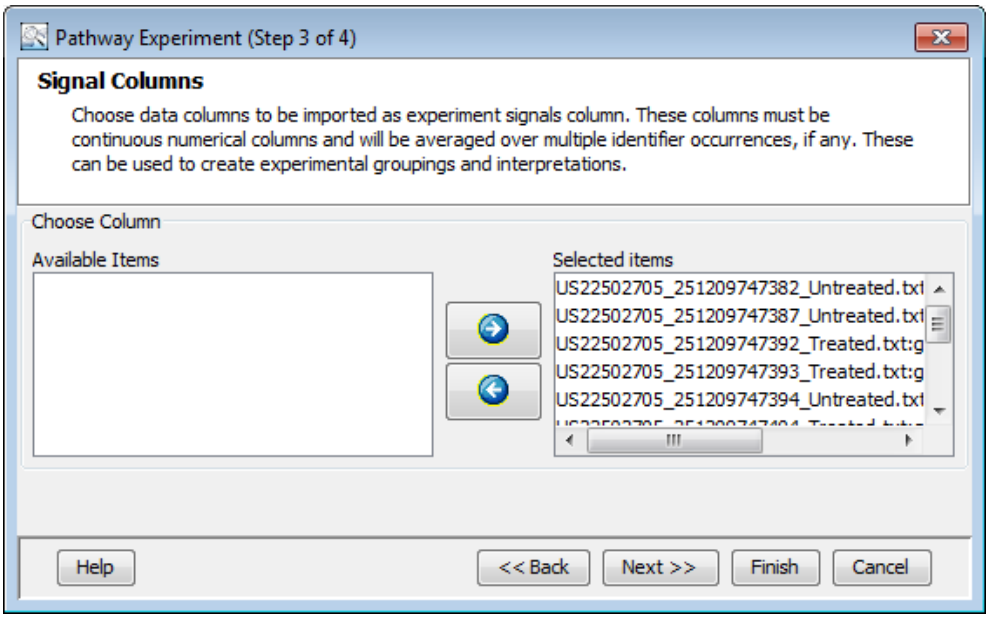


Figure 32.23: Choose signal columns

Step 4 of 4: In this step you can select list associated data columns, if available in your entity list. Identifying this type of data allows you to overlay data in the pathway view. If multiple rows contain the same entity, data in the first row is chosen. This step is also optional. Click Finish to complete the workflow.

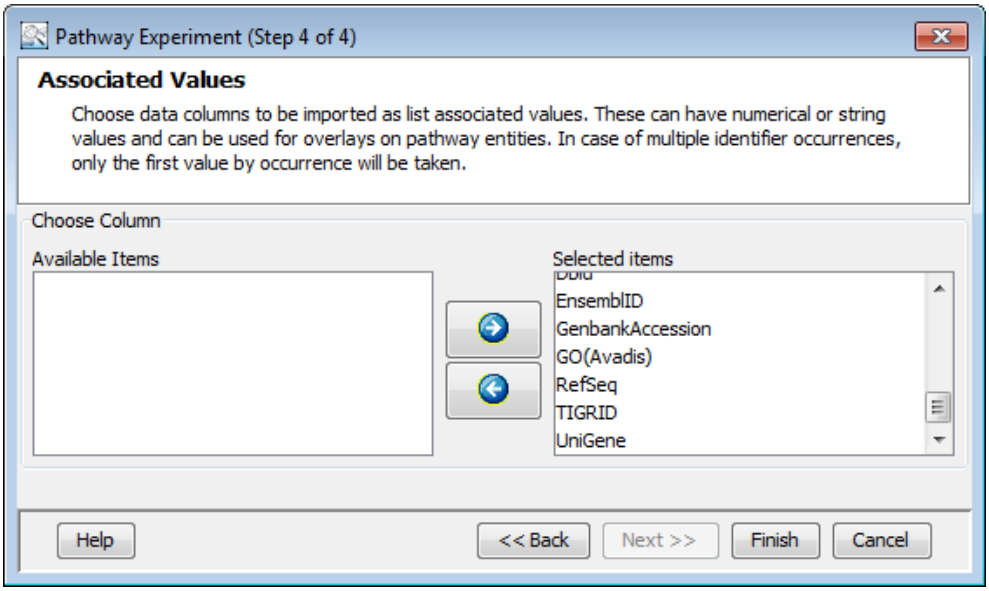


Figure 32.24: Choose extra columns

Spreadsheet: The workflow results in an Entity List inspector that contains a spreadsheet with all the matching entities from the database search. This entity list is saved as **All Entities** in the Project Navigator. You can now perform NLP Network Discovery on this list.

Pathway Experiment Workflow

A typical **GeneSpring** experiment provides a recommended workflow to guide you through the most common analysis steps. A pathway experiment workflow browser contains the following analysis options:

1. **Experiment Setup:** This section allows you to group imported samples using **Experiment Grouping** (for example, samples 1,2,3 belong to a group A and samples 4,5,6 belong to a group B) and also specify an **Interpretation** to compare group A to group B. This grouping and interpretation information can subsequently be used when interpreting the overlaid data. Go to Experiment Setup on page 236 for more details on how to group samples and create interpretations.
2. **Simple (Guided) Analysis:** A two-step guided analysis to create biological pathways. Default network creation settings are used (see NLP Network Discovery on page 1051).
3. **Advanced Analysis:** In this analysis, you can specify filters on entities, relations, and their associated properties to build interaction networks (see NLP Network Discovery on page 1051). You can also provide a term and create a pathway with interactions associated with this term. See MeSH Network Builder on page 1067 for more details.

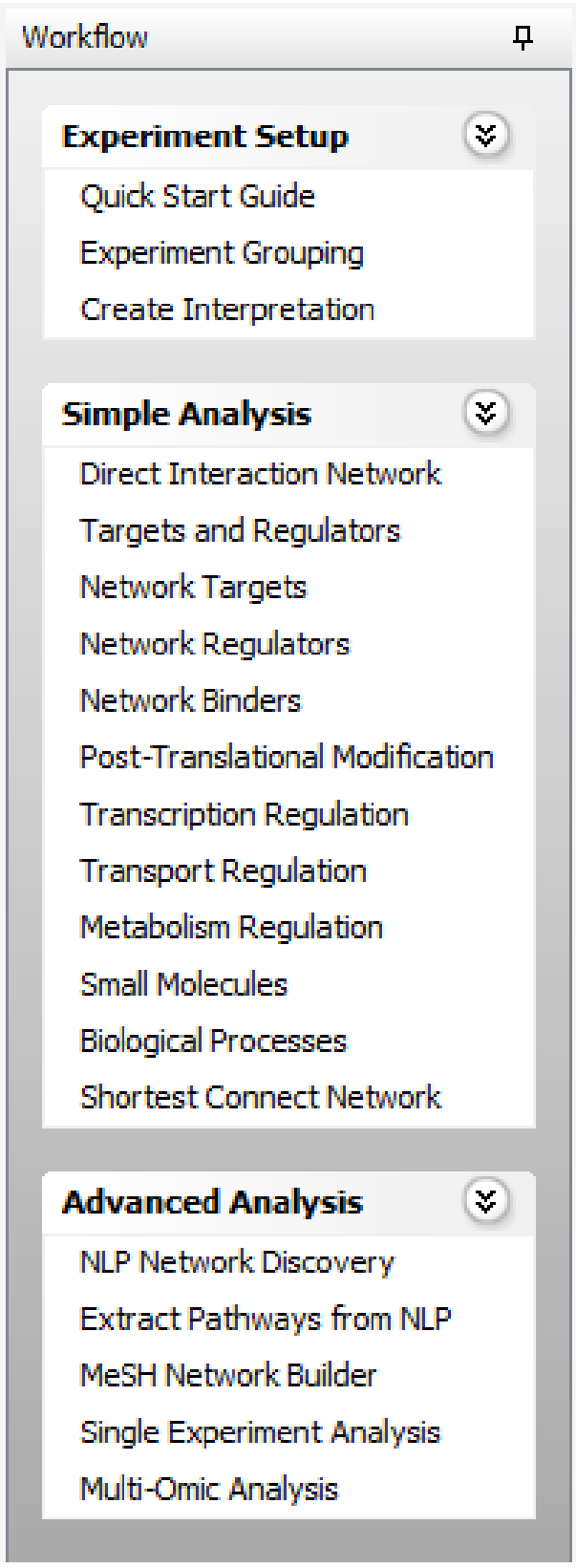


Figure 32.25: Pathway Experiment Workflow

32.6.2 Lassoing

Pathway operations can be run from two places - the workflow browser and the pathway view. A new pathway view can be launched from the toolbar. Workflow operations always run off the selected entity list in the Project Navigator. Operations launched via the pathway viewer's right-click menu run off the selection in the view.

Selected entities in one pathway view will be selected in all other open pathway views within the same experiment. However, not all selected entities in the pathway views will be selected in spreadsheets and other data views (like a Scatter Plot). This process, called lassoing, is based on experiment identifiers (eg. Probe Ids) and means that the selection is limited to entities which have are associated with the same experiment identifier. Clicking on an entity list highlights corresponding objects on the pathway view provided they have a matching experiment identifier. Clicking on a list will restrict all other data views to just entities contained in the chosen list.

The highlighting and restriction effect of clicking on an entity list also applies to other open experiments in the same project. Again this is done based on experiment identifier and behaves as described above.

32.6.3 Simple Analysis

The Simple Analysis guides the user to perform a cursory pathway analysis in two steps: Default parameters are set for the network building algorithms. Common filters that are biologically meaningful to users have been included in this section of the Workflow. They are the same as described in the Simple Analysis workflow of NLP Network Discovery on page 1051.

32.6.4 Advanced Analysis

The advanced analysis section allows you to customize all the algorithms and pathway functions. This section of the workflow browser gives you access to the following analysis features, described in their own sections of this manual:

- Single Experiment Analysis on page 900
- Multi-Omic Analysis on page 900
- NLP Network Discovery on page 1051
- MeSH Network Builder on page 1067
- Extract Relations via NLP on page 1061

Chapter 33

Analyzing CGH Data

This chapter describes in detail how you can import cytogenomics data derived from Comparative Genomic Hybridization (CGH) experiments into **GeneSpring** and visualize or compare results with those of other experiment types like expression or copy number experiments. For supported data types to create a CGH experiment in **GeneSpring**, see Supported data files on page 1080.

CGH is a molecular cytogenetic method for analyzing copy number variations (CNVs) relative to ploidy level in the DNA of a test sample compared to a reference sample. The aim of this technique is to quickly and efficiently compare two genomic DNA samples arising from two sources, which are most often closely related, because it is suspected that they contain differences in terms of either gains or losses of either whole chromosomes or subchromosomal regions (a portion of a whole chromosome).

Through **GeneSpring**'s clustering, correlation, biological contextualization, visualization and multi-omic analysis abilities the aberration data (Loss of heterozygosity - LOH, gain/loss) from a **GeneSpring** CGH experiment can be visualized, analyzed, and compared to results from other experiment types.

33.1 Creating a CGH Experiment

CGH experiments in **GeneSpring** are different compared to other experiment types available in **GeneSpring** in so far as they do not contain normalized signal values as primary data. Instead, visualizations and analyses are performed on average log ratio values and LOH values derived from the analysis workflows in the Agilent Cytogenomics and Agilent Genomic Workbench (AGW) products (see "Supported data files" on page 1080) and the **Call** information derived from this log ratio and LOH data (see "Data processing" on page 1081) when a CGH experiment is created in **GeneSpring**.

We recommend that you read the sections describing these processes, as well as the specific technology creation information (see "Technology Creation in CGH Experiments" on page 1080) before creating your CGH experiment in **GeneSpring**. The experiment creation process is then described under "Experiment Creation" on page 1083.

Once you have created a CGH experiment, the "Experiment Setup" section on page 1085 describes how you can add your experiment parameters and create interpretations used for data analysis and visualization in **GeneSpring**. The "Quality Control" section on page 1086 describes the available summary and filtering steps that allow you to inspect the quality of the imported samples before proceeding to the analysis and visualization workflows.

33.1.1 Supported data files

GeneSpring supports the import of single sample and multi-sample interval-based CGH reports for human, mouse, and rat organisms from Agilent Genomic Workbench (AGW) 7.0 in the form of *.xls files and Agilent Cytogenomics Interval Based Report in the form of *.xls and *.xml files to create CGH experiments.

NOTE

CGH experiment creation requires gene and transcript annotations for the respective organism build and **GeneSpring** provides these annotations for download from the **Annotations Manager** for many builds. If the build you are working with is not available, the Annotations Manager provides the tools to create a custom build.

33.1.2 Technology Creation in CGH Experiments

For CGH experiments, **GeneSpring** creates two technologies during experiment creation:

1 A technology based on the selected Genes and Transcripts model and build that contains all the genes and annotations available from the selected Genes

and Transcripts model. **GeneSpring** creates this technology in the background. **GeneSpring** uses the same technology, once created, for other experiments of the same Genes and Transcripts model and build. If additional annotations are required, these can be updated using the **Update Technology Annotations** > **From file or Biological Genome** function available from the **Annotations** menu.

NOTE This technology defines the number and identity of the entities present in the **All Entities** entity list created by default during experiment creation. However, the number of entities in the experiment is derived from the number of intervals in the input data files and does not match the number of entities in the technology (used for creating the All Entities entity list).

---

2 A technology defining the probe coordinates corresponding to the AMADID number associated with each imported sample. This information is required when translating the imported regions to probes as part of data analysis.

For catalogue arrays, the corresponding technologies are available to download from the Agilent server. **GeneSpring** prompts you to download the correct technology during the experiment creation process if it has not previously been downloaded.

For custom arrays, before creating a new experiment please download the *.tdt file for your custom array from the SureDesign portal to your system. Then go to **Annotations** ⟶**Create Technology** ⟶**Custom Agilent CGH** to create a custom CGH technology (see Creating Custom CGH Technology on page 80).

It is possible to import samples that were analyzed using different AMADIDs. Please specify the corresponding AMADID numbers at sample import.

33.1.3 Data Processing

The CGH report files should be in *.xls or *.xml (Agilent Cytogenomics Interval Based Report only) format.

GeneSpring imports the information contained in the following columns (see Table 33.1) to create the CGH experiment:

Table 33.1: Imported Columns by Source File Type

AGW single sample report	**AGW multi sample report**	**Cytogenomics report**
Chromosome	Chromosome	Chromosome
Start	Start	Start
Stop	Stop	End
Average CGH LR	Average CGH LR	Log ratio
LOH	LOH	LOH
Header information imported as sample annotations	Header information imported and associated for all samples	

GeneSpring generates an additional column **Call** for each sample from the imported data as follows:

Table 33.2: Aberration Type Call Determination by Data Type

Aberration Type	**Avg CGH Log Ratio**	**LOH Score**
Gain	Avg CGH Log Ratio > 0	NULL
Loss	Avg CGH Log Ratio < 0	NULL
LOH	NULL	LOH Score
Normal*	NULL	NULL

* Normal calls are displayed as blank spread sheet columns

This **Call** information is used in **GeneSpring** to filter out samples, find regions with common aberrations across samples, clustering as well as data overlay on

pathways.

33.1.4 Experiment Creation

Use this procedure to create a new CGH experiment in **GeneSpring**.

1. Launch **GeneSpring**.
   The **Startup** dialog is displayed and offers three options:
   - Create new project
   - Open existing project
   - Open recent project
2. Choose to either create a new project that will hold the experiment you are about to create, or open a recent or existing project to add the new experiment to an existing project.
   - If you select **Create new project**, a window appears in which you can record the details (**Name** of the new project and **Notes**). Click **OK** to proceed.
     The **Experiment Selection Dialog** then appears with two options:
     – **Create new experiment**
     – **Open existing experiment**
     Select **Create new experiment** in this dialog to create a new CGH experiment.
   - If you chose to open an existing project and want to add a new CGH experiment to that project, click the **New experiment** icon in the toolbar.

   The **New Experiment** dialog opens. Assign the **Experiment name** and specify the **Analysis type** as **CGH** from the drop-down list. Use the **Experiment notes** field to record any additional information you would like to save along with the experiment for later reference.
3. Click **OK**.
   The New CGH Experiment wizard-driven workflow opens.
4. In Step 1 of this workflow, specify the experiment **Organism**, **Genome Build**, and **Gene Annotation** using the corresponding drop-down lists. The drop-down lists only contain values for annotations that have previously been downloaded from the Annotations Manager (see "Annotations Manager" on page 86). Click **Next** ≫ to proceed to Step 2.
5. In Step 2, select **Choose Files** to browse for and select the input data files saved on your system in the **Open** dialog. Click **Open**.
   Alternatively, if you have already created a CGH experiment in **GeneSpring** and want to use the same data to create another experiment, click **Choose Samples** and identify the existing samples in the **Search** dialog. Click **Finish**.
6. Click **Reorder** to arrange the selected files or samples in the desired order. This is the only time and place in **GeneSpring** where you have the option to determine sample order.
   When you have selected all the desired files or samples and put them in the desired order, click **Next** >>.

7 In Step 3, assign the corresponding AMADID number(s) to each of the files selected for import. It is possible to import samples that were analyzed using different AMADIDs. Select all the rows corresponding to samples that are associated with the same AMADID and click **Assign AMADID Numbers**. Repeat with the remaining samples if they are associated with another AMADID number. You can also double-click the corresponding cells to type in individual AMADID numbers. See "Technology Creation in CGH Experiments" on page 1080 for more information on why AMADID numbers are required for CGH experiment creation.

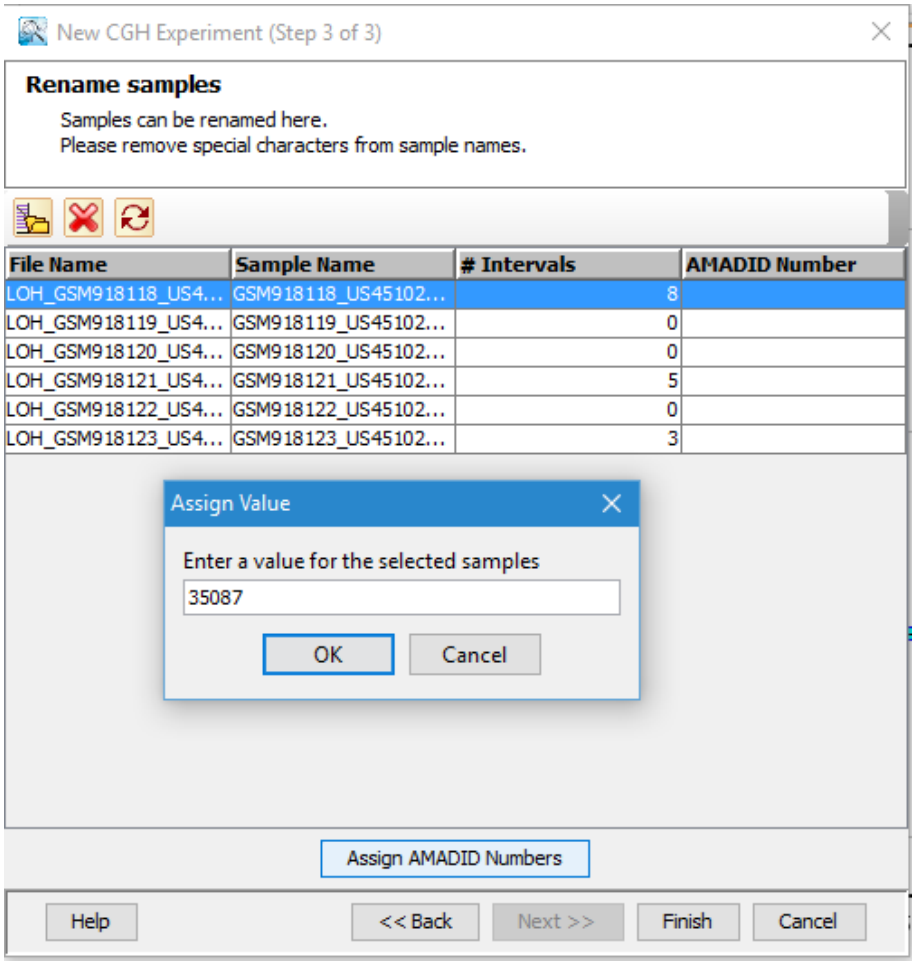


Figure 33.1: Experiment Creation - Rename samples and assign AMADID number

8 Optionally, rename samples by double-clicking the corresponding cell in the Sample Name column and typing the new sample name.

Please note that AMADID assignment and sample renaming are not available when creating a new experiment from existing **GeneSpring** samples.

9 When you have assigned AMADID numbers to all the samples to be imported and named samples as desired, click **Next** >> to create the experiment.

GeneSpring creates the experiment using the specified samples and AMADID information. In the Experiment Navigator, **GeneSpring** creates an **All Entities** entity list that contains all the entities that are part of the **Gene and transcript**

model as well as a **All Regions** region list that contains all the regions found in the imported sample files.

NOTE

Unlike in other **GeneSpring** experiments, quality control, filtering, and analyses will not be performed starting from the **All Entities** entity list since it does not contain experiment specific data. Instead, the QC and filtering steps are performed starting from the **All Regions** region list, with analyses and biological interpretation and contextualization are performed on the entity and probe lists resulting from the respective **Translate Regions** analysis steps described under "Analysis" on page 1090.

The default view of the experiment in the desktop area after experiment creation is a Spreadsheet View of all the regions in the experiment. The cells for different data columns (Average CGH Log Ratio, LOH Value, Call) are colored differently by default.

Chromosome	Start	End	AvgCGHLR_D283	AvgCGHLR_D341	LOH_D283	LOH_D341	Call_D283	Call_D341
chr1	16496	121386673	0.069419	0.01783			Gain	Gain
chr1	106617209	106617268	-0.057008	0.291617			Loss	Gain
chr2	10702	92233767	0.041928	9.02E-4			Gain	Gain
chr2	153179	12720801		-0.002805				Loss
chr2	95344257	153047916	0.028444	-0.008102			Gain	Loss
chr3	62199	90430844	0.061415	0.007574			Gain	Gain
chr3	93538467	102826106	0.051223	0.001825			Gain	Gain
chr4	16138	49581867	0.059781	0.020605			Gain	Gain
chr4	52697788	145301478	0.052005	0.005072			Gain	Gain
chr4	62437318	66697518				6.410177		LOH
chr5	22149	46279735	0.031347	0.015115			Gain	Gain
chr5	1849636	37789193	0.027224				Gain	
chr5	49431141	68791331	0.052297				Gain	
chr5	49431141	141941762		0.006491				Gain
chr5	68812530	68812589	-0.179071				Loss	
chr6	67659	58686125	0.05418	0.527709			Gain	Gain
chr6	11215030	11668363	0.008249				Gain	
chr6	61944991	115563789	0.001778	0.527847			Gain	Gain
chr7	58764	57955498	-0.889155	-0.009126			Loss	Loss
chr7	884743	57560040			296.2112		LOH	
chr7	61059509	99696626	-0.884655	-0.003832			Loss	Loss
chr7	62260197	159100528			296.2112		LOH	
chr8	21227	43738803	0.07829	0.49212			Gain	Gain
chr8	1012663	7858276		0.464693				Gain
chr8	7881477	8460838		0.37368				Gain
chr8	8490446	8952969		0.595873				Gain
chr8	8982291	9578334		0.523187				Gain
chr8	9598226	9819248		0.530167				Gain
chr8	9856474	18483823		0.482248				Gain
chr8	13518932	13946373	0.057154				Gain	
chr8	13978825	14325428	-0.07681				Loss	
chr8	14344377	14498717	0.037313				Gain	
chr8	14525831	31001120	0.085519				Gain	
chr8	46867580	73700950	1.278786				Gain	
chr8	46867580	100464006		0.504926				Gain
chr8	73721985	77775038	1.054508				Gain	
chr8	107852265	111627835				10.680011		LOH
chr9	12078	47239015	0.056251	0.012011			Gain	Gain
chr9	65482565	98197773	0.050822	0.034142			Gain	Gain
chr10	62787	39047586	0.026191	-0.021378			Gain	Loss
chr10	42385323	96972879	0.052028	0.004331			Gain	Gain
chr11	64333	2186034	-0.075056				Loss	

Figure 33.2: Spreadsheet View of CGH experiment

33.1.5 Experiment Setup

Once an experiment is created, the **Data Import Wizard** steps appear on the right hand side. Following is an explanation of the various workflow links

- **Quick Start Guide:** Clicking on this link will open the appropriate chapter of the user manual in your default browser.
- **Experiment Grouping: Experiment Parameters** define the grouping or the replicate structure of the experiment. For details refer to the section on "Experiment Grouping" on page 236.
- **Create Interpretation:** An interpretation specifies how the samples are grouped into experimental conditions for display and used for analysis. For details refer to the section on "Create Interpretation" on page 240.

NOTE You can only create non-averaged interpretations in CGH experiments.

33.1.6 Quality Control

For Quality control purposes, **GeneSpring** offers the option to view a summary of the aberrations present in the imported samples, to filter regions by samples and to filter regions by aberrations.

33.1.6.1 Summary Statistics

The Summary Statistics option provides the following information for each imported sample:

- Number of gains
- Number of losses
- Number of LOH regions
- Total genomic length of gains
- Total genomic length of losses
- Total genomic length of LOH regions

You can use the information provided in this table to decide whether you want to remove whole samples (see Filter by Samples on page 1086) or filter regions by the presence of selected aberrations (Gain, Loss, LOH) (see Filter by Aberrations on page 1087).

33.1.6.2 Filter by Samples

This filter allows you to remove the regions associated with particular samples from a selected region list. **GeneSpring** saves a new filtered region list as a child list of the input region list that only contains the regions found in the remaining samples.

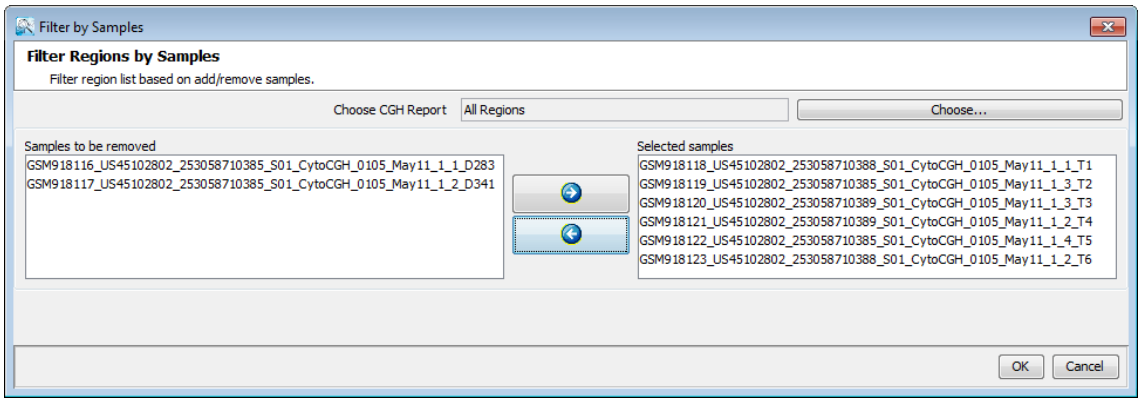


Figure 33.3: Filter by Samples

Use this procedure to filter the regions associated with certain samples from the selected input region list:

1. Select **Filter by Samples** from the **Quality Control** section of the Workflow Browser.
   The **Filter by Samples** dialog box opens (see Figure 33.3).
2. The **Choose CGH Report** field shows the selected input region list. By default **GeneSpring** selects the active region list in the Experiment Navigator (shown in bold). To change the region list, click **Choose...** and select the desired region list in the **Choose CGH Report** dialog box that opens. Click **OK**.
3. Select the samples you want to remove from the **Selected samples** field, which lists all the samples contributing to the selected region list. Use Ctrl-click to select multiple samples spread over the list, or Shift-click to select a range of samples. Click the **Move left** button to move the selected samples to the **Samples to be removed** field.
4. Click **OK** to apply the filter.

GeneSpring saves a new region list that only contains the regions present in the retained samples.

33.1.6.3 Filter by Aberrations

This filter allows you to create a new region list that only contains regions with a certain aberration profile. You can select to retain only regions with Gain, Loss, or LOH aberration types or a combination of these types based on their presence across samples or conditions as defined in the selected interpretation.

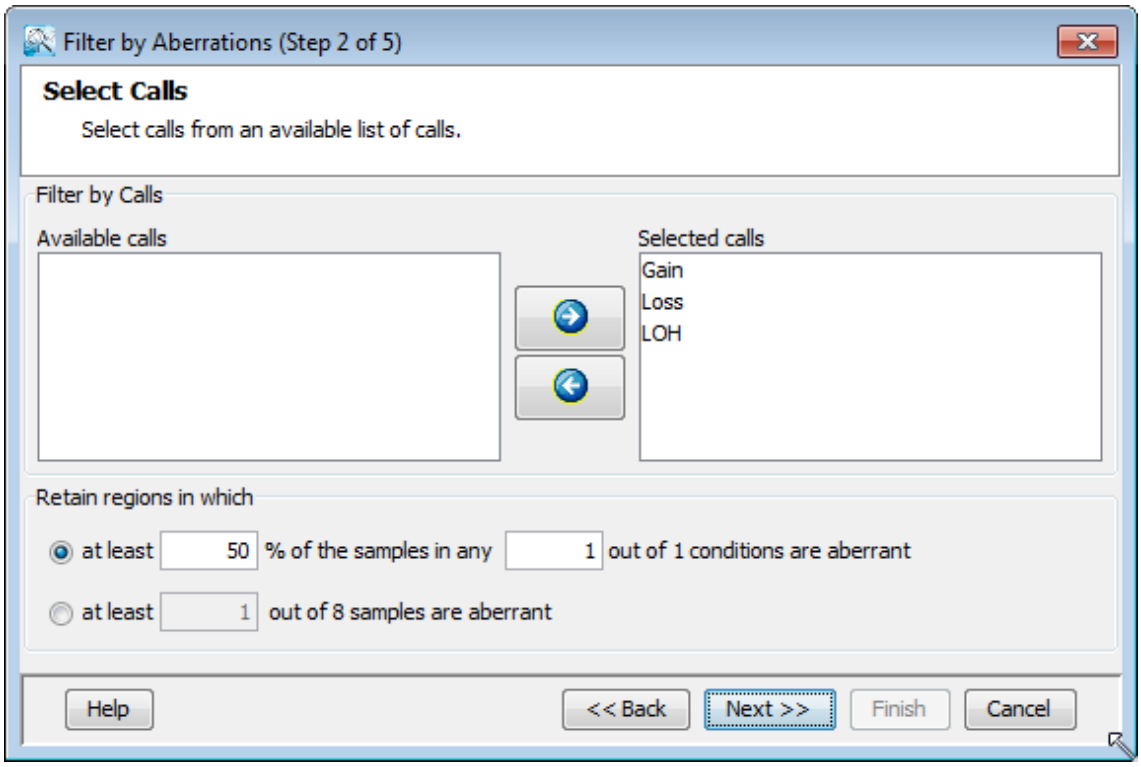


Figure 33.4: Filter by Aberrations Step 2 - Select Calls

Use this procedure to filter the selected input region list by aberration types:

1. Select **Filter by Aberrations** from the **Quality Control** section of the Workflow Browser.

   The **Filter by Aberrations** wizard-driven workflow opens. It contains up to five steps depending on the selected aberrations in Step 2.

2. In Step 1, the **CGH Report** field shows the selected input region list. By default **GeneSpring** selects the active region list (shown in bold) in the Experiment Navigator. To change the region list, click **Choose...** and select the desired region list in the **Choose CGH Report** dialog box that opens. Click **OK**.

3. The **Interpretation** field shows the selected interpretation. By default **GeneSpring** selects the active interpretation (shown in bold) in the Experiment Navigator. To change the interpretation, click **Choose...** and select the desired interpretation in the **Choose Interpretation** dialog box that opens. Click **OK**.

4. Click **Next** >>.

5. In Step 2, define the aberration profile of the regions that you would like to retain in the filtered list by first selecting the calls (aberration types) you want to filter on. Use Ctrl-click to select multiple aberration types. Click the **Move left** button to move the aberration types you do not want in the filtered region list to the **Available calls** field.

   For example, if you want to filter the **All Regions** region list to retain only regions that show a loss of heterozygosity (LOH), select all other aberration

types shown in the **Selected calls** field and move them to the **Available calls** field so that only the LOH type remains under **Selected calls**.

6 Click **Next >>**.

7 Select the **Minimum** and **Maximum cut-offs** for the aberration type that is shown. You can use the slider or type an exact value in the corresponding fields:

- For the **Gain** and **Loss** aberration type, **GeneSpring** filters regions by the specified range of **Average CGH Log Ratio** values in Step 3 of the workflow. Step 3 will only display cut-offs for the aberration types that were selected in Step 2. If neither **Gain** nor **Loss** were selected in Step 2, **GeneSpring** skips Step 3 and proceeds to Step 4.
- For the **LOH** aberration type, you can set the required cut-offs in Step 4 of the workflow. If **LOH** was not selected in Step 2, **GeneSpring** skips Step 4 and proceeds to Step 5.

8 Click **Next >>**.

9 In Step 5, **GeneSpring** displays the details of the regions retained after filtering as well as other details that will be saved along with the new region list, like **Name**, **Organism**, **Notes** (that contain the selected filter settings), **Creation Date**, etc., for later reference.

You can edit the default **Name** and **Notes** fields as required.

For a different result, use the **Back <<** button to change the filter settings in any of the earlier steps.

10 Click **Finish** to save the new region list.

GeneSpring saves a new region list that only contains regions with the specified aberrations as a child list of the selected input region list.

33.2 Analysis

In CGH experiments, **GeneSpring** offers the following analysis options from the **Analysis** section of the Workflow Browser:

- **Translate Regions to Genes**
- **Translate Regions to Probes**
- **Find Common Aberration**
- **Clustering**
- **Correlation**

It is also possible to perform multi-omic pathway or correlation analysis on CGH experiments along with other experiment types in **GeneSpring**. See "Multi-Omic Analysis in **GeneSpring** " on page 603 for more details.

33.2.1 Translate Regions to Genes

In order to perform some of the analyses and make use of the available visualizations, **GeneSpring** requires gene-level information. Use the **Translate Regions to Genes** workflow to translate any regions of interest from the experiment to overlapping genes and create corresponding gene lists for further analysis like **Clustering**, **Correlation**, **GO Analysis**, and **Pathway Analysis**, as well as creating Venn diagrams and visualizing entities in the Genome Browser.

1. Select **Translate Regions to Genes** from the **Analysis** section of the Workflow Browser.

   The **Translate Aberrant Regions to Genes** wizard-driven workflow opens.
2. The **CGH Report** field shows the selected input region list. By default **GeneSpring** selects the active region list (shown in bold) in the Experiment Navigator. To change the region list, click **Choose...** and select the desired region list in the **CGH Report** dialog box that opens. Click **OK**.
3. Click **Next** >>.
4. Step 2 of the workflow displays all the genes overlapping the regions in the input CGH Report by at least 1 base pair. To make the overlap criterion more stringent use the **Change cut-off** button and enter the desired cut-off value in the **Region List Translation parameters** dialog. Click **OK**.
5. Click **Finish**.

NOTE If you know that you want to use a more stringent overlap criterion than 1bp for most of your experiments, you can change the default cut-off at **Tools** > **Options** > **Agilent CGH** > **Translate regions to genes**.

GeneSpring saves a new entity list that contains all the overlapping genes as a child list of the input CGH report and displays the resulting list as a **Spreadsheet** in the desktop area. **GeneSpring** computes the **Call** columns (one for each sample) for each overlapping gene (as well as probe, see "Translate Regions to Probes" on page 1091) based on the logic summarized in Table 33.3.

Table 33.3: Call Determination for Translated Genes and Probes

Aberrations in Overlapping Regions	Call at Gene/Probe Level
Gain	Gain
Loss	Loss
LOH	LOH
Gain, Gain	Complex
Loss, Loss	Complex
LOH, LOH	Complex
Gain, Loss	Complex
Gain, LOH	Gain
Loss, LOH	Loss

You can display any entity list containing genes as a **Spreadsheet** at any time by clicking the **Spreadsheet** toolbar icon when the desired entity list is the active entity list (displayed in bold) in the experiment.

33.2.2 Translate Regions to Probes

To identify the probes that overlap regions of interest, **GeneSpring** offers a way to translate regions to probes using the probe information present in the AMADID technology that was created during CGH experiment creation. You can perform **Clustering**, **Correlation**, **GO Analysis**, and **Pathway Analysis** on the resulting probe lists, as well as create Venn diagrams and visualize probes in the Genome Browser.

1 Select **Translate Regions to Probes** from the **Analysis** section of the Workflow Browser.

   The **Translate Aberrant Regions to Probes** wizard-driven workflow opens.

2 The **CGH Report** field shows the selected input region list. By default **GeneSpring** selects the active region list (shown in bold) in the Experiment Navigator. To change the region list, click **Choose...** and select the desired region list in the **CGH Report** dialog box that opens. Click **OK**.

3 Click **Next** >>.

4 Step 2 of the workflow displays all the probes that show a mid-point overlap with a given aberrant region in the input CGH Report.

5 Click **Finish**.

GeneSpring saves a new entity list that contains all the overlapping probes as a child list of the input CGH report and displays the resulting list as a **Spreadsheet** in the desktop area. In Genome Browser view this entity list is displayed a multi-sample scatter plot. Copy number probes which show a mid-point or complete overlap with a given aberrant region are assigned the average CGH log ratio of the region in the sample while all overlapping SNP probes are assigned a LOH score. Then **GeneSpring** computes the **Call** columns (one for each sample) for each overlapping probe based on the logic summarized in Table 33.2 and Table 33.3.

You can display any probe list or a region list as a **Spreadsheet** in the desktop area at any time by right-clicking over the list icon in the Experiment Navigator and selecting **Spreadsheet View**.

33.2.3 Find Common Aberrations

Use this workflow to identify common aberrations in entities (genes) across two conditions in a selected interpretation. If you know that the aberrations you are looking for are typically present in the different conditions to a lesser (for example normals) or higher degree (for example tumors) you can also specify individual percentage cut-offs for each chosen condition.

NOTE This workflow is only supported for entity lists resulting from the **Translate Regions to Genes** workflow (see "Translate Regions to Genes" on page 1090).

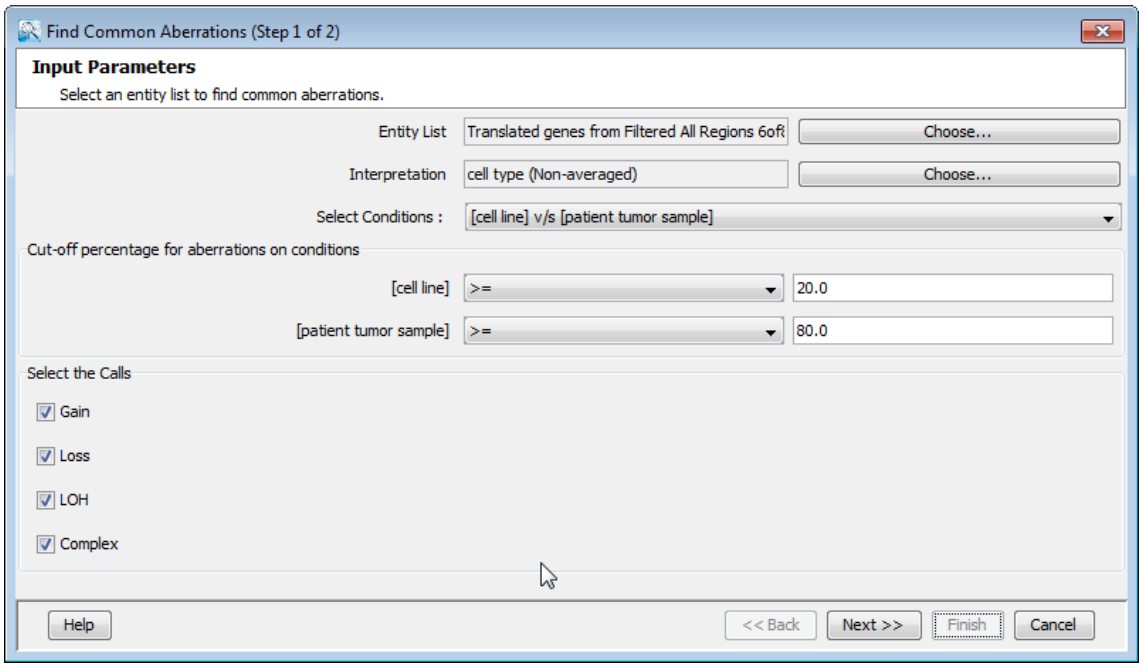


Figure 33.5: Find Common Aberrations

1 Select **Find Common Aberrations** from the **Analysis** section of the Workflow Browser.

   The **Find Common Aberrations** wizard-driven workflow opens.

2 Choose the input **Entity List** in which you want to identify common aberrations. By default, the workflow selects the active entity list (shown in bold) in the Experiment Navigator. Click **Choose...** to select a different entity list. Click **OK**.

3 Choose the input **Interpretation**. By default, the workflow selects the active interpretation (shown in bold) in the Experiment Navigator. Click **Choose...** to select a different interpretation. Click **OK**.

4 If the selected interpretation contains more than two conditions, select the two conditions across which you want to find common aberrations from the **Select Conditions** drop-down list.

5 Determine the **Cut-off percentage for aberrations on conditions** for each condition:

   a Select the required condition from the drop-down, choosing from: <, <=, >, or >=.

   b Define the percentage cut-off in the field corresponding to the named condition from the selected interpretation.

   For example, if you want to specify that a given aberration should only be present in less than or equal to 20% of samples in the first condition, select <= from the drop-down list and type '20' in the provided text field.

6 Select which calls you want **GeneSpring** to look for under **Select the Calls**. You can select any combination of the following calls potentially associated with the input entity list: **Gain**, **Loss**, **LOH**, or **Complex**.

7 Click **Next** >>.

8 Step 2 of the workflow displays the result. **GeneSpring** appends two new columns to each gene containing the exact percentage at which an aberration is present across the samples of each condition.

If the cut-offs you provided result in no entities, **GeneSpring** informs you accordingly and you can return to Step 1 of the workflow using the << **Back** button to change the cut-offs as desired.

9 Review the list of genes with common aberrations across the chosen conditions.

You can go << **Back** to change the cut-offs and rerun the workflow to observe how the number of entities varies.

You can also select a subset of genes from this list by selecting the corresponding rows and save this subset as a custom entity list using the **Save custom list...** button below the table of entities.

10 Click **Finish** when you are satisfied with the result list of genes with common aberrations across the chosen conditions.

GeneSpring saves the result entity list as a child list of the input entity list in the Experiment Navigator. You can use this entity list to perform any of the downstream analysis workflows available for translated entity lists in **GeneSpring**.

33.2.4 Clustering

In CGH experiments you can perform Hierarchical Clustering Analysis on entities, conditions, or both for translated gene lists and probe lists to identify genes, probes or samples with similar aberration profiles based on the associated **Average CGH Log Ratio** data or **Call**.

To enable clustering on **Call** data, each **Call** has been assigned a numerical value as follows:

Table 33.4: Clustering on Calls - Numerical Value Assignment by Call Type

Call	Value for Clustering
Gain	2.0
Loss	-2.0
LOH	1.0
Complex	0.0

These default values are configurable from the **Options** dialog accessible via **Tools** > **Options** > **Agilent CGH** > **Clustering on Calls**.

Use the following procedure to set up and run hierarchical clustering analysis.

1 Select **Clustering** from the **Analysis** section of the **Workflow** browser.
The **Clustering** window opens.

2 Click **Choose...** next to the **Entity List** field and select the translated entity list or probe list that contains the set of genes or probes you want to analyze in the **Choose Entity List** dialog box. Click **OK**. See "Choose Entity List" on page 712.

3 Click **Choose...** next to the **Interpretation** field and select the interpretation that you want to analyze in the **Choose Interpretation** dialog box. Click **OK**. See "Choose Interpretation" on page 713.

NOTE Interpretations, other than the **All Samples** interpretation, have to be created before starting the **Clustering** workflow (see "Create Interpretation") to be available for selection in this workflow step. The selected interpretation in the Experiment Navigator (the one that appears in bold) is selected by default.

---

4 Keep **Hierarchical** selected in the **Clustering Algorithm** drop-down list.

5 Select by which type of data you want to perform clustering under **Cluster by**: **AvgCGHLR** or **Call**.

6 Select one of the following parameters from the **Cluster on** drop-down list: **Conditions**, **Entities**, **Both entities and conditions**.

7 Keep **Eucledian** selected in the **Distance metric** drop-down list.

8 Select one of the following parameters from the **Linkage rule** drop-down list: **Ward's**, **Centroid**, **Median**, **Single**, **Complete**, or **Average**. See "Linkage Rule" on page 744.

9 Select whether you want to **Cluster within conditions** or not.

10 Click **Preview**.
GeneSpring performs clustering analysis according to your chosen parameters and displays the result as a cluster tree in the results pane on the right side of the **Clustering** window.

11 Review the cluster tree.
If you selected **Conditions** or **Entities** from the **Cluster on** drop-down list in Step 6 of this section, a cluster tree is displayed with the following two panes:

- Summary view (to the left for Conditions and on the top for Entities)
- Detailed view (to the right for Conditions and at the bottom for Entities)

If you selected **Both entities and conditions** a combined tree is displayed with two panes:

- Summary view (left pane)
- Detailed view (right pane)

To navigate the combined cluster tree, use the right-click options to switch between the **Condition Tree Navigation Mode** and the **Entity Tree Navigation Mode**.
See "Clustering result visualizations in **GeneSpring**" on page 696.

12 You can change any of the parameter settings in the pane on the left and rerun the analysis by clicking **Preview**.

13 When you are ready to save your clustering result in the experiment, click **OK**.

14 Review the details that will be saved in your experiment along with the cluster tree in the **Inspector** that opens. You can edit the default **Name** and **Notes**.

