

Agilent Genomic Workbench Lite Edition 6.5

CGH Interactive Analysis

User Guide



Agilent Technologies

Notices

© Agilent Technologies, Inc. 2010, 2015

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.

Manual Part Number

G3800-90038

Edition

Revision A0, December 2015

Printed in USA

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

Trademarks

Microsoft ® is a registered trademark of Microsoft Corporation in the United States and/or other countries.

Adobe®, Acrobat® and Adobe®Reader® are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States and/or other countries.

Software Revision

This guide is valid for 6.5 and later revisions of the Agilent Genomic Workbench Lite Edition software, until superseded.

Warranty

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).

Safety Notices

CAUTION

A **CAUTION** notice denotes a hazard. It calls attention to an operating procedure, practice, or the like that, if not correctly performed or adhered to, could result in damage to the product or loss of important data. Do not proceed beyond a **CAUTION** notice until the indicated conditions are fully understood and met.

WARNING

A WARNING notice denotes a hazard. It calls attention to an operating procedure, practice, or the like that, if not correctly performed or adhered to, could result in personal injury or death. Do not proceed beyond a WARNING notice until the indicated conditions are fully understood and met.

In This Guide...

This guide describes how to use the CGH application of Agilent Genomic Workbench Lite Edition to analyze CGH data and to perform comparative genomic hybridization (CGH), copy number variation (CNV), and single nucleotide polymorphism (SNP) studies. It also describes how you can find common aberrant regions, and correlate CGH and expression data.

1 Getting Started

This chapter gives an overview of how to use Agilent Genomic Workbench Lite Edition to import, display, and analyze Comparative Genomic Hybridization (CGH) data, including data from microarrays that contain both CGH and SNP probes (CGH+SNP).

2 Importing, Managing, and Exporting CGH and CGH+SNP Data and Other Content

This chapter describes how to import, organize, manage, and export CGH and CGH+SNP data and other content within the user interface of Agilent Genomic Workbench Lite Edition.

3 Displaying CGH and CGH+SNP Data and Other Content

This chapter shows you how to display log ratio data from imported feature extraction data files and analysis results, as well as gene list and track content, in the Genomic Viewer. It also gives you instructions on how to modify the display to show the data and content the way you prefer.

4 Setting Up CGH Interactive Analysis

This chapter gives instructions on how to set up the interactive analysis functions for CGH and CGH+SNP experiments. These include the Preprocessing, Analysis, Discovery and Reports tabs.

5 CGH Interactive Analysis Reference

This chapter describes the shortcut menus, tab commands and dialog boxes for setting up CGH Interactive Analysis. The last section describes configuration settings.

6 Statistical Algorithms

This chapter describes the algorithms in the CGH module of Agilent Genomic Workbench. The program uses many different algorithms to perform the statistical analyses that are needed for comparative genomic hybridization (CGH), copy number variation (CNV), and single nucleotide polymorphism (SNP) studies. Additional algorithms let you find common aberrant regions, and also correlate CGH and expression data.

Contents

1 Getting Started 17

What is the CGH Application? 19

Using Agilent Genomic Workbench on a Mac 21

Entering a License and Starting the CGH Application 22

- To enter a license for analyzing CGH data interactively 23
- To start the CGH module 24

Using Main Window Components to Display/Analyze Data 25

- What are the main window components? 25
- What can you do with the main components for display of data and results? 27

Switching Applications 29

Using Tabs and Command Ribbons 30

- Tabs 30
- Commands 32

Using the Navigator to Search for Data 33

- To search the Navigator 35

Using the Genomic Viewer to Display Data and Results 36

- What is the Genomic Viewer? 36
- To change the size of and detach panes from the Agilent Genomic Workbench main window 38
- To maximize and reattach panes to the Agilent Genomic Workbench main window 39

General Instructions for Displaying Microarray Data/Results 40

General Instructions for Setting CGH Interactive Analysis Options 41

Getting Help 43

- To get help within Agilent Genomic Workbench 43
- To contact Agilent Technical Support 44

To learn about Agilent products and services 44

2 Importing, Managing, and Exporting CGH and CGH+SNP Data and Other Content 45

Importing Files 46

To select a different location for data files 47
To import Agilent GEML design files 48
To import Axon design files 49
To use eArray to update design files 49
To import a genome build 50
To import Agilent FE or Axon data files 51
To import a UDF file 52
To import tracks 55
To import array attributes 56
To import an experiment file 57
To import filters 58
To import a genotype reference file 58