15 Click **OK** to save the hierarchical clustering tree in your experiment.

The cluster tree is saved as an object in your experiment hierarchy below the selected input entity list and displayed in the desktop area.

A cluster tree object created in a CGH experiment can be further explored and analyzed in the same way as cluster trees in other experiment types, including the option to add sample parameter plots to the dendrogram view. All the available options are described under "Working with a Cluster Tree in the Dendrogram View" on page 683.

33.2.4.1 Region List Clustering

In CGH experiments you can perform clustering on a region list with the available columns.

To perform clustering on a region list, Right-Click on a region list in the **Experiment Navigator** and select **Run Clustering**.

All the available options are described under How to Perform Clustering Analysis on Region Lists on page 678.

33.2.5 Correlation

Correlation Analysis for both single CGH experiments and multi-omic experiments containing a CGH experiment is possible in **GeneSpring**. The analysis takes either a translated entity or probe list as input.

See chapters "Correlation Analysis" on page 747 and "Multi-Omic Analysis" on page 603 for more details on each of these two options.

NOTE For CGH experiments, Correlation Analysis is performed on the average CGH log ratio data, since a CGH experiment does not contain normalized signal values.

33.3 Results Interpretations

For biological interpretation and contextualization of the results, **GeneSpring** provides a number of options:

- Gene Ontology (GO) Analysis is enabled for translated entity and probe lists (see "GO Analysis" on page 1097).
- Venn Diagrams can be created on both entity and probe lists resulting from the corresponding Translate Regions workflows (see "Analysis" on page 1090).

  Use the **Venn Diagram** drop-down icon to launch Venn diagrams on entity lists and select either Gene Lists or Probe Lists in the dialog. For more information about the Venn Diagram visualization in **GeneSpring** and its associated features see "The Venn Diagram" on page 173.
- Visualize CGH data alongside expression or copy number data in the **GeneSpring** Genome Browser. Special visualization options for CGH data beyond the standard Genome Browser features are available. See "Genome Browser Visualization" on page 1097 for more details.
- Import entity lists from other experiments to compare results or visualize entities in **GeneSpring** .

  Use the **Import entity list** toolbar icon to launch the import dialog. For more information about this utility see "Import Entity List from File" on page 289.
- A range of entity list and region list interrogation operations is available from the **Utilities** section of the Workflow Browser. See "Utilities" on page 1100 for more details.

33.3.1 GO Analysis

See chapter "Gene Ontology Analysis" on page 827 for more details.

33.3.2 Genome Browser Visualization

See chapter "Genome Browser" on page 195 for more details.

Specific for CGH experiments, the following Genome Browser visualization options are available:

- Multi-sample tracks for CGH reports, probe lists, and imported probe-level reports.
- Download CGH specific annotations (dGV) from the Annotations Manager and visualize them in the Genome Browser alongside your CGH experiment results.

In the Genome Browser view for CGH experiments you can drag and drop Region Lists, Entity Lists, and Imported Probe Reports. The defaults views are:

- Region Lists are shown as multi-sample interval plots.
- Entiy Lists are shown as region tracks.
- Imported Probe Reports are shown as scatter plots.

Figure 33.6 illustrates the **Genome Browser** view of a CGH experiment with a multi-sample interval track and a multiple-sample scatter plot.

Figure 33.6: Genomer Browser view of a CGH experiment displaying a Multi-sample Interval and Scatter Plot(s).

For multi-sample plots in CGH, you can:

- Right-click on a track $\longrightarrow$**Show/Hide samples** to select or deselect samples to be displayed in the Genome Browser view.
- Customize colors for samples.

33.4 Pathway Analysis

Pathway Analysis for both single CGH experiments (in the form of **Single Experiment Analysis** accessed from the Workflow Browser) and multi-omic experiments containing a CGH experiment is possible in **GeneSpring**.

See chapters "Pathway Analysis" on page 859 and "Multi-Omic Analysis" on page 603 for more details on each of these two options.

NOTE

Unique for **GeneSpring** CGH experiments, the data overlay over pathways for matching entities in the form of quilt plots also contains symbols that indicate the call (Gain, Loss, or LOH) for a particular sample. Heatstrips are not available for CGH experiments. LOH and Complex calls data is not displayed in the pathway spread sheet.

33.4.1 Quilt plots in Pathway View of CGH Experiment

Quilt Plots represent the aberration type call associated with matching entities directly in the Pathway Viewer.

Figure 33.7 illustrates a pathway resulting from SEA in a CGH experiment. The Quilt plots illustrate the aberration type which is detailed in the legend.

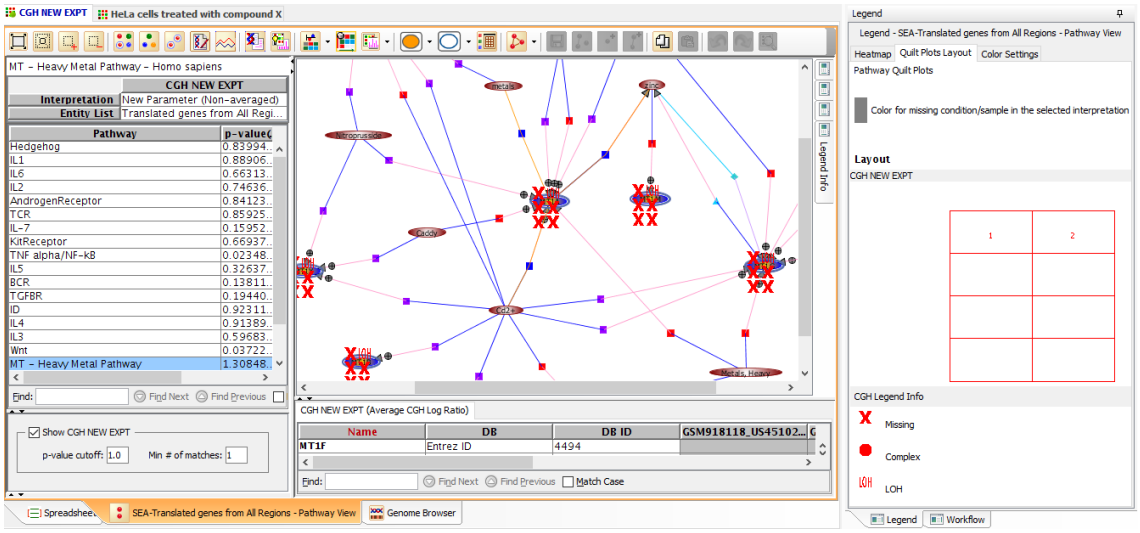


Figure 33.7: Pathway View of a CGH Experiment illustrating the Quilt Plots with different type of Aberrations.

33.5 Utilities

A number of additional utilities are available to work with entity lists and region lists in a CGH experiment in **GeneSpring**. All of them are described in more detail in this section.

33.5.1 Filter on Entity List

This utility allows you to filter the entities in an entity list based on the data in the various columns in the list.

1. Select **Filter on Entity List** from the **Utilities** section of the Workflow Browser.

   The **Filter on Entity List** wizard-driven workflow opens.
2. Select the entity list to be filtered. By default the **Entity List** field contains the active entity list (shown in bold) in the Experiment Navigator. Click **Choose...** to change the default selection.
3. Click **Next** >>.
4. Click the **Search Field** cell to view a list of all the columns present in the selected entity list as a drop-down list. Select the column for which you want to define a filtering parameter from this drop-down list.
5. Click the **Condition** cell to view the drop-down list of available conditions. Choose from these options for:
   - Numerical columns: '=', '≠', '≤', '≥', 'in the range'
   - Categorical columns: 'equals', 'does not equal', 'starts with', 'ends with', 'contains'
6. Double-click the **Value** cell to type in the desired filtering criterion.
7. Click **Add** to add another row and repeat Steps 4 to 6 to define another condition.
8. Combine the defined conditions using the **OR** or **AND** options provided in the **Combine search conditions by** drop-down list.
9. To remove an unwanted condition, select the corresponding row and click **Remove**.
10. Click **Next** >>.
11. The table in Step 3 of the workflow shows all the entities in the selected entity list with the entities that match the filter criteria highlighted as selected.

NOTE This table is interactive. Clicking anywhere on the table will erase the filter-based selection. However, using Ctrl-Click you can select additional rows without losing the filter-based selection.

12. Click **Next** >> to take forward only the selected entities in the table.

13 Review the information saved along with the new entity list containing only the selected entities from Step 3 of the workflow. You can edit the **Name** and **Notes** of the new entity list here.

14 Click **Finish** to save the new filtered entity list.

GeneSpring saves a new entity list that contains all the selected genes as a child list of the input entity list.

33.5.2 Region List Operations

The **Region List Operations** utility allows you to:

- create a new region list that only contains selected columns from the input region list,
- filter the input region list based on specified conditions on the retained columns,
- add columns that are functions of selected existing columns.

In the first step of the **Region List Operations** workflow, the input region list is chosen and the columns which need to be present in the output region list can be selected. It is essential that the output region list contains the columns *Chromosome, Start, End*. Otherwise an error is given. The input region list can be a region list from the Experiment Navigator or an annotation from the Annotations Manager.

The subsequent steps of the wizard depend on whether the input region list is large or small. If the region list contains more regions than the 'Region List Size Threshold' specified in *Tools* $\longrightarrow$*Options* $\longrightarrow$*Miscellaneous* $\longrightarrow$*Region List Size*, it is considered as 'large', otherwise it is considered as 'small'. In case of large region lists, Step 2 of the workflow shows a histogram view or summary statistics view of the region list. But for small region lists, the workflow proceeds directly to Step 3, which shows the complete spreadsheet view of the region list.

In Step 2, the chosen filter options will be applied only upon clicking the 'Apply Filter' button, whereas in Step 3 the region list view is dynamically refreshed whenever a filter option is changed. Also in Step 2, only if the region list resulting out of the chosen filter options becomes 'small', the 'Next' button is enabled to allow you to move to the more interactive filtering options in Step 3.

In both steps 2 and 3, various filtering options on the different columns are shown on the right side. There is a section on the right-side for each of the columns, which can be collapsed or expanded using a triangle-shaped icon on the toolbar of that section. For categorical columns, if the number of the categories is less than or equal to 30, check boxes are provided to select the categories. Also icons are provided to 'select all' or to 'clear' the selection. For other categorical columns, categories can be selected by entering the category value in text boxes. Whether the category should be equal to or not equal to the specified value is chosen using the radio button option. Text boxes can be added or deleted using the '+' / '-' buttons. For other numeric columns, minimum and maximum values can be specified using the slider or the adjacent text-box.

When the filtering options are chosen (or applied in case of Step 2), the result is displayed on the left-side. In Step 2, the default view is the spreadsheet, but various other view options are provided on the toolbar above the spreadsheet.

Apart from these, there is a special toolbar icon ''**Column Operations**'' which is used to add new columns that are functions of the existing columns. This icon has a drop-down menu with a list of standard functions, along with an option ''New Column With Formula'' to specify a new function. This option allows specification of a wide variety of functions.

The output region list which is a result of this utility is saved in the Experiment Navigator as a child list of the input region list, unless the input region list was an annotation, in which case it is saved under a folder called **Created from annotations**.

33.5.3 Import Region List

This option allows you to import relevant region lists in *.bed, *.tsv, *.csv, or *.txt format into the CGH experiment for visualization alongside your CGH results.

1. Click **Import Region List** under the **Utilities** section in the Workflow Browser.
2. Browse for the desired region list on your system in the **File Import (Step 1 of 4)** dialog.
   - If applicable to the region list that you want to import, select **The file has zero-based, half open coordinates**.
3. Click **Next**.

   If you chose a *.bed file, **GeneSpring** imports the region list directly without going through any of the other workflow steps. For all other file formats, the following steps apply.
4. Specify the applicable column **Separator** (for example Tab, Space, or Comma), Text qualifier (if present), Missing value indicator (if present), and Comment indicator (for example #, or *) in the **File Import (Step 2 of 4)** dialog.

   The **Preview** displayed at the bottom of the dialog changes dynamically to help you select the appropriate options.
5. Click **Next**.
6. Specify the correct **Header Row Option** in the **File Import (Step 3 of 4)** dialog.

   The default option is **Take the first row in the selection as the column header**. If the file does not contain a header row, select **There is no row containing column headers**. Use the **Preview** at the top of the dialog to make the appropriate choice.
7. Click **Next**.
8. In the **File Import (Step 4 of 4)** dialog, specify which columns in your region list file correspond to the three mandatory columns for region list import in **GeneSpring**: Chromosome, Start, End.
9. Specify which of the remaining optional columns you want to import by deselecting any unwanted columns (all columns are selected by default).
10. Specify the **Data Type** (for example, integer, float, string) and **Attribute Type** (categorical or continuous) for each optional column using the available options from the drop-down lists that appear on clicking on the corresponding cell in the table at the bottom of the dialog.

11 Click **Finish**.

GeneSpring imports the region list. If this is the first imported region list, **GeneSpring** creates a new folder called **Imported Region Lists** at the bottom of the **All Regions** region list object hierarchy and saves the imported region list under that folder. Any further imported region lists are then saved under the **Imported Region Lists** folder.

33.5.4 Annotate Region List

The **Annotate Region List** utility allows you to annotate the regions in a given region list (the **Source Region List**) with selected columns from another region list (the **Annotation**). You can also annotate region lists in the experiment using generic annotations downloaded from **Annotations Manager**. A **Source Region List** can be a region list from the Experiment Navigator. **Annotation** can be a region list or a continuous data object from the Experiment Navigator. The output region list resulting from this workflow is saved in the Experiment Navigator as a child list of the **Source Region List**.

1 Select **Annotate Region List** from the **Utilities** section of the Workflow Browser.

The **Annotate Region List** wizard-driven workflow opens.

2 In Step 1, choose the **Source Region List** and **Annotation** and specify the following criteria for annotation:

- **a** Annotation columns: Select the columns from **Annotation** that need to be added to the **Source Region List** as annotations.
- **b** Suffix for annotation columns: A string that you want **GeneSpring** to use as suffix in the names for the annotation columns added.
- **c** The criteria for annotating a region R1 from the **Source Region List** with a region R2 from the **Annotation** can be specified either as presence of overlap between R1 and R2 (select **R1 overlaps with R2** or as a maximum distance between R1 and R2 in base pairs(select **Maximum distance between R1 and R2** and specify the distance in base pairs). When the maximum distance option is selected, you can choose to annotate the **Source Region List** with only the closest R2s satisfying the criterion.
- **d** You can choose not to include rows with empty annotations in the output by selecting the corresponding option.

3 Click **Next >>**.

4 Step 2 displays the output regions with the annotations for your review.

5 Click **Finish** to save the annotated region list in the Experiment Navigator.

Each region from the **Source Region List** is then annotated with the selected columns from each of the **Annotation** regions that satisfy the specified criteria. The resulting region list contains for every region in the **Source Region List** as many rows as there are **Annotation** regions satisfying the criteria. In case there are no regions satisfying the criteria, then only one row is shown with empty cells in the annotation columns.

If you chose the option 'R1 overlaps with R2', the amount of overlap as a percentage of the length of R1, and as number of base pairs (bp) is included in the output

region list as two additional columns. This allows you to filter the annotated region list by the amount of overlap using the **Region List Operations** utility (see "Region List Operations" on page 1101).

If you chose the **Maximum distance between R1 and R2** option, the distance between the two regions is included as an additional column in the output region list. This distance is defined as zero if the regions overlap, otherwise it is 1 + the length of the gap between the two regions.

33.5.5 Compare Region Lists

This utility compares a Source Region List (RL_1) with another region list (RL_2) with regard to proximity of the regions in RL_1 with regions in RL_2. Both RL_1 and RL_2 can be a region list from the Experiment Navigator. You can also compare region lists in the experiment using generic annotations downloaded from **Annotations Manager**.

1. Select **Compare Region Lists** from the **Utilities** section of the Workflow Browser.
   The **Compare Two Region Lists** wizard-driven workflow opens.
2. In Step 1 of the workflow, choose the **Source Region List** and the region list to compare it to in the **Compare With** field. By default, **GeneSpring** selects the active region list in the Experiment Navigator for both lists. Click **Choose...** to change either one of the region lists. Click **OK**.
3. Still in Step 1, specify the criterion according to which you want to compare the two region lists:
   - **Source region overlaps with compared region**
     GeneSpring creates the following two region lists:
     - Regions overlapping with RL_2: This contains all the regions of RL_1 which overlap with at least one region from RL_2. For each such region of RL_1 **GeneSpring** also outputs the number of regions from RL_2 that overlap with it.
     - Regions not overlapping with RL_2: This contains all the regions of RL_1 which do not overlap with any of the regions from RL_2.
   - **Distance between source region and compared region <= X bp**
     Specify the maximum distance X (in bp) allowed for two regions to be considered as close to each other. For this option **GeneSpring** creates the following two region lists:
     - Regions close to RL_2: This contains all the regions of RL_1 which are within the specified distance from at least one region of RL_2. For each such region of RL_1 **GeneSpring** also outputs the number of regions from RL_2 that are within the specified distance.
     - Regions not close to RL_2: This contains all the regions of RL_1 which do not have any regions of RL_2 within the specified distance.
4. Click **Next >>**.
5. Step 2 displays the results in the form of three tabs: **Summary**, **Regions overlapping/close to R2**, **Regions not overlapping with/close to R2**. If the size of any of the input region lists is larger than the **Region List Size Threshold** specified in *Tools* $\longrightarrow$ *Options* $\longrightarrow$ *Miscellaneous* $\longrightarrow$ *Region List Size*, then only the summary of the comparison is displayed. Otherwise spreadsheet views of all the resulting region lists are available.

6 Click **Finish**.

GeneSpring saves the result region lists in the Experiment Navigator as child lists of the Source Region List.

33.5.6 Import Probe-level Reports

Both Agilent Cytogenomics and Agilent Genomics Workbench products have probe-level reports in the *.xls format as output. While **GeneSpring** does not support importing these reports for experiment creation and analysis, it is possible to import these reports for visualization in the Genome Browser alongside any previously conducted analyses in a CGH experiment. Use this procedure to import a probe-level import:

1 Select **Import Probe-level Report** from the **Utilities** section of the Workflow Browser.

The **Import Probe-level Report** dialog opens.

2 Click **Choose file(s)** to browse for the probe-based report files that are saved on your system. Select the desired file or multiple files and click **Open**. If the reports you would like to report are in several different locations, repeat this step until you have selected all the desired reports.

3 Review the selected report files.

4 To remove any unwanted reports, select the corresponding rows in the **Import Probe-level Report** dialog and click **Remove file(s)**.

5 Click **OK**.

GeneSpring saves the imported probe-level reports as region lists in a separate folder under the **All Entities** entity list called **Imported Probe-level Reports** and displays the region lists as a scatter plot in the Genome Browser view.

NOTE

If **GeneSpring** detects data for the same sample in multiple files selected for input, the data is merged and saved in **GeneSpring** as a single region list. For example, for particularly large reports, AGW splits the probe-level reports into two files, one file containing all the data up until a certain chromosome number and the second file containing the data associated with the remaining chromosomes. Such files are automatically merged again and reported as a single region list in **GeneSpring**.

Chapter 34

Copy Number Analysis

34.1 Introduction

This chapter introduces Copy Number, Allele Specific Copy Number, Parent Specific Copy Number, LOH, Log Ratio and Common Genomic Variant Regions, which are new features in **GeneSpring**.

It was generally thought that genes were almost always present in two copies in a genome, resulting in a 'Copy Number' of 2. But recent studies have proved that the human genome shows extensive Copy Number variation (CNV). In humans, CNVs encompass more DNA than single nucleotide polymorphisms (SNPs). Like other types of genetic variation, some CNVs have been associated with susceptibility or resistance to disease (Pathogenic). CNV analysis includes detection of regions whose Copy Number varies in comparison to a reference genome.

For each SNP allele, denoted by the letters A and B, the individual Copy Numbers of the two alleles are referred to as Allele Specific Copy Number (ASCN); their sum adds up to the total Copy Number. This can be assigned to the genome from the total Copy Number, signal intensities of the alleles and the genotype call. Another way to split up the total Copy Number into components is based on Copy Numbers in individual chromosomes inherited from each parent; these are called Parent Specific Copy Number (PSCN). This information is derived from the total Copy Number and the Allele Specific Copy Number. Allele Specific Copy Number and the Parent Specific Copy Number help in tracing the chromosomal amplifications/deletions at the allele level. Loss of Heterozygosity (LOH) refers to change from a state of heterozygosity in a normal genome to a homozygous state in a tumor genome. LOH can result from copy-loss events such as hemizygous deletions (Copy Number dependent), and also from copyneutral events such as chromosomal duplications (Copy Number independent).

Log Ratio is a simple ratio of the summarized intensities of the sample versus reference or normal as the case may be, in log (to the base 2) scale.

All the above computations follow either the **Against Reference** method or the **Paired Normal** method. The former method involves comparing specified arrays against a reference file created

from a set of reference individuals. The latter compares each specified array against a corresponding array obtained from a normal tissue sample of the same individual. The sections below explain in detail the various steps of Copy Number analysis which includes computation of Copy Number, ASCN, PSCN, LOH score and Log Ratios.

Determining **Common Genomic Variant Regions** is another feature in **GeneSpring**, which identifies regions of the genome that are significantly amplified or deleted across a set of samples. **GeneSpring** follows the procedure described in this reference on Genomic Identification of Significant Targets in Cancer (GISTIC) [22]. For complete details on implementation, see section Common Genomic Variant Regions.

GeneSpring supports genotyping analysis for Affymetrix and 'Illumina. In case of Illumina, outputs from GenomeStudio are supported.

34.1.1 Terminology in Copy Number analysis

Commonly found terms with respect to Copy Number analysis are explained in Table 34.1

34.2 Technologies supported by GeneSpring

GeneSpring supports Affymetrix CEL files and Illumina output files from GenomeStudio (extracted using a plug-in available for **GeneSpring**) for Copy Number analysis and the computation follows slightly different workflows for the different technologies listed below.

- **Affymetrix Genome-Wide Human SNP Array 6.0, Genome-wide Human SNP array 5.0, and Human Mapping 500K Array Set**
- **Affymetrix Human Mapping 100K Set**
- **Illumina Genotyping output files**

The primary aim of Copy Number analysis is to segment the genome according to the Copy Number. For Affymetrix technology, the following steps broadly constitute the Copy Number analysis.

34.2.1 Experiment Creation

GeneSpring calls Affymetrix Power Tools (APT) (`http://www.affymetrix.com/support/developer/powertools/changelog/apt-probeset-summarize.html`) for carrying out summarization of the CEL files using the PLIER algorithm. See Probe Summarization Algorithms for details on PLIER algorithm. Some of the PLIER parameters are :

Table 34.1: Terminology in Copy Number Analysis

Copy Number	Number of copies of a particular gene within the genome; Common Copy Number is 2, but human genome has been found to show extensive variation in Copy Number.
Copy Number Variants	A variant region is classified as Copy Number Variant (CNV) only if if it is of at least 1Kb in size. CNVs can be Copy Number gains (insertions and duplications) or losses (deletions) relative to the reference genome. CNVs can be de novo or inherited. Most variations are benign; assigning a variation to pathogenicity needs a genome wide association study
Copy Number Variants Region	Copy Number Variants Regions (CNVRs) represent merging of several overlapping, but independently ascertained CNV segments. CNVRs can have large number of potential Copy Number variants within it.
Copy Number Polymorphism	Copy Number Polymorphism (CNP) are subset of CNVs that segment at a frequency of one percent in the population. Large-scale Copy Number polymorphisms are common and widely distributed in the human genome.
Copy Number Confidence	CBS assigns a confidence value for the segments; For each segment with a mean log ratio value, a T test is done against the mean '0' (which indicates Copy Number 2) and a p value is obtained. Negative logarithm to the base 10 of this p value is reported as confidence.
Log Ratio	Log Ratio is the ratio of the summarized intensities of the sample versus the reference or the normal, as the case may be, on log to the base 2 scale.
Mean Log Ratio	For each Copy Number segment, CBS gives the 'Mean Log Ratio'. This value can be viewed in genome browser after running Copy Number analysis.
Copy Neutral LOH	Loss in Heterozygosity (LOH) that results from events which do not involve change in Copy Number; for instance, chromosomal duplication. The LOH score is between 0 and 1; Score over 0.5 is considered to be LOH by default; but the value is configurable.
Homozygous, Heterozygous and Hemizygous condition	If both alleles in a gene are same, it is homozygous; if they are different, it is heterozygous; if one is missing, it is hemizygous
Against Reference Analysis	Compares specified arrays against a reference file created from a set of reference individuals
Paired Normal	Compares each specified array against a corresponding array obtained from a normal tissue sample of the same individual
Allele Specific Copy Number	Assigning allele calls to the SNP using Fawkes Algorithm
Parent specific Copy Number	**GeneSpring** assigns a split in the total Copy Number to infer the parent specific Copy Number, from the Allele Specific Copy Number. Refer to section PSCN for details.
p value in assigning Allele Specific Copy Numbers	It is the p value of the Fawkes result; check p values of fawkes result for details. The lower the value is, the better the confidence in prediction is.
GISTIC	Identifies common genomic variant regions ; commonly called Genomic Identification of Significant Targets in Cancer; see Common Genomic Variant Regions.

- For Quantile Normalization, default of 1 percent or 50,000, whichever is higher, is used.
- Sets all chips median to 1000.
- Simplified Expression Analysis (SEA) optimization method is used.
- Other parameters are set to default in APT.

For Affymetrix Mapping 100k array set, BRLMM algorithm is also run during experiment creation and the outputs are stored. Figure 34.1 explains experiment creation for all Affymetrix technologies.

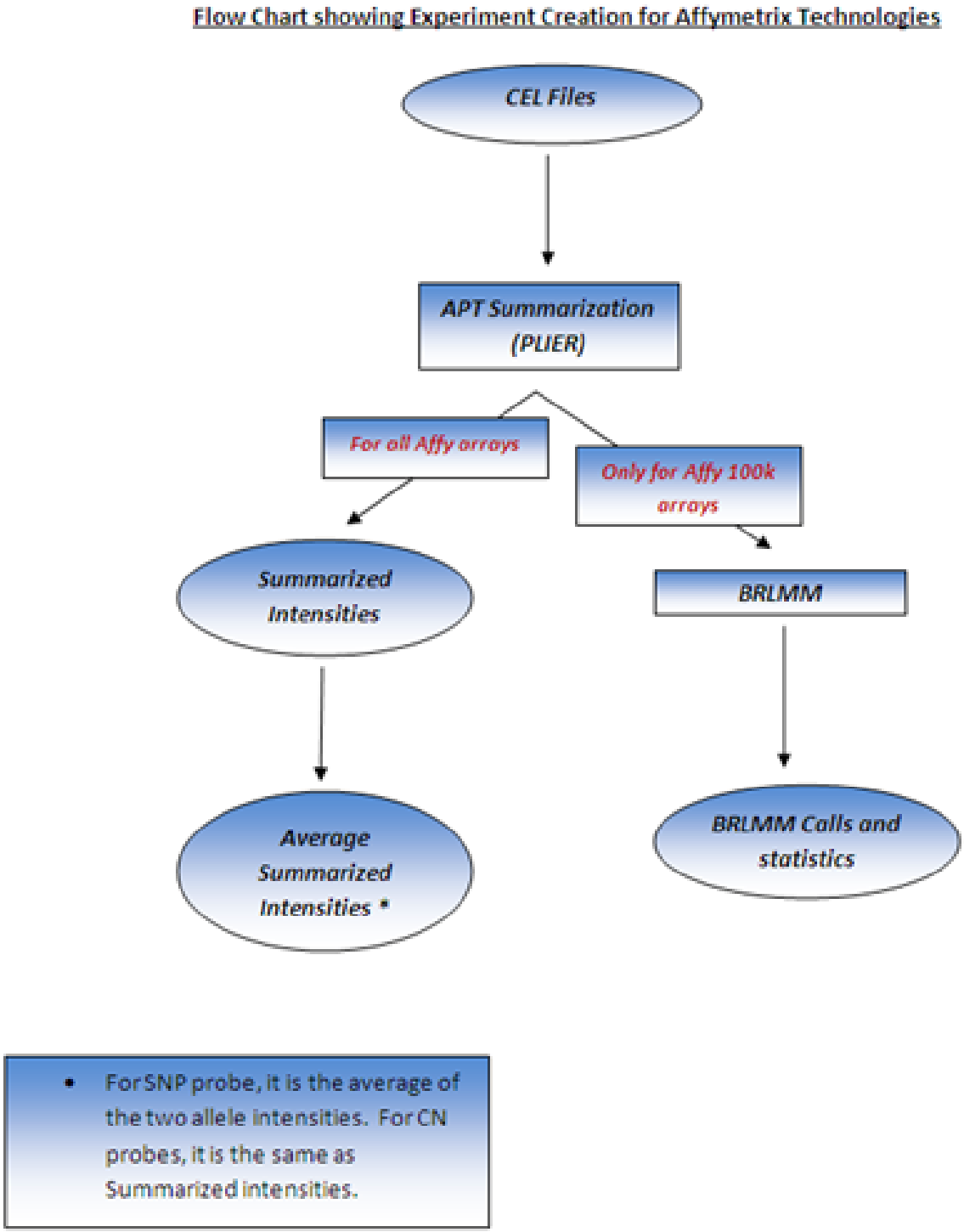


Figure 34.1: Experiment Creation for Affy CEL files

34.2.2 Reference

During experiment creation, there is an option to choose the standard reference prepackaged within the tool, or to use a custom reference created by the user.

GeneSpring comes prepackaged with a reference created using the Phase II 270 HapMap samples, [26], the second generation human haplotype map from The International HapMap Consortium. According to the reference cited above, this Phase II HapMap characterizes over 3.1 million human single nucleotide polymorphisms (SNPs) genotyped in 270 individuals from four geographically diverse populations and includes 25 to 35 percent of common SNP variation in the populations surveyed.

If the user has reasons to believe that the 270 HapMap samples of the standard reference will not be a good reference for his data, and if he has access to data that is better suited for the comparision, then he can create a custom reference with those samples. Note that the samples used for custom reference should belong to one of the standard technologies supported by the product. The reference creation procedure is the same for both standard and custom references and is shown in this figure 34.2 for the Affymetrix Genome-Wide Human SNP Array 6.0, the Genome-wide Human SNP array 5.0, and the Human Mapping 500K Array Set.

Handling sex chromosomes during reference creation: The reference could be a mix of female and male samples and female samples will have a higher signal for X chromosome compared to male samples. To take care of this, signals for chromosome X from male samples are scaled by a factor of 1.4 for for Affymetrix Mapping 100k array set (Hind, Xba and paired). With this scaling, the reference effectively becomes an 'all female' sample set.

For all the other arrays (250k, 500k, V5, V6), the average signals are replaced by the median of the copy number 2 clusters obtained from Birdseed algorithm. This effectively accounts for the scaling and weighting required for male samples. Birdseed algorithm does not run on Affymetrix Mapping 100k array set and hence the explicit scaling for this array set.

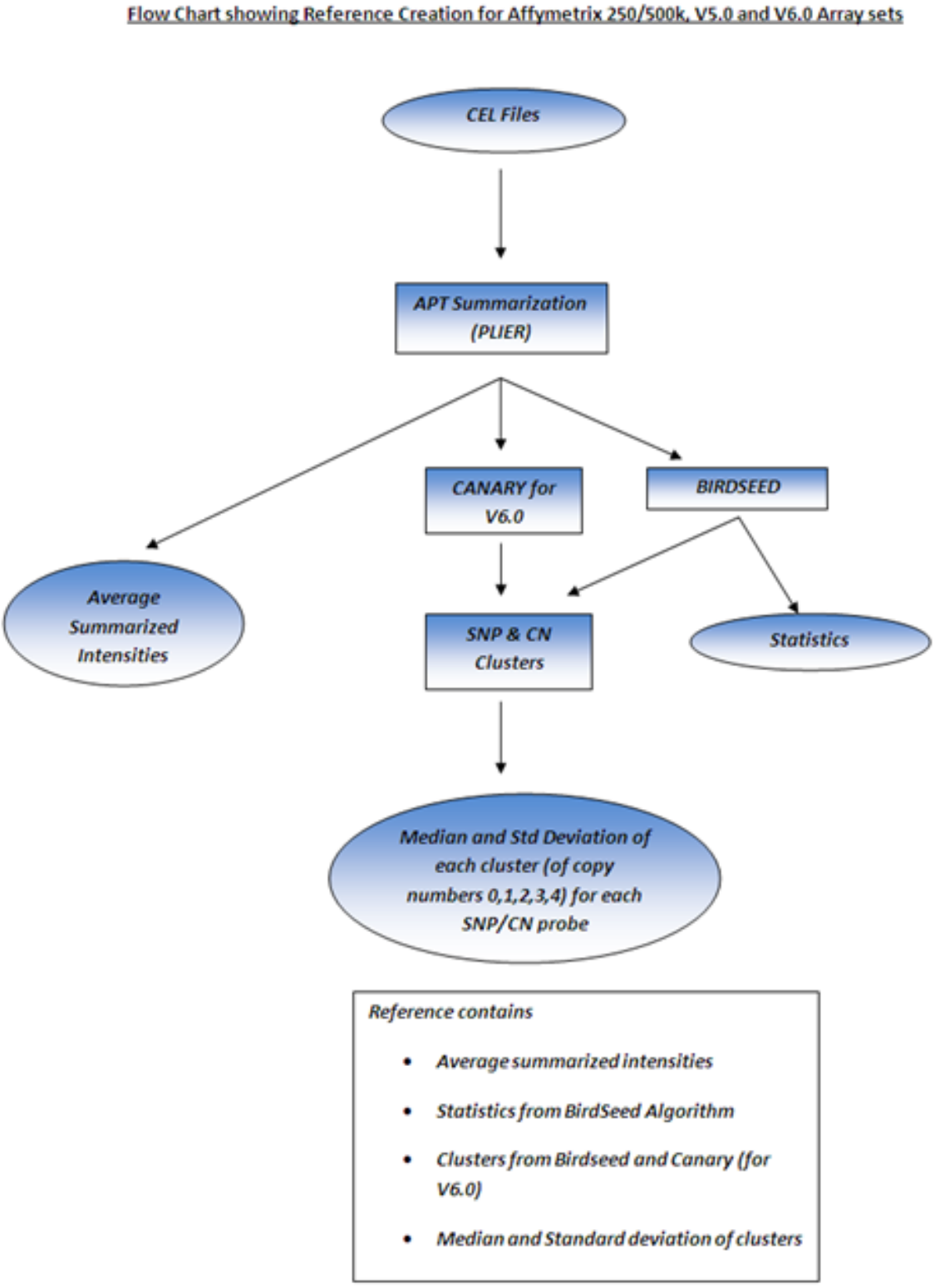


Figure 34.2: Affymetrix Genome-Wide Human SNP Array 6.0, Genome-wide Human SNP array 5.0, and Human Mapping 500K Array Set - Reference Creation

For the Affymetrix Mapping 100k array set, BRLMM is run in place of Birdseed (and Canary) and the procedure is shown in Figure 34.3.

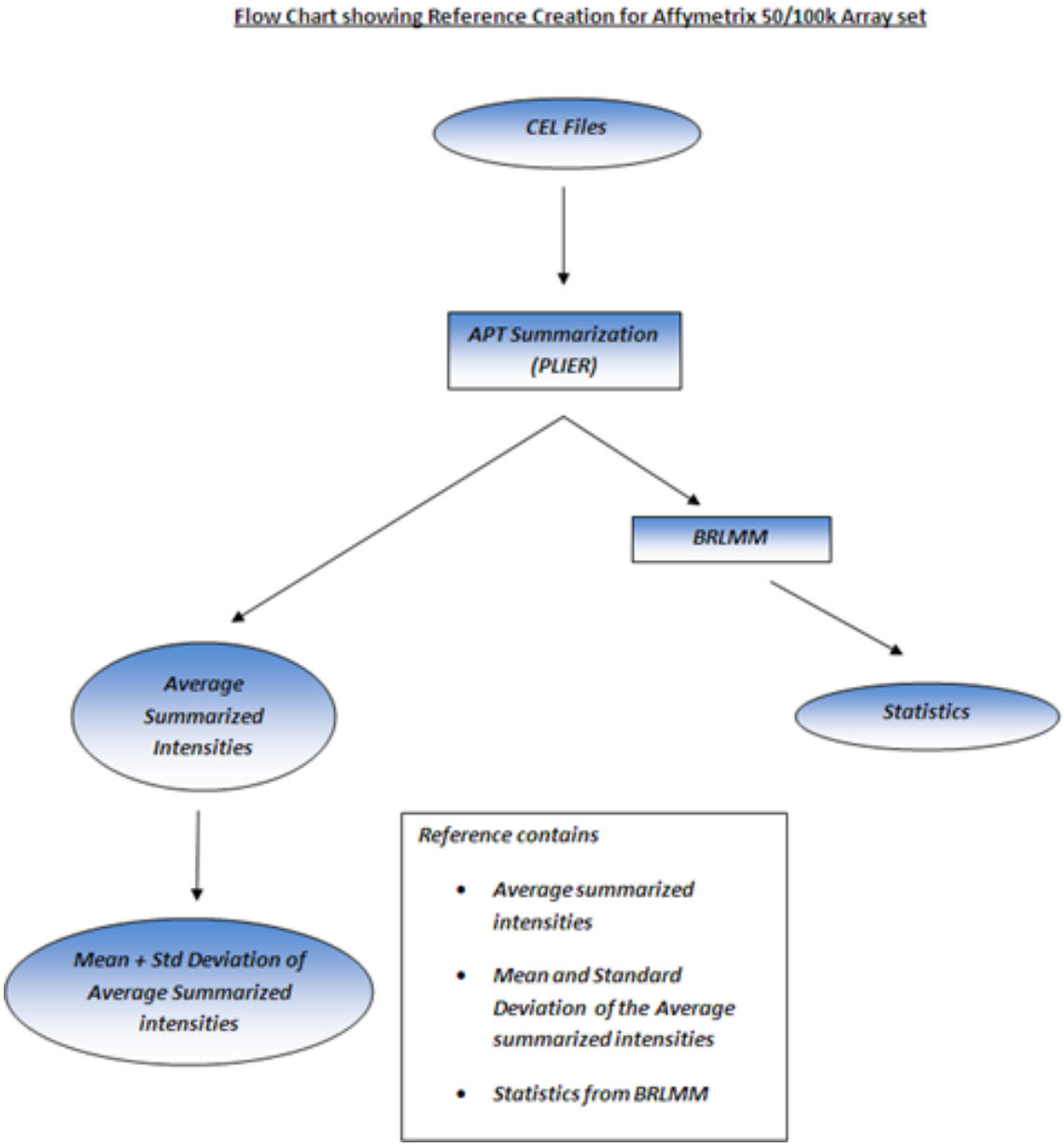


Figure 34.3: Reference Creation for Affy 100K array set

34.2.3 Copy Number Analysis

Once an experiment has been created, Copy Number analysis can be performed. This step uses the following information:

- The Summarized dataset.
- The chosen Reference.
- Interpretation for deducing if it is Against Reference' analysis or 'Paired Normal' analysis. See section Significance of Interpretation in Copy Number analysis for details.

From the above inputs, Log Ratios, Copy Numbers, LOH scores, Allele Specific Copy Numbers and Parent Specific Copy Numbers are calculated.

Log ratios:
- **Against Reference:** The log ratio is the ratio of sample probeset intensity to reference probeset intensity (i.e., the average of the probeset intensities over all the arrays in the reference), transformed by logarithm to base 2. For SNP probes, the intensity used is the average of the individual A and B intensities. By way of normalization, the median of the sample probeset intensities is scaled to the median of the reference probeset intensities before log ratios are computed.
- **Paired Normal:** For disease/tumor samples, the log ratio is the ratio of tumor probeset intensity to the normal probeset intensity, transformed by logarithm to base 2. For SNP probes, the intensity used is the average of the A and B intensities. Log ratios are then median shifted to 0 for each disease/tumor sample. For normal samples, the log ratio is generated using the Against Reference computation above.

Copy Number: Log Ratios are fed to a CBS segmentation routine for segmentation into regions of roughly equal log ratios. For each resulting segment, a confidence value (-log10(pvalue)) is computed; higher confidences indicate segments in which the mean log ratios are well away from 0 and the variance in the log ratios is small. For each segment, we also compute a copy number value, using a calibration method that associates the mean log ratios in that segment with a copy number value. The copy nunber values output are continuous values up to the first decimal digit and range from 0 to 4.

See sections Circular Binary Segmentation algorithm (CBS) for segmenting the genome on the basis of Copy Number. CBS outputs segment start and end points along with the Mean Log Ratios for each segment. Copy Numbers are assigned to segments based on a mapping explained in section Post Processing to assign Copy Numbers to segments from CBS.

Allele Specific Copy Number: For SNP probesets, the Total Copy Number is split into its A and B allele specific components using the Fawkes algorithm. This computation is not performed for the Affymetrix Mapping 100K technology. The total copy number is rounded off to the nearest integer before being fed to Fawkes and the ASCN's output by Fawkes are always integral.

Parent Specific Copy Numbers (PSCN): Total Copy Numbers are split into Parent Specific Copy Numbers, i.e., individual copy numbers for each of the 2 chromosomes. The intuition behind this computation is as follows. Consider a copy number segment with say copy number 3, and consider all SNPs in this segment. If all of these have ASCN A=2,B=1, we can conclude that one of the chromosomes has 2 copies and the other has 1 copy. On the other hand, if all of these have ASCN A=3,B=0, then it is not possible to unambiguously determine the split into PSCNs (all 3 A's could be on one chromosome or 2 A's on one and 1 A on the other). So PSCN inference is based on heterozygous SNPs and nearby homozygous SNPs inherit PSCN's inferred from hereozygous SNPs. Long stretches of homozygous SNPs lead to inconclusive PSCNs and are given missing values. The stretch to be called as homozygous can be set from the configuration options from the 'Tool' menu (default value is 5). For example, if there are 3 probe-sets with 2,0 ASCN split and the neighbouring ASCN values

are 1,1 then the entire region is assigned a 'diff' value of 1 (even for the 2,0 region where the diff is expected to be 2).

Inputs are the total Copy Number values for segments and the Allele Specific Copy Number. A 'diff' which is the difference between the ASCN values is assigned; the 'diff' values are a measure of 'Allelic Imbalance'. Along with the 'diff' value, the minimum (min) and maximum (max) values of the ASCN split can be viewed in the genome browser.

- For Copy Number 2, the PSCN possibilities are 2,0 (homozygous) and 1,1. For this segment of total Copy Number 2, **GeneSpring** detects and displays segments with predominantly 1,1 split by ignoring the homozygous points in between.
- For total Copy Number 3, the possibilities are 3,0 and 2,1. Homozygous regions (3,0) are ignored and the regions with predominantly 2,1 split are displayed with minimum and maximum.
- For total Copy Number 4, the possibilities are 4,0 and 3,1 and 2,2. Homozygous regions (4,0) are ignored; in the remaining regions, the homozygous points are again removed and CBS is run on this. The homozygous points that are ignored should be atleast 5 SNPs long, by default; this value is configurable. On the segments that are returned by CBS (which include blanks for homozygous points), we get the median and populate the mim and max using this value.

LOH Scores HMM Implementation computes LOH scores. The inputs are BRLMM calls in case of Human Mapping 100K Set and Fawkes outputs in case of Genome-Wide Human SNP Array 6.0, Genome-Wide Human SNP Array 5.0 and Human Mapping 500K Array Set.

The possible Fawkes states and their mappings are presented in the table 34.2

Fawkes State	**Mapping for LOH**
A0B0	No Call
A0B1	BB
A1B0	AA
A0B2	BB
A2B0	AA
A1B1	AB
A0B3	BB
A1B2	AB
A2B1	AB
A3B0	AA
A0B4	BB
A1B3	AB
A2B2	AB
A3B1	AB
A4B0	AA

Table 34.2: Mapping Fawkes state to LOH

- **Against Reference:** For computing LOH scores for all samples, HMM (Against Reference version) is run using the genotype calls of the samples generated using Fawkes and some statistics about the the standard reference genotype calls dataset saved in the standard reference packaged with the technology.

  In human genome, the normal level of homozygosity is around 70 %.When doing an 'Against Reference' analysis, any region with more than 70 % homozygosity can be safely

assumed to be showing LOH. This combined with some meaningful filter conditions under 'Identify Copy Neutral LOH' can reduce false positives and remove small pockets of high homozygosity while picking up LOH regions.

- **Paired Normal:**
  - In Fawkes, the calls for Normal samples are set to 2 before sending to LOH computation.
  - The Strand LOH HMM (Paired Normal version) is run to calculate the LOH scores for the tumor samples
  - The Strand LOH HMM (Against Reference version) is run to calculate the LOH scores for the normal samples

34.2.4 Special mention for Affymetrix Mapping 100k Array

- BRLMM needs at least 6 distinct CEL file pairs or at least 6 distinct CEL files. Hence for Affymetrix Mapping 100k Array set, if the number of file pairs (or files) is fewer than 6, experiment creation is aborted.
- Allele Specific Copy Number and hence Parent Specific Copy Number cannot be generated for Affymetrix Mapping 100k Array.
- Post CBS, SNP/CN clusters from reference are used for the mapping to assign Copy Numbers to segment. For Affymetrix Mapping 100k array set, since there are no CN clusters (only BRLMM runs on this array), clusters generated from Affymetrix Genome-Wide Human SNP Array 6.0 technology of the reference are used here.

The workflow and utilities common across all the technologies are described in the sections Workflow description for Affymetrix Technology and Workflow description for Illumina. The algorithms used for various computations are explained in the section Copy Number Algorithms.

34.3 Workflow description for Affymetrix files

34.3.1 Create Technology

Before starting Copy Number analysis, ensure that the required technology is downloaded. Go to the menu **Annotations → Create Technology → From Agilent Server**. From the list of available updates, look for those with the type 'Affymetrix GenoChip' along with the name of the exact technology. Technology downloads are available for Affymetrix Genome-Wide Human SNP Array 6.0, Genome-wide Human SNP array 5.0, and Human Mapping 500K Array Set (individual and paired arrays) and Affymetrix Human Mapping 100K Set (individual and paired arrays). The download size is listed along side and the tool gives an estimate of the actual space required before starting the download. See Figure 34.4.

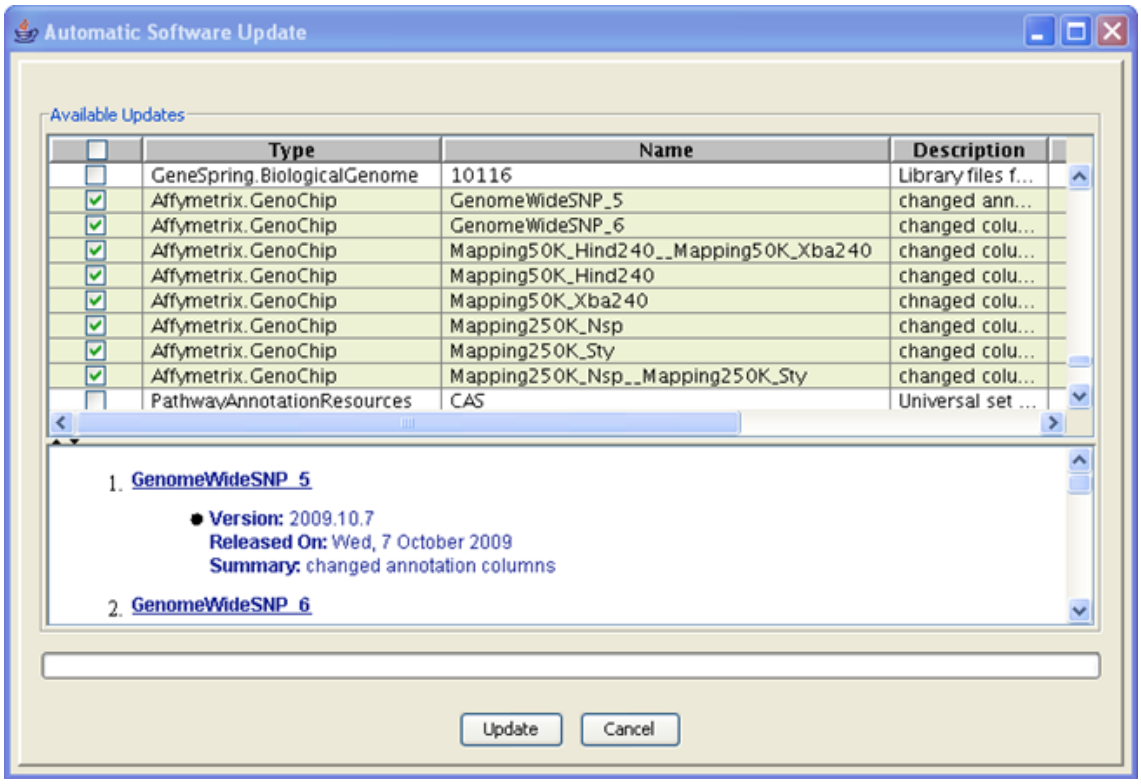


Figure 34.4: Create Technology for Copy Number Analysis - Affymetrix technology

Note:Appropriate technology should be created before creating custom reference.

34.3.2 Creating a Copy Number experiment

To start the Copy Number analysis, an experiment has to be created with the Affymetrix CEL files in **GeneSpring** . This can be done by creating a new experiment within any existing project or by creating a new experiment in a new project.

1. **Creating new experiment in existing project:**
   Go to menu *Project* $\longrightarrow$ *New Experiment*. Give a name for the experiment, choose the analysis type as *Copy Number* and the experiment type as *Affymetrix Copy Number*. Note that for Copy Number analysis, only the Data Import Wizard workflow is supported. You can define experiment notes if any and click 'OK'. More details on creating new experiment are described below.

2. **Creating new experiment in new project:**
   Alternately, close the existing project and go to menu *Project* $\longrightarrow$ *New Project*. A window comes up prompting user to define a name for the project, as well as any notes. On clicking 'OK' here, an *Experiment Selection dialog* window appears with two options:

- **Create new experiment**: This allows the user to create a new experiment (steps described below).
- **Open existing experiment**: This allows the user to use existing experiments from previous projects for further analysis.

Clicking on **Create new experiment** opens up a 'New Experiment' dialog in which 'Experiment name' can be assigned. Under analysis type, choose *Copy Number* from the drop down menu and then select the experiment type as *Affymetrix Copy Number*. Note that in Copy Number analysis, only *Data Import Wizard analyis* is supported. Additional notes can be defined by the user in this window. Click 'OK' to begin the experiment creation.

A 3 step wizard opens up for creating a new Affymetrix Copy Number Experiment.

Step 1 of 3 - Load Data: This step allows loading of files/samples. *Choose Files* brings up a file chooser to choose 'CEL' files. *Choose Samples* brings up the sample search wizard. Refer Search for details on how to perform search. Figure 34.5 shows 'Load Data' step.

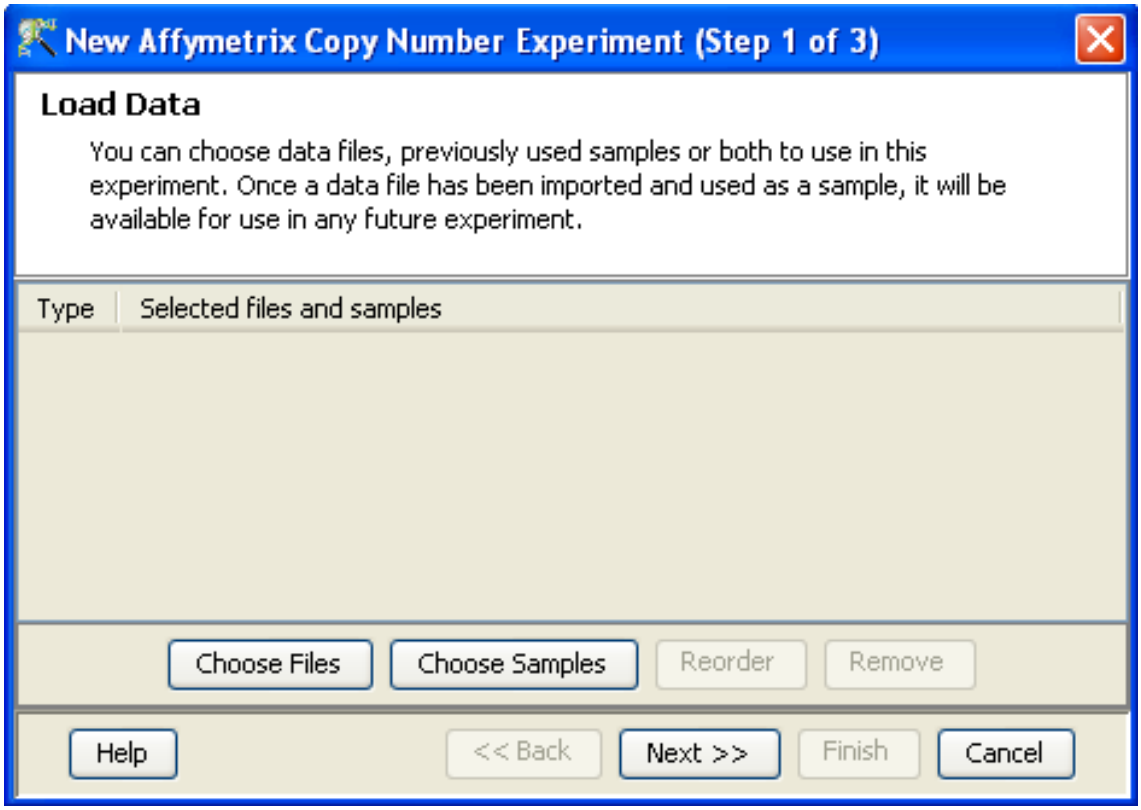


Figure 34.5: Step 1: Load Data

Once done, click *Next* to proceed.

Step 2 of 3 - Pair CEL files: This step provides the file pairing interface, and is shown only when CEL files from Hind and Xba arrays are input together. See Figure 34.6.

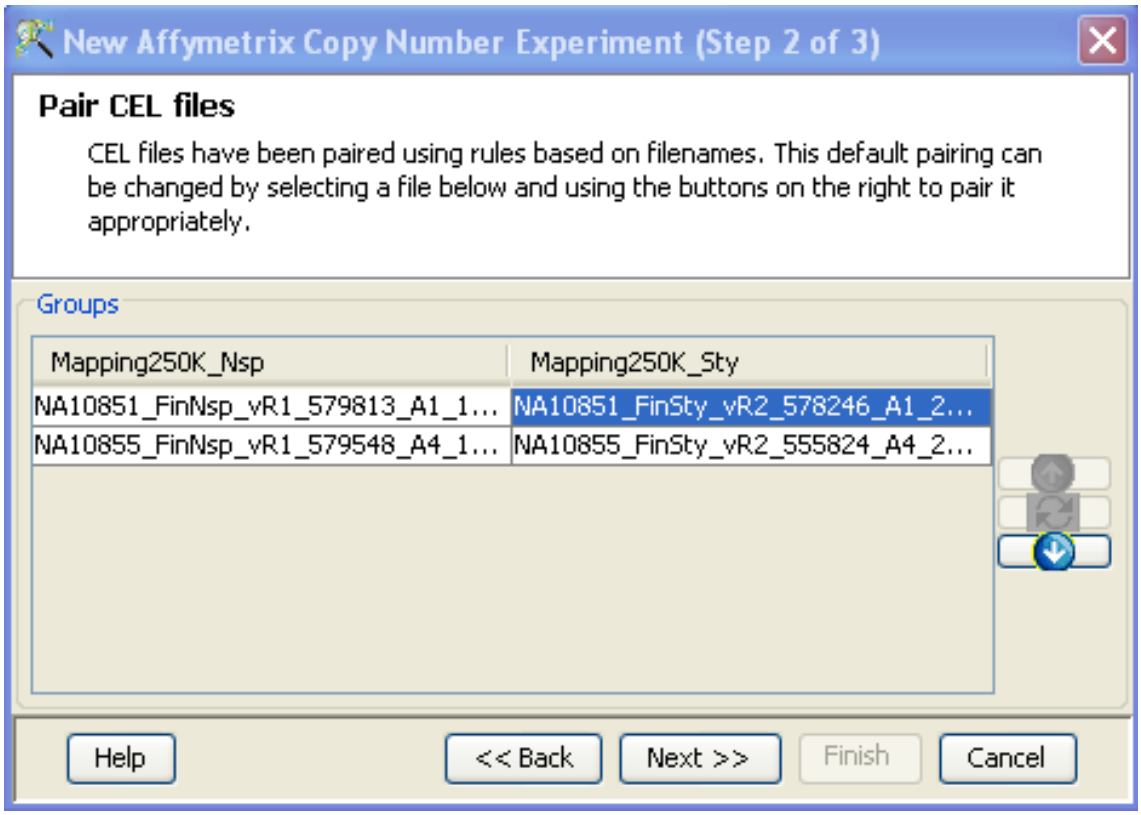


Figure 34.6: Step 2: Pair CEL files

This step is skipped if CEL files from *Genome wide Human SNP array 6.0* or *Genome wide Human SNP array 5.0* are used; it is also skipped when Hind and Xba files are separetely input.

Both the 100K arrays and the 500K arrays comprise of two arrays of half the size each (the 100K arrays have Xba and Hind arrays of size 50K each and the 500K arrays have NSP and STY arrays of size 250K each). **GeneSpring** will attempt to automatically pair up the arrays based on naming rules. You can change the pairing or the order, by clicking on the appropriate cell and clicking on appropriate arrows.

NOTE: The order of the columns in the dataset will be the same as the order in which they occur in the selection interface. If you want the columns in the dataset to be in any specific order, you should order them here appropriately.

Step 3 of 3 - Choose Copy Number / LOH Analysis Type: This step allows the user to define options for carrying out Copy Number computation. **Choose Reference Type** option gives a choice to select pairing against standard reference or custom reference. See section Technologies supported in **GeneSpring** for details.

Figure 34.7 illustrates Step 3.

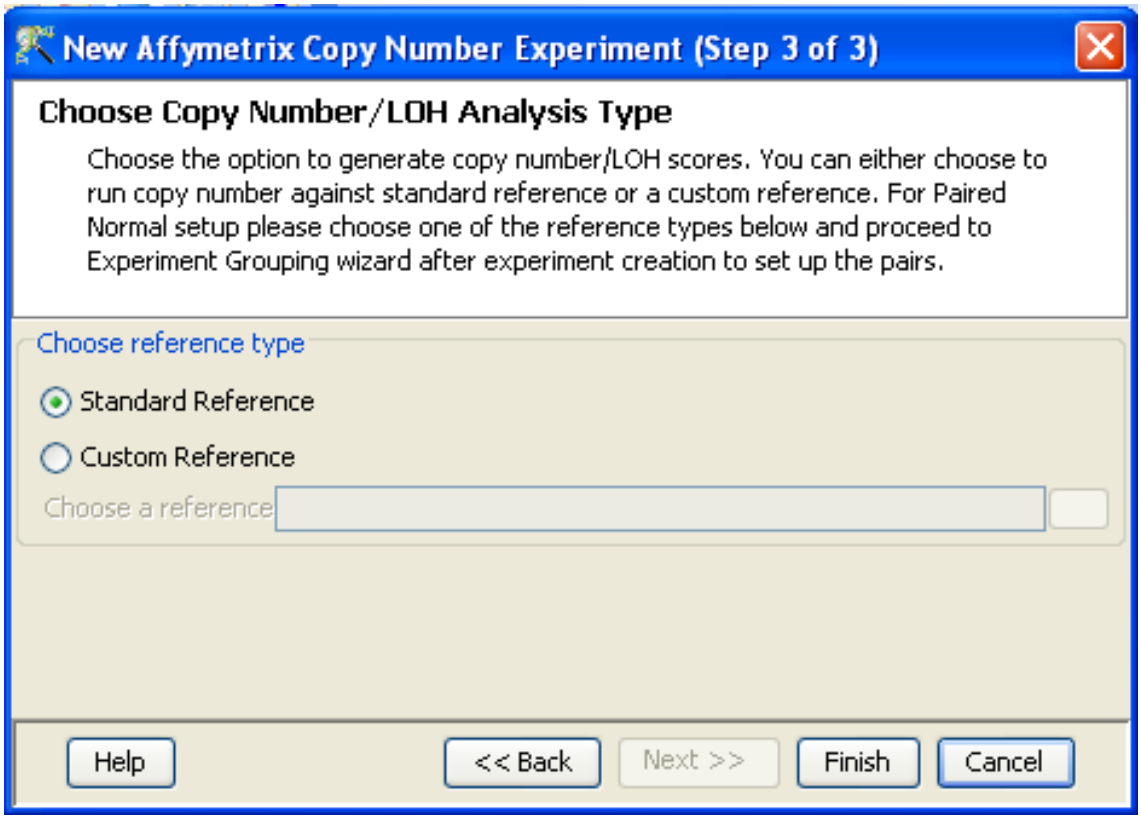


Figure 34.7: Step 3: Choose Copy Number/LOH Analysis Type

See section Custom Reference Creation for details on creating custom reference. Once custom references are created using the precedure explained here, user can pick them for using here. Choose the required one and click *Finish* to exit the experiment creation wizard.

Technical details on what happens during experiment creation are explained in section Experiment Creation.

34.3.3 Experiment Setup

Quick Start Guide: Clicking on this link will take you to the appropriate chapter in the on-line manual of **GeneSpring** .

Experiment Grouping: Experiment parameters defines the grouping or the replicate structure of the experiment. For details refer to the section on Experiment Grouping. Please note that the 'Parameter Type' during Experiment grouping is used for profile plots. In CN analysis we do not have profile plots. Hence, the 'Parameter Type'(as appears during experimental grouping) is not relevant to the Copy Number Analysis experiments.

Create Interpretation: An interpretation specifies how the samples should be grouped into experimental conditions both for visualization purposes and for analysis. For instance, if

you have different types of tumours and you want view these in batches in the Genome Browser, you can create appropriate interpretations. For details refer to the section on Create Interpretation. See section Significance of Interpretation in Copy Number analysis on how to define the interpretation to run Paired normal analysis.

34.3.4 Quality Control

34.3.4.1 Quality control on samples

Quality Control or the Sample QC lets the user decide which samples are ambiguous and which are passing the quality criteria. Based upon the QC results, the unreliable samples can be removed from the analysis. The QC view shows three tiled windows:

- 3D PCA scores
- 2D PCA scores, Experiment Grouping and Eigen values
- Legend

The *Add/Remove Samples* button allows the user to remove the unsatisfactory samples and to add the samples back if required. Note that the summarization is not re-done upon sample removal.

The QC view can be launched by using the workflow handle **Quality Control → Quality Control on Samples**. See Figure 34.8.

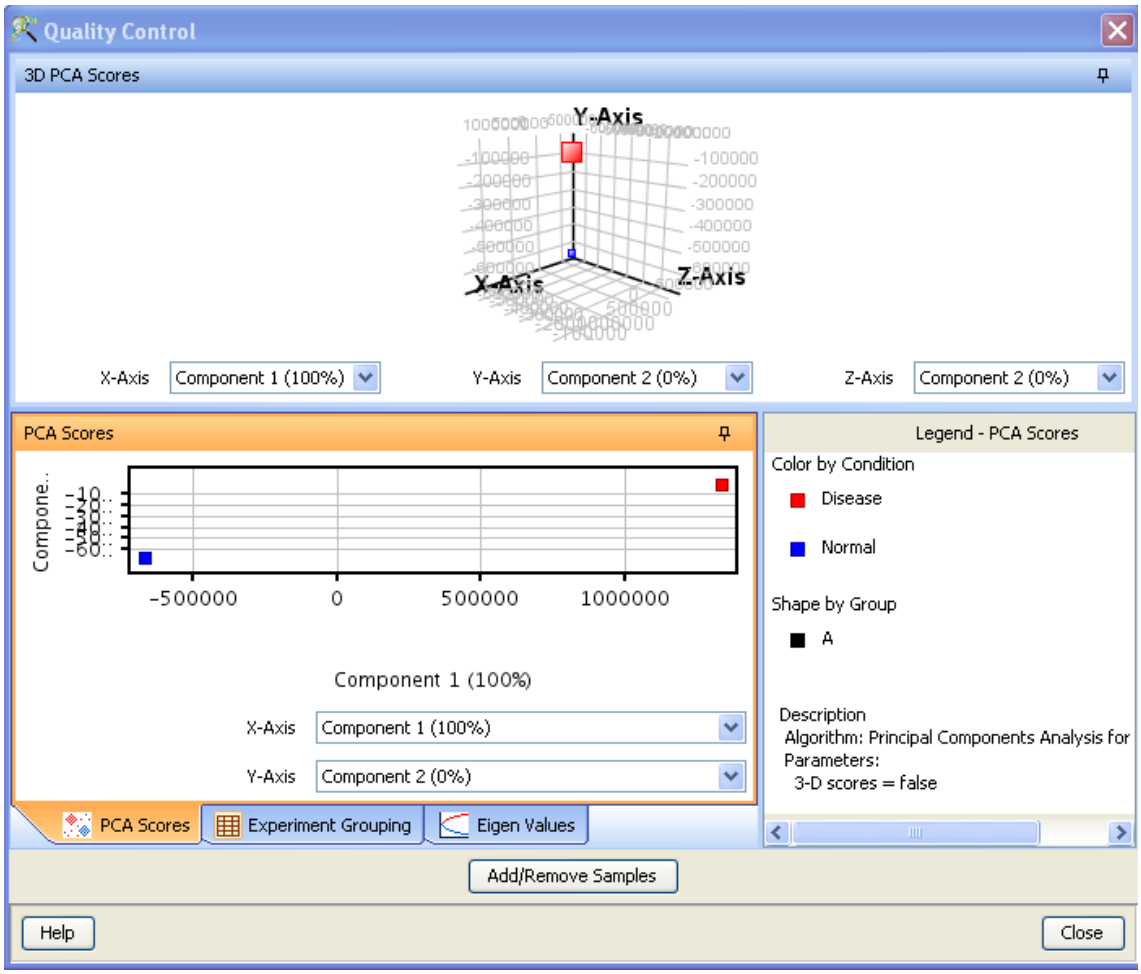


Figure 34.8: QC views for Copy Number Experiment

34.3.4.2 Batch Effect Correction

Batch effect is an example of systematic error which arises when different samples are run under slightly different conditions, for instance on different days. This step identifies and corrects markers that show consistent signal within batches but large variations between batches.

Batches are defined by conditions in the selected interpretation. Batches below a certain threshold number of samples are ignored in this procedure. Each remaining batch is T-tested against a pool of all the remaining batches. An entity is flagged for correction if it obtains a p-value below the specified threshold in the specified fraction of batches.

Process:

1. For every marker that has been marked for correction in a batch, b1, calculate the mean of the values for that marker in all the samples in all the batches that are NOT marked for

correction. Let us call this ref-mean.

2. Calculate the mean of the values for that marker in all the samples in b1 only. Let us call this b1-mean.
3. Get the difference of the two means, (ref-mean - b1-mean); let us call this difference as mean-scale.
4. Add this mean-scale to every sample of b1 for that marker to make b1-mean equal to the ref-mean.

Table 34.3 explains the rest of the procedure for each marker (or probeset) and for every batch.

Points to Note:

- Un-averaged condition information is used, even if the user has checked the option to 'Average over replicates in condition' during **Create Interpretation**.
- Batch effect correction runs on summarized data set only.

Performing batch effect correction in GeneSpring

Start the batch effect correction by using the workflow handle **Quality Control** → **Batch Effect Correction**. A two-step wizard opens up.

Step 1 of 2 : Batch Effect Correction

Apart from selecting the interpretation, the user can feed the other parameters for batch correction in this step. See Figure 34.9.

Table 34.3: Batch Effect Correction

T test	Distribution of samples in a batch against the distribution of the rest of the samples in other batches. Default parameters of 'not against zero, equal variance, un-paired, do not remove outliers' for T test.
P Value	For a batch, the number of p values obtained are equal to the number of markers.
Corrected P Value	Apply Benjamini Hochberg correction to correct for False Discovery Rate (FDR). This yields the corrected p value or q value.
Choosing for correction	For a marker, if the corrected p value is ≤ 0.001 or the specified cut-off, then the batch in which it is present is a candidate for batch correction.
No Correction done if	1. If the batch has less than the defined *Minimum Samples per batch* 2. If the total number of batches requiring correction is less than *'Percentage of bad batches allowed'*, then none of the batches are subject to correction.
Bad and Good batches	For every marker, get the number of batches where it needs correction ('bad' batches') and the number where it does not need correction (after discounting as per the above rules).
Reference Batch	• If, for a marker, all the batches require correction, then the batch with the least number of bad markers is used as the reference batch for correction. In other words, the batch with minimum number of markers with p values ≤ 0.001 or the specified cut-off, is chosen as the reference batch. • In case of only two batches or if the p-value cutoff is chosen as 1, the batch coming alphabetically first will always be chosen as the reference batch, and all the other batches will be corrected taking this batch as the reference. • In case a particular probe needs correction in all the batches, then the first batch among the input batches (sorted alphabetically) with minimum number of corrections will be chosen as the reference batch.

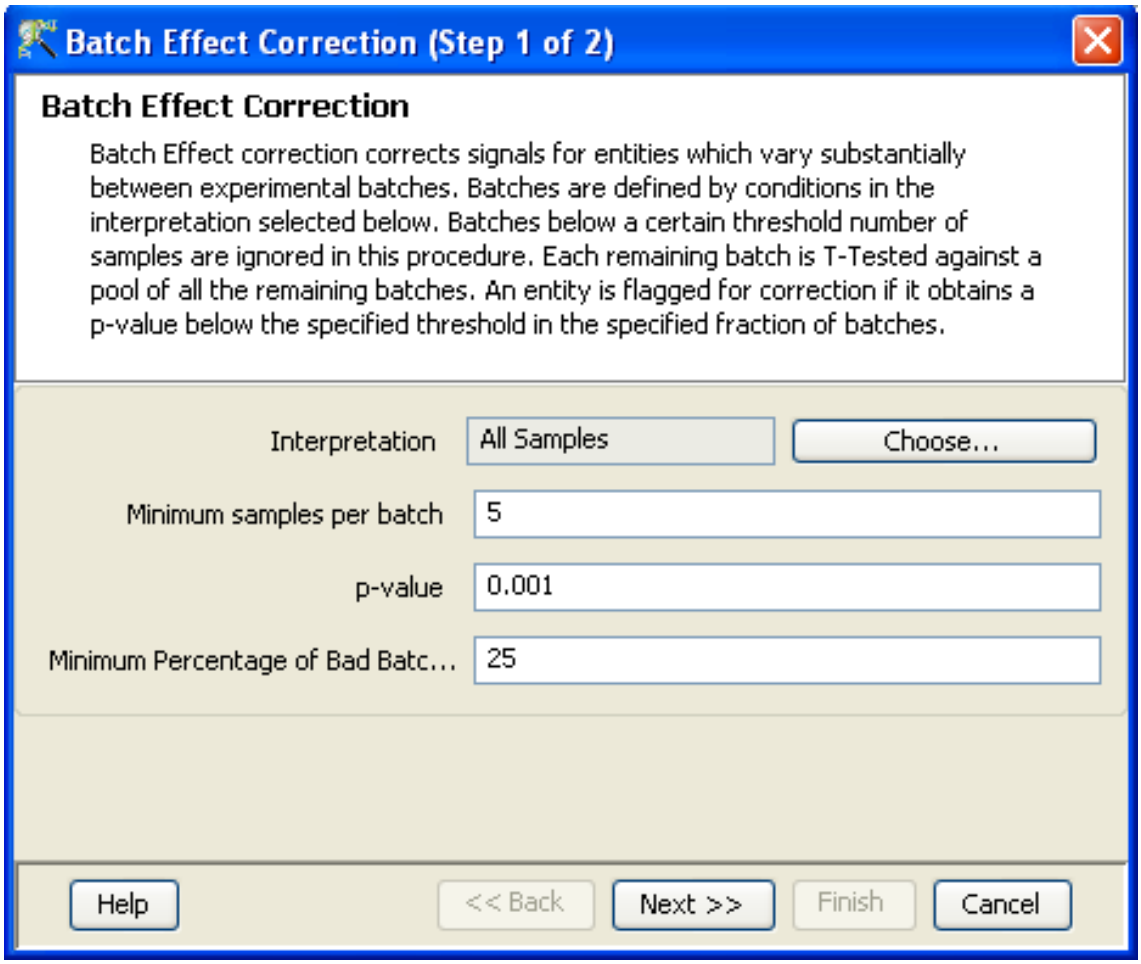


Figure 34.9: Batch Effect Correction - Step 1

The parameters required for batch effect correction are

- *Minimum Samples per batch* Default value is 5; user can change this number, based on the data in the experiment.
- *P value* Default value is 0.001; user can change based on requirement.
- *Percentage of bad Batches Allowed* Default value is 25 percent.

On clicking *Next*, the batch effect correction procedure starts; it proceeds probeset by probeset. If at the end there are no batches that qualify for correction, based on the input parameters, a message is shown that "None of the features was corrected. Try increasing p-value". At this stage, the user can change the parameters in step 1 and try again, or cancel the batch effect correction step.

Step 2 of 2 : Output Views

The result of the batch effect correction is displayed in Spreadsheet, Summary Statistics and Histogram views and is illustrated in Figure 34.10.

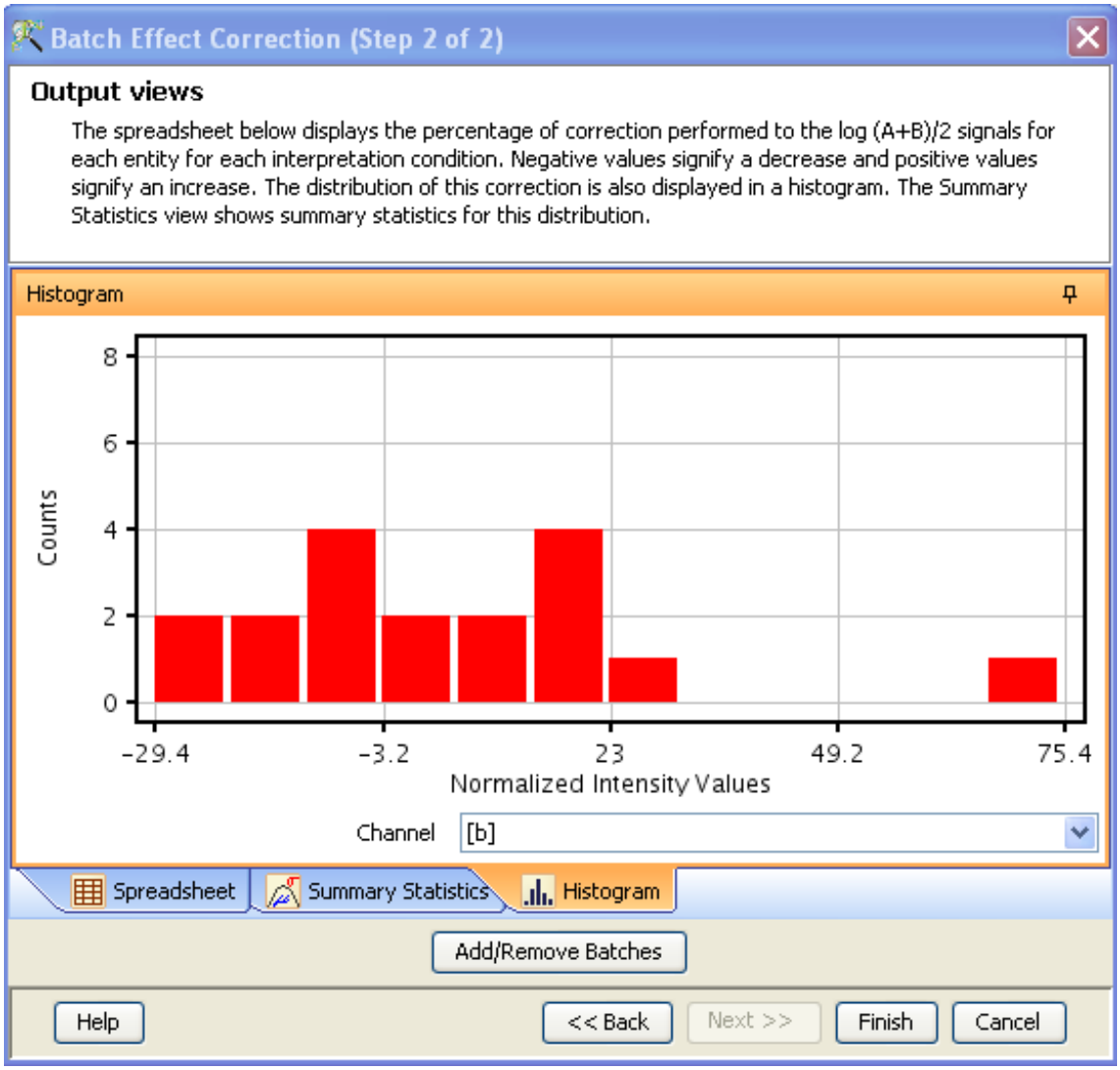


Figure 34.10: Batch Effect Correction - Step 2

- **Spreadsheet:** Displays the percentage of correction performed to the log(A+B)/2 signals for each entity for each interpretation condition. Negative values signify a decrease and positive values signify an increase. Note that a value of '0.0' indicates 'NO BATCH EFFECT CORRECTION'.
- **Summary Statistics:** For each of the batch, summary statistics are shown here.
- **Histogram:** Distribution of correction is displayed in histogram.

Clicking on the *Add/Remove Batches* brings up a window where batches can be pushed between the two options - 'Batches not to be considered' and 'Batches to Keep'. This is an explicit handle to remove some batches from batch correction procedure, after viewing the results.

Click *Finish* to complete the batch effect correction process. Note that batch effect correction will overwrite any previous results after showing a warning message.

34.3.5 Copy Number Analysis

Copy Number analysis includes computation of Log Ratios, Copy Number, Allele Specific Copy Number, Parent Specific Copy Number and LOH scores and is applicable only for Affymetrix technologies. See section Technologies supported by **GeneSpring** for details on how the computations are done for various technologies. Copy Number analysis can go through the 'Against Reference' method or the 'Paired Normal' method and this information is deduced from the interpretation.

34.3.5.1 Inferring method of analysis from Interpretation in Copy Number Analysis

If the user has normal and tumour samples obtained from the same individual, then he will want to run a 'Paired Normal' method of analysis where the tumour samples are compared against the normal samples. If not, the user can decide to run the 'Against Reference' method of analysis where the samples are compared against the standard or custom reference.

To run 'Paired Normal Analysis':

- Create experiment grouping with the parameters 'Condition' and 'Group'
- Under 'Condition' assign samples to be considered as normal as 'Normal'; assign the rest of the samples with any term other than 'Normal'.
- Under 'Group' give identifiers for each group ensuring each group contains at least and at the most, one 'Normal' sample.
- Create an Interpretation with both 'Condition' and 'Group' as experiment parameters and run 'Copy Number Analysis' from this interpretation.

'Against Reference Analysis' will be run if:

- Chosen Interpretation does not contain 'Condition' and 'Group' as parameters.
- Even if the chosen Interpretation has 'Condition' and 'Group' as parameters, if there are no samples assigned as 'Normal' (case sensitive).

NOTE: Note that you can create as many interpretations as you want with varying parameters and choose to run Copy Number analysis on them. Please note that Copy Number Analysis always runs on all samples of the experiment irrespective of the interpretation chosen. Choosing an interpretation in Copy Number Analysis helps to run either a Paired Normal or an Unpaired analysis. However, downstream analysis which use copy number results run on specific interpretation chosen. Make sure that the analysis is run from that particular interpretation that you are interested in. Also remember that the Copy Number results will be overwritten if the analysis is run again.

The computations take different course depending on whether it is 'Against Reference' or 'Paired Normal' type and also depending on the array type and is explained in the section Technologies supported in **GeneSpring** .

Copy Number Analysis can be started from the workflow menu **Copy Number Analysis**.

A 2 step wizard comes up.

Step 1: Interpretation Chooser Choose the interpretation that you want the Copy Number analysis to be run on. See section Significance of Interpretation in Copy Number analysis for detailson how to define interpretation for running 'Against Reference' analysis and 'Paired Normal' analysis.

Step 2: Pairing Information Based on the interpretation, **GeneSpring** infers if the analysis is against reference or paired normal and shows the pairing information here. User can go back and choose a different interpretation if this is what he wants.

The Figure 34.11 shows an instance where 'Paired Normal' method of analysis is inferred. Note that in this type of analysis, the tumour samples are compared against the normal samples while the normal samples are compared against the standard / custom reference.

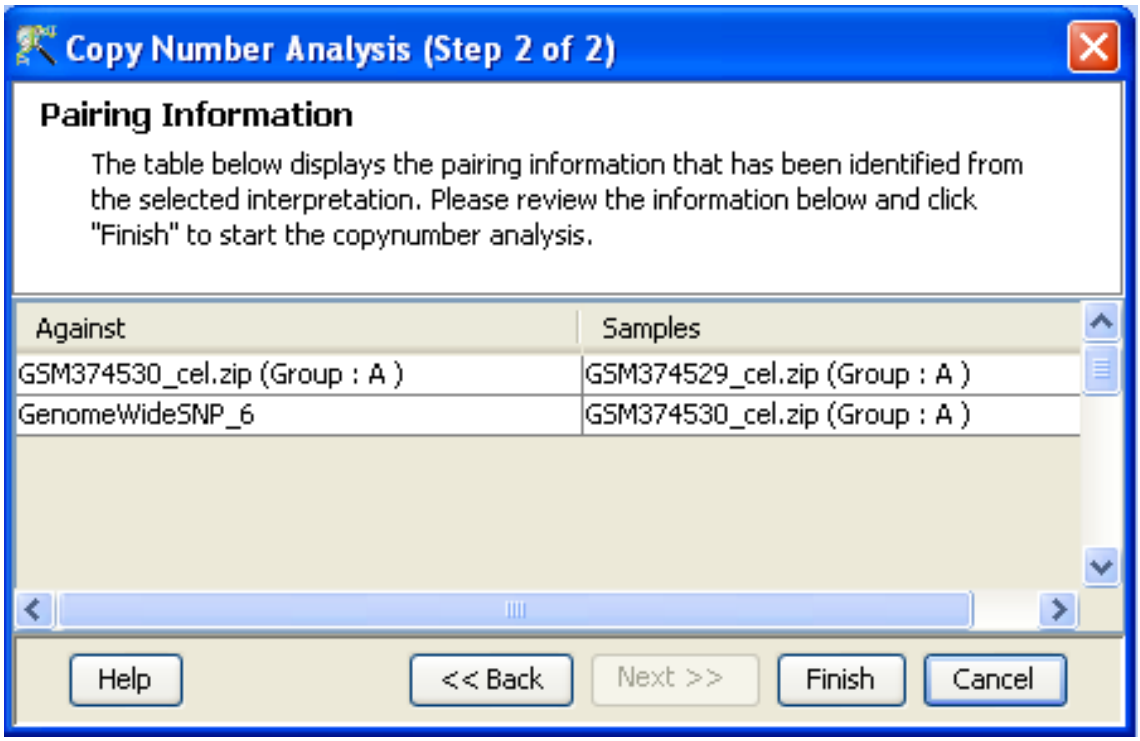


Figure 34.11: Copy Number Analysis - Paired Normal Method

The Figure 34.12 shows an instance where 'Against Reference' method of analysis is inferred. Whatever reference was chosen during experiment creation (standard or custom) will be used here.

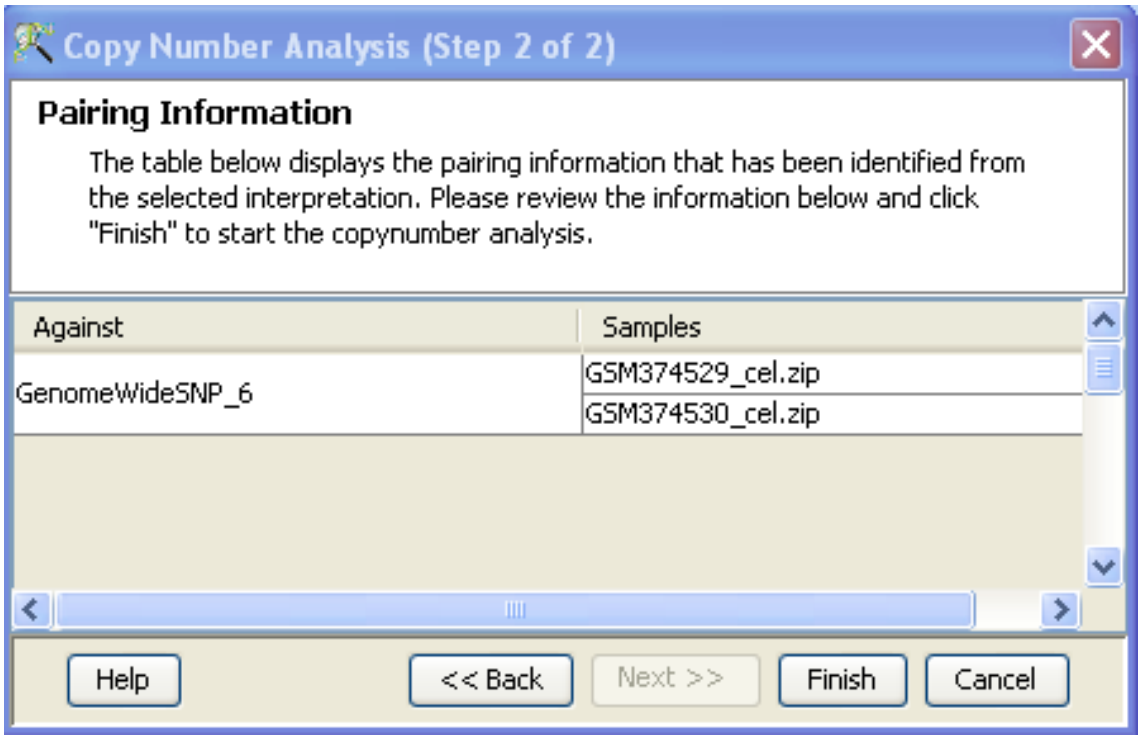


Figure 34.12: Copy Number Analysis - Against Reference Method

NOTE: Only one Copy Number / LOH value will be saved per experiment. If the user runs the analysis again, the previous values will be over-written.

The Copy Number analysis results can be viewed as a Heat map or in the Genome Browser. Further analysis like Common Genomic Variant Regions or Filters or Results Analysis can be carried out.

34.3.6 Common Genomic Variant Regions

This involves identification of regions of the genome that are significantly amplified or deleted across a set of samples. The ultimate motive is to relate these regions of aberration to cancer pathogenesis and the method is commonly referred to as **Genomic Identification of Significant Targets in Cancer (GISTIC)**.

GISTIC analysis runs on the outputs of Copy Number Analysis; it runs independently on the regions of Copy Number Amplification and Copy Number Deletion. At the end, two genomic regions which show aberrations with respect to Copy Number amplification and deletion are identified. Table 34.4 gives a snapshot of GISTIC process.

Table 34.4: snapshot of 'Common Genomic Variant Region' Detection Algorithm

Inputs	Genome of each sample with Copy Number data and Log Ratios; Illumina outputs can be directly used for GISTIC analysis. For Affymetrix CEL files, run Copy Number analysis and then run GISTIC.
Modes of running GISTIC	GISTIC can run in Coarse or Fine mode. Coarse mode applies Q value cut-off filter to identify regions of aberration. In Fine mode, after identifying peaks, aberrant segments are assigned to each peak; then all aberrant segments corresponding to this peak are set to '0' and iteratively, we detect further peaks till we reach the threshold for the peaks or till we fail to detect any more peaks. (Q value obtained by this process is called Peeled Q value.)
Output	1. A gene list called 'Overlapping genes with identified regions of aberration at the specified Q value cut-off'. This gene list will be absent if no genes are found in those regions. In Fine Mode, the region that is passed for identifying overlapping genes is defined by the peak position. In Coarse mode, the entire region identified as region of aberration after applying Q value cut-off filter is passed for identifying overlapping genes. 2. Probe set wise Q value list (unfiltered) 3. Probe sets in identified regions of aberration (after applying the Q value cut-off filter); the Q value listed here is the peeled Q value.
Configurable Options	Many GISTIC parameters can be set from the menu **Tools** → **Options** → **Copy Number Algorithms** → **GISTIC**. See section Configuration options for Copy Number analysis.
Reference	Broad Institute's Cancer Program Publication [22]

Process:

Obtaining Mean Log Ratios Segment the genome based on Copy Number; for each segment with same Copy Number, get the mean of the log ratios and replace the log ratio value of each probe with this mean log ratio value.

Identifying Regions of Amplification Look at the Mean Log Ratio value for each segment for each sample; by default, any value above 0.9 is retained as it is and all those below 0.1 are thresholded at 0. The lower and upper cut-off values are configurable.

Identifying Regions of Deletion Look at the Mean Log Ratio value for each segment for each sample; any value below -1.3 is retained as it is and all those above -0.1 are thresholded at 0. The lower and upper cut-off values are configurable.

Obtaining G score Get the average of sum of the log ratios for each segment (done individually for regions of amplification, deletion and LOH) to get G score for each probeset across all samples.

Obtaining P value From the log ratios for all probe sets for all samples, find frequency by dividing by binning interval. Multiple this frequency across samples and this will give a distribution curve. Each G score is mapped to an index on the frequency curve. The area of the curve beyond this index divided by the total area under the curve gives the P value for each score.

Obtaining Q value and Peeled Q value Apply Benjamin Hocheberg Correction to obtain Q value which is corrected P value. Peeled Q value is the minimum Q value in the Region considering only the aberrant segments associated with this Peak.

Assigning Peak Q value Peak Q value is the minimum Q value in the Region considering only the aberrant segments associated with the Peak. This is shown for both Coarse and Fine mode ; Fine mode shows Peeled Q value. When there are multiple segments associated with a probeset, one Peak Q value corresponding to the most significant peak is shown.

Handling X and Y Chromosomes Due to Copy Number corrections, it is possible that large number of deletions are detected in X and Y chromosomes; hence these are not considered by default for GISTIC analysis in both Coarse and Fine mode.

34.3.6.1 Running Common Genomic Variant Region analysis

Common Genomic Variant Regions analysis is enabled only for Copy Number experiments and can be called from the workflow menu **Analysis** → **Common Genomic Variant Regions**. Note that the inputs required for the analysis are Log Ratios and Copy Number segmented Genomes. The analysis is supported for both Affymetrix technology and Illumina technology. In the case of Affymetrix technology, it can only run after running Copy Number analysis as the required inputs are generated only after Copy Number analysis.

> **NOTE:** For Affymetrix technology, Common Genomic Variant Regions needs Copy Number analysis results to proceed. For Illumina technology, after creating the experiment, it can be directly called.

A 4-step wizard opens up on clicking .

Step 1: Interpretation Chooser Choose an interpretation; samples for which the analysis will run will be determined by the interpretation and it will be shown in this step.

Step 2: Q Values This step gives the unfiltered probe sets along with Q values for all the regions of Copy Number amplification and deletion. There are tabs at the bottom of the table for viewing 'Amplification Q values' and 'Deletion Q Values'. These lists are saved and shown as 'Unfiltered Q value' in a folder called 'GISTIC results' in the experiment.

In this step, there is an option to choose 'Coarse mode' or 'Fine mode' for identifying regions of aberration. Coarse mode applies just the Q value cut-off filter; In Fine mode, after identifying peaks, aberrant segments are assigned to each peak; then all aberrant segments corresponding to this peak are set to '0' and iteratively, we detect further peaks till we reach the threshold for the peaks or till we fail to detect any more peaks. This is called Peeled Q value.