Working with Experiments to Organize Imported Data 59

To display the array designs and data in the program 59
To create a new experiment 60
To add arrays to an experiment 62
To associate a genotype reference with a CGH+SNP array 63
To change the order of arrays in an experiment 64
To change the display names for arrays in an experiment 65
To rename an array in an experiment 65
To remove arrays from an experiment 66
To select or remove calibration array(s) 66
To show or hide array attributes in an experiment 67
To display or edit array attributes in an experiment 68
To display or edit the attribute values of a specific array 69

Managing Content 70

To display a list of the content stored in the program 70

To find specific items in the Navigator 71
To display the properties of a specific design 72
To update probe annotation in design files 72
To rename an array in the Data pane 73
To remove data or design files from the program 73
To create a gene list 74
To import a gene list 75
To display the genes in a gene list 75
To add one gene list to another 76
To rename a gene list 76
To delete gene list(s) 77
To create a track 77
To display the details of a track 78
To combine tracks 78
To rename a track 79
To delete tracks 80
To display genotype reference details 80
To rename a genotype reference 80
To delete a genotype reference 80

Exporting and Saving Content 82
To export array attributes 82
To export experiments 83
To export a gene list 84
To export tracks 84
To export filters 85
To copy what you see in the main window 85
To copy the list of array colors for an experiment 86
To save data and design information from an experiment 86
To save SNP data and design information from an experiment 86

3 Displaying CGH and CGH+SNP Data and Other Content 89

Selecting an Experiment for Analysis 90

To select an experiment 90
To select or deselect arrays in the experiment 91
To change the display color of an array 93
To analyze an experiment one sample at a time 94

Displaying Array Data 96
To display the scatter plots 96
To show or hide data in the scatter plots 98
To turn on or off data visibility in scatter plots 99
To customize scatter plot ranges and colors 99
To change scatter plot appearance 101
To print the scatter plot 102
To create custom scales for Views 102
To locate and display data (or results) within the Views 103
To create a Signal Intensity Bar Chart 104
To smooth and plot CGH log ratio data 105
To produce an echo example plot 106
To produce a MovAvg Example plot 107

Displaying Gene Lists and Tracks 109
To show gene lists in Gene View 109
To select gene list display color 110
To display a gene list as a table 110
To change the appearance of genes in Gene View 111
To show tracks in Gene View 111
To change the appearance of tracks 112
To show track information in reports 113
To limit aberration calls to the genomic boundaries of the track 114
To display tracks in UCSC Browser 114
To change the graphical display to a different genome build 115

Displaying CGH+SNP Content and Results 117
To show SNP data panel 117
To show LOH regions in genomic viewer 118

Searching for Probe and Gene Information 119
To search Tab View for specific probe information 119
To search Agilent eArray for probe information 120
To search the Web for information on probes in Tab View 121
To create a custom Web search link 121
To update or delete a custom Web search link 122

4 Setting Up CGH Interactive Analysis 123

Working with Interactive Analysis Options 124

Changing Preprocessing Options 126
To create or modify an array filter 126
To create or modify a feature filter 127
To create or modify a design filter 127
To apply a filter 128
To remove a filter 129
To apply GC Correction 129
To display a plot of GC corrected data 130
To apply centralization 130
To combine (fuse) arrays 130
To combine intra-array replicates 132
To combine interarray replicates 133
To display QC metrics of arrays and set array QC status 134
To display SNP QC Metrics 135

Changing Analysis Options 136
To set up a moving average calculation to smooth the data 136
To apply the aberration algorithm 137
To apply fuzzy zero correction 138
To apply the nesting filter 139
To calculate SNP copy number 139
To manually reassign peaks 140
To calculate LOH 141

Displaying Results Before Selecting Discovery Options 142

To show results of analysis 142
To move through View displays to observe results 143
To save a result 145
To restore a saved result 146
To print moving average and aberration calculation results 146
To customize the appearance of analysis results 146