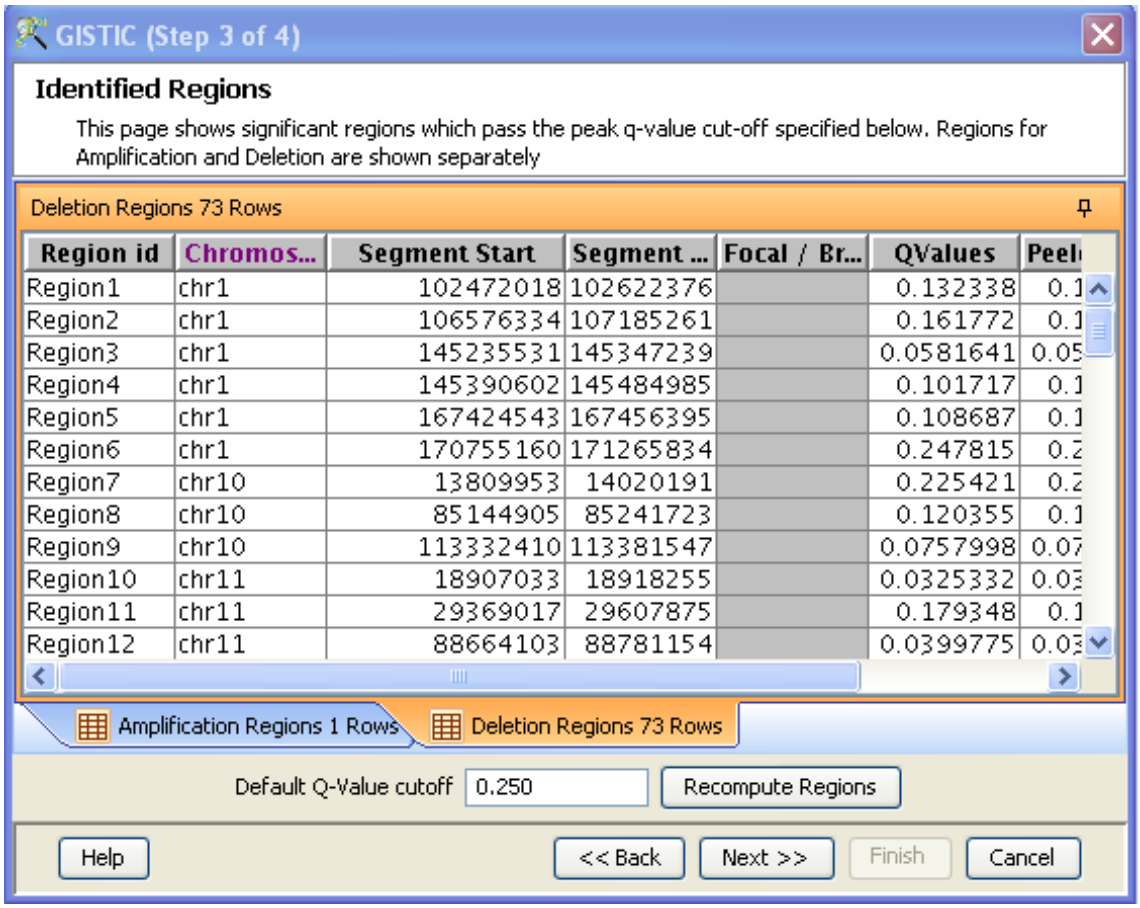


Figure 34.13: Common Genomic Variant Regions - Step 3

Step 3: Identified Regions Figure 34.13 shows the identified regions with a default cut-off value of 0.25 for Amplification, Deletion and LOH scores. The result is in the form of a table giving information on Chromosome number, Segment Start and End, whether it is Focal or Broad region along with the Q value and Peak position for that region. These are basically regions of aberration listed for amplified/deleted Copy Number segments.

In Fine Mode, the peak region is identified considering only the aberrant segments associated with the Peak. Peeled Q value is the minimum Q value in the region and this peeled Q value is reported as the peak Q value for fine mode.

It is possible to iteratively redo the computation with a different Q value cut-off value from this step.

Step 4: Overlapping Regions The probes corresponding to the identified regions of aberration are identified and the genes containing these probes are then output as 'Overlapping genes' provided such genes exist. This gene list is created for each of amplification and deletion provided such genes exist, along with probe-wise Q value list. See Figure 34.14.

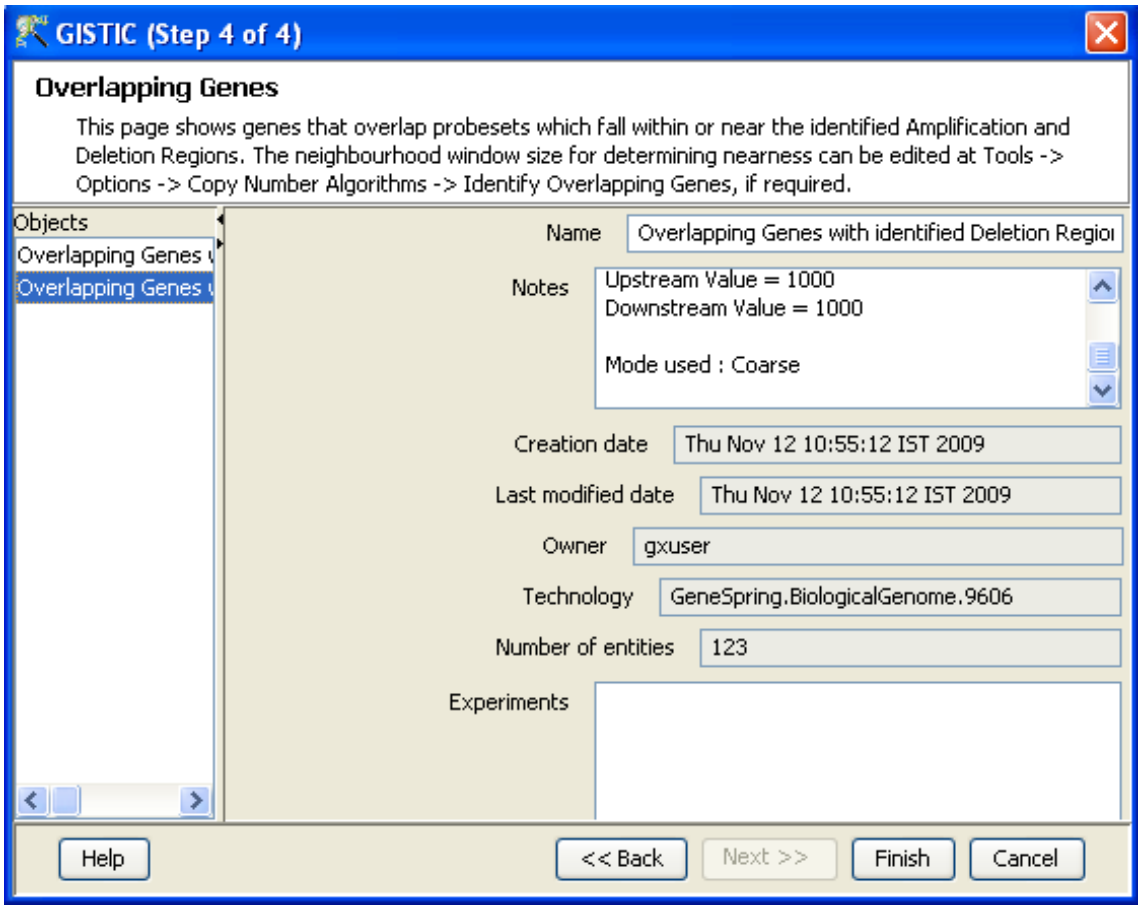


Figure 34.14: Common Genomic Variant Regions - Step 4

Note: Rerunning with same parameters overwrites results.

34.3.6.2 Note on Identifying Overlapping Regions in GISTIC

Given an entity list, **GeneSpring** identifies overlapping genes using a default value of 1000 units to detect genes upstream and downstream, as explained in section Identify Overlapping Genes. When we run 'Identify Overlapping Genes' from within GISTIC, the entity list that is input is slightly different for Fine Mode, compared to the Coarse Mode. In Fine mode, the entity list corresponding to only the peak region is used for identifying overlapping genes. Whereas in Coarse mode, the entity list corresponding to the entire region identified as aberration is used for identifying overlapping genes.

34.3.7 Filters

Once the Copy Number analysis is done, various filters can be applied to pull out data satisfying certain specified criteria; typically the results of filtering are entity lists. Filtering is possible on

Log ratio, Copynumber and Copy Number confidence, LOH score, Parent Specific Copy Number and Known Genomic Variant Regions (based on DGV).

Workflow options include:

1. Filter By Regions
2. Identify Copy Neutral LOH
3. Filter by Parent Specific Copy Number
4. Filter by known CNVs

34.3.7.1 Filter By Regions

'Filter by regions' is a top level filter where ranges of data can be specified for Copy Number (CN), Copy Number Confidence, LOH and Log ratios.

Inputs:
1. Entity list - 'All Entities' by default
2. Interpretation to derive the condition information
3. Sample aggregation information
   - x out of y samples should satisfy the filter criteria
   - x % of samples in any y out of z conditions should satisfy the filter criteria
4. Minimum number of probesets in the region
5. Filter criteria on the various fields with 'And' condition.

Process:
1. Go through each probeset in sequence along the genome. At each probeset, determine if the probeset satisfies the specified Copy Number (includes Copy Number, Copy Number confidence, log ratios) and LOH conditions. If the probeset does not have values for a particular condition (e.g., LOH for CN probesets) then inherit the value for that flag from the previously measured value. This is applicable only for LOH - any other missing value is ignored.
2. For each sample, consider only those probesets which satisfy both the CN and LOH filter conditions.
3. Aggregate the probe sets across samples and apply the sample aggregation condition on probesets resulting from step 2. For each such probeset, we now get a pass/fail indication depending upon whether or not the sample aggregation condition is satisfied.
4. On the aggregated list, apply the region size threshold and output only those probesets, which obey this filter.

Outputs:
1. **Output view:** A searchable spreadsheet with the entities that passed the filter criteria will be displayed
2. **Entity list:** An entity list containing all SNPs which pass the filter criteria is saved.

Performing 'Filter by Regions' in GeneSpring : From the Copy Number experiment, launch 'Filter by Regions' from the workflow browser → Filter → Filter by Regions. A 3 step wizard opens up.

Step 1: Input Parameters Takes the 'All Entities' list by default and the selected interpretation for filtering.

Step 2: Filter Conditions Takes the sample aggregation information criteria and the filter conditions.
For sample aggregation, choose to retain entities where

- A defined number of samples out of the total samples meet the filter conditions.
- Or a defined percentage of samples in the specified number of interpretations meet the filter conditions.

Option to define minimum probesets per regions meeting the filter conditions is also there.

Filtering itself can be done on Copy Number values, Copy Number Confidence, Log ratios and LOH scores. Check the required fields and define the condition; Conditions include 'equal to, Not equal to, Less than, Greater than, Less than or equal to, Greater than or equal to, in the range, and outside the range'. Define the relevant Filter value after choosing the condition. Remember to 'Enter' after giving the filter value.

See Figure 34.15.

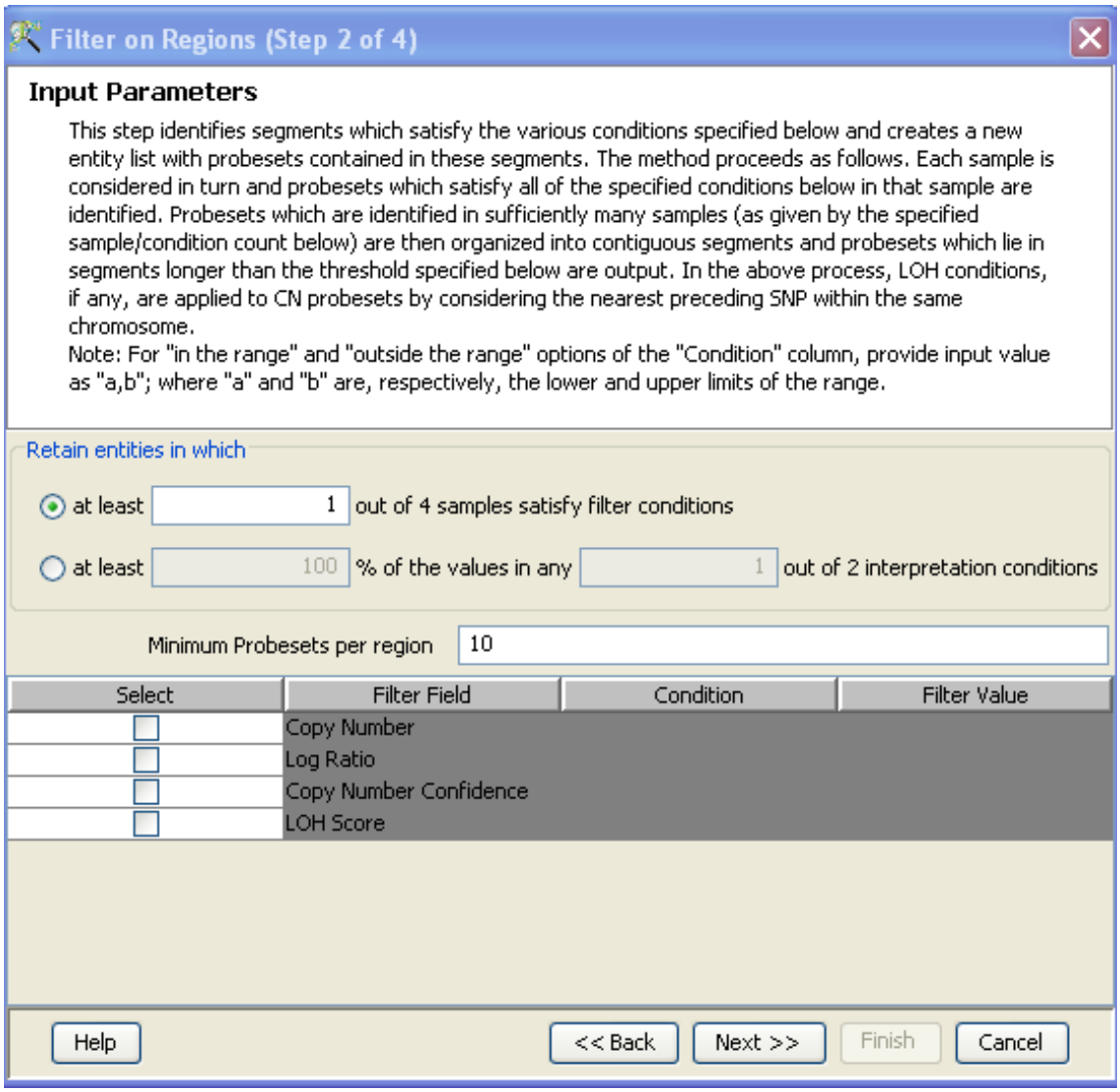


Figure 34.15: Step 2: Filter Conditions for Filter by Region

Step 3: Output Views of Filter by Region Shows the probe set IDs and the number of entities that passed the filtering criteria.

Step 4: Save Entity List Displays the filtered entity list along with details like probe set ID, Affy SNP ID, dbSNP RS ID, Chromosome number, Physical position, and Strand information. *Finish* will exit the wizard and create an entity list called 'Filter on Regions' under 'All Entities' in the experiment.

34.3.7.2 Identify Copy Neutral LOH

This filter is applied only on the copy neutral regions of the genome within the specified LOH threshold.

Process

1. For each sample, go over all probesets to determine regions of contiguous probesets which have LOH greater than the specified threshold. This threshold can be set from the menu **Tools** → **Options** → **Copy Number Algorithms** → **Segment Thresholds** and the default value is 0.5 (Value above 0.5 indicates Loss of Heterozygosity).
2. If the probeset does not have LOH value (e.g., LOH for CN probesets), then inherit the value from the previously measured value.
3. For each of the regions identified using the above process, find out the average copy number. The average copy number of the region is the mean of the copy number values of all the probesets in the region.
4. If this average copy number is in the range of 1.5 to 2.5, then retain all the probesets. Else the enitre region is said to have failed the filter.
5. Aggregate the probe sets across samples and apply the sample aggregation condition on probesets. For each such probeset, we now get a pass/fail indication depending upon whether or not the sample aggregation condition is satisfied.
6. On the aggregated list, apply the region size threshold and output only those probesets, which obey this filter.

In human genome, the normal level of homozygosity is around 70 % and are confined to less than 100 kb. When doing an 'Against Reference' analysis, any region with more than 70 % homozygosity can be safely assumed to be showing LOH. This combined with some meaningful filter conditions (using the option 'Minimum probesets per region') can reduce false positives and remove small pockets of high homozygosity while picking up LOH regions.

In V6 experiments, SNP-SNP distance is typically 1.3kb and this translates to a minimum of 75 probesets if one wants to apply the 100kb filter.

34.3.7.3 Parent Specific Copy Number Filter

Filter criteria is applied on parent specific Copy Number; the user can define the desired Copy Number value/range for allele A and allele B in the filtering criteria in addition to the sample aggregation information. Both SNP and CN probes are considered for PSCN filter.

A probe-set may show missing PSCN value due to the following reasons:

1. It is a CN probeset.
2. SNPs may show missing PSCN because they were part of a long homozygous stretch and hence ignored while assigning PSCN values.
3. SNPs may also show missing PSCN because its copy number and subsequently ASCN was not calculated for reasons like missing birdseed clusters.

These missing PSCN values are inherited from previously available values, as explained in section Filter by Regions.

Note that this filter is not supported for Affymetrix 50/100k array set.

Go to workflow **Filter** → **Filter by Parent Specific Copy Number**; a 4 step wizard will open up.

Step 1: Interpretation Chooser Select an interpretation for the filtering.

Step 2: Input Parameters Sample aggregation information can be defined in terms of:

- Number of samples satisfying the filter criteria (OR)
- Percentage of samples in a defined number of interpretations satisfying the filter criteria

Minimum probesets per region can also be defined.

See Figure 34.16.

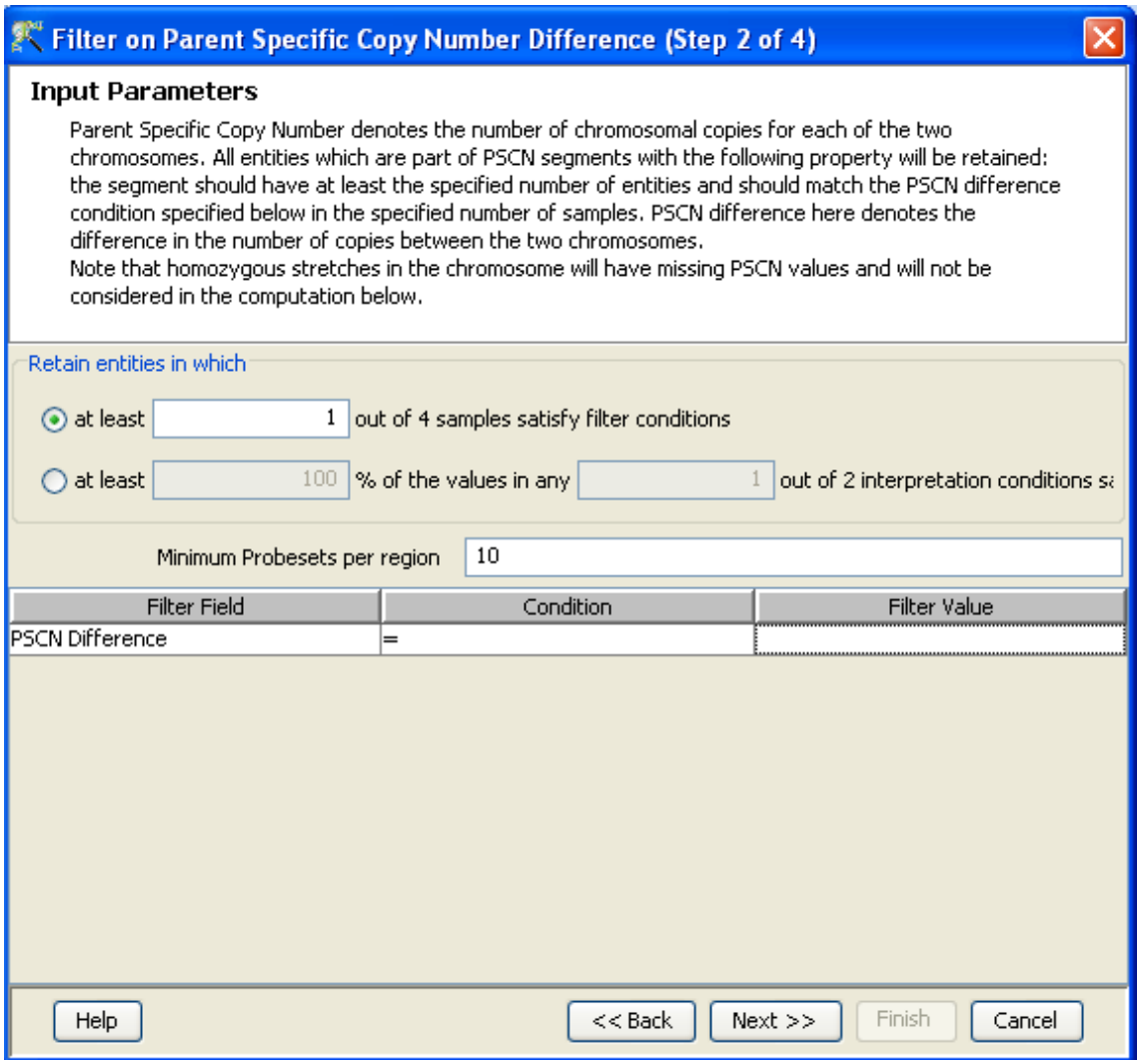


Figure 34.16: Step 2: Input parameters for PSCN Filter

Step 3: Output Views on Parent Specific Copy Number calls Gives the Probe set IDs and the number of samples for which the entity passed the filter criteria.

Step 4: Save entity list Details of the entity list created after applying filter conditions are shown here. Clicking *Finish* will create a new entity list called 'Parent Specific Copy Number Filter' in the experiment navigator and exit the wizard.

34.3.7.4 Filter By Known CNVs

Process is explained in Table 34.5

On clicking the workflow menu **Filter** → **Filter by Known CGVs**, a 2 step wizard opens up.

CGVs	Common Genomic Variant Regions.
Inputs	Entity list and the Database of Genomic Variants (DGV) packaged within **GeneSpring** . See http://projects.tcag.ca/variation/ for information on DGV.
Process	For each probeset in the input entity list, identify if it overlaps any of the known genomic variant region.
Outputs	Entity list with all entities that overlap any of the known genomic variant regions

Table 34.5: Filter by CGVs

Step1: Entity List Chooser Pass an entity list here

Step 2: Results Shows the number of entities in known genomic variant regions as well as the number of entities outside this region. Clicking *Finish* will create two entity lists in the experiment navigator - 'Entities in known genomic variant regions' and 'Entities outside known genomic variant region'.

34.3.8 Views

For Copy Number analysis, views are supported in the form of heat map and the Genome Browser.

34.3.8.1 Heatmap

For the given entity list and interpretation, heat map can be drawn for Copy Number and LOH scores for each sample.

See Figure 34.17 for a Heap Map view for a Copy Number Experiment.

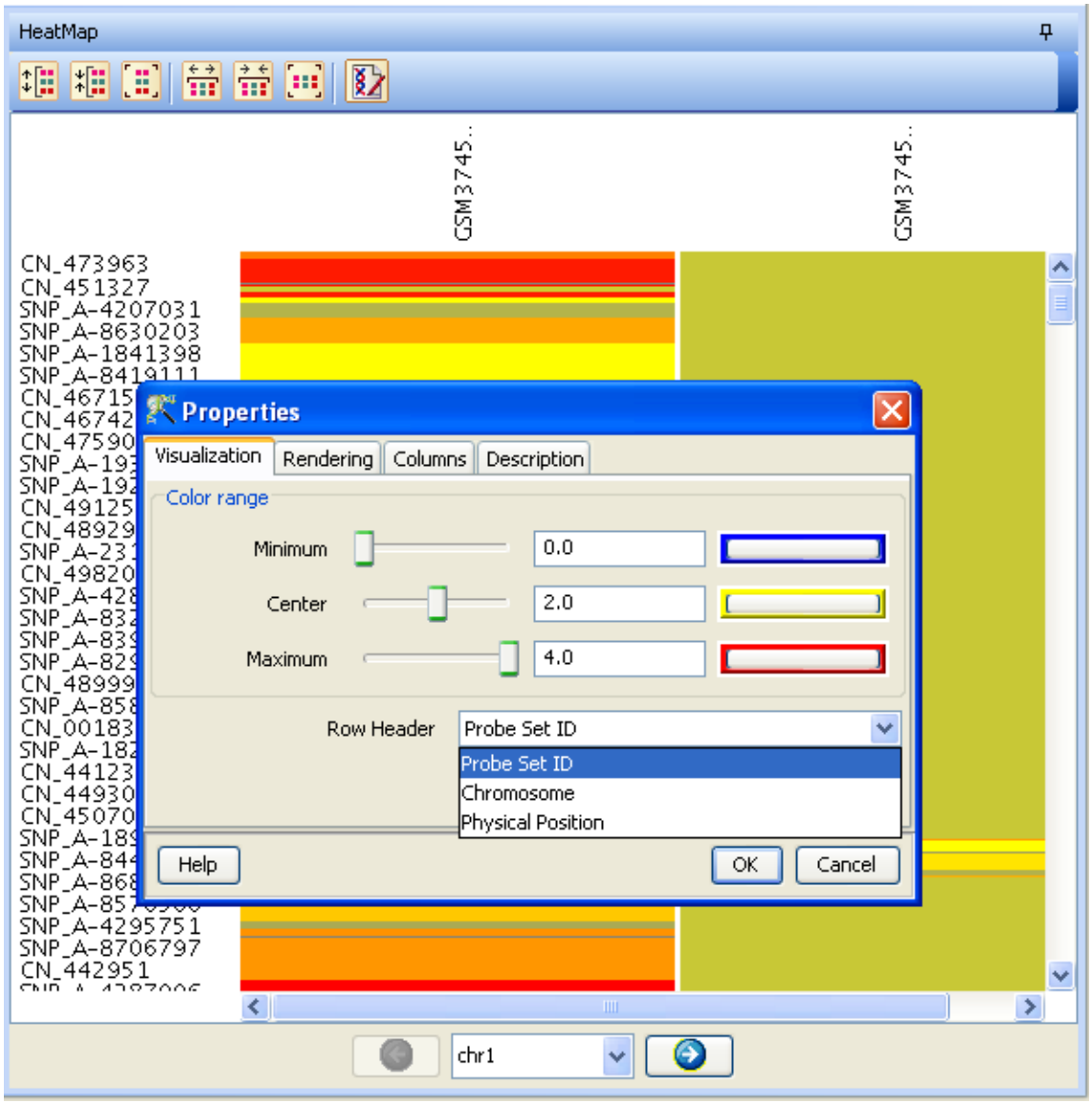


Figure 34.17: Heap Map View for a Copy Number Experiment

An additional feature on the heat map for Copy Number results is the functionality to create entity lists from selection. Click the icon **Create Entity list** in the toolbar; it creates an entity list with the selection on the heat map. The entity list called 'Region map entity list' is added to the experiment navigator.

For Copy Number analysis, the heat map is actually a region map as it shows the Copy Number or LOH regions. On **Right click** → **Properties** from the heat map, there is an option to choose 'Probeset ID, Chromosome or Physical Position' under 'visualization' tab. Each probe or chromosome or physical position is samples and the view can be collapsed / expanded to fit all. 'Column' tab allows choosing samples to view. There is a chromosome browser at the bottom.

34.3.8.2 Genome Browser

Genome Browser allows viewing Copy Number data at the genome level, and in combination with gene/transcript annotation data of the organism. The data can be viewed along with data from other technologies (expression, etc) for the same organism. See section Viewing Copy Number Experiments in Genome Browser for further details.

34.3.9 Results Analysis

34.3.9.1 Identify Overlapping Genes

Helps to identify the genes present in regions of interest (detected from Copy Number analysis); Process explained in Table 34.6.

Start 'Identify Overlapping Genes' from the workflow **Results Analysis** → **Identify Overlapping Genes**. A 2 step wizard opens up.

Step 1: Entity List Chooser Pass the entity list here.

Step 2: Output View Gives a list of the overlapping gene IDs and the chromosomes they are located on; the upstream and downstream region to search for genes is a configurable option from the menu **Tools** → **Options** → **Copy Number Algorithms**. Clicking *Finish* will create the entity list in the experiment navigator called 'Overlapping Genes' and exit the wizard.

For description of 'Identifying Overlapping Genes' with GISTIC result, see section GISTIC - Identify Overlapping Genes.

34.3.9.2 GO Analysis

Once a gene list is available from Copy Number analysis (generated by running 'Identify Overlapping Genes' OR 'GISTIC'), one can do GO analysis. For details, see section Gene Ontology Analysis.

34.3.9.3 NLP Network Discovery

NLP Network Discovery is possible with a gene list. See section NLP Network Discovery for details.

Table 34.6: Identify Overlapping Genes

Inputs	1. An entity list. 2. Upstream and downstream distance to look for genes in the chromosome. Default value is 1000 units on both sides (start and end locations); this value can be changed from the menu **Tools** → **Options** → **Copy Number Algorithms** → **Identify Overlapped genes - Upstream / Downstream Values**.
Process	1. Identify the valid probesets from the selected entity list. A valid probeset is defined as one which has both the chromosome and physical position information. 2. For all probesets belonging to a chromosome, chr1, get the position map from the interval group map. Using the chromosome, start and end position columns in the biological genome, an interval group map is generated on the fly. 3. For each probeset use the physical position information to get a list of indices of overlapping genes, by applying the upstream and downstream distances. 4. Use these indices to get the gene-ids from the biological genome. 5. These gene-ids are the overlapping genes.
Outputs	1. For the 'Identify Overlapping Genes' step, a spreadsheet with one column as the gene-id and another column with the chromosome number is shown to the user. 2. An entity list where each entity is a gene is saved at the end.

34.3.10 Utilities

The utilities section in Copy Number analysis allows user to perform some tasks relevant to downstream operations of Copy Number analysis quickly.

Import Entity List from File	A simple import functionality by matching column names. See section Import Entity List from File for details.
Create Probe List	This utility allows searching an existing entity list for certain positions of chromosomes. The result is a SNP list which match the search criteria, based on physical coordinates.
Filter on Entity List	This utility allows user to filter an Entity list using its annotations and list associated values. See section Filter on Entity List for details.
Export Segmentation Results	User can choose to export Copy Number or LOH segments as also segments identified to have aberrations (amplifications and deletions). **GeneSpring** automatically gives the name of the cytoband the segments overlap with. To get the cytoband overlap information, user should have downloaded the annotation data using the menu **Annotation** → **Update Genome Browser Data** → **From Agilent Server**. The tool prompts for choosing an organism build to obtain this information. The exported text file can then be dragged and dropped into Genome Browser for viewing. See chapter Viewing Copy Number Experiments in Genome Browser for details.

Table 34.7: Utilities in Copy Number Analysis

34.4 Copy Number analysis of Illumina

Users can directly bring in Copy Number results from Illumina's GenomeStudio within **GeneSpring** and do analysis. **GeneSpring** comes packaged with a plug-in to extract output from Illumina; this plug-in needs to be copied from **GeneSpring** into GenomeStudio for activation as described here .

34.4.1 Obtaining Data from Illumina

The following steps will guide in importing genotype information from GenomeStudio:

1. Copy the `GX.Genotyping.Export.dll` plug-in
   from `GENESPRING INSTALLATION DIRECTORY\app\Illumina\GX.Genotyping.Export.dll`
   to `GENOMESTUDIO INSTALLATION DIRECTORY\Modules\BSGT\ReportPlugins`
   (Over-write the file if there is already one with the same name in the latter directory.)
2. Restart GenomeStudio.

3. Open the Project.
4. Select `Reports` option from the the `Analysis` menu.
5. Click `Report Wizard` option from the `Report` sub menu. This launches the `Report Type` wizard page.
6. Select `Custom Report` radio button, and `GeneSpring Exporter 1.2 from Strand Life Sciences` from the drop-down list, and click *Next*. This launches the second page of the `Sample Groups` wizard page.
7. Select the sample groups you want to include in the report, and click *Next*. This launches the `Destination` wizard page.
8. Provide an `Output Path` (you can use the `Browse` button), and `Report Name`, and click Finish. This launches a `Progress bat` with information about the incremental progress in generating the Report. After the `Report` is written completely, you will see a message with an option to view the `Report`.

The plug-in will extract the following information relevant to the technology:

- dbSNPID
- Chromosome Name
- Chromosome Position
- HWE (for association experiments)
- CNV regions

From the Illumina output files, **GeneSpring** directly uses the following values for each sample.

- log R - Provides the log (base 2) ratio of the normalized R value for the SNP divided by the expected normalized R value.
- BAF (B allele Frequency)
- Copy Number values and their confidences
- Genotype of the subject SNP for the sample along with the score.

NOTE: Ensure that Copy Number analysis is done in GenomeStudio before running the plug-in. All the columns required for analyzing Illumina data will be available only after running the Copy Number analysis within GenomeStudio.

Please note that sample names should not be numeric and contain special characters such as `! * # ; ? \ / : " < > |` and space.

Allele specific Copy Numbers are computed for Illumina outputs within **GeneSpring** from BAF and Copy Number values. Parent Specific Copy Numbers are computed as explained in section PSCN.

$$BAF = \frac{B}{A+B}$$

B is the B allele. In place of A+B, the Copy Number value is substituted and the BAF value is used on the left hand side of the equation to obtain the value of 'B'. A is obtained from the total Copy Number and the B value. Thus the allele specific Copy Number values are obtained from Illumina output.

Note that log R is used directly for Illumina wherever log Ratio is mentioned.

34.4.2 Handling Missing Values

- If any missing values are present either in Chr or Position columns, those entities won't considered in copynumber datasets/analysis.
- If any missing values are present in any sample columns, those values are treated as missing values.

34.4.3 Workflow description for Illumina Outputs

To start an Illumina experiment, user needs to have run the plug-in and extracted output from GenomeStudio as explained in the section above on **'Obtaining Data from Illumina'**.

Workflow remains almost the same for Illumina outputs as with Affymetrix files explained above. Some significant differences are:

1. Technology is created on the fly based on the samples.
2. QC step is skipped for Illumina.
3. Since Copy Number information is already present in Illumina, the analysis part skips this step and instead calculates only 'Allele specific Copy Number'.
4. There is no filter for 'Copy Neutral LOH' as LOH values are not present.
5. It is not possible to filter on entity list.

The table below summarizes the steps in Illumina workflow highlighting how it is different the Affymetrix workflow explained in Table 34.8.

Create Technology	Technology is created on the fly based on the samples used to create the experiment.
Create Experiment	Experiment creation is similar to Affymetrix; use the output files generated from GenomeStudio Refer section Create Experiment for other details.
Experiment Set up	As explained for Affymetrix files in section Experiment Set up.
QC	QC step is skipped for Illumina as **GeneSpring** works with outputs from Illumina's GenomeStudio only.
Analysis	Computes Allele specific Copy Number as explained in section Allele Specific Copy Number for Illumina Outputs. Common Genomic Variant Region analysis (GISTIC) also runs as explained for Affymetrix files in section Common Genomic Variant Regions.
Filters	Other than the filter for identifying copy neutral LOH regions and general filtering on LOH, all filters that run on Affymetrix files work on Illumina outputs; refer section Filter.
Views	Heap Map and Genome Browser
Results analysis	As explained for Affymetrix files in section Results Analysis.
Utilities	Remain the same as in Affymetrix files Utilities except for the utility to filter on entity list. This is not supported for Illumina

Table 34.8: Workflow for Illumina output files

34.5 Create Custom Reference

GeneSpring allows advanced user to choose their own references for comparison instead of using the standard 270 HapMap samples, as explained in section Technologies supported in **GeneSpring** .

Users can create their own custom references for running Copy Number analysis within **GeneSpring** from the menu **Tools** → **Create Custom Reference**. Ensure that you have downloaded the required technology before proceeding for custom reference; See Create Technology for details. Note that custom reference is relevant only for Affymetrix technologies and not for Illumina Copy Number outputs.

See section Technologies supported by **GeneSpring** for details on how the reference gets created for various Affymetrix technologies.

On clicking **Create Custom Reference**, a 3 step wizard opens up.

Step 1: Load Data Choose files/samples to be used for custom creation; note that you cannot mix files from different technologies.

Step 2: Pair CEL files Shows interface for pairing CEL files.

Step 3: Reference Description Allows user to define a name and notes for the custom reference.

34.6 Useful information

34.6.1 Using disc cache

Under the menu **Tools** → **Options** → **Copy Number Algorithms** → **APT Execution Options**, 'Use Disc Cache' is an option. It is advisable to use the disc cache in Windows system. Note that it will fail with Linux Operating System with more than 90 samples. When disc cache is not used, the execution process creates batches of probesets and this can slow down the process. In worst situation, the process can quit if there is memory outage. Using disc cache helps to avoid these problems and uses the disc memory while running processes.

34.6.2 Entity Lists and Translation rules in Copy Number

Entity lists in Copy Number analysis behave as normal entity lists generated elsewhere in the tool, for example, in expression experiments. Right click operations on entity lists are valid except for 'Search' option in the entity list inspector; there is no 'search' option on entity lists created in Copy Number experiments.

Translation of entity lists created in Copy Number and association experiments is possible only within these technologies. It is not possible to translate entity lists from expression to Copy Number experiments or vice versa. By default, Translation is done on the basis of gene IDs; in this case, it will be possible to run translation only between experiments of same technology within Copy Number (for example, only between two V6 experiments). There is also an option to do translation on the basis of **dbSNP RSP IDs** which will allow translation between experiments of all types in Copy Number and even from association experiments, if dbSNP RSP ID column is present.

Note that the explicit translation mapping from the menu **Tools** → **Miscellaneous** → **Translation Mapping** does not work with Copy Number and association experiments.

34.6.3 Configuration options for Copy Number analysis

From the menu **Tools** → **Options** → **Copy Number Algorithms**, various configuration options are available for Copy Number analysis.

GISTIC:

1. **Maximum peaks/chromosome:** This defines the number of peaks to be taken per chromosome starting with the minimum Q value. Default value is 250 for amplification and 250 for deletion regions.
2. **Lower Amplification Cut-off:** Set to 0.1; those below 0.1 would be taken as 0.

3. **Upper Amplification Cut-off:** Set to 0.6; those above 0.6 would be taken as 0.6; those in-between 0.1 and 0.6 would be retained as it is.
4. **Lower Deletion Cut-off:** Set to -1.0; those below -1.0 would be taken as -1.0
5. **Upper Deletion Cut-off:** Set to -0.1; those above -0.1 would be taken as 0.
6. **Use only segments with Copy Number confidence:** Set to 2.0 by default, which indicates low confidence.
7. **Q value cut-off for peak detection:** Set to 0.25 by default. It can also be changed while running GISTIC in step 3.
8. **Interval for binning:** Set to 0.001 by default. This can be increased based on the number of samples.
9. **Normalization-Validation: reduce each sample median to zero:** By default, this is checked.

Options for running CBS:

1. **Validate Change Points using :** Gives option to choose T test or Permutation test for validation the segmentation. Permutation test is slower and takes about 30 minutes per sample. T test is faster.
2. **Number of Permutations:** Applicable only if permutation test is chosen.
3. **p-value cut-off:** Applicable for T test where the change points are validated. Default value is 0.002 but can be changed by the user from this option.
4. **Mean Difference between Segments:** User can also set a mean log ratio difference threshold between segments. Segments which differ in their mean log ratio by a value lesser than this cut-off will be joined.
5. **Minimum length of the Segment** User can set the minimum length of the segment that is acceptable as output from CBS.
6. **Offset to jump during optimization:** In each set of log ratios, algorithm searches for the maximum statistic by considering the data as a circle and by iterating within arcs of length 2 to (n-1)/2. The offset defines the arc interval for the search; for the default value of 3, the algorithm will go in steps of 2, 5, 8, ... After it identifies an arc of length l with the maximum statistic, it checks for (offset - 1) values above and below this length to get the best estimate of the maximum statistic.
7. **Size of segment to use jump offset:** Defines size of segment for the above condition.

Segment Thresholds: Allows defining the LOH threshold to be considered while applying filtering and GISTIC analysis.

Fawkes Options:

1. **Number of parallel threads:** For optimization in multi core machines.

2. **Number of probes in the homozygous stretch:** For Parent Specific Copy Number computation, user can define the number of probes to be present in a segment to qualify as homozygous stretch.
3. **Minimum Number of probesets in a segment:** For Parent Specific Copy Number computation, this sets the number of probes to qualify as a heterozygous segment.

Identifying Overlapping Genes - Up Stream and Downstream values: Set to 1000 units.

APT Execution Options: Allows user to 'Use Disc Cache'. It is advisable to use the disc cache in Windows system. Note that it will fail with Linux Operating System with more than 90 samples. When disc cache is not used, the execution process creates batches of probesets and this can slow down the process. In worst situation, the process can quit if there is memory outage. Using disc cache helps to avoid these problems and uses the disc memory while running processes.

34.6.4 Performance Statistics for Copy Number Analysis

For Copy Number and Association Experiments with more than 100 samples, a 64-bit QuadCore Desktop with more than 4 GB RAM is recommended.

Approximate statistics for 75 samples of Affymetrix Genome-Wide Human SNP Array 6.0 on a 32-bit, QuadCore Desktop with 2 GB RAM are:

- Experiment Creation takes 1.5 hours
- Combined time for the following operations for 'Against Standard Reference' analysis is 7-8 hours:
  - Log Ratio
  - CBS segmentation to identify Copy Number segments
  - Copy Number computation which is post processing of CBS segments
  - Parent Specific Copy Number Computation
  - LOH analysis
- Paired Normal analysis for the same experiment could go upto 14-15 hours depending on the pairings.

34.7 Copy Number Algorithms

34.7.1 BRLMM

GeneSpring calls the BRLMM algorithm from the Affymetrix Power Tools (APT) for extracting genotyping calls for Affymetrix 50/100k array set. For complete details on BRLMM, refer to the

Affymetrix white paper [12]. The flow chart in Figure 34.18 reproduced from the white paper details the tasks carried out within BRLMM.

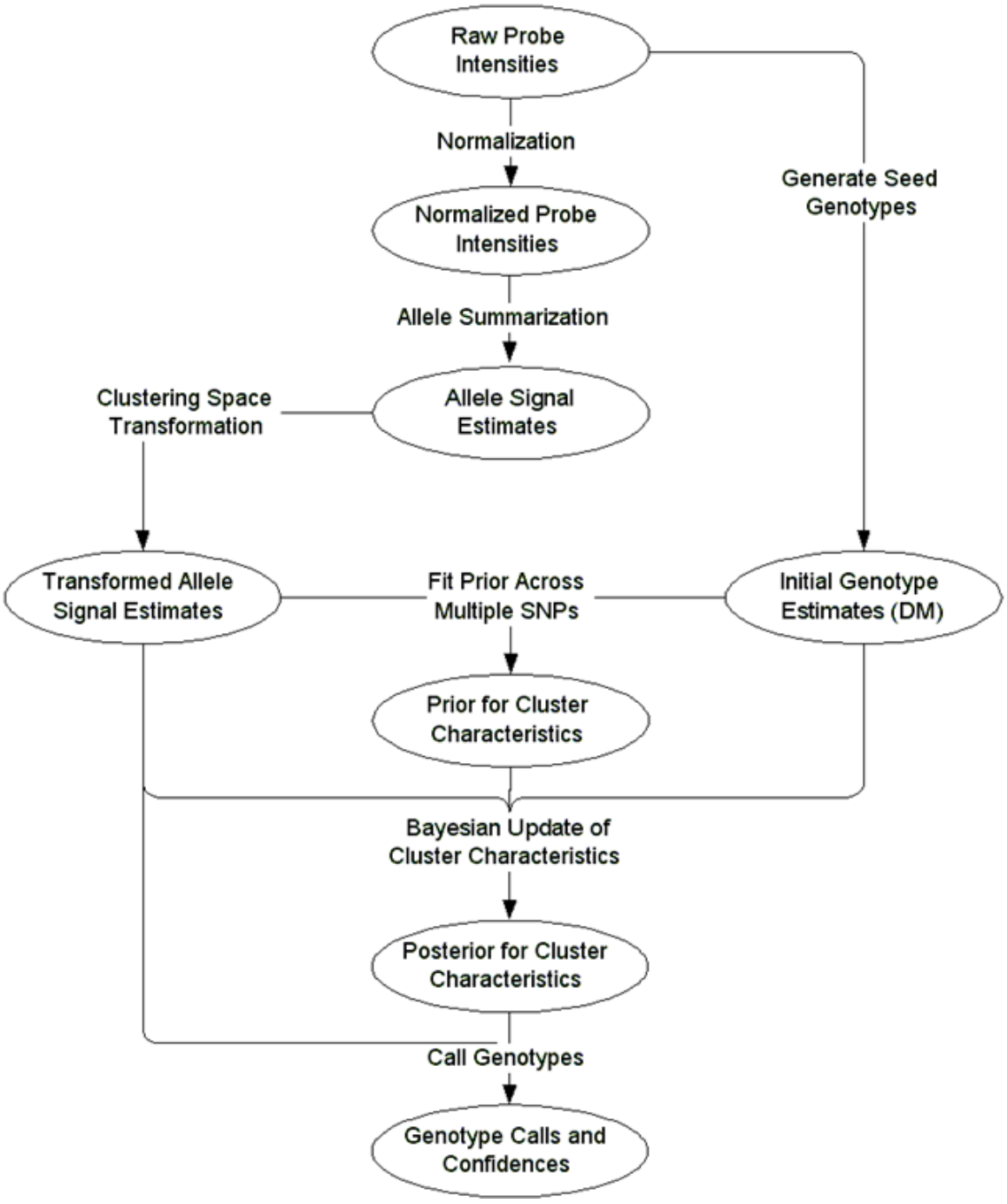


Figure 34.18: BRLMM-Flow chart

Extracting Allele Signals: The first step in analyzing the 100K arrays is the generation of signals for each SNP and each allele (A,B). Signal generation uses Quantile Normalization followed by RMA. No background correction is performed (the signal values are typically

large and mileage obtained from Background Correction is expected to be limited). Quantile normalization is run on all SNPs together (i.e., when Hind and Xba arrays are analyzed together, SNPs from both are pooled together for quantile normalization.) Signal Generation involves running RMA twice, once each on the A and B alleles; these are the allele specific signals. The combined signal is the average of these two signals.

Deriving the prototypes: In parallel, BRLMM derives an initial guess for each SNP?s genotype using the DM approach (with confidence threshold set to 0.17 for high stringency). It then looks across SNPs to identify cases where there are at least a certain minimum number of examples of each of the 3 genotypes according to the initial guesses. This subset of SNPs is used to estimate a prior distribution on the typical cluster centers and variance-covariance matrices. Each SNP is then visited in turn and the cluster centers and variances implied by the initial genotype guesses are combined with the prior information in an ad-hoc Bayesian procedure to derive a posterior estimate of cluster centers and variances (it is principally this step that distinguishes BRLMM from RLMM). Finally, a genotype and confidence score is assigned for each observation according to its Mahalanobis distance from the three cluster centers.

Extracting Genotype Calls: The next step is to generate Genotype Calls. BRLMM calls genotypes by a template-matching procedure comparing the transformed allele signal values observed in an experiment to the typical values (prototype) expected for each genotype. The genotype that is closest in typical value is the one that is assigned (a minimum distance classifier). The approximate confidence in that call is based on the ratio of the nearest prototype to the second nearest prototype.

Some important points concerning BRLMM are listed in Table 34.9.

Minimum Samples	BRLMM requires at least two observations of each genotype to build the prior, so the absolute minimum number of samples required is 6, though running with this small a number is not advised.
Batch Size	While more samples will generally lead to better performance, it is found that for good datasets performance reaches a plateau with as few as 50 samples, whereas for lower-quality datasets it can take as many as 100.

Table 34.9: Additional notes on BRLMM

34.7.2 Hidden Markov Model (HMM)

Computing LOH scores requires a smoothing procedure because signals and ratios are quite noisy and do not indicate clear boundaries for regions with LOH. Hidden Markov Models serve as good rounding procedures in this context, because they can account for locality-based dependencies along a chromosome (i.e., if a particular SNP is in a region of LOH then the likelihood of a neighboring SNP being in a region of LOH also goes up). HMM implementation in **GeneSpring** is based on the reference [48].

LOH scores for analysis against a reference are generated from genotype calls using an HMM with 3 states, representing Loss of Heterozygosity (L), Retention of heterozygosity (R-HET), and Retention of Homozygosity (R-HOM), respectively. The emission probabilities at L and R-HOM are set to .99 for Homozygous and 0.01 for Heterozygous. The emission probabilities at R-Het are set to .989 for

Heterozygous, 0.001 for No Calls and 0.01 for Homozygous. Transition probabilities are defined exactly as detailed in [42] and very similar to the dChip paper http://compbiol.plosjournals.org/perlserv/?request=get-document&doi=10.1371/journal.pcbi.0020041 and are recapitulated in the image shown in figure Figure 34.19.

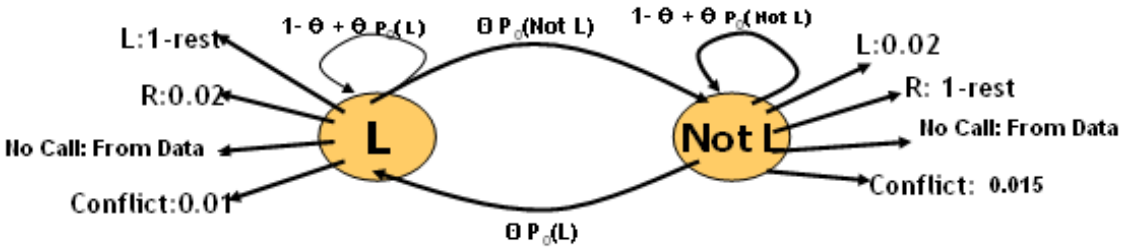
To↓, From→	LOH	Hom-R	Het-R
LOH	$(1-\theta)+\theta\, P_o(L)$	$\theta P_o(L)$	$\theta P_o(L)$
Hom-R	$\theta P_o(R)P(Hom)$	$[(1-\theta+\theta P_o(R)]P_R(Hom\|Hom)$	$[(1-\theta)+\theta P_o(R)]P_R(Hom\|Het)$
Het-R	$\theta P_o(R)P(Het)$	$[(1-\theta+\theta P_o(R)]P_R(Het\|Hom)$	$[(1-\theta)+\theta P_o(R)]P_R(Het\|Het)$

Figure 34.19: Transition Probabilities for LOH analysis against Reference HMM

Here, $P_0(L) = .01$, $P_0(R) = 0.99$ and θ is set to $1 - e^{-2d}$ where d is the distance between the current and previous SNPs in units of 100MB. A higher value would increase the number of LOH regions detected but also increase false positives.

For analysis against reference, all the probabilities mentioned in this image above are computed from reference CEL files.

For paired normal analysis, a different (simpler) HMM shown in the figure fig 34.20 is used; the emission alphabet is no longer genotype calls but a Loss(L), Retention(R), Conflict(C) or Non-Informative call(N) computed from the paired samples as indicated in the figure. A smaller value would lead to fewer LOH calls.



		Tumor			
		A	B	AB	No Call
Normal	A	N	N	C	N
	B	N	N	C	N
	AB	L	L	R	N
	No Call	N	N	R	N

Figure 34.20: The Paired Normal HMM

Note that the L,C,N,R calls are not explicitly output in the spreadsheet.

Sex chromosomes are ignored for LOH computation.

34.7.3 Canary algorithm

V6 affymetrix uses CANARY (Copy Number Analysis Routine) to genotype 1320 regions of CNP (Copy Number Polymorphism). CANARY is a one dimensional Gaussian Mixture Model (GMM) to cluster samples into discrete Copy Number classes; these clusters are used by **GeneSpring** while assigning Copy Numbers to segments identified by CBS.

GeneSpring calls Affymetrix Power Tools to run Canary algorithm. Note that **GeneSpring** runs CANARY only during reference creationfrom 270 HapMap samples or user-defined custom reference samples.

Inputs:
1. CNP region file - contains CNP boundaries relative to the genome and a list of the probes used to calculate the intensity for each region
2. Raw CEL files - Probe summarization and scaling are done within Canary.
3. A 'prior' Model file which contains gaussian cluster parameters of discrete genotype classes observed in the same region. The priors are generated from 270 HapMap samples processed independently at the Broad Institute and Affymetrix. The prior model file contains the cluster mean, variance, and the expected cluster frequencies.

Outputs: Copy Number clusters for CN probes.

References:
1. http://www.affymetrix.com/support/developer/powertools/changelog/apt-canary.html for complete information on Canary algorithm implementation.
2. http://www.affymetrix.com/support/technical/whitepapers/canary_algorithm_whitepaper.pdf for Affymetrix White Paper on Canary Algorithm
3. Reference [10].

34.7.4 Birdseed algorithm

Birdseed is the algorithm used to extract genotype calls from Affymetrix Genome-Wide Human SNP Array 6.0, Genome-wide Human SNP array 5.0, and Human Mapping 500K Array Set. Birdseeed uses a customized Expectation-Maximization (EM) algorithm to fit two dimensional Gaussians to SNP data, producing genotypes and confidence scores for every individual at every SNP. **GeneSpring** calls Birdseed from Affymetrix Power Tools.

GeneSpring runs Birdseed algorithm to create references from HapMap samples as well as to create custom references, from user inputs. Table 34.10 gives an overview.

Algorithm For the most part, this algorithm is a 2 dimensional GMM analogous to the 1 dimensional Canary; the two dimensions are summarized probe intensities for each of the two alleles.

Uses Expectation-Maximization algorithm as follows. The models are used for initialization. Each sample is then assigned a probability of belonging to each cluster (Estimation). Next, each cluster is redefined based off the samples that belong to it (Maximization), as well as

Inputs	1. Intensities from a set of CEL files for each SNP allele. 2. A 'priors' model file of gaussian clusters representing copy neutral SNPs (either from HapMap samples or from user specified reference samples).
Outputs	1. Genotype calls as 0(AA), 1(AB), 2(BB) or -1(No call) at each SNP. 2. Confidence scores ranging from 0 (Least confident) to 1 (Most confident) 3. Posterior Model file containing posterior models that characterize genotype and allele specific probe responses for each SNP. This file is then used by the CBS and Fawkes algorithm.
Useful Tips	1. Because it is a clustering algorithm, Birdseed should typically be run on many samples at a time (50 or more). 2. Internally, Birdseed determines the gender of each sample.
References	1. http://www.affymetrix.com/support/developer/powertools/changelog/apt-probeset-genotype.html#quickstartbirdseed for more details on implementation. 2. http://www.broadinstitute.org/mpg/birdsuite/birdseed.html 3. [10] for technical information.

Table 34.10: Snap-shot of Birdseed Algorithm

being tethered to the expected location of the model. The expectation and maximization steps are iterated until convergence, at which point one can assign the likelihood of the model (gaussian parameters). The likelihood of the model is dependent of how well the model explains the observed as well as how the model fits certain expectations (for example, that the clusters are evenly spaced).

Birdseed chooses between models built from different initializations and between 1,2 and 3 clusters explaining the data. If the best model has fewer than 3 clusters, genotype classes corresponding to clusters not in the model are imputed to increase sensitivity to rare genotypes. The resulting 3 clusters represent the probe responses to each genotype class on the particular batch being run.

Genotyping and confidence call Let the sample i be genotyped by the cluster j (i.e, the cluster j maximised the probability when substituted by the intensities of sample i). The formulae of the weighted gaussian cluster is:

$$P(\frac{j}{i}) = (\frac{wj}{2\pi \times |\sum j|^{\frac{1}{2}}})e^{(}(-\frac{1}{2}(X_i - {}_j{}^T))(\sum j{-}1(X_i - {}_j)))$$

Global confidence score The relative likelihood of belonging to the next cluster

$$conf_{(global-i)} = \frac{P_(j_{(second-best)}|i)}{P(j_{best}|i)}$$

Local confidence score

$$q_i = \sqrt{[(-\frac{1}{2} \times (X_i - {}_j)^T)(\sum j{-}1(X_i - {}_j))]}$$

$$conf_{local-i} = \frac{1}{1 + e^{(}p - q_i)} - \frac{1}{1 + e^p}$$

where q is the number of standard deviations a sample is from its assigned cluster and p is the number of standard deviations beyond which the local confidence score quickly increases. Note that the default $p = 4.0$.

The overall confidence score Affymetrix Birdseed 'Overall confidence' score is then

$$Confidence = (cconf_{global-i} + (1 - c)(conf_{local-i})$$

where c is the relative confidence contribution parameter, $0 \leq c \leq 1$, DEFAULT $c = 0.8$.

34.7.5 CBS for segmenting genome with respect to Copy Number

The Circular Binary Segmentation (CBS) algorithm finds the break points in a sequence of log ratios into segments with significantly different distributions. The break points are found in a recursive way till no further significant segmentation is possible.

Input to CBS are the Log ratios while the outputs are Segment Break Points and Mean Log Ratios of each segment. Table 34.11 gives an overview of the process.

Table 34.11: Snap-shot of CBS Algorithm

Smooth outliers:	The log ratios need to be smoothed by removing outliers.
Segmentation:	Finds change points in genome using a statistic to identify a ternary break as explained in the reference [15].
Validate Change Points:	Gives option to choose T test or Permutation test for validation the segmentation. Permutation test is slower and takes about 30 minutes per sample. T test is faster. Default option is T test and can be changed from the configuration options. Number of permutation or the p value cut off can also be set from here. Default number of permutations is 5000 while default p value cut-off is 0.002.
Configuration options:	Many other parameters in CBS are configurable as explained in section Configuration Options for Copy Number Analysis.

References for CBS:

1. *Circular Binary Segmentation for the Analysis of Array-based DNA Copy Number Data* [14] for basic technical information on CBS.
2. For latest updates on CBS algorithm, check the website quoted in this reference [15].

34.7.6 Post Processing to assign Copy Numbers to segments created by CBS

Once the segments are identified and validated by CBS, Copy Numbers have to be assigned to the segments.

Inputs:
- Reference clusters from Birdseed and Canary algorithms.
- Copy Number segments from CBS.
- Mean log ratios of segments from CBS.

Process:
- For each of the Copy Number clusters (0,1,2,3,4) obtained from Birdseed and Canary algorithms, there is a mean value assigned to each probe. Call this as m0,m1,m2,m3,and m4.
- Get the Log Mean Ratio by taking the ratio m0/m2, m1/m2, m2/m2, m3/m2, m4/m2 and then the logarithm to the base of 2 of this ratio.

- There will be one Log Mean Ratio for each of the Copy Number values for each probe. ('0' value for Copy Number 2)
- Get a median of the log mean ratios across all the probes for each Copy Number value to get 5 medians, M0, M1, M2, M3 and M4. This is the 'Median Map'. Standard deviation value is also available for each cluster.
- For Affymetrix Mapping 100k array set, since there are no clusters (only BRLMM runs on this array), clusters generated from Genome-Wide Human SNP Array 6.0 technology of the reference are used here.
- Take the Mean Log Ratio of each segment from CBS and check with the 'Median Map' described above
- A fitting is done as explained below.
  - Using the standard deviation associated with each cluster, a distribution curve is created (assuming Gaussian distribution) for each cluster and the overlap point of consecutive clusters are identified.
  - The overlap point between Copy Number 0 and 1 is assigned a value of 0.5; similarly 1.5 for overlap between 1 and 2, and so on for the other clusters.
  - The mean log ratio is mapped to this distribution and values are assigned.
  - If the standard deviation is '0' for any cluster, then a fresh value is assigned by taking the mean of the standard deviation values of adjacent clusters and dividing it by the value of the Copy Number of the cluster which is in-between.

Outputs:
- Copy Numbers for each segment
- Confidence:
  - For each segment, if the Mean Log Ratio corresponds to Copy Number in the range (Normal +/- 0.5), a p value of '1' is assigned.
  - For segments with Mean Log Ratio mapping to Copy Numbers out of the range 1.5 to 2.5, a T test is done against '0' and a p value is obtained.
  - Multiple testing correction is applied to this p value by multiplying it by the number of segments which are NOT of Copy Numbers 1.5 to 2.5 to get final p value.
  - Negative logarithm to the base 10 of the final p value reported as confidence.
  - In Paired Normal Analysis, for X and Y chromosomes of Normal Male samples, the confidence is set to 0 if the Copy Number is between 0.5 and 1.5.

Corrections for Sex chromosomes: 'Male' is inferred for Copy Number less than '1.5'. In Paired Normal Analysis, for males, for both X and Y chromosomes, '1' is subtracted as correction. For females, there is no correction for Chromosome X while Y chromosome is set to '0'.

34.7.7 Fawkes algorithm

Fawkes (Fast Analysis With Kopy-number et SNPs) produces SNP genotypes of the form n.m,c where
n is the non negative integer number of the copies of the A allele
m is the non negative number of copies of the B allele
c is the floating point confidence of this call, where 0 is most confident and 1 is least confident.

Table 34.12 gives an overview.

Inputs	1. Total Copy Number 2. Gaussian cluster data from Birdseed algorithm 3. Allele intensities at each SNP
Outputs	1. Allele Specific Copy Number 2. Confidence associated with the above
Notes	The Allele Specific Copy Number generated by Fawkes is the input for the LOH analysis using HMM. For Paired Normal analysis, the input to LOH are the genotype calls generated by Fawkes by setting the Copy Number of all 'Normal' samples to '2'.
References	1. [10] for technical information.

Table 34.12: Snap-shot of Fawkes Algorithm

Algorithm Briefly, the gaussian cluster outputs of Birdseed (associated with copy neutral calls) are extrapolated to generate gaussian clusters for Copy Number 0, 1, 3, 4 ,etc. The intensities of the SNP (X_A, X_B) are substituted into each cluster and the cluster with the maximum probability decides the genotype call.

- For example, extrapolation of clusters for Copy Number 3 would imply generating 4 clusters, A3B0, A2B1, A1B2, A0B3.
- The 'Means; of Copy Numbers 0, 1 , 2 are extracted from the Birdseed clusters and both the 'Means' and 'variances' for all other Copy Numbers are linearly extrapolated as a function of 'Means'.

The underlying A and B allele frequencies w_a and w_b are estimated using the frequencies of the AA, AB and BB clusters.

$$\begin{aligned} w_a &= 1 - w_b \\ &= \frac{2w_{AA} + w_{AB}}{2} w_j = w_a^N * w_b^M * \frac{N + M - 1!}{N! * M!} \end{aligned}$$

For genotype cluster j, with N copies of A allele and M copies of B allele.

Confidence This is exactly as per Birdseed Overall Confidence Score

34.8 Tutorials for Copy Number Analysis

A recorded webinar on Copy Number Analysis in **GeneSpring** is available at
http://genespring-support.com/support/genespring-eseminars

Chapter 35

Association Analysis

This chapter provides instructions to run an Association experiment using **GeneSpring** .

35.1 Introduction

GeneSpring provides a toolkit for Genome-Wide Association Analysis (GWAS). The focus of the Association Analysis lies on the analysis of SNP and phenotype data for studies that can involve thousands of unrelated samples genotyped at hundreds of thousands of SNPs.

It includes a variety of features:

- Quality Control (QC) methods to identify anomalous samples which would confound the analysis.
- Filters to remove undesirable SNPs from the analysis.
- Rigorous Principal Components Analysis (PCA)-based methods for identifying and removing sources of stratification in the experiment.
- Statistical tests to identify SNPs that are associated with disease incidence, and methods to eliminate `false alarm` results.
- Haplotype inference and analysis methods, which can detect more subtle disease-causing groups of SNPs.
- A wide range of publication-quality intuitive visualization options, which include LD Plot, Haplo block view, etc.
- Full integration into other features of GeneSpring GX, such as Pathway Analysis and Genome Browser.

In **GeneSpring** you can create Association Analysis experiments for the Affymetrix and Illumina chip technologies, as well as import genotype data in a text file.

Note: Illumina Association Analysis experiments do not require any algorithm to fetch the genotype calls.

35.2 Technology

GeneSpring supports the following technologies:

Technology	Algorithm
Affymetrix Mapping 50K Xba240	BRLMM
Affymetrix Mapping 50K Hind240	BRLMM
50K Xba240 and 50K Hind240	BRLMM
Affymetrix Mapping 250K Nsp	Birdseed
Affymetrix Mapping 250K Sty	Birdseed
250K Nsp and 250K Sty	Birdseed
Affymetrix GenomeWide SNP5	Birdseed
Affymetrix GenomeWide SNP6	Birdseed
Illumina (all)	None; genotype calls are generated in GenomeStudio.
Custom (Illumina Format)	None; genotype calls are present in the input file.

Table 35.1: Technologies and Genotype Call Algorithms for Association Analysis Experiments

Notes:

- You need at least six distinct CEL files for Hind or Xba experiments.
- You need at least six distinct file pairs for combined Hind and Xba experiments.

35.3 Experiment Creation

GeneSpring allows you to create Affymetrix, Illumina, and Custom Association experiments. You can run Affymetrix Association experiments on CEL files (refer to 35.2 section for information on the technologies and Genotype Call Algorithms).

You can also run Illumina Association experiments on .txt files containing genotype information created using Illumina's GenomeStudio. You can import genotype information as a .txt file from GenomeStudio using a plug-in (refer to section 35.3.1.2).

You can also run Custom Association experiments on .txt files containing genotype information in the Illumina GenomeStudio format (refer to section 35.3.1.1).

35.3.1 Illumina Association Analysis Experiment

You can create an Illumina Association Experiment from the Project menu (*Project* menu $\longrightarrow$ *New Experiment*. In this window, choose the analysis type as *Association* and the experiment type as Illumina Association Analysis.

The Illumina Association Analysis Experiment wizard has one step.

On the Wizard page, click on Choose Files or Samples button to open the file browser or the Sample/RawFile Search Wizard. You can select the files and samples using the browser and the wizard respectively.

The Sample/RawFile Search wizard has two steps.

Step 1: Allows you to Add, Remove, and Combine Search Conditions (by ‘‘OR” or ‘‘AND”). You can select ‘‘Show User Attributes” check box to open the search results with the user attributes, and enter an appropriate value for the ‘‘Max Results per page” (default value: 100).

Step 2: Launches the search results of samples (select and double click to launch the Sample Inspector). Select the Samples you want to use in the Experiment, and click ‘‘Finish”.

- Click on ‘‘Reorder” button to change the ordering of the samples.
- Click ‘‘Finish”.

35.3.1.1 Illumina File Format

This section provides an overview of the File (.txt) Format for Illumina Association Experiments.

The .txt files have five technology columns:

- **Name:** Provides the name of the probe.
- **Chr:** Provides the Chromosome number.
- **Position:** Provides the physical position of the probe.
- **HW Equil:** Provides the Hardy--Weinberg (HW) Equilibrium score for the SNP.
- **CNV Region:** Provides information on any nonpolymorphic probes falling in the region. This column is automatically populated with information, and may not be current because the number of known Copy Number Variant(CNV) regions is constantly changing. This column is for informational purposes only.

The.txt files also have six sample-specific columns:

- **GType-Sample:** Provides the genotype of the subject SNP for the sample.
- **Score:** Provides Genotype Call score of the subject SNP for the sample.
- **B Allele Frequency:** Provides the theta value for a SNP, corrected for cluster position. Cluster positions are generated from a large set of normal individuals.
- **Log R Ratio:** Provides the log (base 2) ratio of the normalized R value for the SNP divided by the expected normalized R value.
- **CNV Value:** Provides an estimate of Copy Number at individual locus.
- **CNV Confidence:** Provides the level of confidence that the CNV value is correct, based on the CNV algorithm used.

35.3.1.2 Importing Illumina Files

You can export genotype information as a .txt file from GenomeStudio. This would require you to register the GeneSpring plug-in in the GenomeStudio. The following instructions will guide you in exporting genotype information from GenomeStudio:

1. Copy the plug-in from `INSTALLATION_DIRECTORY\app\Illumina\GX.Genotyping.Export.dll` to `Genomestudio\modules\BSGT\ReportPlugins\`. Installation Directory is the directory where GeneSpring has been installed.
2. Restart GenomeStudio.
3. Double click on a Project from the `Recent Projects` pane in the GenomeStudio `start page`. You can also create a project from the `New Projects` pane, running the project creation wizards. You can use these wizards to create different types of projects, e.g., Genotyping, Gene Expression, etc. (refer to GenomeStudio User Manual for more information).

   Note: The GeneSpring plug-in is only meant for `Genotyping` Projects.
4. Select the `Reports` option from the the `Analysis` menu.
5. Click `Report Wizard` option from the `Report` sub menu. This launches the `Report Type` wizard page.
6. Select `Custom Report` radio button, and `GeneSpring Exporter 1.0 from Strand Life Sciences` from the drop-down list, and click Next. This launches the second page of the `Sample Groups` wizard.
7. Select the sample groups you want to include in the report, and click Next. This launches the `Destination` wizard page.
8. Provide an `Output Path` (you can use the `Browse` button), and `Report Name`, and click Finish. This launches a `Progress bar` with information about the incremental progress in generating the Report. After the `Report` is written completely, you will see a message with an option to view the `Report`.

35.3.1.3 Custom Technology

You can create a technology name from the "New Illumina Association Analysis" wizard.

- Select `EnterNew` option from the `Select a technology name` drop-down box.
- Enter a name relevant for the experiment.

If you want to create an experiment with a Custom Technology, you can create a sample .txt file and run an "Illumina Association Analysis" Experiment. All technology and sample-specific columns should be present in this sample file (refer to section 35.3.1.1). The GType-Sample column is the only one which will be used hereafter in this experiment.

35.3.2 Affymetrix Association Analysis Experiment

You can create an Affymetrix Association Experiment from the Project menu (*Project* menu $\longrightarrow$ *New Experiment*. In this window, choose the analysis type as *Association* and the experiment type as Affymetrix Association Analysis

Refer to Creating a Copy Number Experiment section for the steps involved in the process.

Please note that the 'Parameter Type' during Experiment grouping is used for profile plots. In Association analysis we do not have profile plots. Hence, the 'Parameter Type' (as appears during experimental grouping) is not relevant to the Association Analysis experiments. The statistical analyses and SNP regression analyses only use the parameter values of the input interpretation and not the 'Parameter Type' of the input interpretation.

35.4 Quality Control

In **GeneSpring** you can filter samples for missing genotype calls, and run the birdseed report and EIGENSTRAT filter on samples.

- Filter samples by Missing values: Removes samples for which too many of the SNPs did not return a valid genotype.
- Run "Birdseed report" on samples: Displays a summary of results if the Birdseed genotype call algorithm is being used (for certain Affymetrix technologies; refer to Table Technology).
- EIGENSTRAT filter on samples: Identifies samples that are contributing heavily to stratification in the experiment.

35.4.1 Filter Samples by Missing Values

You can identify samples that have a higher proportion of missing genotype calls, across SNPs. A high proportion of missed calls for a given sample can indicate an anomalous or inferior genotyping process for the sample, and can make later analyses less valuable. It is therefore customary to exclude these samples in further analysis. You can set a cutoff for missingness(default value= 0.1), and the samples that have missingness rates greater than the cutoff value are removed.

Notes:

- You should choose an appropriate cutoff value to filter out samples that have higher missing genotype calls, and are less relevant for the experiment.
- You can remove or add back samples from the experiment using the Include/Exclude Samples button.

Input	• Active entitylist • Cutoff for missingness (default value= 0.1)
Process	• Calculates the proportion of missing genotype calls, combined across all SNPs for each sample. • Identifies the samples for which the proportion of missing genotype calls exceeds the specified cutoff.
Output	In the output, samples for which the proportion of missing genotype calls exceeds the specified cutoff are highlighted. These samples could potentially be excluded from further analysis.

Table 35.2: Summary of Steps: Filter Samples by Missing Values

Note: It is not advisable to apply this filter with settings that result in a very unequal number of cases and controls. For case--control association studies, best results are achieved with an approximately equal number of cases and controls.

35.4.2 Birdseed Report

You can launch a summary of the Birdseed algorithm's results if the experiment technology uses the algorithm (refer to Table Technology). It provides a quick glance at the input CEL files, at a more detailed level than just the genotype calls, which are used exclusively by the rest of the workflow. Refer to the table below for more information on each column.

Refer to the section on Birdseed Algorithm for more information on the Algorithm.

35.4.3 EIGENSTRAT Filter on Samples

The EIGENSTRAT filter allows you to remove the samples contributing heavily to the existing stratification in the experiment.

Stratification is an undesirable phenomenon that occurs when the samples used in the experiment can be bracketed into distinct groups that show different characteristics with respect to the condition (e.g., disease) being studied.

```
Suppose all cases were race A and all controls were race B, and race A was known
to have a far greater susceptibility  to the disease than race B. This situation
would confound the analysis, because any association could simply be attributed
to the race differences between cases and controls.  We say in this case that
the samples exhibit high stratification, and the racial difference is a source
of stratification.
```

A well-designed experiment would ideally have little to no stratification, but sources of stratification are often almost impossible to identify samples which contribute disproportionately to the experiment's stratification. If they are found, it is often advisable to exclude these samples from further analysis to drastically reduce the confounding effects of stratification.

The filter identifies the principal components that contribute most to the covariance between the samples. Those that contribute significantly more than expected are interpreted as `axes of stratification`. A hypothesis test called the Tracy-Widom (TW) test is used to measure whether the covariance contribution of a PC is significantly more than that of the succeeding PCs (arranged in order of decreasing eigenvalues).

The filter shows the z-score for each sample along the axes of stratification. The z-score is a measure of the outlier-ness of the sample. You can filter out the samples that have z-score magnitudes greater than a certain cutoff value. section).

The following steps will guide you to filter out samples that contribute heavily to the existing stratification:

1. Check if PC_1 is significant (i.e., p-value less than the cutoff). There is no stratification if PC_1 is not significant.
2. If PC_1 is significant, check PC_2, PC_3, PC_4 etc., in sequence for significance. Continue till you come across an insignificant PC.
3. If the first k PCs (i.e., $PC_1, PC_2, \ldots, PC_{k-1}, PC_k$) in step 2 are significant, but $(k+1)^{th}$ is not, then these PCs (i.e. PC1, PC2 ... PCk-1, PCk) are the axes of stratification.

   The EIGENSTRAT Filter parameters can be configured to:

   - Adjust the `Top PCs to display` slider to launch the plot with only the first k PCs.
   - The `TW p-value cutoff` slider can be moved appropriately to adjust the threshold for PC significance. If the TW p-values for all the PCs are above the cutoff, then

an information dialog box is displayed and the `3D PCA scores` and `z-scores` tabs are removed from the 'EIGENSTRAT filter' dialog. This is because there will be no `significant PCs` over which z-scores can be calculated. However, the cutoff can be raised to select the PCs and display computed z-scores for those PCs.

- In the `z-scores` tab of the bottom pane, the samples with z-scores whose magnitudes are greater than the z-score cutoff along any significant PC (in this case, the first k PCs) are highlighted. In **GeneSpring** the default z-score cutoff is set to the highest z-score among all samples for the significant PC. Because of this, no samples get highlighted in the default view. However, the user has the option to change the z-score cutoff value as desired. This would then highlight the outlier samples which can be considered for exclusion from further analysis. The z-score cutoff can be changed using the "Change cutoff" button. Z-score is a measure of the outlier-ness of the sample. Suppose the cutoff is 2, and the z-score of a certain sample along a certain PC is greater than 2, then that sample is an outlier.

  Note: A higher z-score implies a greater contribution to the existing stratification.

Restoring removed samples:
Run one of the QC filters (e.g., Filter samples by missing values, EIGENSTRAT filter on samples), and use the "Include/Exclude samples" button and then select the samples you want to restore.

EIGENSTRAT notes on minimum number of samples/SNPs for successful operation

- A very low number of samples (less than 30, in many situations) may lead to inferior performance. It is advisable to use Genomic Control in such situations.
- A relatively small number of SNPs (e.g. 5k) are enough to detect a stratification between significantly distinct sub-populations within a given sample set. However, a larger number of SNPs (e.g. 100k) would be required to differentiate closely related populations and around 20k SNPs are good enough for most studies.

35.4.3.1 EIGENSTRAT Filter dialog

The result screen has two panes (Refer to Figure 35.1). The top pane has three tab options, viz., Sample Covariances, Sample Covariance Plot, and 3D PCA Scores (default view). The bottom pane has four tab options, viz., Experiment Grouping, Eigenvalue vs PC plot, -log (TW p-value) vs PC plot (default view), and Z-scores.

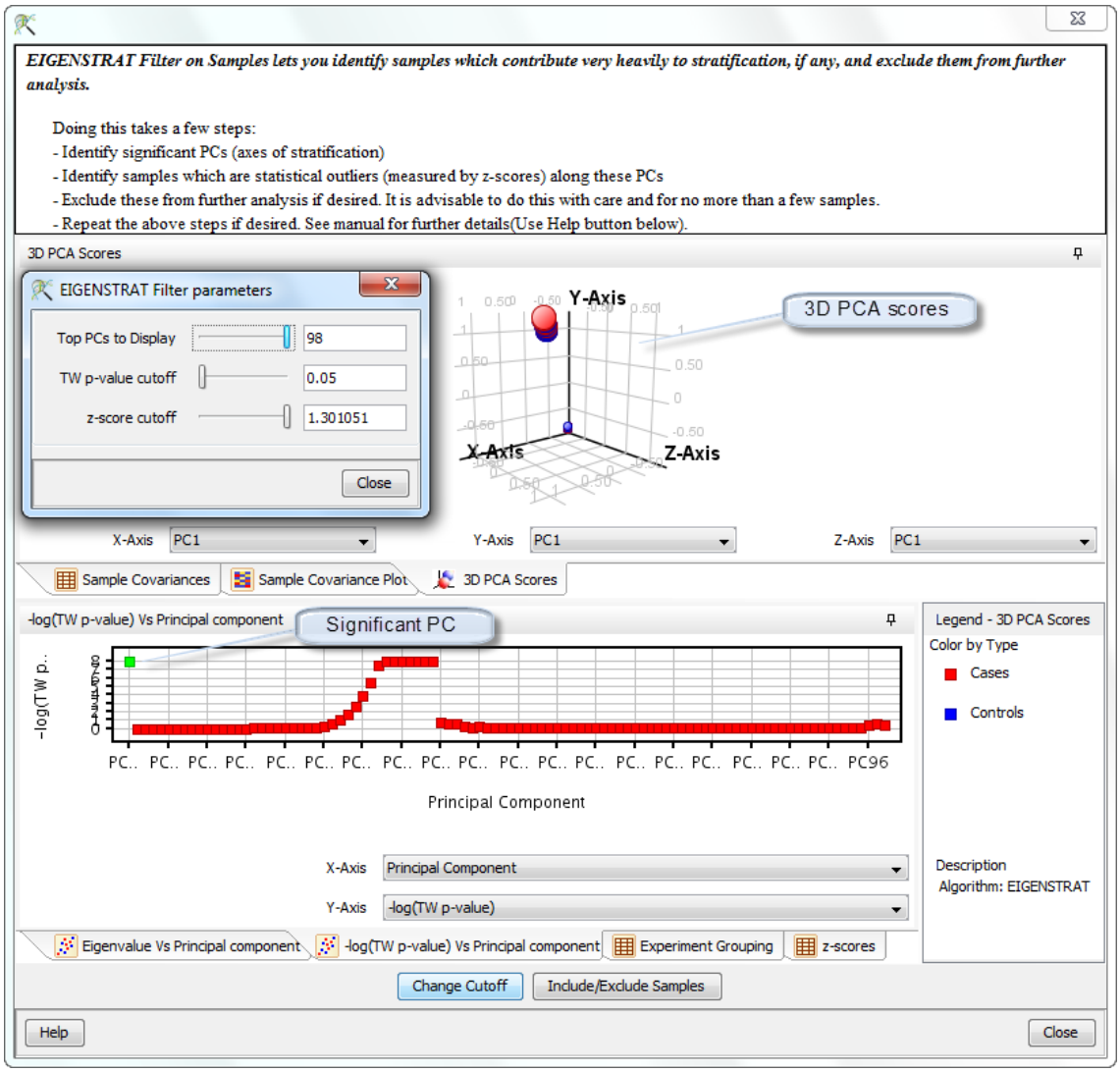


Figure 35.1: EIGENSTRAT Filter

Sample Covariance: Contains the covariance matrix ($X^T X$) of normalized genotype calls.

The covariance matrix is the product of the normalized genotype calls matrix (X) for n SNPs across p samples and its transpose matrix (X^T).

Sample Covariance Plot: Launches a heatmap of the covariance matrix. You can configure the color range from the Properties option on the context menu.

3D PCA Scores: Plots each sample as per the z-scores along the three chosen PCs.

Notes:

- You can select the PCs from the X-, Y-, and Z-axis drop-down boxes.
- You can select any sample from the plot canvas, and then filter it out using the **Include/Exclude Samples** option.
- You can zoom into the plot canvas using Shift + left-click + mouse hover from bottom to top. Press Shift + left-click + mouse hover from top to bottom to zoom out.
- You can rotate the plot canvas using Ctrl + left-click + mouse hover. The direction of mouse hover determines the axis, and rotation (clockwise/anti-clockwise).

Experiment Grouping: Shows the experiment grouping details.

Eigenvalue vs PC plot: Plots the eigenvalue for each PC.

$-\mathbf{log_{10}}(\mathbf{TWp-value})$ **vs PC plot:** Plots the transformed Tracy--Widom p-value for each PC.

z-score: Provides the z-score for each sample across all the significant PCs.

35.5 Filters

In **GeneSpring** you can Filter SNPs by Missing Value, Identify SNPs with Differential Missingness, and Filter SNPs by HWE p-value and MAF.

GeneSpring offers several filters for removing SNPs which are undesirable in some way from the analysis:

Filter SNPs by Missing Value:
Allows you to remove SNPs which show 'missing' genotype calls for an unacceptably high proportion of samples.

Identify SNPs with Differential Missingness:
Identifies SNPs for which missingness patterns seem to significantly differ between cases and controls.

Filter SNPs by HWE p-value:
Allows removal of SNPs whose genotype frequencies deviate too much from Hardy-Weinberg Equilibrium, a standard genotype frequency baseline.

Filter SNPs by MAF:
Allows removal of SNPs for which the genotype calls (across samples) are dominated too much by one allele to the exclusion of the other.

35.5.1 Filter SNPs by Missing Value

If a SNP has too many missing calls (i.e., the proportion is above a set cutoff, default value= 0.1), it is not worth analyzing. Here, you can filter out such SNPs. Once you set a cutoff value (default value= 0.1), the SNPs with missingness rates greater than the cutoff are identified and filtered out.

35.5.2 Identify SNPs with Differential Missingness

In **GeneSpring** you can identify SNPs that show a strong correlation between missingness and case--control status. A χ^2 (chi-square) test quantifies the significance of this correlation. This step displays SNPs with a p-value less than a specified cutoff (default value= 0.05).

`For a given SNP, if 95% of the missing genotypes occurred among the cases and only 5% among the controls, there could be good reason to suspect a correlation. Such SNPs could show a disease-dependent characteristic which interfered with the genotyping process.`

The significance of any correlation is measured using a χ^2 test. Refer to the table below.

Calculation

- Suppose the missing and non-missing genotype call counts for Cases and Controls are N_{MC}, N_{PC}, N_{MCont}, and N_{PCont} respectively.
- The expected occurrences of missing and non-missing genotype calls are given by the relations:

  $E_{MC} = (N_M \times N_C)/N$, $E_{PC} = (N_P \times N_C)/N$, $E_{MCont} = (N_M \times N_{Cont})/N$, and $E_{PCont} = (N_P \times N_{Cont})/N$.
- $\chi^2_{MC} = (N_{MC} - E_{MC})^2/E_{MC}$. Similary, χ^2_{MCont}, χ^2_{PC}, and χ^2_{PCont} are calculated.
- $\chi^2 = \chi^2_{MC} + \chi^2_{MCont} + \chi^2_{PC} + \chi^2_{PCont}$.

Note: 'Identify SNPs with Differential Missingness' establishes correlation between missingness and disease status; whereas 'Filter SNPs by Missing Value' calculates the aggregate missingness rate for each SNP, ignoring the disease status.

35.5.3 Filter SNPs by HWE p-value

In **GeneSpring** you can filter out SNPs that show a strong deviation from Hardy-Weinberg Equilibrium (HWE). A chi-square (χ^2) test quantifies the significance of this deviation. This filter

removes SNPs with a p-value less than the cutoff (default value: 0.001). You can configure the p-value cutoff from the **Change cutoff** button.

The HardyWeinberg principle states that both allele and genotype frequencies in a population remain constant, i.e., they are in equilibrium, from generation to generation unless specific disturbing influences are introduced. Many tests assume HWE or perform best when it is followed, and SNPs which show large deviations from it are often pathological or anomalous. Most SNPs in large experiments will not deviate significantly from HWE, and any SNP which does is generally considered unsuitable for further analysis. In this step, you can filter out such SNPs.

A χ^2 test is used to measure the significance of any deviation. The details are below.

Notes: When a SNP has the same genotype for all the samples HWE filter automatically sets the p-value to 0; the SNP clearly violates HWE and should always be filtered out.

35.5.3.1 Calculation:

- Suppose the three genotype call counts are N_{AA}, N_{AB}, and N_{BB} respectively.
- Let,

$$f_A = \frac{2N_{AA} + N_{AB}}{2(N_{AA} + N_{AB} + N_{BB})}.$$

- Then the expected occurrences of the each of the three genotypes are given by the relations: $E_{AA} = f_A^2(N_{AA} + N_{AB} + N_{BB})$, $E_{AB} = 2f_A(1 - f_A)(N_{AA} + N_{AB} + N_{BB})$, and $E_{BB} = (1 - f_A)^2(N_{AA} + N_{AB} + N_{BB})$.
- $$\chi^2_{AA} = \frac{(N_{AA} - E_{AA})^2}{E_{AA}}.$$

  Similarly, χ^2_{AB} and χ^2_{BB} are calculated.
- $$\chi^2 = \chi^2_{AA} + \chi^2_{AB} + \chi^2_{BB}.$$

35.5.4 Filter SNPs by MAF

The MAF (Minor Allele Frequency) filter removes the SNPs with minor allele frequencies less than a specified cutoff value (default value= 0.01).

For a given biallelic SNP, it is desirable not to have a very large imbalance between the frequencies of the two alleles. If for instance the lower of the two frequencies (i.e., the minor allele frequency) is too low, many tests may give untrustworthy results. Note that we use no external definitions to select the minor allele, instead defining it as the allele with the lower frequency using only the data in the experiment (ignoring missing values).

Calculation:

- The sum of the two counts is an even number because the missing calls occur in pairs.
- The frequency (f_A) of A is $N_A/(N_A + N_B)$, and B (f_B) is $N_B/(N_A + N_B)$; where N_A and N_B are the two counts.
- Min(f_A, f_B) is the MAF.

Recommendations: MAF is automatically run for SNP tagging and LD analysis. You are advised to run the MAF filter before any of the analysis steps.

35.6 Analysis

In **GeneSpring** you can perform EIGENSTRAT Correction on Samples, Statistical Analysis, SNP Tagging, SNP Regression, Haplotype Trend Regression, and LD Analysis.

35.6.1 EIGENSTRAT Correction on Samples

EIGENSTRAT Correction on Samples allows you to correct for stratification (refer to section 35.4.3) in the experiment.

The method is very similar to that used in 35.4.3 section; the matrix of genotype calls (dimension: number of SNPs × number of samples) is subjected to PCA similar to section 35.4.3 The EIGEN values are found using the same method as in section 35.4.3, but are then used to modify the genotype data. The modified data will contain all salient features of the original genotype data, except that it will not show significant stratification. It can therefore be used in later analysis steps without fear of stratification-related confounding effects.

Notes:

- The EIGENSTRAT corrected data generated here cannot be used in some of the statistical tests offered under the Association Analysis workflow. This is because the genotype calls have been changed to decimal numbers and therefore cannot be treated as conventional genotypes in some ways.
- The phenotypes are also modified by this method, so that they become floating-point values, like the modified genotypes.

The following steps will guide you in using the EIGENSTRAT Correction on Samples:

Input Parameters (Step 1) Select the Mode of Inheritance (default: Additive; refer to section 35.6.1.1) for EIGENSTRAT Correction.

- You can select any of the three modes, viz., Additive, Recessive, and Dominant from the drop-down box.
- You can add a new mode, and then enter the weight factors for each of the genotype calls, viz., AA, AB, and BB. (*Tools* $\longrightarrow$ *Options* $\longrightarrow$ *Miscellaneous* $\longrightarrow$ *Mode of Inheritance*; refer to section 35.6.1.1)

PC Selection (Step 2) The screen has two panes: refer to figure below. The top pane has two options, viz., Eigenvalue vs PC and TW p-value vs PC Plots (default view).

The bottom pane has three options, viz., Sample Covariances, Sample Covariance Plot, and Experiment Grouping (default view).

The EIGENSTRAT correction follows the same procedure as EIGENSTRAT filter for selecting significant PCs, please refer to 35.4.3 section.

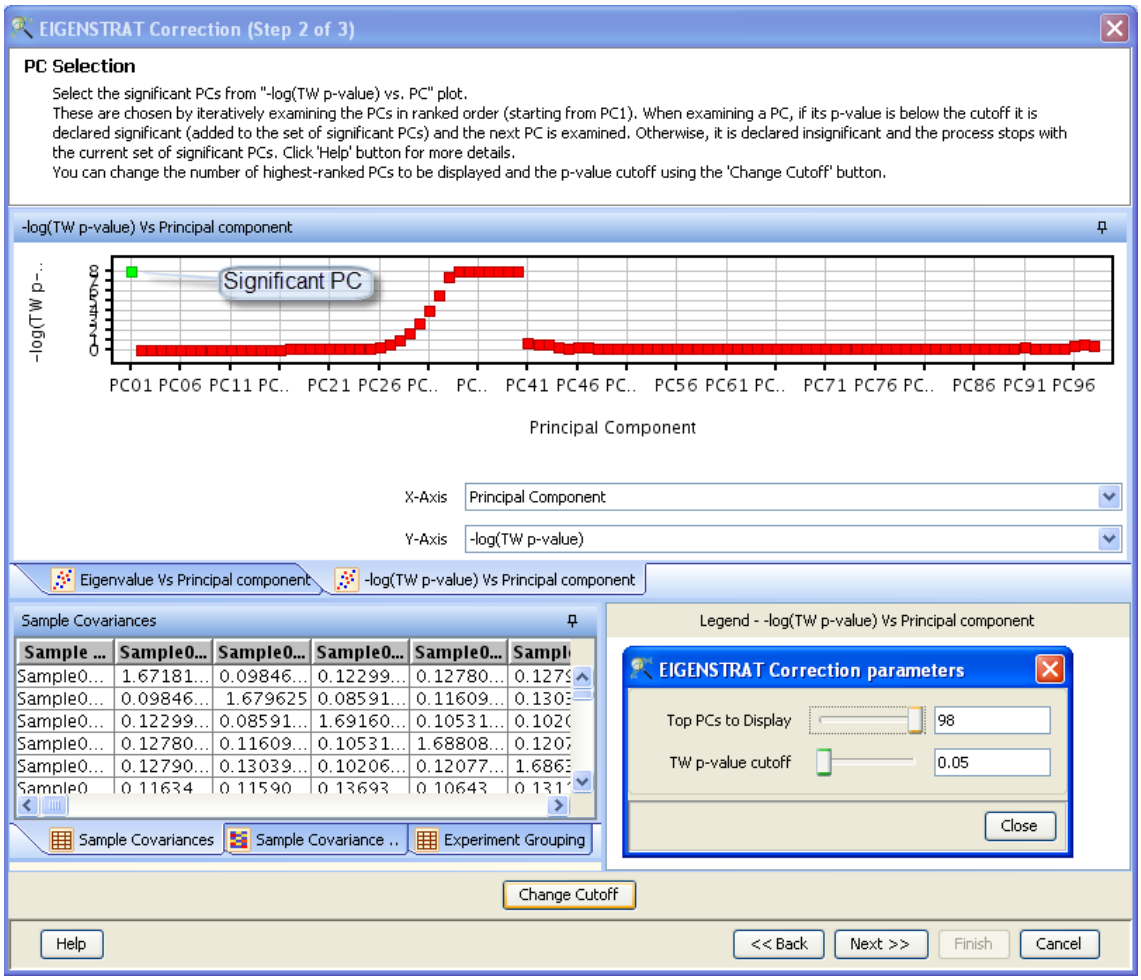


Figure 35.2: EIGENSTRAT Correction View

Eigenvalue vs PC plot:
Plots the eigenvalue for each PC.

$-\log_{10}(\mathbf{TWp-value})$ vs PC plot:
Plots the transformed Tracy--Widom p-value for each PC.

Sample Covariances:

Contains the covariance matrix ($X^T X$) of normalized genotype calls.

The covariance matrix is the product of the normalized genotype calls matrix (X) for n SNPs across p samples and its transpose matrix (X^T).

Sample Covariance plot:

Launches a heatmap of the covariance matrix. You can configure the color range using *Context (right-click)* menu $\longrightarrow$ *Properties*.

Experiment Grouping:

Shows the experiment grouping details.

Results (Step 3) The window has two panes (Figure 35.3); the top pane launches the 'Corrected Dataset' of genotype calls, and the bottom panes display SNP-wise and Sample-wise correction. These two are summary statistics measuring the magnitude of the correction on the data, viewed SNP-wise and sample-wise respectively.

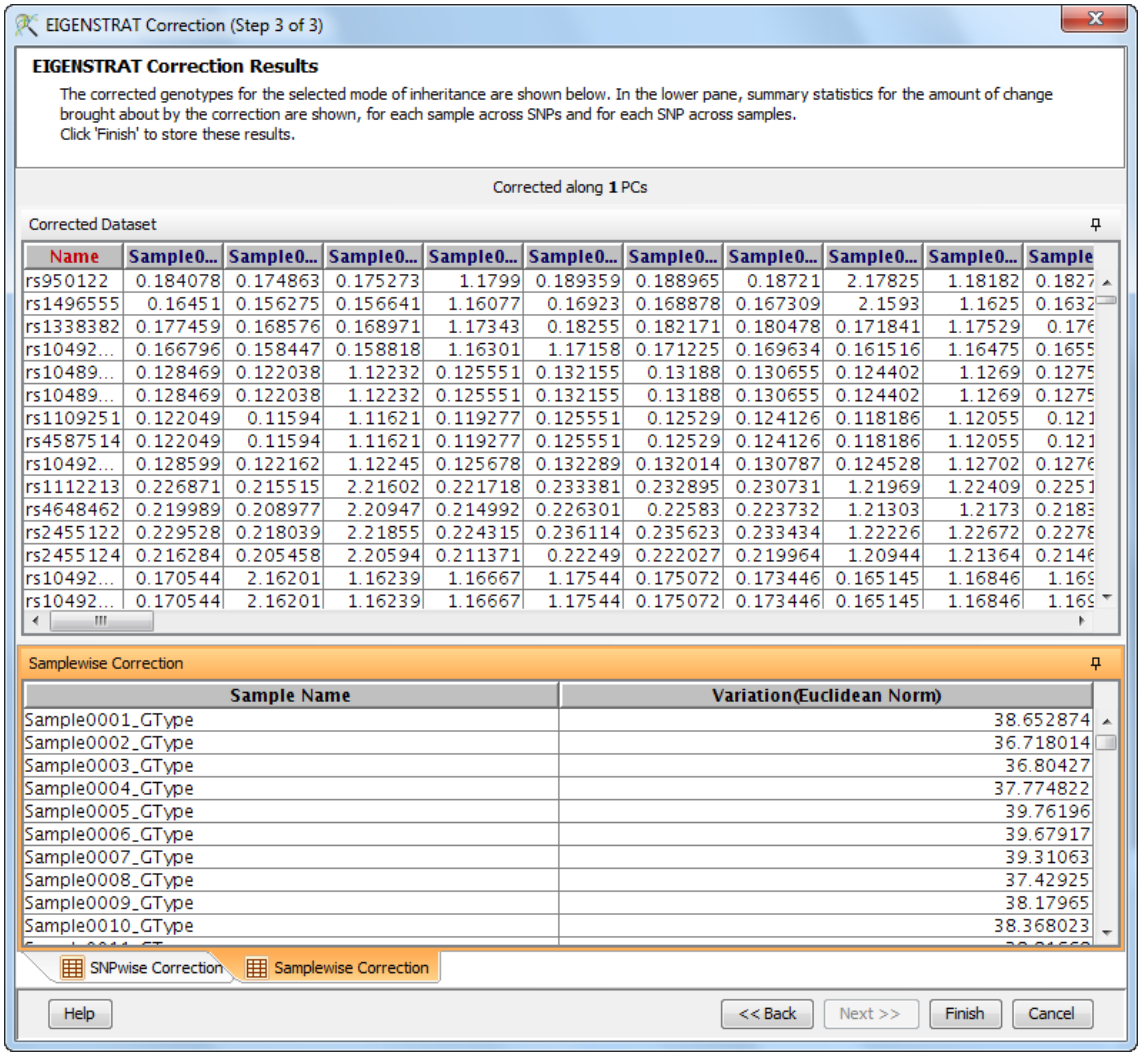


Figure 35.3: EIGENSTRAT Correction Results

Calculation:

- Let $G_{1k}, G_{2k}, \ldots, G_{nk}$ be the genotype calls for n SNPs in the k^{th} sample.
- Let $C_{1k}, C_{2k}, \ldots, C_{nk}$ be the corrected genotype calls for n SNPs in the k^{th} sample.
- Then, (Sample-wise) correction for the k^{th} sample $= \sqrt{\sum_{i=1}^{n}(C_{ik} - G_{ik})^2}$.
- Similarly,(SNP-wise) correction for the j^{th} SNP $= \sqrt{\sum_{i=1}^{p}(C_{ji} - G_{ji})^2}$.

35.6.1.1 Mode of Inheritance

The mode of inheritance for the condition being studied is the relationship between the expressed phenotype and the (possibly unknown) phenotype-associated SNP(s) genotype. For example, in a "dominant" mode of inheritance, the phenotype would be expressed whenever the causal allele was present in the associated SNP, regardless of homo/heterozygosity. In contrast, a recessive mode would result in an expressed phenotype only when the SNP was homozygous in the causal allele, and an additive mode would result in a phenotype expression proportional to the number of causal alleles present. All this can be quantified using relative weights as in Table 35.12. This table can be augmented with your own custom modes (*Tools* $\longrightarrow$ *Options* $\longrightarrow$ *Miscellaneous* $\longrightarrow$ *Mode of Inheritance*)

35.6.2 Statistical Analysis

The tests in this section aim to answer the central question of association analysis: *which SNPs are associated with expression of the phenotype being studied?* They are all statistical hypothesis tests. The calculation details are in the tables below. Some tests have additional options which may be selected:

- Mode of Inheritance: Can be selected only if the test's calculation provides for a flexible assignment of numbers to genotypes
- EIGENSTRAT corrected data: Possible only if the test makes no assumption that genotypes/phenotypes must be integer-valued
- Apply Genomic Control: Possible only if the test is a χ^2 test
- Permutative p-value computation: Possible for all those that are not exact tests
- Multiple testing correction: Always possible, because every test generates a p-value

Brief descriptions of the tests are as follows.

Pearson's χ^2 test:
Tests whether there is a significant correlation between the allele counts of a SNP and case--control status. It can only be applied to case--control experiments, and cannot be applied to EIGENSTRAT-corrected data. For some SNPs, this automatically gets replaced by Fisher's exact test (refer to note in section 35.6.2.7, and section 35.6.2.6). It is possible to run genomic control on the χ^2 test.

Note: If any of the counts, $E_{ij} \leq 10$ (What is E_{ij}?) for a SNP, **GeneSpring** calculates the p-value using the Fisher's exact test.

Fisher's exact test:
A more accurate version of Pearson's χ^2 test retaining all the characteristics of that test. Genomic control cannot be run on this. Neither can permutation testing.

Cochran-Armitage test:
Tests whether there is a significant correlation between the genotype counts of a SNP and case--control status. It can only be applied to case--control experiments, and cannot be applied to EIGENSTRAT-corrected data. It is possible to run genomic control on this.

χ^2 correlation test:
Tests whether there is a significant correlation between the genotypes and phenotypes across samples at a given SNP. It can be applied to any data, including EIGENSTRAT-corrected data. It is possible to run genomic control on this.

Notes: When a SNP has the same genotype for all the samples all the statistical analysis tests would set the p-value to 1; such a SNP cannot be associated with any trait

- **Entity List and Interpretation (Step 1)** association.filterhwe

  Select an Entity List and Interpretation from the Entity List and Interpretation dialog, and click Next. This opens the 'Select Test' dialog.

- **Select Test (Step 2)**

  Select a test, the cases, and controls from the respective drop-down lists, and click Next. This opens the Select Inputs dialog.

- **Select Inputs (Step 3)**

  - Select a Mode of Inheritance (default option: Additive; refer to Table 35.12) from the drop-down list.

    Notes: You cannot select the Mode of Inheritance for χ^2 and Fisher's Exact tests.

  - You can select the 'Use EIGENSTRAT Corrected Data' check box to run χ^2 Correlation test with the data corrected for stratification.

    Notes: This option is only available for χ^2 Correlation test.

  - Select a method for p-value Computation (default option: Asymptotic).

    Notes: You can enter the Number of Permutations (default value: 100) for the Permutative option.

  - You can select a Multiple Testing Correction method (default: No Correction; refer to section 35.6.2.3), and click Next. This opens a list of significant SNPs and the respective p-values [6], [60].

– You can select the Genomic Control (refer to section 35.6.2.1) check box for the Pearson's χ^2, χ^2 Correlation, and Cochran-Armitage tests [49].

Notes: This option is not available for Fisher's exact test.

- **Significant SNPs (Step 4)** Significant SNPs dialog launches a list of Significant SNPs with their respective p-values.

Notes: Click on the Change Cutoff button to change the p-value cutoff from the slider or text box.

35.6.2.1 Genomic Control

Genomic Control [49] is a standard method to compensate for population stratification (refer to section 35.4). It is a modification to the calculation of a χ^2 test. When there is stratification present in the samples, the test statistics (χ^2 tests) for all the SNPs get inflated by some constant factor called the Variance Inflation Factor (VIF) (`> 1`); the less the stratification, the closer the factor is to 1. Genomic control attempts to measure this factor and downscale all calculated χ^2 statistics by this factor to mimic what the results would have been without stratification.

The χ^2 statistic is calculated for each SNP, and we take the mean of these values to evaluate the VIF. Since many of the SNPs may have χ^2 statistics `< 1`, and so the mean can be `< 1`, there can and will be situations when Eigenstrat will identify stratification, but GC will not.

Given an entity list, a VIF is calculated and the SAME factor is applied to all the loci in the list,i.e., the VIF is constant across the genome or entity list, whichever was chosen. We do not assume that the VIF follows a distribution.

GeneSpring offers two approaches to deal with stratification: EIGENSTRAT correction and Genomic Control. EIGENSTRAT is the more rigorous and preferable method in most GWAS situations. However, there are two situations in which Genomic Control may be preferred:

- If the amount of data is too small to confidently run EIGENSTRAT (refer to section 35.4.3).
- If the Pearson's χ^2 or Cochran-Armitage test is desired (since these do not run on EIGENSTRAT corrected data, but do accept Genomic Control).

35.6.2.2 Permutative p-value Computation

All the tests (except for Fisher's exact test) offered here make standard statistical assumptions that approximate the exact answer better when run with more data. However, the results are only asymptotically accurate. The accuracy is comprised in the case of small- to moderate-sized datasets. Permutative testing solves this; it does not hinge on statistical assumptions and essentially estimates the probabilities involved through repeated simulation. The simulation can sacrifice computing power for accuracy.

Number of Permutations: It determines how accurate the answer is. The default (100) is generally acceptable. The higher the setting, the more accurate the answer, but longer is the time taken on a machine.

35.6.2.3 Multiple Testing Correction

Multiple Testing Correction (MTC) is a statistical problem in GWAS that is addressed using one of a family of methods that **GeneSpring** provides.

- More stringent Multiple Testing Correction (MTC) methods declare fewer SNPs significant, but have a lower risk of false positives. **GeneSpring** offers MTC methods with varying levels of stringency, which depends on the data to some extent. A popular ordering of these methods from most to least conservative is as hereunder:
  1. Benjamini-Yekutieli (BY)
  2. Bonferroni
  3. Benjamini-Hochberg (BH)
  4. Storey's q-value
- Less stringent methods are recommended for first-stage analysis to avoid false negatives. BY, BH, and Storey's (Storey's in particular) perform better when there are more truly significant SNPs.

35.6.2.4 Trait Types: Quantitative and Categorical

GeneSpring implicitly classifies a trait as either categorical or quantitative. Quantitative traits are those for which the interpretation parameter value is numeric for all samples. Categorical traits are all others.

The categorical trait is further classified as Binary, Nominal and Ordinal traits. Before running certain statistical tests like χ^2 Correlation, Categorical traits require a Nominal/Ordinal trait type selection to be done. Depending on the chosen trait type, numerical values are assigned to the input conditions.

When the input trait type is Categorical and there are more than 2 conditions for a parameter, then the user is required to select either of the two trait types before running any further analysis. However, if there are only two conditions, for instance, population A and B (Cases and Controls), it is taken automatically as a binary trait and the option of other trait types do not show up. A binary trait type will arbitrarily assign a value of '0' and '1' to the cases and controls.

35.6.2.5 Nominal and Ordinal Trait Types

Nominal:

Nominal trait type is where the conditions of a given trait cannot ordered. For example,

eye color is a trait with multiple labels: black, brown, hazel, green, blue etc. that cannot be ordered as each of the colors represent an independent condition. However, a standard phenotype or a reference category can be chosen from any one of the above conditions to which the remaining conditions can be compared with. For instance, if black color is chosen as a reference phenotype, it will be assigned a value of '0' and for the remaining conditions(colors), the values would be arbitrarily assigned as '1','2' and '3'. Suppose there are only two conditions i.e. cases and controls, then the nominal trait type will behave like a binary trait type.

Ordinal:

In this trait type, the numbers assigned to objects or events represent the rank order (1st, 2nd, 3rd, etc.) of the entities assessed. An example of ordinal measurement would be any disease state classified as Mild, Moderate, or Severe. If the user selects the Moderate $\longrightarrow$ Mild $\longrightarrow$ Severe as the order from top to bottom, then Moderate will be assigned the value '0', Mild will be given the value '1' and Severe will be assigned '2'. Note that in the nominal scale there is no order implied. A ''Moderate-Mild-Severe" attribute would be best assigned as Ordinal because reordering the attribute values(e.g., to ''Mild-Severe-Moderate") makes a difference.

35.6.2.6 Pearson's χ^2 Test

GeneSpring allows you to perform χ^2 test based on disease status of each sample, and genotype information for all SNPs, across samples.

Calculation

- Suppose the allele A and B counts for Cases and Controls are N_{ACase}, N_{BCase}, N_{ACont}, and N_{BCont} respectively.
- The expected occurrences of of the allele counts for A and B alleles are given by the relations: $E_{ACase} = (N_A \times N_{Case})/N$, $E_{BCase} = (N_B \times N_{Case})/N$, $E_{ACont} = (N_A \times N_{Cont})/N$, and $E_{BCont} = (N_B \times N_{Cont})/N$.
- $\chi^2_{ACase} = (N_{ACase} - E_{ACase})^2/E_{ACase}$. Similarly, χ^2_{ACont}, χ^2_{BCase}, and χ^2_{BCont} are calculated.
- $\chi^2 = \chi^2_{ACase} + \chi^2_{ACont} + \chi^2_{BCase} = \chi^2_{BCont}$

Note: If any of the counts, $E_{ij} \leq 10$ for a SNP, **GeneSpring** calculates the p-value using the Fisher's exact test.

35.6.2.7 Fisher's Exact Test

GeneSpring allows you to perform Fisher's exact test based on disease status of each sample, and genotype information for all SNPs, across samples. Fisher's exact test is typically not used for a relatively large number of samples. There are two reasons for this:

- Fisher's exact test can take a significant amount of computing power if the number of samples is very high. The definition of ''very high" depends on your machine specification.
- Pearson's χ^2 test becomes asymptotically identical to Fisher's exact test for a large number of samples

35.6.2.8 Cochran-Armitage Test

GeneSpring allows you to perform the Cochran-Armitage test for trend. This tests if there is a correlation between genotype counts for a given SNP and case--control status. More specifically, it tests for a linear trend between the genotype and phenotype.

35.6.2.9 χ^2 Correlation Test

GeneSpring allows you to perform χ^2 correlation test. This statistically tests the value of Pearson's correlation coefficient (r) between genotypes and phenotypes.

Column	Description
cel files	CEL file name.
computed gender	Estimated gender.
call rate	Genotype call rate at the default or user-specified threshold.
het rate	Percentage of "AB" genotype call (i.e., the heterozygosity).
hom rate	Percentage of "AA" or "BB" genotype call (i.e. the homozygosity).
cluster distance mean	Average distance to the cluster center for the genotype.
cluster distance stdev	Standard deviation of the distance to the cluster center for the genotype.
raw intensity mean	Average of the raw PM probe intensities.
raw intensity stdev	Standard deviation of the raw PM probe intensities.
allele summarization mean	Average of the allele signal estimates (log2 scale).
allele summarization stdev	Standard deviation of the allele signal estimates (log2 scale).
allele deviation mean	Average of the absolute difference between the log (base 2) transformed allele signal estimate and its median across all chips.
allele deviation stdev	Standard deviation of the absolute difference between the log transformed (base 2) allele signal estimate and its median across all chips.
allele mad residuals mean	Average of the median absolute deviation (MAD) between observed probe intensities and probe intensities fitted by the model.
allele mad residuals stdev	Standard deviation of the median absolute deviation (MAD) between observed probe intensities and probe intensities fitted by the model.
em-cluster-chrX-het-contrast gender	Gender estimate based on estimated heterozygosity on chrX.
em-cluster-chrX-het-contrast gender chrX het rate	Estimated heterozygosity on chrX.
cn-probe-chrXY-ratio gender meanX	Average intensity of chrX CN probes.
cn-probe-chrXY-ratio gender meanY	Average intensity of chrY CN probes.
cn-probe-chrXY-ratio gender ratio	Ratio of average chrY CN probe intensity to average chrX CN probe intensity.
cn-probe-chrXY-ratio gender	Gender estimate based on ratio of chrY to chrX average CN probe intensities.

Table 35.3: Birdseed Report

Input	• Active entitylist • TW P-value cutoff for the Principal Components • Z-score cutoff for the samples (default value: maximum z-score across all samples and PCs) • Number of top PCs to display (default value: p-2; where p is the number of samples)
Process	• Identifies the directions of stratification, and groups the samples. • Finds samples that contribute heavily to the existing stratification. Computes the Z-scores for each sample along the axes of stratification.
Output	Samples which have absolute value of Z-score higher than the cutoff are highlighted.

Table 35.4: Summary of Steps: EIGENSTRAT Filter

Input	• Entity List • Cutoff for missingness (default value= 0.1) **Note:** You should choose an appropriate cutoff value to filter out SNPs that have higher missing genotype calls, and are less relevant for the experiment.
Process	• Calculates the proportion of missing genotypes (combined across samples). • Counts the number of missing genotype calls added over all samples. **Note:** The missing calls occur in pairs, i.e., either both or none of the alleles for a SNP are present. Only one cannot be missing. The number of missing calls is an even number. • Divides the missing calls by twice the number of samples, which is the total number of allele calls for the SNP. The result is the missing call proportion. • Filters out the SNP if the proportion is above the cutoff.
Output	List of SNPs to be considered for further analysis.

Table 35.5: Summary of Steps: Filter SNPs by Missing Value

Input	• Genotype information for all SNPs, across samples • Disease Status of each sample • P-value cutoff (default value= 0.05)
Process	• Calculates the missing genotypes for each SNP, combined across samples. • Counts the number of missing genotype calls added over all samples. • Calculates the χ^2 statistic for each SNP. • Finds the p-values, given the χ^2 statistic is distributed as chi-square with one degree of freedom. • Identifies the SNP if the p-value is less than the cutoff.
Output	List of SNPs where there is a correlation between missingness and case--control status.

Table 35.6: Summary of Steps: Identify SNPs by Differential Missingness

	Missing	**Non-Missing**	**Total**
Cases	N_{MC}	N_{PC}	N_C
Controls	N_{MCont}	N_{PCont}	N_{Cont}
Total	N_M	N_P	N

Table 35.7: Contingency Table for Differential Missingness

Input	• Genotype information for all SNPs, across samples • p-value cutoff (default value= 0.0010)
Process	• Separately counts the number of occurrences of the genotypes AA, AB, and BB, across all the samples. • Calculates the χ^2 test statistic for each SNP. • Finds the p-values (χ^2 distribution with one degree of freedom). • Filters out the SNP if the p-value is below the cutoff.
Output	List of SNPs to be considered for further analysis.

Table 35.8: Summary of Steps: Filter SNPs by HWE p-value

Input	• Genotype information for all SNPs, across samples • MAF cutoff (default value= 0.01)
Process	• Calculates the MAF across all the samples. • Counts the occurrences of A and B separately across all the samples; A and B are the alleles for the SNP. • Filters out SNPs with MAF less than the cutoff.
Output	List of SNPs to be considered for further analysis.

Table 35.9: Summary of Steps: Filter SNPs by MAF

Input	• Active entity list • Mode of Inheritance (default option: Additive; refer to section 35.6.1.1) for details
Process	• Identifies the principal axes of variation for the data. • Corrects the data to reduce the variations along the principal axes.
Output	Corrected genotype information for all the SNPs, across samples.

Table 35.10: Summary of Steps: EIGENSTRAT Correction on Samples

Corrected Dataset	Provides the genotype calls matrix with the corrected data.
Sample-wise Correction	Calculates the differences between the uncorrected and corrected datasets for each sample, and then quantifies the extent of correction. This is done by taking the Euclidean norm of the difference between the uncorrected and corrected vectors for each sample.
SNP-wise Correction	Calculates the differences between the uncorrected and corrected datasets for each SNP, and then quantifies the extent of correction. This is done by taking the Euclidean norm of the difference between the uncorrected and corrected vectors for each SNP.

Table 35.11: EIGENSTRAT Correction Result Screen

Mode	**AA**	**AB**	**BB**
Additive	0	1	2
Recessive	0	0	1
Dominant	0	1	1

Table 35.12: Mode of Inheritance

Input	• Genotype information for all SNPs in the given (input) entity list, across samples • Interpretation with two conditions • Genomic Control (optional; refer to section 35.6.2.1) • p-value Computation method (default option: Asymptotic) • Multiple Testing Correction (optional; refer to section 35.6.2.3)
Process (for each SNP)	• Calculates the allele counts for each SNP, combined across samples. • Calculates the χ^2 statistic for each SNP. • Finds the p-values, given the χ^2 statistic is calculated with one degree of freedom. • Filters out the SNP if the p-value is above the cutoff.
Output	List of SNPs to be kept with respective p-values to be considered for further analysis.

Table 35.13: Summary of Steps: Pearson's χ^2 Test

	Allele A	**Allele B**	**Total**
Cases	N_{AC}	N_{BC}	N_C
Controls	N_{ACont}	N_{BCont}	N_{Cont}
Total	N_A	N_B	N

Table 35.14: Contingency Table for Pearson's χ^2 Test

Input	• Genotype information for all SNPs in the given (input) entity list, across samples • Interpretation with two conditions • Multiple Testing Correction (optional; refer to section 35.6.2.3)
Process (for each SNP)	• Calculates the allele counts for each SNP, combined across samples. • Calculates the allele count for each SNP, combined across samples. • Filters out the SNP if the p-value is above the cutoff.
Output	List of SNPs to be kept with respective p-values to be considered for further analysis.

Table 35.15: Summary of Steps: Fisher's Exact Test

Input	• Genotype information for all SNPs in the given (input) entity list, across samples • Interpretation with two conditions • Mode of Inheritance (default option: Additive; refer to section 35.6.1.1 section) • Genomic Control (optional; refer to section 35.6.2.1) • Multiple Testing Correction (optional; refer to section 35.6.2.3)
Process (for each SNP)	• Calculates the Genotype call counts for each SNP, combined across samples. • Calculates the test statistic for each SNP using χ^2 distribution (one degree of freedom). • Finds the p-value for each SNP. • Filters out the SNP if the p-value is above the cutoff.
Output	List of SNPs to be kept with respective p-values to be considered for further analysis.

Table 35.16: Summary of Steps: Cochran-Armitage Test

	Genotype AA	Genotype AB	Genotype BB	Total
Cases	r_0	r_1	r_2	r_+
Controls	s_0	s_1	s_2	s_+
Total	n_0	n_1	n_2	N

Table 35.17: Contingency Table for Cochran-Armitage Test

Mode	Genotype AA (d_0)	Genotype AB (d_1)	Genotype BB (d_2)
Additive	0	1	2
Dominant	0	1	1
Recessive	0	0	1

Table 35.18: Weights (d_i) for Cochran-Armitage Test

Input	• Genotype information for all SNPs in the given (input) entity list, across samples • Interpretation with two or more conditions • If trait type is categorical, then ordinal or nominal option can be selected (refer to section 35.6.2.4). • Mode of Inheritance (default option: Additive; refer to section 35.6.1.1) • Genomic Control (optional; refer to section 35.6.2.1) • Multiple Testing Correction (optional; refer to section 35.6.2.3) • EIGENSTRAT corrected data (optional)
Process (for each SNP)	• Performs correlation test between the SNP data and the traits for each SNP, across samples. • Generates a $\chi^2 statistic$. • Finds the p-value for each SNP. • Filters out the SNP if the p-value is above the cutoff.
Output	List of SNPs with respective p-values to be considered for further analysis.

Table 35.19: Summary of Steps: χ^2 Correlation Test

Calculation:
Let x_i and y_i be the genotype and trait for the i^{th} sample.
Test statistic (uncorrected data) =

$$(n-1)\frac{\left(\sum_{i=1}^{n} x_i y_i - \left(\frac{\sum_{i=1}^{n} x_i \sum_{j=1}^{n} y_j}{n}\right)\right)^2}{\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)\left(\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right)}$$

Let k be the number of PCs along which the data has been corrected.
Test statistic (corrected data) =

$$(n-k-1)\frac{\left(\sum_{i=1}^{n} x_i y_i - \left(\frac{\sum_{i=1}^{n} x_i \sum_{j=1}^{n} y_j}{n}\right)\right)^2}{\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)\left(\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right)}$$

35.6.3 SNP Tagging

Tag SNPs are representative SNPs in a region of the genome with high linkage disequilibrium. They serve as maximally informative markers of the input entity list (SNPs) and often only comprise of a small fraction of the originally input SNP list. Tag SNPs can be used on behalf of the original set of SNPs in other analyses without losing much accuracy. Also, using the genotype value of the tag SNPs, the genotype values of the remaining SNPs called the non tag SNPs can be predicted with good accuracy.

Carlson's greedy algorithm [13] is being run to identify the tag SNPs. The main parameters associated with this algorithm are described hereunder:

MAF Cutoff:
The tagging algorithm can give unpredictable results if any allele frequency is too low. To prevent this, the algorithm has a self-contained MAF filter exactly like the one mentioned in the "Filters" section (35.5.4).

LD cutoff:
r^2 is a commonly used measure of correlation between two SNPs. It can range from 0 to 1, with better correlation being a higher r^2. Our tagging method can guarantee that any non-tag SNP in the input list will have an r^2 with some tag SNP of at least this LD cutoff. So every non-tag SNP will be guaranteed to be represented in the tag SNP list, to an extent measured by this cutoff. A higher cutoff results in more tag SNPs (higher cost of analysis), which represent the data better, i.e., the information loss from excluding the non-tag SNPs will be lower. A lower cutoff causes the opposite.

Given a pair of biallelic markers A (with alleles A_o and A_1) and B (with alleles B_o and B_1),

let f_{ij} denote the frequency of the haplotype A_iB_j, where, i, j $\in$ {0,1}. Then, $f_{i.}$=$\sum_{j=0}^{1} f_{ij}$ and $f_{.j}$=$\sum_{i=0}^{1} f_{ij}$ denote the marginal frequencies of allele A_i at locus A and allele B_j at locus B, respectively. The LD measure r^2 is defined by [28]

$$r^2 = \frac{(f_{00} - f_{0.}f_{.0})^2}{f_{0.}f_{.0}f_{1.}f_{.1}}$$

Distance Threshold:

Decides how far away a non-tag SNP can be from the tag SNP that represents it. A higher value here can reduce the number of tag SNPs marginally but will increase the time taken, sometimes drastically. A lower value does the opposite. A distance of 50 represents the number of SNPs on either side of the SNP being queried. It takes into account the 50 SNPs as per the entity list being used for the tagging, since those are the SNPs it is using to compute the correlation. Thus, if the input list for tagging is $(e_1, e_2, e_3, \ldots, e_100)$,then for e_1, r^2 is computed with a list of SNPs, $E_1(e_2, e_3, \ldots, e_50)$; similarly for e_50, r^2 is computed with a list of SNPs, $E_50(e_1, e_2, \ldots, e_49 and e_51, e_52, \ldots, e_99)$.

35.6.3.1 Summary of Steps

This section provides a summary of steps involved in SNP Tagging.

Input

- Entity List of SNPs
- MAF cutoff (default value= 0.01)
- LD (r^2 cutoff (default value= 0.8)
- Distance Threshold (default value= 50)

Process

- Calculates the MAF for each SNP across all samples.
- Filters out SNPs with MAF below the cutoff.
- Computes r^2 between the SNPs for the distance threshold specified.
- Executes $Carlson's$ algorithm to find the bins and their representative Tag SNPs [13].

Output

The tag SNPs, a subset of all input SNPs, and their corresponding bins (non-tag SNPs).

Notes:

- SNP Tagging doesn't run on sex chromosomes.
- SNP Tagging ignores SNPs that do not have chromosomal location(unknown chromosomes).

35.6.4 SNP Regression

GeneSpring allows you to perform linear and logistic regression of sample genotypes against phenotypes. You can regress each SNP independently(single SNP) or multiple SNPs simultaneously (none of the individual SNPs in a gene might meet the significance threshold, for example, but the block might when tested as one block).

35.6.4.1 Summary of Steps:

This section provides a summary of steps involved in SNP Regression.

Input

- Entity List of SNPs
- Trait type: Quantitative [5] or Categorical (Binary [19], Nominal, or Ordinal [45]; refer to sections Quantitative and Categorical Trait Types and Nominal and Ordinal Trait Types)
- Mode of Inheritance (default option: Additive; refer to section 35.6.1.1)
- Option for Single or Multiple SNPs
- Multiple Testing Correction (optional; refer to section on 35.6.2.3)
- EIGENSTRAT corrected data (optional)
- p-value Computation method (default option: Asymptotic)

Notes: GeneSpring does not support Permutative method for p-value Computation for Categorical traits.

Process (for each SNP)

Quantitative:
Runs linear regression.

Categorical:
Runs logistic regression. Single SNP Regression will be run between trait and individual SNPs while multiple SNP Regression will be run between trait and the whole block of SNPs.

Notes:"Reference Phenotype" refers to the de facto `control`-like phenotype value (for a nominal trait), with respect to which all tests are run. For example, if a disease is being studied and there are 6 phenotype values for a nominal trait, 5 will be considered as different phenotypic expressions of disease-affected status, and the other one will be considered as the phenotypic expression of disease-unaffected status.

Output
List of SNPs with χ^2 and T-Statistics p-values if the trait is categorical. If the trait is quantitative, then F-statistic is used instead of χ^2. Please refer to table given below.

Trait Type	Statistic Computed for SNP blocks	Statistic Computed for individual SNPs
Quantitative	F	T
Categorical: Binary	χ^2	χ^2
Categorical: Nominal	χ^2	T
Categorical: Ordinal	χ^2	T

Table 35.20: Statistic Computed for SNP blocks and individual SNPs

Notes:

- The p-value is an indicator of the strength of correlation between the disease status and the SNP or multiple SNPs.
- The F- and T-Statistics p-values are equal for single SNP regression.

35.6.4.2 Assumption on Multiple SNP regression

You know a group of SNPs (Gene or some specific region on genome) that may have relation with disease status. You can use these SNPs as input for multiple SNP regression to test *whether those SNPs as a block are related to disease status.* For instance, if quantitative trait is input and Multiple-SNP Regression is selected, then F-statistic p-value is computed for the SNP block and T-statistic p-value is computed for the individual SNPs that contribute to this relation.

The F-statistics p-value indicates relationship between whole block and disease status. The T-statistics p-value tells the contribution of each SNP in the block in that relationship. So, if F-statistic p-value is not significant then you may not be interested in individual SNPs T-statistics p-value (which is why the cut-off is based on F-statistic p-value), but the inverse may not be true i.e. if F-statistic p-value is significant then it is not necessary that any/all T-statistics p-values has to be significant.

A Note on Multiple SNP Regression: Maximum number of SNPs that can be handled is dependent on the amount of RAM and the value of XmxFraction in the properties.txt file. The memory(in bytes) required to store input matrix = 4 × Number of SNPs × Number of Samples. Also, during regression we create 4-5 matrices of similar order for intermediate calculation.

35.6.4.3 Summary of Multiple Linear and Logistic Regression:

This section provides a summary of steps involved in Multiple Linear and Logistic Regression.

Input
Genotype information for all SNPs, across samples

Process (for each SNP)
Performs Multiple Linear Regression if the trait type is quantitative and Multiple Logistic Regression if the trait type is Categorical, using each sample as a data point

- Independent variables: SNP genotypes
- Dependent variable: trait being studied

Output

- p-value for each SNP.
- Regression coefficients and associated likelihood statistics for each SNP.

35.6.5 Haplotype Trend Regression (HTR)

In a diploid organism, there are two non-identical copies of each autosomal chromosome. A combination of alleles at different loci on the same chromosome is referred to as Haplotype. Using HTR, haplotypes that are significantly associated with the sample phenotypes can be identified [61]. HTR takes an entity list, an interpretation, and a number of loci as input for haplotyping. A haplotype block is defined as consecutive set of SNPs on a genome. Block size refers to the number of loci. Suppose the input entity list, E = e1, e2, e3,. . . ,e8 with haplotype block size (number of loci) = 3 is input for HTR. From this input list, 8 haplotype blocks are possible, and entities of the input list are the starting SNPs of each haplotype block. The next 2 SNPs of every block are the consecutive SNPs from the genome.

35.6.5.1 Summary of Steps

This section provides a summary of steps involved in haplotyping trend regression.

Input

- Entity List
- Interpretation with trait. You can select binary, ordinal, or nominal option for categorical traits.
- Number of Loci (default value: 3)

Process (for each SNP)

1. Identifies the possible haplotype pairs for each sample in the given haplotype block.
2. Computes Haplotype frequencies using the Expectation-maximization (EM) algorithm.
3. Filters out haplotypes with frequencies less than 0.01.
4. Computes the probability of a particular haplotype(s) to occur in each of the samples.
5. Uses a regression based approach to relate the inferred haplotype probabilities with the corresponding phenotypes.
6. Computes F Statistic p-value for each haplo block, and t-statistic p-value for each haplotype if the trait is quantitative. If the trait is categorical, then instead of F-statistic, χ^2 statistic p-value is computed.

Output

- Provides a Haplo block list. The first SNP of the block is stored in the list.
- Provides the statistic P-values for each Haplo block and individual Haplotype as illustrated in the table below.

Trait Type	Statistic Computed for Haplo blocks	Statistic Computed for individual Haplotypes
Quantitative	F	T
Categorical: Binary	χ^2	χ^2
Categorical: Nominal	χ^2	T
Categorical: Ordinal	χ^2	T

Table 35.21: Statistic Computed For Haplo blocks and individual Haplotypes.

Notes: you can configure the Haplotyping Parameters from the tools menu. (*Tools* $\longrightarrow$ *Options* $\longrightarrow$ *Miscellaneous* $\longrightarrow$ *Haplotyping Parameters*)

Notes: HTR entity List is a different entity list that can be exported only as text file. Contact GeneSpring support for a script to convert an HTR list to a 'regular' entity list.

You can perform common tabular operations for all the analyses results using the context (right-click) menu options (please refer to: Right-click Operations - Context Menu)

35.6.6 LD Analysis

Linkage Disequilibrium (LD) refers to the non-independence of alleles at different sites. It is the non random association of two or more loci (SNPs/biallelic markers) in a population **GeneSpring** allows you to perform Linkage Disequilibrium (LD) analysis as a method for genetic mapping. The analysis computes two measures of LD, i.e., D-prime (D') and r^2 between pairs of SNPs per Chromosome.

35.6.6.1 Summary of Steps

This section provides a summary of steps involved in LD Analysis.

Input

- Entity List of SNPs
- Interpretation

Process

- Filters out the SNPs with MAF less than 0.05.
- Computes individual allele frequencies and Haplotype frequencies of the input SNPs using the EM algorithm.
- Computes the coefficients (r^2 and D-prime (D') of linkage disequilibrium between each pair of SNPs per Chromosome.

Output

The output of LD analysis is a set of LD plot nodes created as children under the input entity list in the experiment navigator. When the MAF output is spread across multiple chromosomes, separate LD plot nodes are created for each Chromosome (Figure 35.4).

Notes:

- If the number of SNPs for any chromosome in the MAF output is equal to one or greater than the "Maximum number of SNPs per chromosome" then that chromosome is not considered further in LD analysis. You can configure the "Maximum no. of SNPs per Chromosome" value from *Tools* $\longrightarrow$ *Options* $\longrightarrow$ *Data Analysis Algorithms* $\longrightarrow$ *LD*.
- LD Analysis doesn't run on sex chromosomes.
- LD Analysis ignores SNPs that do not have chromosomal location (unknown chromosomes).

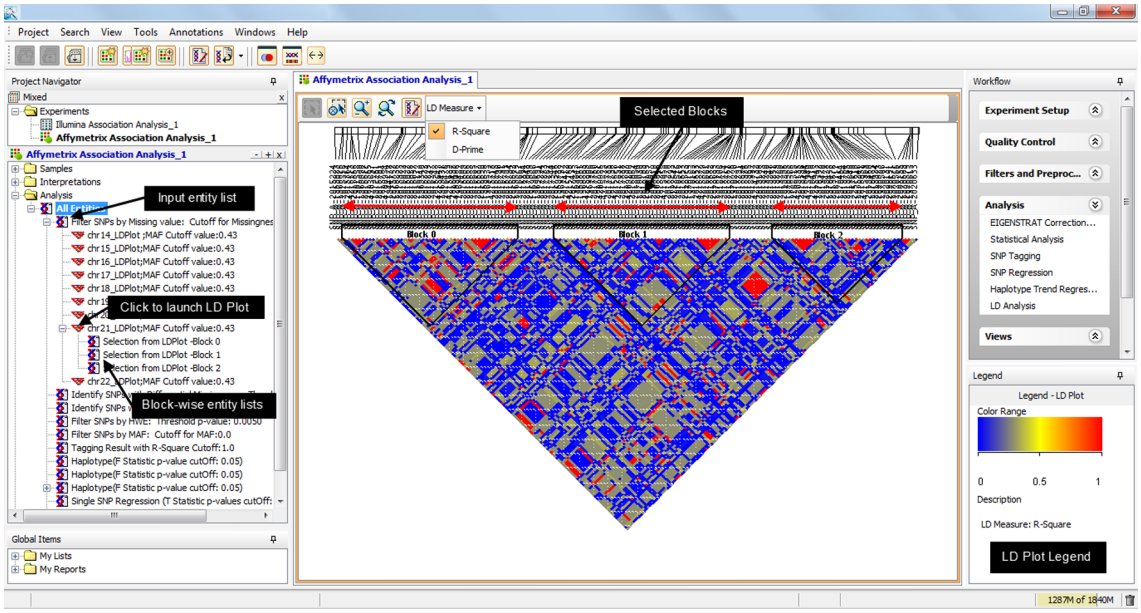


Figure 35.4: LD Plot

Launching LD Plot

- Click on an LD Plot node in the experiment navigator to launch the LD Plot.
- Select an option from the LD Measure menu options: r^2 or $D - prime$ (default option).
- Drag the mouse pointer over the plot to select the blocks of interest (Figure 35.4).
  Refer to LD Plot section for more information.

35.7 Views

GeneSpring provides an addition tool to view the results of association analysis experiements. You can launch the Genome Browser from the main toolbar or the Workflow link, and drag and drop the results of an association analysis experiment into it.

35.7.1 Genome Browser

You can launch the genotype calls or allele frequencies on the Genome Browser (Figure 35.5). Genotype call to Allele mapping is as illustrated in the table below:

Genotype Call	Allele
-1	No Call
0	AA
1	AB
2	BB

Table 35.22: Mapping for Genotype Calls

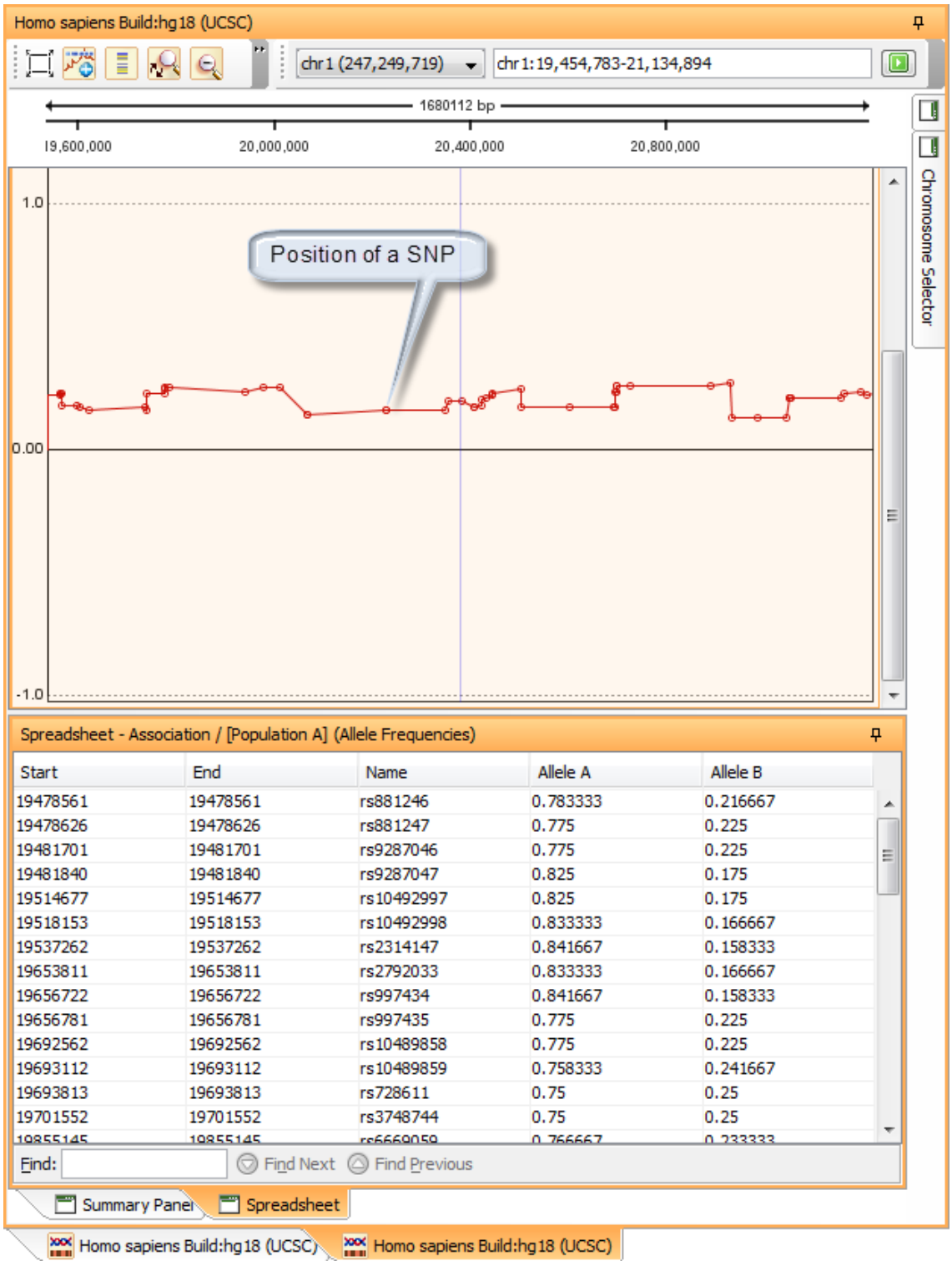


Figure 35.5: Allele Frequencies on Genome Browser

Notes:

- Run the `Calculate allele frequency` utility on an interpretation to view allele frequencies of the samples on the Genome Browser.
- You can run the `Calculate allele frequency` utility on `All samples`; however, it cannot be launched on the Genome Browser.

Refer to Genome Browser chapter for more information.

35.8 Results Interpretations

GeneSpring allows you to Identify Overlapping Genes, and run Gene Ontology and Pathway Analyses.

35.8.1 Identify Overlapping Genes

GeneSpring allows you to identify adjacent genes whose coding regions are partially overlapping.

Input
Entity List of SNPs

Process (for each SNP)

- Identifies a list of genes that are associated with the given SNPs.
- Creates a list of genes that show overlapping.

Output
A list of overlapped genes.

35.9 Utilities

GeneSpring provides additional utilities to Import Entity List from a File, Create a Probe List, Calculate Allele Frequency, and Filter on an Entity List.

35.9.1 Using disc cache

Under the menu **Tools** → **Options** → **Copy Number Algorithms** → **APT Execution Options**, 'Use Disc Cache' is an option. It is advisable to use the disc cache in Windows system.

Import Entity List from File	You can import a file as an Entity List by matching column names. Refer to section Import Entity List from File for details.
Create Probe List	You can create Subset Entity Lists from existing entity list. You can add (and remove) multiple conditions search the existing entity list. Each condition requires Chromosome number, Physical Position, Value, and Metrics. The result is a Probe list matching the conditions.
Calculate Allele Frequency	You can run this utility to calculate Allele frequencies on an interpretation. You can launch the Genome Browser, and then drag and drop the data. This opens the Select Data dialog, which allows you to launch Allele Frequencies along the genome.
Filter on Entity List	You can filter an Entity list using its annotations and list associated values. See section Filter on Entity List for details.

Table 35.23: Utilities in Association Analysis

Note that it will fail with Linux Operating System with more than 90 samples. When disc cache is not used, the execution process creates batches of probesets and this can slow down the process. In worst situation, the process can quit if there is memory outage. Using disc cache helps to avoid these problems and uses the disc memory while running processes.

Chapter 36

GeneSpring WorkGroupPlus

GeneSpring is available in two modes, namely, Desktop mode and **WorkGroupPlus** mode. The Desktop mode allows you to perform the analysis including the creation of objects locally on the machine. In the **WorkGroupPlus** mode, analysis objects can be shared among the users and groups. The **WorkGroupPlus** mode requires **GeneSpring WorkGroupPlus** Server and the client to be installed in one physical system.

36.1 GeneSpring WorkGroupPlus setup

For **GeneSpring WorkGroupPlus** set-up, install **GeneSpring WorkGroupPlus** Server and **GeneSpring Client** in one physical system. Prior to the initial launch of **GeneSpring Client** complete the setup of **GeneSpring WorkGroupPlus** Server, and **Floating License Server**using the **GeneSpring** Manager (Refer **GeneSpring WorkGroupPlus** manual for further details on how to perform various set-up actions using **GeneSpring** Manager). An administrator has all the privileges for the **GeneSpring WorkGroupPlus** and can manage multiple administrative tasks using the **GeneSpring** Manager. All users or group administrators can login to the **GeneSpring Client** through remote access and can use **GeneSpring** Manager to perform user specific operations (Further details are described in User and Groups).

36.1.1 GeneSpring WorkGroupPlus Administrator

Administrator privilege is required to perform multiple actions for **GeneSpring WorkGroupPlus** like;

- Install, activate, and manage the **Floating License Server**.
- Add, Edit, Disable, or Delete users
- Assign Group administrator privileges to an user

- Assign or Remove permissions to users
- Schedule Backup and Restore
- Download from the Annotations Manager or **GeneSpring WorkGroupPlus** Server homepage
- Update Pathway Interactions and the BrigeDb Databases
- Import Biocyc pathways, KEGG pathways, or KEGG demo pathways
- Workgroup data migration from **GeneSpring** Workgroup 11.0
- Pathway data migration from IKB server 11.0

36.1.2 GeneSpring Client License Manager

The **GeneSpring** Client License Manager (**Floating License Server**) can be activated and managed only by the **GeneSpring WorkGroupPlus** Administrator.

As a **GeneSpring** user:

- Login to **GeneSpring** using the Shared License Option of the License Activation window (Refer Figure 36.1).
- Enter the License Server Address with the valid port (Refer Figure 36.1).
- Select the license modules as required (Refer Figure 36.2) and click *OK* .

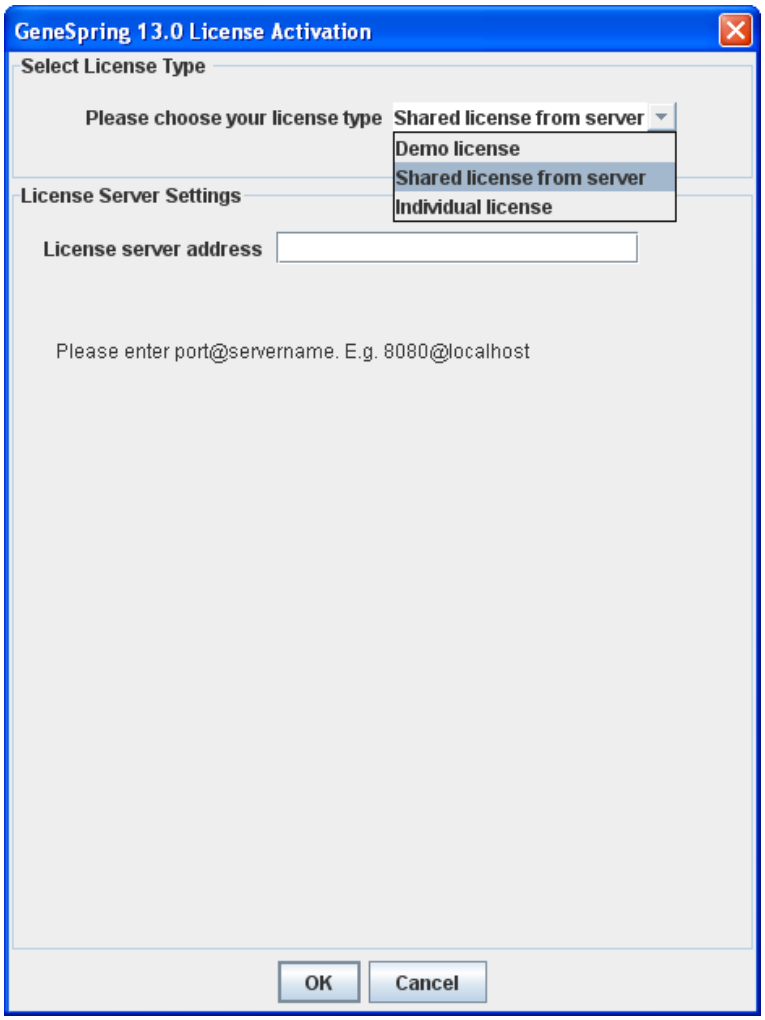


Figure 36.1: Shared License from server window

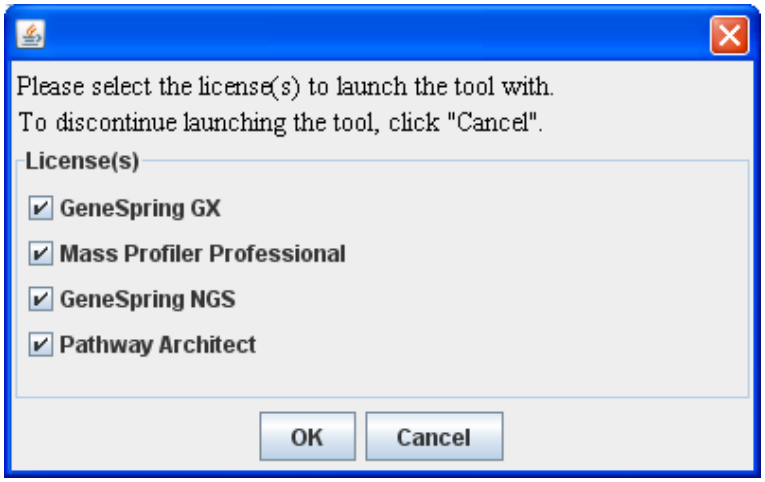


Figure 36.2: Select License Modules to connect to **WorkGroupPlus**

36.1.3 Login

After activating **GeneSpring** client, the tool launches and opens up the login dialog.

The following connection parameters to the **GeneSpring WorkGroupPlus** Server needs to be specified in the login dialog:

- Username - The user account which the administrator has created for you to connect to the server.
- Password - The password for your user account used for authentication.

For every launch of **GeneSpring** Client, you have to provide the Username and the password.

36.2 GeneSpring Manager

GeneSpring Manager is an applet through which the administrator, or group administrator, or users can perform multiple actions.

You can access the **GeneSpring Manager** via a web browser from the main page of the **GeneSpring**. Supported browsers are Internet Explorer on Windows, Mozilla on Linux, Safari on MacOSX, and Google Chrome.

To launch the **GeneSpring Manager**, in the browser run `http://localhost:8080/WorkgroupServer/` and click on the **GeneSpring Manager** link under the Server Management section.

The **GeneSpring Manager** applet requires JAVA 1.6 (or later) Runtime or a later version to be installed on the system.

On this web-page:

- Login to the **GeneSpring Manager** using your user account credentials. After a successful login, the browser will display the user interface of the **GeneSpring Manager**.
- You can view the entire backup history by selecting **View Backup History**.

  This utility is enabled for the administrator and all the users of **GeneSpring WorkGroupPlus**. If the backup is in progress, an attempt to access the **GeneSpring Manager** will open the **View Backup History** page.

  In the **Backup History** page, you can view:

  - **Date** - The date and time when the Backup process started.
  - **Status** - The status of the Backup process:
    - **Completed** - Backup completed successfully.

* **In Progress** - Backup is in process and you cannot access the function of **GeneSpring**.
* **Failed**- Backup failed to complete.
* **Not Started**- Backup did not start.

36.2.1 Users and Groups

GeneSpring WorkGroupPlus has a notion of user accounts to maintain privacy of data and a notion of groups and permissions to facilitate collaboration with fellow researchers.

Refer Operations on **GeneSpring** Objects for further details on how to share objects or other resources.

36.2.2 GeneSpring Manager User Interface

36.2.2.1 Menus and Toolbar

The following are the main menus in the **GeneSpring Manager** User Interface:

- **User:** This menu has the utilities to login, logout, and change password.
- **Admin:** This menu will be enabled and available only for administrators and group administrators.
- **Utilities:** This menu will be enabled and available only to the administrator.
- **Search:** This menu provides utilities to search and retrieve **GeneSpring** resources to which you have access privileges.
- **Tasks:** This menu provides utilities to monitor and search tasks that have been submitted to the **GeneSpring WorkGroupPlus** Server.
- **About:** This menu provides details of the **GeneSpring Manager** with version and build number.

The functions of the various toolbar icons in the **GeneSpring Manager** User Interface are:

- **Navigation icons:** These icons provide convenient navigation to the previous or next visited location in the navigation tree.
- **Refresh icon:** This icon will refresh the navigation tree to its current state.
- **Home icon:** This icon sets the navigator on the home folder of the user who is logged in.

- **Search icon:** This icon opens the *Advanced Search* dialog that has an interface to construct complex queries for searching resources in the **GeneSpring WorkGroupPlus** Server.
- **Simple Search:** The textbox on the right of the toolbar provides a utility to conduct a simple search and retrieve resources. Enter a text in the textbox and click the search icon next to it. This will carry out a case insensitive search on resources.

36.2.2.2 The Navigation Tree

The **GeneSpring Manager** has the navigation tree on the left and the details table on the right. It provides an explorer like interface for the resources of the **GeneSpring WorkGroupPlus** Server. The navigation tree displays a hierarchy of folders and resources to which the logged in user has access. Each resource appears in a folder with the same name as the owner of the resource.

There are two types of resources in the navigator tree:

- **GeneSpring Resources:** These are resources that have been created by you.
- **Generic Document Resources:** These are generic files that have been uploaded onto the **GeneSpring WorkGroupPlus** Server. These are stored in the **My Documents** folder.

Selecting a node in the navigation tree will display the details of the resources under this node on the right. For example, selecting a folder in the navigator will display the details of the resources in this folder; selecting a project will display the experiments in the project; and selecting an experiment will display the samples and analysis folder.

The table on the right gives details of the folders and resources. The resources in the table are sorted by the name of the resource. To sort by any other column in the table, click on the column header.

Selecting a **GeneSpring** resource will display a table with the name, technology, owner, creator, resource type, creation date, modified by, last modified date, and permissions.

You can right-click on any resource and either **Inspect** it or **Share** with other users.

Selecting a **My Documents** resource in the navigator displays a table with the name, type, owner, creator, permissions, size, versions, creation date, last modified date, and modified by. Right-click on an object and select **Show Version** to access specific versions of these resources and any other previous versions.

36.3 Operations on GeneSpring Objects

Objects - Projects, Experiments, Samples, or all of these, and other analysis objects created by an user are stored on the **GeneSpring WorkGroupPlus** Server and can be shared among different users and groups as per the permissions granted by the owner, group administrator, or the administrator. Refer section on objects.

36.3.1 Object ownership

What is ownership: Every object has a creator and an owner associated with it. The creator of an object is the user who created it and can never be changed.

Who can change ownership: A group administrator can change the ownership of any object owned by a member of his group. **GeneSpring WorkGroupPlus** administrator can change the ownership of any object in the system. The ownership of an object can be changed from one user to another any number of times. At any instance though, there will be only one user who is the owner of an object. Objects cannot be owned by groups and when an object is initially created, the creator and the owner are the same user.

How can ownership be changed: The ownership of an object can be changed via **Change owner** option in the right click menu of the object. **Change Owner** dialog displays the members of the group in the drop down menu. Select a user and click **OK** to change the ownership of the object.

The *Change Owner* option is available only for independent objects. While changing ownership one can choose to propagate the ownership change hierarchically to all the children of the object as well. A Project's ownership can be changed from **Project** menu.

Delete object: Only the owner, group administrator, or the **GeneSpring WorkGroupPlus** administrator can delete objects. Objects can be deleted using the **Delete** option in the right click menu of the object or from the **Delete** icon in the toolbar menu of the search wizard.

36.3.2 Object permissions

What are permissions: For sharing, objects can be given two kinds of permissions;

1. **Read**

   Objects with read permission to a group are accessible to the members of the group. The members can search for these objects, inspect them, open them in the case of Projects and Experiments, or use them in case of Models, Scripts, Pathways, Samples, and Entity lists. However, these objects cannot be modified.

2. **Write**

   Objects with write permission for a group can be accessed and also modified by the group members. For example, modifications can include changing object attributes, adding children (Experiments to Projects or Entity Lists to Experiments), edit attachments to Samples, or changing contents of Scripts.

No permission

If an object has neither read nor write permissions for a group, then the group members cannot access the object. Permissions can also be explicitly set for an individual user. If an user has explicit read or write permission for an object, then they can access or modify that object accordingly.

Who can change the permissions: Only the owner of the object, group administrator, or the **GeneSpring WorkGroupPlus** administrator can alter the permissions.

How can permissions be changed: The permissions can be changed from the **Share** option in the right click menu of the objects or the from **Change Permissions** icon in the search wizard.

How to share an object:

- To share an object, right-click on the object and select **Share**.
- This opens the **Permissions** dialog (Refer Figure 36.3).
  - The upper panel has the details of the selected object and the lower panel has the current state of permissions of the object.
  - The **Group** tab shows the groups to which the user belongs and the permissions available to each of the groups.
  - The **User** tab shows all the users in the groups that the owner belongs and the explicit permissions available to each of the users.

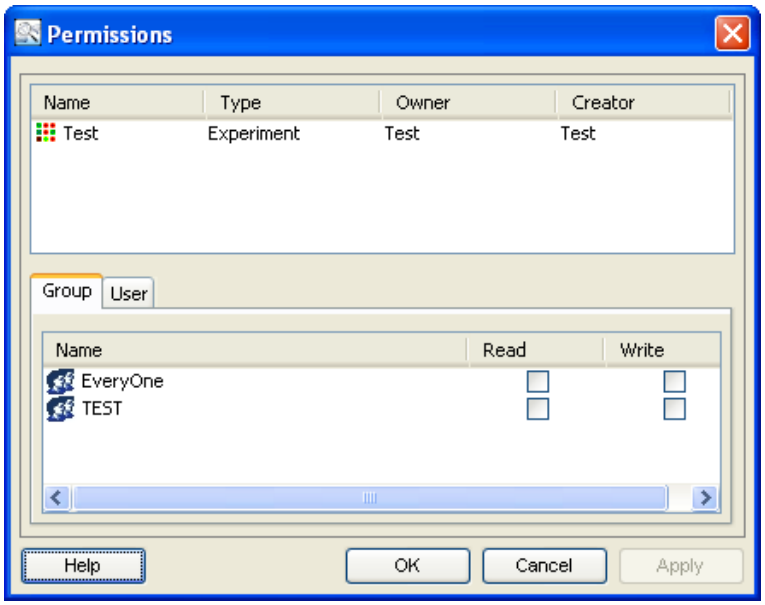


Figure 36.3: Permission Dialog

- Multiple objects can be shared by selecting multiple objects and click **Share Objects** from the right click menu or **Change Permissions** icon in the toolbar menu of the search wizard.
  - Here, the upper panel has the details of all the independent objects chosen and the lower panel has the details of the permissions on these objects for each group/user (Refer Figure 36.4).

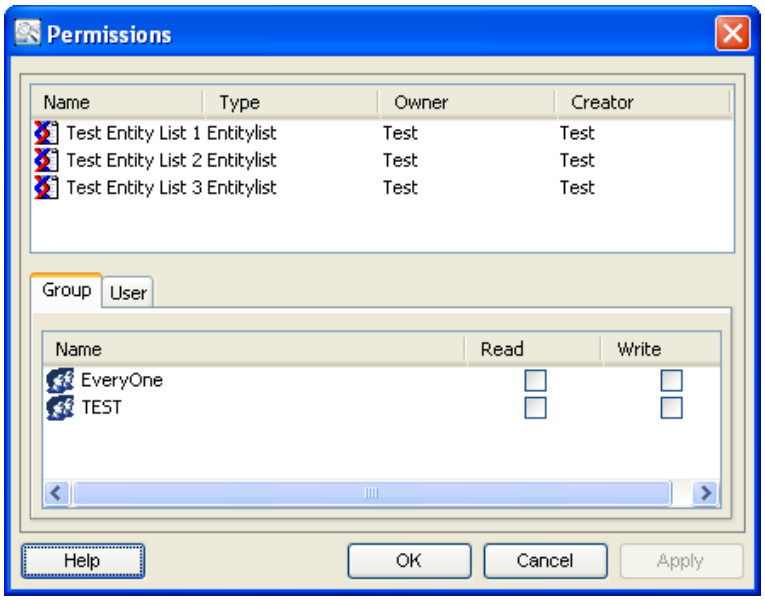


Figure 36.4: Permission Dialog for multiple objects

- You can choose to propagate the permissions to Folders, Projects, Experiments, and analysis objects to all their descendants in the hierarchy.
  - Select an object with descendant to **Share**
  - A confirmation message will appear as illustrated in Figure 36.5.
  - On confirmation all the descendants will have the same permission as the object.

Figure 36.5: Confirmation Dialog for sharing descendants of an object

NOTE

The Permissions interface will display the permissions granted by the owner of the object. By default, if a group has explicit permissions to read/write on an object, all users in that group will also have read/write permissions implicitly. However, if a user of group (having write permission) has only read permission, the same will be displayed for the user.

- Permission for an object can be given to all users (even if they are not part of a group) by sharing with **Everyone** group. **GeneSpring WorkGroupPlus** administrator can allow or disallow sharing of objects to **Everyone** through **GeneSpring** manager.

36.3.3 Conflicts with permissions

Permissions can have two kinds of conflicts:

1. **User-Group**

   For an object, if the owner gives write permission to a group and only read permission to one of the group member, then the permission granted explicitly for that particular user will take precedence while for other members the write permission will be retained.

2. **Group-Group**

   If an user belongs to two groups and an object has read permission to one group and write permission to the other, then the most liberal permission, i.e. write permission will take precedence for the user on that object.

36.3.4 Propagating permissions

- When an object is shared using right click option or search menu, share dialog opens up displaying the users and groups. Clicking OK on the **Permissions** dialog prompts a window asking whether to propagate the permissions of the object to its descendants or not. If the permissions of the object are not propagated to its descendants then they retain their original permissions.
- When permissions are propagated to descendants, it is the change in permission that gets propagated.
  - For example, consider a Project with permissions for Group A:read, that contains an Experiment with Group A:read; and Group B:write.
  - If the permission on the Project is changed to Group A:write allowing the propagation of permission to the descendants, the experiment permission will change to Group A:write; and will retain Group B:write.
  - If the descendants have an alternate permission to the parent object, then the same is retained.
- While propagating permissions, there might be some descendants on whom permissions cannot be altered by the user.
  - Consider a Project owned by user A with two Experiments, one owned by user A and another by user B.
  - When user A changes permissions on the Project and propagate permissions, then the propagation applies only to Experiment A (user A is the owner).
  - The permissions on Experiment B(owned by user B) are not altered as user A does not have the required privileges.

36.3.5 Inheriting Permissions

- If existing objects are added to a parent, then the permissions on that object do not change. Thus, if an existing Experiment is added to a Project or if an Entity List is added to an

Experiment, the original permissions of the Experiment on the Entity List are retained.

- When a new object is created and added to a parent, the object inherits the set of permissions from the parent.
  - Consider a Project (owner:user A), having permissions, Group A:write; Group B:read.
  - If user A creates a new Experiment and adds it to the Project, then the Experiment will inherit the permissions of the Project namely; Group A:write and Group B:read.
  - The same is true if another user B who also belongs to both groups Group A and Group B creates and adds a new Experiment.
  - If a user C belonging to Group B creates and adds a new Experiment to the Project, then the Experiment will retain only related permissions of the Project; Group B:read.
  - Group A:write permission is not inherited as user C does not belong to Group A.

36.4 Run in background

In **GeneSpring WorkGroupPlus**, you can execute certain resource intensive tasks in the background through a **Compute server**. This allows you to simultaneously perform other analysis. **GeneSpring WorkGroupPlus** Server has designated Compute server(s) and can be configured by the **WorkGroupPlus** administrator to enable the execution of several such tasks simultaneously.

36.4.1 Task Manager

The **Task Manager** acts as an interface between the **Compute server** and **GeneSpring** by providing a status and other details of the tasks scheduled to be run in the background. It can be accessed from *Tools* $\longrightarrow$ *Task Manager*. Task Manager displays a table with the following columns: Position in queue, Task name, Owner, Status, Submitted on, Scheduled on, Ended on, Compute server IP, Application context, and Task ID. If the task execution has started, the status message is **Executing**. If the task is scheduled for later, the status message is **Scheduled** and if the **Compute server** is not available to start a task at the specified time, the status shows **Queued**. The **Position in queue** for all **Scheduled** tasks is '0' and the **Queued** tasks have numbers starting from 1. On initiation of the task execution, the position in queue moves up for the remaining **Queued** tasks. Once a remote task is executed, the status will change to **Successful** or **Failed**. All tasks go through the following order: Scheduled $\longrightarrow$ Queued $\longrightarrow$ Executing $\longrightarrow$ Successful. The Task ID is a unique number assigned to each task and is used in the internal operations of the **Compute Server**.

Depending on the status of a task, you can perform multiple operations on the scheduled tasks from the **Task Manager**.

NOTE Click **Refresh** after any change in the options of the Task Manager to view them.

From the Task Manager;

- You can alter the Scheduled tasks with these options:
  - **Suspend/ Resume**: You can pause a scheduled task using **Suspend** option and re-start it using the **Resume** option.
  - **Reschedule**: Reschedule a task to a different date and time.
  - **Delete/ Terminate**: Delete a scheduled task or Terminate an ongoing task.
- For all successfully completed or failed scheduled tasks, you can view the logs from these options:
  - **View Error log**
  - **View Output log**
  - **View Status log**
- Select the task(s) and use **Clear** option to remove any Successful, Failed, Deleted, and Terminated tasks from the Task Manager.
- **Refresh**: After every action in the Task Manager, click **Refresh** to view the changes.
- **View Status**: This allows you to view the Compute server memory usage in executing the task.

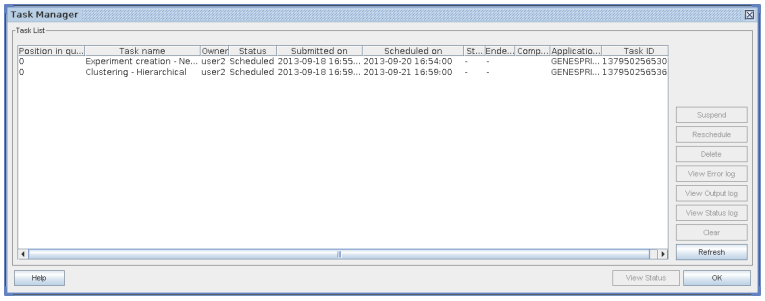


Figure 36.6: Task Manager

36.4.2 Operations with an option to Run in background

- **Experiment Creation**: You can run perform Experiment Creation task in the background by selecting the **Run in background** option in the last step of the experiment creation wizard in Data Import Wizard workflow mode. If the experiment is being created from raw files, they are first uploaded from the local machine as samples and then the experiment creation task is scheduled to run in background.

  When the task is successful, view the result of the task (a new experiment) by refreshing the project.
- **Add/Remove Samples**: You can access Add/Remove Samples option from **Quality Control** $\longrightarrow$ **Quality Control on Samples** link in the workflow browser in the Data Import Wizard. In the Analysis: Biological Significance Workflow, you can access it from the options for Quality Control (Step 3).

Select Add/Remove Samples and choose samples for addition/removal. Select **Run in background** option to schedule and perform this task on the **Compute Server**.

When the task is successful, view the result of the task by refreshing the experiment.

- **Clustering**: Run Clustering in the Data Import Wizard mode from **Analysis** $\longrightarrow$ **Clustering** in the workflow browser. After selecting the options to performing Clustering, the second step has the option to run clustering in the background.

  When the task is successful, view the result of the task by refreshing the experiment.

- **SNP Tagging**: You can run SNP Tagging in the background from the first page of the SNP Tagging Wizard. Configure the SNP Tagging parameters (see SNP Tagging) from this page, and select Run in background to perform the analysis in the Compute server.

  1. Select **Run in background** check box. This gives you two options:
     - Select the **Schedule Now** radio button to start the analysis immediately. If there is a pre-existing queue of analysis, then the process will start immediately after the execution of last analysis in the queue.
     - Select the **Schedule task at** radio button, and then select a time and date from the drop-down option to run the analysis at a desired time.
  2. Click **Next**. This will open an "Information" message confirming the task schedule. You can check the status (refer to Task Manager section) of the task from **Tools** $\longrightarrow$ **Task Manager**.
  3. Once the task is completed, refresh the project. You will be able to see a new entity list of **Tagged SNPs** under the **All Entities** list.

36.4.3 Interpreting Task Logs

- The Compute Server runs out of memory during the execution of any of the tasks to be run in background:
  The task is not executed and the task status in the Task Manager displays **Failed** and the **View Status log** displays that the Compute Server has run out of memory. This can be modified by the administrator of the **GeneSpring WorkGroupPlus**.

- The project is deleted before or during the run in background task of experiment creation:
  The task status in the Task Manager displays successful and the **View Status log** displays a message that the experiment creation was successful but the experiment was not added to the project. You can view the created experiment from **Search** $\longrightarrow$ **Experiments** and add the experiment to a project (it has to be open) using the icon.

- The permission to the shared project is revoked by the owner and another user has scheduled a remote experiment creation on the same project:
  The Task Manager displays that the task was successful and the **View Status log** displays that the task was successful but the experiment is not added to the project. You can view the created experiment from **Search** $\longrightarrow$ **Experiments** and add the experiment to a project (it has to be open) using the icon.

- Permission for shared samples is revoked or the owner deletes the samples while another user is creating an experiment using these samples with run in background option:
  The task status in the Task Manager displays **Failed** and the **View Status log** displays that there was an error in reading the sample.

- Experiment gets deleted when the task 'Add/Remove Samples' is run in background:
  The task status displays **Failed** and the **View Status log** displays that the operation failed unexpectedly.

- The task 'Add/Remove Samples' is scheduled with run in background option on a shared experiment for which the user has only **Read** permission:
  The task status displays `Failed` and the **View Status log** displays that there was an error in reading the experiment.

- Clustering task in scheduled to run in background with 2 samples using K Means and with the following parameters- Cluster on Conditions and Number of Clusters is greater than 2:
  The task status displays **Failed** and the **View Status log** displays that the number of clusters are same or more than the number of columns.

- The experiment is deleted after scheduling Clustering to run in background:
  The task Status displays **Failed** and the **View Status log** displays that the particular Entity List on which Clustering is to be performed could not be found and might be deleted.

- The task 'Clustering' is scheduled to run in background on a shared experiment for which the user has only **Read** permission:
  The task status displays **Failed** and the**View Status log** displays that there was an error in adding classification to the experiment

Chapter 37

Writing Scripts in GeneSpring

GeneSpring offers users the ability to write their own scripts, execute these within the current context and save them. The scripting utility allows users to combine operations in **GeneSpring** with a more general Jython programming framework to yield automated scripts. Using these scripts, one can run transformation operations on data, automatically pull up views of data etc.

37.1 The Script Editor

Scripts can be typed, uploaded or copied and pasted, into the *Script Editor* Window. To run a script, go to *Tools* $\longrightarrow$ *Script Editor*. This opens up the *Script Editor* window. Write your python script into this window and click on the Run icon to execute the script. Alternatively, if you have a script in a file, open the file in any editor, copy and paste the script into the script editor or if you have it as a python script file, then you can simply upload it using the 'Upload' icon. Then click on Run icon to execute the script. Errors in the execution of this script, if any, will be recorded in the log window, which acts as the default standard output location. These logs can be cleared using the 'Clear' icon.

Once the script is in the window, it can be saved using the 'Save as' icon. The scripts are saved into the database and can be retrieved later using the Open icon. Scripts can now be stored in a hierarchy of folders. Subsequent changes can be saved using the Save icon.

Clicking on the 'Inspect' icon opens up a Script inspector where the name of the script and notes can be given. It also contains the creation date, modification date and the owner.

The *Script Editor* also contains other text editor features like undo, redo, go to line, search, replace etc.

This chapter provides API documentation for some of the basic operations needed to access, manipulate and save data in **GeneSpring**, generate customized views and save these results. We

also have a section detailing how you can interface these options with R packages available outside the tool.

Please note that the Script Editor in **GeneSpring** expects Jython (Python with Java class import capabilities) scripts. Therefore, to write meaningful scripts using the functions described in the documentation below, you will need some knowledge of the Python programming language. See https://docs.python.org/3/tutorial/ for a basic Python tutorial. Knowledge of Java will help you write more powerful scripts. See http://java.sun.com/j2se/1.5.0/docs/api/ for the Java API.

37.2 Hierarchy of data organization in GeneSpring

In order to access data within **GeneSpring**, it is essential to understand the data organization hierarchy. Refer to the section on Data Organization. At any point in time, **GeneSpring** can have only ONE active project, which can contain one or more experiments, out of which only ONE will be the active experiment. Each experiment will have its own Samples, Interpretations and Analysis.

37.2.1 Accessing Projects, Experiments and their Constituent Elements

Usage	Notes

# Project variable 	
`# Get the current active project`
`proj = script.marray.project.getActiveProject()`
	The active project (object) gets stored in the *proj* variable

# Experiment variable 	
`# Get the currently active experiment`
`expt = script.marray.project.getActiveExperiment()`
	The active experiment (object) gets stored in the *expt* variable

`# Get the list of samples in this experiment` `samples = expt.getSamples()`	*samples* variable will contain a list of the samples using which this experiment was created. The *expt* variable should contain an experiment object as shown above
`# Get the parameters for a sample` `params  = sample.getParameters()`	*params* (a java LinkedHashMap) contains details of the sample. For example, it contains the parameters associated with the sample (obtained from the experiment grouping information of the experiment that the sample belongs to)
`# Get the default interpretaion for this experiment` `interpretation = expt.getDefaultInterpretation()`	The *expt* variable should contain an experiment object as shown above

Continued on Next Page...

Table 37.1 -- Continued

Usage	Notes
`# Get the active interpretation for this experiment` `interpretation = expt.getActiveInterpretation()`	The *expt* variable should contain an experiment object as shown above

`# Get the active technology for this experiment`
`techId = expt.getTechnology()`
`technology = script.marray.project.createTechnology(techId)`
	The *expt* variable should contain an experiment object as shown above. The *technology* variable will contain the technology object for the experiment *expt*

`# Get all entity lists in the experiment` `entityLists = expt.getAllEntityLists()`	The *expt* variable should contain an experiment object as shown above
`# Get the current active entity list in the experiment` `entityList = proj.getActiveEntityList()`	The *proj* variable should contain a project object as shown above
`# Get the "All Entities" entity list` `entityList = expt.getAllEntitiesEntityList()`	The *expt* variable should contain an experiment object as shown above
`# Get the Experiment Grouping` `exptGrouping = expt.getExperimentGrouping()`	The *expt* variable should contain an experiment object as shown above
`# Get the conditions map from the experiment grouping` `conditionsMap = exptGrouping.getConditionsMap()`	*conditionsMap* (a java Map object) contains the conditions corresponding to the experiment grouping *exptGrouping*. You can iterate through it as follows: for key in conditionsMap.keySet(): print key, " : ", conditionsMap.get(key)

Table 37.1: Accessing Projects and Experiments

Please note that there are multiple ways to access a function that belong to a module in Python. A few of these are:

1. Importing the function from the module in the script:

```
# include the getActiveProject function from the module
# script.marray.project in this script
from script.marray.project import getActiveProject

proj = getActiveProject()
```

2. Importing all the functions from the module in the script:

```
# include all the functions from the module
# script.marray.project in this script
from script.marray.project import *

proj = getActiveProject()
```

3. Referring to a function in the context of its module:

```
proj = script.marray.project.getActiveProject()
```

37.2.2 Accessing the Experiment Dataset

The data seen in the Spreadsheet view when the ''All Entities'' Entity list is active in the Project Navigator) shows the master dataset for that experiment, also known as the normalized dataset. This normalized dataset can be accessed as follows:

```
expt = script.marray.project.getActiveExperiment()
analysis = expt.getAnalysis()
node = script.marray.project.analysis.getNormalisedDatasetNode(analysis)
ds = node.getDataset()
```

The variable *ds* now contains the normalized dataset for the entire experiment (You can also replace `getNormalisedDatasetNode` by `getRawDatasetNode` to get the raw dataset). The dataset contains columns and rows. A number of operations can be carried on a dataset object *ds*, as shown below:

Usage	Notes
`# Get the number of columns in the dataset` `numCols = ds.getColumnCount()`	The variable *numCols* will contain number of columns (both data and annotation columns) in the dataset *ds*
`# Get the number of rows in the dataset` `numRows =  ds.getRowCount()`	The variable *numRows* will contain number of columns in the dataset *ds*

`# Get the ith column in the dataset`
`column = ds.getColumn(i)`
`pythonCol = script.coercion.to_py(column)`
	Here *ds* is a valid dataset. *i* must be an integer value between 0 and $n-1$ where n is the number of columns. This will get the data from the column into a python list

`# Get the contents of the ith col, jth row in the dataset ds` `data = ds.getColumn(i).get(j)`	Gets the data from the ith column, jth row of the dataset *ds*

Continued on Next Page...

Table 37.2 -- Continued

Usage	Notes

`# Create new integer column`
`data = [1,2,5,6] # data can be a python list`
`name  = "Counts"`
`intCol = script.dataset.createIntColumn(name, data)`
	Creates a Integer column with the specified name having the given data as values. Here, *name* is a string, *data* is an python list containing integer values

`# Create new float column`
`data = [1.1, 2.2, 5.5, 6.6]`
`name = "Intensity"`
`floatCol = script.dataset.createFloatColumn(name, data)`
	Creates a Float column with the specified name having the given data as values. Here, *name* is a string, *data* is a python list containing float values

`# Create new string column`
`data = ["Gene A", "Gene B", "", "Gene C"]`
`name = "Genes"`
`stringCol = script.dataset.createStringColumn(name, data)`
	Creates a String column with the specified name having the given data as values. Here, *name* is a string, *data* is a python list containing string values

`# Create a new dataset`
`cols = [intCol, floatCol, stringCol]`
 	
`d = script.dataset.createDataset("New Dataset",cols)`
	Each line of the code snippet is explained below:
	(1) Create an array of columns (see above)
	(2) Create a new dataset consisting of those columns

`# Add a new column to the dataset ds` `ds.addColumn(col)`	Adds the *col* column to the dataset *ds*

`# Transpose the dataset ds`
`from com.strandgenomics.cube.dataset import DatasetUtil`
`transDs = DatasetUtil.transpose(ds)`
	transDs will contain the transpose of *ds*

`# Get all the data from a column into a float array (float[])`
`from com.strandgenomics.cube.dataset import DatasetUtil`
`from com.strandgenomics.cube.framework.data import ArrayUtil`
`col = ds.getColumn(n)`
`strandFloatArray = DatasetUtil.getColumnData(col)`
`floatArray = ArrayUtil.getContents(strandFloatArray)`
	The variable *floatArray* will contain all the data in the nth column in the dataset *ds*. n must be an integer. The nth column in the dataset must contain only floating point values

`# Get all the data from a column into a integer array (int[])`
`from com.strandgenomics.cube.dataset import DatasetUtil`
`col = ds.getColumn(n)`
`intArray = DatasetUtil.getIntegerArray(col)`
	The variable *intArray* will contain all the data in the nth column in the dataset *ds*. n must be an integer. The nth column in the dataset must contain only integer values

Continued on Next Page...

Table 37.2 -- Continued

Usage	Notes

`# Get all the data from a column into a string array`
`from com.strandgenomics.cube.dataset import DatasetUtil`
`col = ds.getColumn(n)`
`stringArray = DatasetUtil.getStringArray(col)`
	The variable *stringArray* will contain all the data the nth column in the dataset *ds*. n must be an integer. The nth column in the dataset must contain only string values

Table 37.2: Accessing Experiment Dataset

37.2.3 Some More Useful Functions

Usage	Notes

`# Create Entitylist of non missing signal values`
`script.marray.project.\`
`createNonMissingSignalsEntityList(expt)`
	A new entity list, ''Entities without any missing signal values'' is added to the experiment *expt* and will be seen in the Project Navigator

`# Create New Entity List`
`entityList = \`
`script.marray.project.createNewEntityList\`
`(name, notes, entityListType,\`
`entityListTechnology, ds)`
	A new entity list is created and stored in the variable *entityList*. *name* and *notes* are string variables of your choice. *entityListType* and *entityListTechnology* can be obtained from the current experiment *expt* as follows:
	`entityListType = \`
	`expt.getExperimentType()`
	`entityListTechnology = \`
	`expt.getTechnology()`
	ds is the dataset you want to save in your entity list and can be created as shown previously. *ds* should have the first column as the identifier column with identifiers from the technology

`# Add an entity list to the experiment`
`parentList = expt.getAllEntitiesEntityList()`
`parentList.addChild(entityList)`
	Here, instead of the ''AllEntities'' List, you could choose to have some other list as *parentList*. See `expt.getAllEntityLists()` above

`# Get the flag dataset`
`expt = script.marray.project.getActiveExperiment()`
`analysis = expt.getAnalysis()`
`node = script.marray.project.analysis.getFlagDatasetNode(analysis)`
`flagDs = node.getDataset()`
	This gives flagging information about each data point in the dataset

Continued on Next Page...

Table 37.3 -- Continued

Usage	Notes

`# Get the indices of all signal columns in the dataset`
`sigColIndices = script.marray.project.analysis.\`
`getSignalColumnIndices(ds)`
	`sigColIndices` will contain an array giving column indices of the signal columns in the *ds* dataset.

`# Get the column having a certain mark`
`# (such as GO ID column).`
`# Each column in a dataset will have a`
`# particular mark, which identifies its type`
`# i.e. Identifier column, signal values, annotation etc.`
 	
`from com.strandgenomics.spring.project.mark \`
`import MarkManager`
 	
`marks = MarkManager().getMarks()`
 	
`markType = ""`
 	
`for mark in marks:`
`if mark.toString() == "Gene Ontology accession":`
`markType = mark.getType()`
`break`
 	
`col = script.marray.project.analysis.\`
`getColumnOfMarkType (ds, markType)`
	Each line of the code snippet is explained below:
	(1) Required imports
	(2) Get the list of all marks. You could print this to see which mark matches your needs, and use it for the match in (4)
	(3) Create a variable markType.
	(4) If the mark equals ''Gene Ontology accession'', get the corresponding markType
	(5) Get the column with the correct markType for ''Gene Ontology accession'', i.e get the Column with GO ID values in the dataset ds. (See table below for some common marks*)

`# Get column indices for an interpretation`
`colIndices = script.marray.project.\`
`translate.getColumnIndices(interpretation)`
	Gets the column indices (with respect to the normalized dataset node) of the conditions in the given interpretation

`# Get annotation columns for a technology`
`names = technology.getAnnotationColumnNames()`
`cols = script.marray.project.ui.\`
`getAnnotationColumns(technology, names)`
	technology is a variable containing the technology

`# Get the default annotations dataset`
`techId = expt.getTechnology()`
`techDataset = script.marray.project.ui.\`
`getDefaultAnnotationsDataset(techId)`
	techDataset will contain the default annotations for the technology with technology ID *techId*

Continued on Next Page...

Table 37.3 -- Continued

Usage	Notes

`# Show a profile plot for the dataset ds`
`script.view.ProfilePlot(dataset = ds,\`
`title = "My Profile Plot",\`
`columnIndices = colIndices).show()`
	ds and *colIndices* should be appropriately populated, as shown in the examples above. You can also manually create a colIndices array as follows:
	`colIndices = [1,2,5,8,10]`

`# Show a histogram for the dataset ds`
`script.view.Histogram(dataset = ds,\`
`title = 'My Histogram',\`
`column = 2).show()`
	ds should be appropriately populated

`# Show a box whisker plot for the dataset ds`
`script.view.BoxWhisker(dataset = ds,\`
`title = 'My Box Whisker Plot',\`
`columnIndices = colIndices).show()`
	ds and *colIndices* should be appropriately populated, as shown in the examples above. For a list of all available views, please refer to the ''script.view'' section in the **GeneSpring** Scripting manual

Table 37.3: Some Useful Functions

37.2.4 Some Common Marks

Mark name	Signifies
Agilent Probe ID, Identifier, Probe Set ID, Transcript Cluster ID, Exon ID, Affymetrix ProbeSet Id	The identifier for that particular technology.
Normalized Signal, Raw Signal, Background Corrected Signal	Marks for columns containing signal values.
Gene Name, Gene Symbol	Marks for name of the gene.
Gene Ontology accession	GO ID column
Entrez Gene Id	Entrez Gene Id column
Ensembl	Ensembl Id column
Unigene Id	Unigene Id column
SwissProt	SwissProt Id column
Flag	Flag column

Table 37.4: Some Common Marks

Note: Marks are case-sensitive. You can get a full list of available marks by doing the following:

```
from com.strandgenomics.spring.project.mark import MarkManager
marks = MarkManager().getMarks()
print marks
```

Now that you can access the dataset of your choice, or create it using the available datasets within your project, it can be presented using a number of pre-created views, such as Scatter Plots, Box Plots, Histograms etc. The API for these functionalities are available in the Scripting Manual here.

37.2.5 Creating UI Components

Using the `createComponent` function, a UI dialog can be presented to a user so he or she can provide inputs at program runtime. The types of components that you can create are shown below. Please note that in order to use any of these, you need to include the following line in your script:

```
from script.omega import createComponent, showSimpleDialog, showDialog
```

Usage	Notes

`# UI Component Template`
`from script.omega import createComponent, showDialog`
 	
`c = createComponent(type="float",\`
`id="float", description="Enter p-value",\`
`value=0.05)`
 	
`input = showDialog(c)`
 	
`print 'User entered: ', input`
	Each line in the example is explained below:
	(1) Required imports
	(2) Creates a simple UI component *c* that takes floating point input values. In this case, the default value is 0.05
	(3) Shows a dialog to the user, where the user can enter another value
	(4) Prints input obtained from the user (You can use the *input* variable in your program)

`# Text Box`
`c = createComponent(type="text",\`
`id="input", description="TextArea",\`
`value="Enter text here")`
	Creates a component for integer, text or floating point input. For the *type* argument, you can replace *text* by *int* or *float* or *string* to get different input types

`# Drop-down box`
`c = createComponent(type="enum",\`
`id="enum", description="Enumeration",\`
`options=["option 1","option 2","option 3"],\`
`value="option 2")`
	Creates a Drop-down box component with 3 options. The user initially sees *option*2, but can choose any option among the 3 options

`# Check Box`
`c = createComponent(type="boolean",\`
`id="boolean", description="Check Box")`
	Creates a Check Box component

Continued on Next Page...

Table 37.5 -- Continued

Usage	Notes

`# Radio Button`
`c = createComponent(type="radio",\`
`id="radio", description="Radio",\`
`options=["option 1","option 2","option 3"])`
	Creates a Radio component, to select only one input from a defined list of inputs

`# File chooser`
`c = createComponent(type="file",\`
`id="file", description="File Chooser")`
	Creates a component to select a file from disk

`# Single column chooser`
`c = createComponent(type="column",\`
`id="column", description="Single Column Chooser",\`
`dataset=ds)`
	Creates a component to select one column from the dataset *ds*

`# Multiple column chooser`
`c = createComponent(type="columnlist",\`
`id="columnlist", description="Multiple Column Chooser",\`
`dataset=ds)`
	Creates a component to select multiple columns from the dataset *ds*

`# Entity List chooser`
`c = createComponent(type="EntityList",\`
`id="entityList", description="Entity List")`
	Creates a component to select an Entity List

`# Group 2 or more of the previous components`
`groupComponent = createComponent(type="group",\`
`id="group", description="Inputs",\`
`components=[c1,c2])`
 	
`result = showDialog(groupComponent)`
	Creates a component that combines two or more components. The 2 components, $c1$ and $c2$, which can be examples of any of the components shown above, are grouped together and shown in the same dialog to the user. *result* is a map whose keys correspond to the *id* argument of the components in the group. The corresponding values are again the values of the constituent components. For example, if component $c1$ has $id = id1$, then $result['id1']$ is the value of component $c1$

Continued on Next Page...

Table 37.5 -- Continued

Usage	Notes

`# Example of a dialog with a banner message`
`from script.omega import createComponent, showSimpleDialog`
 	
`c = createComponent(type="group",\`
`id="group", description="Choose inputs",\`
`components=[c1,c2,c3,c4])`
 	
`props = {'title' : 'Title for the dialog',\`
`'hasBanner' : java.lang.Boolean(1),\`
`'bannerDescription' : 'Detailed banner description',\`
`'cancelLabel' : 'Close',\`
`'hasOK' : java.lang.Boolean(1)}`
 	
`result = showSimpleDialog(c, properties=props)`
 	
`print result`
	Each line in the example is explained below:
	(1) Required imports
	(2) Creates a component c combining components $c1$, $c2$, $c3$, $c4$ (which can be examples of any of the components shown above)
	(3) Creates a properties map for the dialog that will show up
	(4) Shows the dialog to the user
	(5) Prints result obtained

Table 37.5: Creating UI Components

37.2.6 Example Scripts

Example 1 : Display MVA Plot

```
from script.view import ScatterPlot
from script.omega import createComponent, showSimpleDialog
from script.marray.project.analysis import getNormalisedDatasetNode

# define a function to return the nomalized dataset
def getDataset():
exp = script.marray.project.getActiveExperiment()
analysis = exp.getAnalysis()
node = getNormalisedDatasetNode(analysis)
return node.getDataset()

#  define a function for opening a dialog, given dataset d
def openDialog(d):
A = createComponent(type='column', id='Col A', dataset=d)
B = createComponent(type='column', id='Col B', dataset=d)
C = createComponent(type='column', id='Color By', dataset=d)
g = createComponent(type='group',id='g',description='Inputs',\
components=[A, B, C])
r = showSimpleDialog(g)
return r and (r['Col A'], r['Col B'], r['Color By']) or (0,0,0)

# function to show the plot with two columns and color by column
def showPlot(avg, diff, color):
plot = script.view.ScatterPlot(title = "MVA Plot",\
xaxis=avg,yaxis=diff)
plot.colorBy.columnIndex = color
plot.show()

# ------------ Main -------------
# open a dialog, and take inputs
d = getDataset()
a, b, color = openDialog(d)
if a and b:
# Compute the average and difference
avg = (d[a] + d[b])/2
avg.setName('Average')
diff = d[a] - d[b]
diff.setName('Difference')

# Append the columns to the dataset
d.addColumn(avg)
d.addColumn(diff)

# Show the Plot
x = d.indexOf(avg)
y = d.indexOf(diff)

showPlot(x, y, color) # set color = -1 for no color
```

Example 2 : Create Entity List

```
# This script will create a new sub-entitylist with those
# entities from "All Entities" who have present flag ("P")
# in at least 1 out of all samples.

# get the active experiment
activeExperiment = script.marray.project.getActiveExperiment()

# get the analysis object from active experiment
analysis = activeExperiment.getAnalysis()

# columns of flag dataset are the flag columns from all the samples
flagDataset = \
script.marray.project.analysis.getFlagDatasetNode(analysis).getDataset()

# find row indices where at least 1 of the flag columns have "P" flag
from com.strandgenomics.cube.framework.data import DefaultIntArray
indices = DefaultIntArray()
for row in range(flagDataset.getRowCount()):
flag = 1
for col in range(1, flagDataset.getColumnCount()):
if flag and flagDataset[col].get(row) == "P":
indices.add(row)
flag = 0

# get "All Entities" list from active experiment
allEntities = activeExperiment.getAllEntitiesEntityList()

# create new sub-entitylist
# createNewSubEntityList(parent, name, notes, indices, columns)
# where, parent = parent entitylist
# name = name of the new entity list
# notes = notes for the new entity list
# indices = subset of row indices of parent entity list
# columns = columns to add to the new entity list;
# size of column should be equal to size of new entity list

columns = []
subEntityList = \
script.marray.project.elist.createNewSubEntityList(\
allEntities,"presentProbes", "notes", indices, columns)

# add new sub-entitylist to All Entities list
allEntities.addChild(subEntityList)
```

Example 3 : Accessing entity lists, samples etc.

```
# get active entity list of the project
proj = script.marray.project.getActiveProject()
entitylist = proj.getActiveEntityList()

# get identifier column of the entity list
identifierColumn = entitylist.getIdentifierColumn()
print "Identifier Column : ", identifierColumn.getName()

# get all columns in the entitylist as an array
columns = entitylist.getColumns()
print "Columns:"
for i in xrange(len(columns)):
print "\tCol[%d] : %s" % (i, columns[i].getName())

# get technology of entitylist
technology = entitylist.getTechnology()
print "Technology of active entity list : ", technology

# get primary mark of entity list
mark = entitylist.getPrimaryMark()
print "Primary Mark of active entity list : ", mark

# get size of the entity list
size = entitylist.getSize()
print "Size of entity list : ", size

# get the active experiment
activeExperiment = script.marray.project.getActiveExperiment()

# get the array of all the samples in the active experiment
samples = activeExperiment.getSamples()

# get the number of samples
numSamples = len(samples)
print "Number of samples in the active experiment: ", numSamples
```

Example 4 : Obtain dataset associated with entity list

```
from com.strandgenomics.marray.project.translator \
import EntityListTranslator
from com.strandgenomics.cube.framework.data import ArrayUtil
from com.strandgenomics.cube.dataset import DatasetUtil
from script.marray.project.analysis import getNormalisedDatasetNode

# function to get entity list dataset of particular entity list
def _getEntityListDataset0(entityList, expt):
analysis = expt.getAnalysis()
node = getNormalisedDatasetNode(analysis)
dataset = node.getDataset()
idColumn = dataset.getColumn(0)
eidColumn = entityList.getIdentifierColumn()
indices = EntityListTranslator.getMatchingRowIndices(\
eidColumn, idColumn)
indices = ArrayUtil.createIndexedIntArray(indices)
return DatasetUtil.getRowSubsetDataset(dataset, indices)

# function to return entity list named 'name
def getEntityListDataset(name):
expt = script.marray.project.getActiveExperiment()
lists = expt.getAllEntityLists()
for entityList in lists:
if entityList.getName() == name:
return _getEntityListDataset0(entityList, expt)

# ------------ Main ---------
entityListName = "Entitylist of selection"
d = getEntityListDataset(entityListName)
if d:
print "Number of rows : ", d.getRowCount()
print "Number of columns :", d.getColumnCount()
else:
print "No such entity list found!"
```

Example 5

```
# Creates a dataset containing identifier, signal and annotation columns
# associated wih the active entitylist and active interpretation

v=script.marray.view.Table()
d=v.dataset
script.view.Table(dataset=d).show()
```

Example 6 : Demonstration of Search API

```
from java.util import *

# function to define searchable object types
def getObjectTypes():
objTypes = ArrayList()

# Search for all objects of type Entitylist
objTypes.add("Entitylist")

# You can search for multiple object types simultaneously
# objTypes.add("Sample")
# objTypes.add("Experiment")

return objTypes

# function to define search conditions
def getConditions():
conditions = ArrayList()

# Each condition is represented using a java HashMap.
# The HashMap has 3 keys:
# "Search Field" : The value for this key must be a string;
# "Search Value" : The value for this must be either
# a string or a java Date object, as appropriate;
# "Condition" :
# - If  "Search Value" contains a string, then "Condition"
# can be one of "equals", "includes", "starts with" or "ends with"
# - If  "Search Value" contains a Date or a number,then "Condition"
# can be one of "=", "<=" or ">="
# Please go to the Advanced Search option in Search in the Tool Bar.
# The drop down boxes there give a list of values
# that are applicable to each field.

condition_1 =  HashMap() #first condition
condition_1.put("Search Field", "Name")
condition_1.put("Condition", "includes")
condition_1.put("Search Value", "Filtered on Flags")

condition_2 =  HashMap() #second condition
condition_2.put("Search Field", "Number of entities")
condition_2.put("Condition", ">=")
condition_2.put("Search Value", "3000")

# you can add more conditions if you wish.
conditions.add(condition_1)
conditions.add(condition_2)
return conditions
```

```
# function to define how to combine conditions
def getJoinByType():
# The conditions mentioned in getConditions
# can be joined using "OR" or "AND"

return "OR"
#return "AND"

# function to define search complexity
def getIsSimple():
# 0 indicates advanced search,
# 1 indicates simple search (no conditions)

return 0
#return 1

# main search function
def runSearch(objTypes, conditions, joinBy, isSimple):
from com.strandgenomics.marray.project.search import *

searchManager = SearchHandler.getSearchManager()
return searchManager.search(\
objTypes, conditions, joinBy,isSimple)

# function to parse search results
def parseResults(results):
from com.strandgenomics.marray.project.search import *
from script.marray.eproject.search import createDataNodes
from com.strandgenomics.enterprise.client.filesystem import *

dataset = SearchUtil.getResultsDataset(results)
if not dataset:
print "No results found!"
return

numRows = dataset.getRowCount()
objectToTypeMap = LinkedHashMap()
for index in range(0, numRows):
objPath = dataset.getColumn("Path").get(index)
object = FileObject(objPath)
type = dataset.getColumn("Type").get(index)
objectToTypeMap.put(object, type)

dataNodes = createDataNodes(objectToTypeMap)

return dataNodes
```

```
# --------- Main ------------
objTypes = getObjectTypes()
conditions = getConditions()
joinBy = getJoinByType()
isSimple = getIsSimple()

results = runSearch(objTypes, conditions, joinBy, isSimple)
# Alternatively, for a simple search without filtering based on conditions,
# you can try:
# results = runSearch(objTypes)
if results:
resultNodes = parseResults(results)
# illustrates result
print "Search Results : "
for node in resultNodes:
print node.getName()
```

37.3 The R Editor

37.3.1 Setting up R Editor in GeneSpring

Following these steps to setup and use R Editor in **GeneSpring**:

1. Close **GeneSpring** if it is running.
2. Go to `< INSTALL_DIR >\bin\packages\external\jython\2.1\lib`
3. Open the file named ‘‘registry” present at this location.
4. Edit the line ‘‘`#python.os=None`” to read ‘‘`python.os=nt`” and save and close the file.
5. Now the batch scripts can be run.
6. Please edit the file and add a ‘‘#” before the line ‘‘`python.os=nt`” before restarting **GeneSpring**.

R scripts can be called from **GeneSpring** via *Tools* $\longrightarrow$ *R Editor*. You will need to first set the path to the R executable in *Tools* $\longrightarrow$ *Options* $\longrightarrow$ *Miscellaneous* $\longrightarrow$ *R Path*, then write or open an R script in this R Editor, and then click on the Run icon. If this path is incorrect, a dialog saying ‘‘Unable to find R on this system.’’ will pop up.

Example R scripts are available in the *samples/rScripts* sub-folder of the installation directory; these show how the **GeneSpring** dataset can be accessed and sent to R for processing and how the results can be fetched back. Access is provided to the normalized intensity values and the raw intensity values for all experiment types except;

- Copy Number,
- Association, and

Help on writing R scripts in **GeneSpring** appears below. Broadly, the input is;

- the active entity list,
- active interpretation in the currently active experiment.

The output is an entity list that is automatically added as a child of the input entity list. Additionally, this child entity list can have associated data that is generated by the R script.

37.3.2 Commands related to R input from GeneSpring

1. Create a data frame in R and populate it with data from **GeneSpring** . The data can either be the raw or normalized intensity values.

```
eg<-getDataset()

# Rows in data frame : all rows corresponding to the active entity list
# Columns in data frame :
#  First column : The Identifier column
#  Subsequent columns : the normalized intensity data columns
#                           corresponding to the active interpretation

eg<-getRawDataset()

# Rows in data frame : all rows corresponding to the active entity list
# Columns in data frame :
#  First column : The Identifier column
# Subsequent columns : the raw intensity data columns
#                                corresponding to the active interpretation
```

2. A data frame in R, populated with the experimental grouping information for the active interpretation

```
dg<-getExperimentGroupingDataset()

# the first column contains sample names for the active interpretation
# the second column contains condition names

dg<-getExperimentGroupingDataFrame()

# the first column contains sample names for the active interpretation
# subsequent columns contain the parameters in the active interpretation,
# one parameter per column
```

3. Get column index

```
ci<-getColumnIndices(condition_name, eg=getExperimentGroupingDataset())

# This is with respect to the data.frame returned by getDataset()
# 'condition_name' is one of the values in the second column of
# getExperimentGroupingDataset()
```

4. To obtain user input from a dialog

```
# to ask for a value "FDR CutOff" (a decimal number), with a default value of 0.05 :
   fdr<-askUser("FDR CutOff",float,.05)

# to ask for a value "Num Permutations" (a integer), with a default value of 100 :
   numPerm<-askUser("Num Permutations",int,100)
```

37.3.3 Commands related to R output to GeneSpring

To send results back to **GeneSpring**:

```
addResultColumns(data.frame(row_index, sum), c("Sum"))

# use the above function with 2 arguments :
# a dataframe and a list of output column names.

# Data frame first argument : row indices from the input entity list,
# for instance, if the input entity list has 100 entities, and you want
# the output to have the 1st, 10th and 15th items from the input entity list;
# ensure that rowindex contains just 1,10,15.

# Data frame subsequent arguments : any data columns associated with
# the output entity list; note that these must have exactly the same
# length as the output entity list, i.e., the length of rowindex.

# Output Column Names :
# a vector containing the name of each data column to be added to the output entity list
```

37.3.4 Debugging a Script

1. While debugging an R script, it is useful to print out the values of the important variables. This is done in R, by simply mentioning the variable name (as you would when working in R). For example, you can print out the value in sum as follows:

   ```
   # assigning value to sum
   sum <- sum[sum > 0]

   # printing value in sum
   sum
   ```

   The value in the variable sum will be printed in the log window panel ('Console') of the R Editor window. Any errors occurring in R will also appear here.

2. If an error ''Unable to parse R generated output file `<filename>`'' occurs in the Console window, it means that the execution process in R did not complete successfully and therefore **GeneSpring** could not retrieve a valid result.

   Check that the function addResultColumns(..) has been called correctly. (The function writes data from the R program to a temporary file on the disk, which **GeneSpring** then parses to generate the child entity list)

3. If the R script fails to execute successfully, check log files in the `<INSTALLATION_DIRECTORY>/app/tmp/R_SCRIPT` folder

37.3.5 Example R scripts

Example 1

```
#The actual R script content appears below between the ‘‘hash lines’’
#The following simple example shows how to compute the sum and
#standard deviation of the data columns

# data.frame
df<-getDataset()

numrows <- nrow(df)
numcols <- ncol(df)

# first column is identifier column.
# data starts from second column.
data <- df[2:numcols]

sum <- (1:numrows)
row_index <- (1:numrows)

for (i in 1:numrows) {
s <- sum(data[i,])
if (s <= 0) {
sum[i] <- -1
row_index[i] <- -1
       }
       else {
       sum[i] <- s
       row_index[i] <- i
       }
      }

sum <- sum[sum > 0]
row_index <- row_index[row_index > 0]
addResultColumns(data.frame(row_index, sum), c("Sum"))
```

Example 2

R Script to run SAM in R

```
# Choose an entity list and an appropriate interpretation in the navigator
# (currently working only for non-averaged interpretations)
# Run the script
# Choose an experiment parameter with 2 conditions
# Provide FDR CutOff value and Number of Permutations
# The result will be a new child entity list with
# SAM q-values and regulations stored as list associated values
-------------------------------------------------------------------------------

#user inputs, possible types are int, float, string
-------------------------------------------------------------------------------
fdr<-askUser("FDR CutOff",float,.05)
numPerm<-askUser("Number of Permutations",int,100)


# get normalized data corresponding to the interpretation
-------------------------------------------------------------------------------
df<-getDataset()
numrows <- nrow(df)
numcols <- ncol(df)
data_names <- names(df)

# get conditions for the chosen interpretation
-------------------------------------------------------------------------------
eg<-getExperimentGroupingDataset()
condition_names <- unique(eg[2])
num_conditions <- length(condition_names[,1])

# The core R Script
-------------------------------------------------------------------------------
library(Biobase)
library(samr)

# do only if number of conditions for chosen interpretation is exactly 2
if (num_conditions==2) {

    numSamples = length(data_names) - 1
    # subtract 1 because first column in data_names is the ID column
    # assign class labels
    c1 <- getColumnIndices(condition_names[1,1], eg)
    c2 <- getColumnIndices(condition_names[2,1], eg)
    sample.class<-c(rep(1, numSamples))
    for (i in 1:length(c1)) {sample.class[c1[i]-1]=1}
    for (i in 1:length(c2)) {sample.class[c2[i]-1]=2}

    # set up sam
    sam.data <-list(x=df[,2:numcols],y=sample.class,geneid=rownames(df),
```

```
    genenames=df[,1],logged2=TRUE)
    samr.obj <- samr(sam.data,  resp.type="Two class unpaired", nperms=numPerm)

    # determine delta based on fdr
    delta.table <- samr.compute.delta.table(samr.obj)
    delta = delta.table[1,1]
    for (i in 1:nrow(delta.table)) {
if (is.nan(delta.table[i,5]) || (fdr>delta.table[i,5])) {delta=delta.table[i,1];break}
    }

    # filter on delta
    siggenes.table <- samr.compute.siggenes.table(samr.obj,delta, sam.data, delta.table)
    up.genes <- siggenes.table$genes.up
    down.genes <- siggenes.table$genes.lo

    # set up q value and fold change vectors
    rowindex <- (1:numrows)
    for (i in 1:numrows) {rowindex[i] <- i}
    qval <- rep('NA',numrows)
    qval[as.numeric(up.genes[,3])] <- as.numeric(up.genes[,8])
    qval[as.numeric(down.genes[,3])] <- as.numeric(down.genes[,8])
    dir <- rep('NA',numrows)
    dir[as.numeric(up.genes[,3])] <- 'UP'
    dir[as.numeric(down.genes[,3])] <- 'DOWN'

    # identify number of non-missing q vals, ugly, needs to be made more elegant
    nonMissing=0
    for (i in 1:numrows) {if (qval[i]!='NA') {nonMissing<-nonMissing+1}}
    if (nonMissing>0) {
      qvalResult<-rep('NA',nonMissing)
      dirResult<-rep('NA',nonMissing)
      rowindexResult=rep(1,nonMissing)
      j=1
      for (i in 1:numrows)
        {if (qval[i]!='NA') {
          rowindexResult[j]=rowindex[i]
          qvalResult[j]=qval[i]
          dirResult[j]=dir[i]
          j=j+1
         }
        }
      # put back the entity list
      addResultColumns(data.frame(rowindexResult, qvalResult,dirResult),
      c("SAMR Q-value","SAMR REGULATION"))
   }
}

if (num_conditions!=2) {"The Interpretation does not have exactly 2 conditions"}
```

37.4 GeneSpring-R Integration

37.4.1 Old paradigm

The previous section dealt with R Scripting via the following paradigm: A new R process is launched from **GeneSpring** and data is exchanged between **GeneSpring** and R via the disk. To elaborate, the following steps take place when a R script is run from the R Editor

1. Relevant **GeneSpring** data (raw or normalized intensity data, experiment grouping information) is written to disk.
2. R functions to read data from disk and write data to disk are added to the R script. (This is how R is able to interpret the API. For example the `getDataset()` function basically consists of a *read.table* call)
3. This modified R script is now executed as a R batch command.
4. Upon completion of this R process, data from R is written to disk.
5. This R-written data is finally uploaded back into **GeneSpring** as an entity list.

37.4.2 GeneSpring-R Integration : New paradigm

The earlier approach, while quite simple, had several limitations, chief among them being non-interactiveness. Secondly, one was limited to creating only entity lists (and only one during one run) in **GeneSpring** . Further, experiment data (raw and normalized values) are always written out to disk, irrespective of whether they are required or not.

These limitations are overcome by the following new paradigm: Use of a library package for R (*genespring*) to enable dynamic data access. Note that now, programmatic access into and out of **GeneSpring** can be done from R itself. In other words, users can program entirely in R, as long as the **GeneSpring** tool is running. The R Editor can still be used to save and execute the final developed script.

37.4.3 Package Installation Instructions

To install the R library *genespring*, execute the following script from Tools $\longrightarrow$Script Editor.

```
# begin script
script.marray.product.rPackage.install()
# end script
```

The above script does the following :

1. Installs (or updates) the latest rJava library package in R.
2. Creates and configures the R library **GeneSpring** according to your installation directory.
3. Installs the library **GeneSpring** in R.

Additional details can be found in `<INSTALLATION_DIRECTORY>/app/GeneSpring-R/README.txt`

37.4.4 R Application Programming Interface

The API allows you to retrieve data from **GeneSpring** and set data in **GeneSpring**.

1. Search **GeneSpring** for Projects, Experiments, Entity Lists, and Samples
2. Download **GeneSpring** data like:
   (a) Samples including sample attributes and sample attachments.
   (b) Normalized/Raw/Control intensity data and flag information.
   (c) Entity lists including list-associated values.
3. Data Setting (For Agilent Single Color, Agilent Two Color, Affymetrix, and Illumina Expression experiments only):
   (a) Upload a file as a sample attachment
   (b) Create a new experiment in **GeneSpring**.
   (c) Create a new entity list in **GeneSpring**.

The below table lists the important API calls in the R library package *genespring*. Further help and documentation can be obtained in R in the usual manner. (i.e. *genespring* in R will list out all the API calls and further help on each function can be obtained through *function*

R API call	Brief Description
initGS	Initializes R session to enable use of GeneSpring-R API
searchGS	**GeneSpring** Search for projects, experiments, samples, entity lists
createGSDataFrame	Creates a new **GeneSpring** Data Frame object from a regular data frame (Used in the ''setDataset'' API calls)
createNewEntityList	Creates a new Entity List in a **GeneSpring** experiment
createNewExperiment	Creates a new experiment in **GeneSpring** (Currently, experiment creation is possible only for Affymetrix, Agilent and Illumina Expression technologies)

Continued on Next Page...

Table 37.6 -- Continued

R API call	Brief Description
createEmptyExperimentObject	Creates an empty **GeneSpring** experiment object The following ''set'' API calls must be called on this object in order to initialize the experiment, before a createNewExperiment call can be made. • setExperimentName (sets experiment name) • setExperimentNotes (sets experiment notes) • setExperimentType (sets experiment type, one of Affymetrix/Agilent Single Color/Agilent Two Color/Illumina Expression) • setFlagDataset (sets flag dataset) • setNormalizedDataset (sets normalized dataset) • setRawDataset (sets raw dataset) • setSampleFileFilter (sets file filter) • setSampleFolder (sets sample folder) • setSampleSequenceInDataset (maps sample names to the corresponding data columns in a **GeneSpring** experiment object)
getChildren	Gets the children of an entity list, an aligned read or a folder in **GeneSpring**
downloadGSDataset	Downloads the specified dataset of a **GeneSpring** experiment (Can be any of raw, control, flag or normalized datasets)
downloadReadListAsSAM	Downloads reads in a **GeneSpring** read list (in SAM format)
downloadRegionList	Downloads **GeneSpring** region list in TSV format
downloadSample	Downloads a specified sample from **GeneSpring** (Can be any of raw, control, flag or normalized datasets)
downloadSamples	Downloads all the samples from the specified **GeneSpring** experiment
downloadSampleAttachment	Downloads a specified sample attachment from **GeneSpring**
uploadSampleAttachment	Uploads a file as a sample attachment in **GeneSpring**
getGSInstDir	Gets the **GeneSpring** installation directory
getAllEntitiesEntityList	Gets the ''All Entities" entity list of an experiment in **GeneSpring**
getChildren	Gets the children of an entity list or a folder in **GeneSpring**
getEntityListData	Gets entity list data for a specified entity list in **GeneSpring**
getExperimentGrouping	Gets the experiment grouping information of a **GeneSpring** experiment
getExperiments	Gets all experiments for a given project in **GeneSpring**
getGSDataset	Gets the specified type of dataset of a **GeneSpring** experiment (Can be any of raw, control, flag or normalized datasets)
getSampleAttachments	Gets all sample attachments for a given sample in **GeneSpring**
getSampleAttributes	Gets all the attributes of a sample in **GeneSpring**
getSamples	Gets all samples for a given experiment in **GeneSpring**
printConstants	Prints all **GeneSpring** package-specific constants
printExperimentDetails	Prints information about a **GeneSpring** experiment
getDemoExperimentId	(Utility function) Returns the ExperimentId of the Demo HeLa cells experiment in **GeneSpring**
getDemoProjectId	(Utility function) Returns the ProjectId of the 'Demo Project' in **GeneSpring**
readFromFile	(Utility function used in the **GeneSpring**-R API examples) Reads a data frame from disk.

Table 37.6: R API calls

37.4.5 Examples

An example R script which uses all the above API calls above can be found at
`<INSTALLATION_DIRECTORY>/app/GeneSpring-R/example.R`

Chapter 38

Table of Key Bindings and Mouse Clicks

All menus and dialogs in **GeneSpring** adhere to standard conventions on key bindings and mouse clicks. In particular, menus can be invoked using *Alt* keys, dialogs can be disposed using the *Escape* key, etc. On Mac **GeneSpring** confirms to the standard native mouse clicks.

38.1 Mouse Clicks and their actions

38.1.1 Global Mouse Clicks and their actions

Mouse clicks in different views in **GeneSpring** perform multiple functions as detailed in the table below:

Mouse Clicks	Action
Left-Click	Brings the view in focus
Left-Click	Selects a row or column or element
Left-Click + Drag	Draws a rectangle and performs selection or zooms into the area as appropriate
Shift + Left-Click	Selects contiguous areas with last selection, where contiguity is well defined
Control + Left Click	Toggles selection in the region
Right-Click	Bring up the context specific menu

Table 38.1: Mouse Clicks and their Action

Mouse Clicks	Action
Shift + Left-Click	Draw Irregular area to select

Table 38.2: Scatter Plot Mouse Clicks

Mouse Clicks	Action
Shift + Left-Click + Move	Rotate the axes of 3D
Shift + Middle-Click + Move up and down	Zoom in and out of 3D
Shift + Right-Click + Move	Translate the axes of 3D

Table 38.3: 3D Mouse Clicks

38.1.2 Some View Specific Mouse Clicks and their Actions

38.1.3 Mouse Click Mappings for Mac

Mac Mouse Clicks	Equivalent Action in Windows/Linux
Click	Left-Click
Apple + Click	Control + Left-Click
Shift + Click	Shift + Left-Click
Control + Click	Right-Click
Alt + Click	Middle-Click

Table 38.4: Mouse Click Mappings for Mac

38.2 Key Bindings

These key bindings are effective at all times when the **GeneSpring** main window is in focus.

38.2.1 Global Key Bindings

Key Binding	Action
Ctrl-N	New Project
Ctrl-O	Open Project
Ctrl-X	Quit **GeneSpring**
Ctrl-F	Search on entities
Ctrl-I	Shows Entity Inspector
Ctrl-R	Brings up 'Properties' window

Table 38.5: Global Key Bindings

Bibliography

[1] Affymetrix Latin Square Data. http://www.affymetrix.com/support/technical/sample_data/datasets.affx

[2] Identifying and Validating Alternative Splicing Events. http://www.affymetrix.com/support/technical/technotes/id_altsplicingevents_technote.pdf

[3] Reiner A, Yekutieli D and Benjamini Y: Identifying differentially expressed genes using false discovery rate controlling procedures, Bioinformatics, 19, 3, (368-375), 2003.

[4] Alternative Transcript Analysis Methods for Exon Arrays. http://www.affymetrix.com/support/technical/whitepapers/exon_alt_transcript_analysis_whitepaper.pdf

[5] Belsley DA, Kuh E, and Welsch RE, Regression Diagnostics. Hoboken, NJ: John Wiley & Sons, Inc., 1980.

[6] Benjamini B and Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Statist. Soc. B. 57, 289-300.

[7] Y.Benjamini and D.Yekutieli (2001): The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, **29**, 1165--1188.

[8] The Bioconductor Webpage. http://www.bioconductor.org.

Validation of Sequence-Optimized 70 Base Oligonucleotides for Use on DNA Microarrays, Poster at http://www.academia.edu/537989/Validation_of_Sequence-Optimized_70_Base_Oligonucleotides_for_Use_on_DNA_Microarrays.

[9] Caspi, R., Altman, T., Dreher, K., Fulcher, C.A., Subhraveti, P., Keseler, I.M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L.A., Ong, Q., Paley, S., Popescu, L., Pujar, A., Shearer, A., Travers, M., Weerasinghe, D., Zhang, P. and Karp, P.D. (2012) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases, *Nucleic Acids Research* **40(D1)**: D742-D753.

and

Keseler, I.M., Collado-Vides, J., Santos-Zavaleta, A., Peralta-Gil, M., Gama-Castro, S., Muniz-Rascado, L., Bonavides-Martinez, C., Paley, S., Krummenacker, M., Altman, T., Kaipa, P., Spaulding, A., Pacheco, J., Latendresse, M., Fulcher, C., Sarker, M., Shearer, A.G., Mackie, A., Paulsen, I., Gunsalus, R.P., and Karp, P.D. (2011) EcoCyc: a comprehensive database of Escherichia coli biology *Nucleic Acids Research* **39**: D583-D590.

[10] Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K, Lee C, Nizzari MM, Gabriel SB, Purcell S, Daly

MJ and Altshuler D: Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare NVs, Nature Genetics, 40, 10, 2008 http://www.nature.com/ng/journal/v40/n10/abs/ng.237.html

[11] Bolstad BM, Irizarry RA, Astrand M, and Speed, TP: A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. Bioinformatics 19(2):185-193 Supplemental information, 2003.

[12] BRLMM: an Improved Genotype Calling Method for the GeneChip Human Mapping 500K Array Set http://www.affymetrix.com/support/technical/whitepapers/brlmm_whitepaper.pdf

[13] Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, and Nickerson DA: Selecting a Maximally Informative Set of Single-Nucleotide Polymorphisms for Association Analyses Using Linkage Disequilibrium, Am. J. Hum. Genet., 74, 106-120, 2004. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1181897

[14] Olshen AB, Venkatraman ES, Lucito R, Wigler M.: Circular binary segmentation for the analysis of array-based DNA copy number data, Biostatistics, Oct;5(4):557-572, 2004. http://www.ncbi.nlm.nih.gov/pubmed/15475419

[15] DNA Copy Number Data Analysis. http://bioconductor.org/packages/2.4/bioc/html/DNAcopy.html

[16] dChip Software: Analysis and visualization of gene expression and SNP microarrays. https://sites.google.com/site/dchipsoft/.

[17] Dempster AP, Laird,NM, and Rubin DB: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. B 39, 1-38, 1977.

[18] Devlin B and Risch N: A Comparison of Linkage Disequilibrium Measures for Fine-Scale Mapping, Genomics, 29, 311-322, 1995.

[19] Dobson AJ: An Introduction to Generalized Linear Models. New York: Chapman & Hall, 1990.

[20] Dudoit S, Yang H, Callow MJ, and Speed TP: Statistical Methods for identifying genes with differential expression in replicated cDNA experiments, Stat. Sin. 12, 1, 11-139, 2000.

[21] Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ and Altshuler D: The Structure of Haplotype Blocks in the Human Genome, Science, 296(5576): 2225 - 2229, 2002. http://www.sciencemag.org/cgi/content/abstract/296/5576/2225

[22] GISTIC documentation http://www.broad.mit.edu/cancer/pub/GISTIC/

[23] Glantz S: Primer of Biostatistics, 5th edition, McGraw-Hill, 2002.

[24] GeneLogic Spike-In Study. http://www.genelogic.com/support/scientific-studies

[25] Gordon K. Smyth: Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments, 2004.

[26] The International HapMap Consortium, A second generation human haplotype map of over 3.1 million SNPs, Nature, October 18; 449(7164): 851-861, 2007. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2689609/?tool=pubmed

[27] Hill AA, Brown EL, Whitley MZ, Tucker-Kellog G, Hunter CP, and Slonim DK: Evaluation of normalization procedures for Oligonucleotide array data based on spiked cRNA controls, Genome Biology, 2, 0055.1-0055.13, 2001.

[28] Hill WG, Robertson A: Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* 38:226-231, 1968.

[29] Hoffmann R, Seidl T, and Dugas M: Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis, Genome Biology 3(7): 0033.1-0033.11, 2002.

[30] Hubbell E, Liu W and Mei R: Robust estimators for expression analysis. Bioinformatics. 18(12):1585-1592, 2002.

[31] Hubbell, E: Designing Estimators for Low Level Expression Analysis. http://mbi.osu.edu/event/?id=28#abstracts_Hubbell

[32] Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Hinz U, Jandrasits C, Jimenez RC, Khadake J, Mahadevan U, Masson P, Pedruzzi I, Pfeiffenberger E, Porras P, Raghunath A, Roechert B, Orchard S and Hermjakob H. The IntAct molecular interaction database in 2012 Nucl. Acids Res. 40(D1): D841 - D846, 2011.

[33] Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B and Speed TP: Summaries of Affymetrix GeneChip probe level data Nucleic Acids Research 31(4):e15, 2003.

[34] Irizarry, RA, Hobbs, B, Collin, F, Beazer-Barclay, YD, Antonellis, KJ, Scherf, U, Speed, TP: Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. Biostatistics 4(2): 249-264, [Abstract, PDF, PS, Complementary Color Figures-PDF, Software], 2003.

[35] John D. Storey: A direct approach to false discovery rates, 2001.

[36] John D. Storey and Robert Tibshirani: Statistical Significance for Genome-Wide Experiments, 2003.

[37] Kanehisa M, Goto S: KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research 28: 27-30, 2000.

and

Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M and Tanabe M: Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Research. 42, D199-D205, 2014.

[38] Lewontin RC. On measures of gametic disequilibrium. Genetics 120: 849-852, 1988.

[39] Li C and Wong WH: Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application, Genome Biology. 2(8), 0032.1-0032.11, 2001.

[40] Li C and Wong WH: Model based analysis of oligonucleotide arrays: Expression index computation and outlier detection, Proc Natl Acad Sci USA. 98: 31-36, 2001.

[41] The Lowess method. http://www.itl.nist.gov/div898/handbook/pmd/section1/pmd144.htm.

[42] Loman, N. Detection of Allelic Deletion via Inference of Loss of Heterozygosity in Acute Lymphoblastic Leukemia by use of a Hidden Markov Model. M.Sc Thesis, Department of Statistics, University of Chicago. 2006. http://www.stat.uchicago.edu/students/seminars/2005-2006/Loman_Nathaniel_080206.pdf.

[43] MAGE-TAB format. http://tab2mage.sourceforge.net/docs/magetab_docs.html

[44] Statistical Algorithms Description Document, Affymetrix Inc. http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf.

[45] McCullagh, P., and J. A. Nelder. Generalized Linear Models. New York: Chapman & Hall, 1990.

[46] Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR and Evelo C. WikiPathways: Pathway Editing for the People. PLoS Biol 6(7), 2008.

[47] Pritchard JK and Przeworski M. Linkage disequilibrium in humans: models and data. Am. J. Hum. Genet. 69: 1-14. 2001.

[48] Rabiner LR. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE 77, 2, Feb 1989. http://www.cs.ubc.ca/~murphyk/Bayes/rabiner.pdf.

[49] Reich DE and Goldstein DB: Detecting Association in a Case-Control Study While Correcting for Population Stratification, Genetic Epidemiology 20:416, 2001.

[50] SGD project. ‘‘Saccharomyces Genome Database - YeastMine Interactions Database” http://yeastmine.yeastgenome.org/yeastmine/aspect.do?name=Interactions (accessed January 2012).

[51] Shaw RG and Olds TM: ANOVA for Unbalanced Data: An overview, Ecology, 74, 6, (1638-1645), 1993.

[52] Speed T: Always log spot intensities and ratios, Speed Group Microarray Page. http://www.stat.berkeley.edu/~terry/zarray/Html/log.html.

[53] Speed FM, Hocking RR, and Hackney OP: Methods of Analysis of Linear Models with Unbalanced Data, J. Am Stat Assoc, 73, 361, (105-112), 1978.

[54] Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, and Reich D: Principal components analysis corrects for stratification in genome-wide association studies, Nature Genetics, 38 (8), 904-909, 2006. http://www.nature.com/ng/journal/v38/n8/abs/ng1847.html

[55] Patterson N, Price AL, and Reich D: Population Structure and Eigenanalysis, PLoS Genet 2(12): e190. doi:10.1371/journal.pgen.0020190 http://www.plosgenetics.org/article/info:doi/10.1371/journal.pgen.0020190

[56] Strand Life Sciences **GeneSpring**. http://www.strandls.com/products/genespring

[57] van Iersel MP, Pico AR, Kelder T, Gao J, Ho I, Hanspers K, Conklin BR and Evelo CT. The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. BMC Bioinformatics, 11(1):5, 2010.

[58] Westfall PH, Young SS: Resampling based multiple testing. John Wiley and Sons. New York, 1993.

[59] Wu Z, Irizarry RA, Gentleman R, Murillo FM, and Spencer F: A Model Based Background Adjustment for Oligonucleotide Expression Arrays. Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 1 (May 28, 2004).

[60] Benjamini Y, and Yekutieli D: The control of false discovery rate under dependency, Ann Stat, 29, (1165-1188), 2001.

[61] Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, and Ehm MG Testing Association of Statistically Inferred Haplotypes with Discrete and Continuous Traits in Samples of Unrelated Individuals Hum Hered 2002;53:7991

[62] http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient

[63] http://en.wikipedia.org/wiki/Spearman's_rank_correlation_coefficient

Index

χ^2 correlation test, 1183
χ^2 test, 1182
t-test
    Welch, 650
GeneSpring Manager
    navigation tree, 1208
    user interface, 1207

activation, 4, 7, 10
advanced workflow
    create gene level experiment, 244
    create interpretation, 240
    experiment grouping, 236
    experiment setup, 236
advancedwf
    utilities, 289
affy exon expression, 533
affymetrix, 361
affymetrix exon
    technology, 587
agilent exon, 483
    data processing, 485
    experiment creation, 484
    expression analysis, 503
    overview, 484
    quality control, 496
    splicing analysis, 503
    splicing visualization, 518
    technology, 529
agilent mirna, 433
    advanced workflow, 460
    backward compatibility, 442
    guided workflow, 446
    sample validation, 441
    targetscan, 479
    translation, 481
annotation
    Gene Ontology Browsing, 827
ANOVA
    F-ratio, 654
    Friedman, 656
    Kruskal-Wallis, 655
    N-way, 657
    N-way
        Type III SS, 657
    one-way, 653
    Repeated Measures, 656
    sum of square, 653
    Welch, 655
association
    views, 1198
association analysis, 1163
    χ^2 correlation test, 1183
    χ^2 test, 1182
    analysis, 1175
    Birdseed Report, 1168
    cochran-armitage test, 1183
    EIGENSTRAT correction, 1175
    EIGENSTRAT filter, 1169
    filter samples by missing values, 1168
    filter snps by hwe p-value, 1173
    filter snps by maf, 1174
    filter snps by minor allele frequency, 1174
    filter snps by missing value, 1173
    filters, 1172
    find significant snps, 1193
    fisher's exact test, 1182
    genome browser, 1198
    genomic control, 1180
    haplotypes, 1195
    haplotyping trend regression, 1195
    hardy-weinberg equilibrium, 1173
    Identify overlapping genes, 1200
    identify snps with differential missingness, 1173
    ld analysis, 1196
    linkage disequilibrium, 1196
    mlr, 1194
    mode of inheritance, 1178
    mtc, 1181
    multiple linear regression, 1194
    multiple logistic regression, 1194
    multiple testing correction, 1181
    permutative p-value computation, 1180
    qc, 1167
    quality control, 1167

results analysis, 1200
snp regression, 1193
snp tagging, 1191
statistical analysis, 1178
technology, 1164
utilities, 1200

BioCyc pathways
import, 866
BridgeDb, 1012

cgh, 1079
classification, 801
build prediction model, 804
confusion matrix, 824
decision trees, 818
lorenz curve, 825
naive bayesian, 816
neural network, 821
pipeline, 802
plsd, 812
predicting outcomes, 823
report, 824
support vector machines, 813
training, 803
validation, 803
viewing results, 823
classification report, 824
classify
decision trees, 823
neural network, 823
SVM, 823
clustering, 669
cochran-armitage test, 1183
confusion matrix, 824
copynumber, 1107
affybatcheffectcorrection, 1123
affyworkflow, 1117
algorithms, 1152
analysis, 1128
configuration, 1150
create custom reference, 1149
disc cache, 1150
filters, 1135
gistic, 1131
gistic overlapping genes, 1135
illumina, 1146
performance statistics, 1152
resultsanalysis, 1144
technology, 1108
terminology, 1108
translation, 1150
tutorials, 1162
utilities, 1146
views, 1142
correlation, 747
custom affy using cdf, 358
custom agilent arrays, 333
custom illumina technology, 382

decision trees, 818
classify, 823
model, 820
training, 819
Differential Expression Analysis, 643
distance measures
angular, 742
canberra, 743
chebychev, 743
differential, 743
euclidean, 742
harmonic, 743
manhattan, 742
minimum, 743
pearson centered, 743
pearson centered abslte, 744
pearson uncentered abslte, 744
pearson uncentered cosine, 744
squared euclidean, 742

egb, 214
EIGENSTRAT correction, 1175
entitylist, 32
es value, 845, 855
experiment, 27

find significant haplotypes, 1195
find significant snps, 1193
fisher's exact test, 1182

gb
chromosomeview, 206
generic single color, 383
generic two color, 411
genes et, 839, 849
genome browser, 195, 1198
genomebrowser
annotationstab, 207
dragentitylists, 204
dragexperiments, 200
dragfiles, 204
edittrackproperties, 210
genomeview, 205
gettingstarted, 196
regionview, 199
rightclickoptions, 213
searchresult, 209
spreadsheet, 207

trackoperations, 214
utilitiescopynumberexperiments, 218
viewcustomizations, 209
visualization, 198
genomic control, 1180
geoimport, 625
experiment parameters, 627
load dataset, 625
possible error messages, 630
getting started
help, 72
script, 62
Tools Options, 62
ui, 21
GO Analysis, 828
GO Terms, 827
goanalysis, 827
groups, 1207
gsa, 849
genes et, 849
introduction, 849
gsea, 839
genes et, 839
introduction, 839
gswgclient, 1203
guided workflow, 222, 435
affymetrix, 338
agilent single color, 300
agilent two color, 319
exonexpression, 537
illumina, 370

haplotypes, 1195
haplotyping trend regression, 1195
hardy-weinberg equilibrium, 1173

identify snps with differential missingness, 1173
inheriting permissions, 1212
installation, 1
activation, 4, 7, 10
copy number and association experiments, 17
Linux, 6
mac, 9
upgrade, 17
windows, 2
interpreting task logs, 1215

ld analysis, 1196
license manager, 13
linkage disequilibrium, 1196
linkage rule
average, 745
centroid, 745
complete, 745
median, 745
single, 745
ward's, 744
Linux
install, 7
uninstall, 8
login dialog, 1206
lorenz curve, 825

Mac
uninstall, 11
mac
install, 9
MAGE-TAB, 633
Points to note, 641
Validation, 641
Mann-Whitney, 652
Mann-Whitney
paired, 652
migration, 105
minor allele frequency, 1174
mlr, 1194
MOA
experiment, 931
workflow, 896
MOA Experiment, 603
mode of inheritance, 1178
model
decision trees, 820
naive bayesian, 817
neural network, 822
SVM, 815
multiple linear regression, 1194
multiple logistic regression, 1194
Multiple Testing Correction, 660
Multiple Testing Correction
Benjamini-Hochberg method, 663
Benjamini-Yekutieli method, 662
Bonferroni, 661
Holm method, 661
Westfall-Young method, 662

naive bayesian, 816
model, 817
train, 817
nes value, 845, 855
neural network, 821
classify, 823
model, 822
training, 821
ngsbridge, 613
NLP, 1037
Normalization

Lowess, 646
Quantile, 645

object ownership, 1209
object permissions, 1209
objects, 1209

pathway, 859
analysis, 859
components, 949
create, 878
edit, 878
export, 885
heatmap, 962
heatstrip, 960
import, 670, 863
list, 912
quilt plot, 960
transfer, 885
view, 940
viewer, 943
permission conflicts, 1212
permutative p-value computation, 1180
pipelines
classification, 802
platforms, 1
Post Hoc Testing, 654
Post-Hoc
SNK, 654
Tukey HSD, 654
prediction
what is, 801
project, 26
propagating permissions, 1212

quicktour
backup, 67

remote execution, 1213
snp tagging, 1215
rtpcr, 593
run in background
add or remove samples, 1214
clustering, 1215
experiment creation, 1214

sample, 28
scatterplot
visualization, 133
scripts, 1217
SEA, 897
significant haplotypes, 1195
significant snps, 1193
similarity measures
pearson, 799
spearman, 799
snp regression, 1193
summarization, 361
summarization
GCRMA, 363
Invariant Set, 363
Li-Wong, 363
MAS5
Absolute Calls, 365
median polish, 363
PLIER, 364
quantile, 362
RMA, 362
Tukey-BiWeight, 364
support vector machines, 813
SVM, 813
classify, 823
model, 815
training, 814

t-test, 648
t-test
paired, 649
tag snps, 1191
task manager, 1213
technology
association analysis, 1164
standard technology creation, 76
technology on the fly, 80
Thresholding, 643
train
naive bayesian, 817
training
classification, 803
decision trees, 819
neural network, 821
SVM, 814

ui, 21
desktop, 22
desktop navigator, 23
legend, 25
workflow, 24
update, 70
users, 1207

validation
for classification, 803
visualization, 115
time series plot, 138

WikiPathways import, 864
windows
install, 3
uninstall, 5