Displaying Results with Discovery Options 150
To create or modify an aberration filter 150
To create common aberration summaries 151
To display common aberrations 152
To display graphical probe penetrance summaries 154
To create interval penetrance summaries 154
To visualize a CNVR (copy number variant region) 156
To set up and perform a joint analysis on CGH and expression data 156
To set up a differential aberration analysis 158
To create, modify, or apply a differential aberration filter 160
To set up a clustering analysis 161
To display results as a heatmap 162
To generate a genotype reference file 164

Creating Reports 166
To create and export common aberration text reports 166
To create CGH aberration text reports 166
To generate graphical CGH aberration reports 167
To generate a SNP Genotype Report 167
To generate a SNP Aberration & LOH Report 168
To generate probe penetrance text reports 169
To create a new Cyto Report template 169
To open an existing Cyto Report template 170
To generate a cytogenetic report 170

5 CGH Interactive Analysis Reference 173

Agilent Genomic Workbench Main Window 174

Switch Application Menu 175

Command Ribbons 176

- Home command ribbon 176
- Preprocessing command ribbon 179
- Analysis command ribbon 183
- Discovery command ribbon 189
- Reports command ribbon 195
- View command ribbon 197
- Tool command ribbon 198
- Help command ribbon 200

Navigator 202

- Search pane 204
- Data pane – icons, special text, and buttons 206
- Data pane – actions and shortcut menus 207
- Experiment pane – icons, special text, and buttons 209
- Experiment pane – actions and shortcut menus 210
- My Entity List pane – Icons, buttons, and special text 220
- My Entity List pane – actions and shortcut menus 220
- Genotypes pane 223
- Genotypes pane – Actions and shortcut menus 223

Genomic Viewer 224

- Genome View 224
- Chromosome View 226
- Gene View 228
- The View Cursor 232
- Tab View 233

Status Bar 238

Dialog Boxes 239

- Aberration & LOH Report Setup 239
- Aberration Filters 240
- Aberration Heatmap 242

Add Gene List <name> to 243
Agilent Feature Extraction Importer 245
Array Level Filters 247
Array Set 249
CGH Aberration Report Setup 250
CGHSmooth Parameters 251
CGHSmooth Plot 253
Chart Properties 255
Chromosome Selection 260
Combine Tracks 261
Compare CNVRs 263
Configure Coloring Ranges and Shades 264
Confirm 267
Confirm Overwrite 268
Copy Number Variant Region 269
Correlation Analysis Setup 270
Correlation Results 272
Create CNVR 274
Create Common Aberration 276
Create Experiment 278
Create Gene List 279
Create Gene List (from CNVR results) 281
Create Gene List (from common or differential aberration results) 282
Create Signal Bar Chart 283
Create Track 284
Customize Search Link 286
Define Sets 287
Design/Edit Cytogenetic Report Templates 288
Design Level Filters 297
Design Properties 299
Differential Aberration Setup 303
Echo Example Plot 305
Edit Array Color 308

Edit Array Order 309
Enrichment Analysis Setup 310
Experiment Properties 313
Export 315
Export Array Attributes 316
Export Experiments 320
Export Filters 321
Export Tracks 322
Feature Level Filters 323
Find in column 325
Font Selection 327
Gene List 328
Generating Genotype Reference File 329
Genotype Reference Details 331
Genotype Reference Importer 332
Go To Gene/Genomic Location 334
Graphical Aberration Summary (report) 335
Graphical Aberration Summary (for common aberration analysis) 337
Graphical Common Aberration Setup 339
Graphical Common Aberration Summary 341
Graphical Differential Aberration Summary 344
Graphical Penetrance Summary 349
Graphical Interval Penetrance Summary 351
Import 354
Import (filters) 356
Import (experiments) 357
Import GEML design files 358
Import Genome Build 359
Import Track 360
Interval Filter 361
Manually Reassign Peaks 364
Matched Sample 365
Microarray Properties 367

MovAvg Example Parameters 371
MovAvg Example Plot 373
Properties (for common aberration result) 375
Properties (of an experiment result) 377
QC Metrics Table 378
QC Metrics – Frequency Distribution 381
QC Metrics Graph 382
Report Access Details 385
Report Management 386
Run Cytogenetic Report 388
Sample Attributes 390
Scroll to Column 391
Search Probes in eArray 392
Select Chromosome Intervals 393
Select Color 394
Select data type for experiments 397
Set genome build and species for Axon design files 398
Setting Clustering Method and Parameters 400
Show/Hide Columns 403
Show Intensity Bar Charts 404
SNP CN QC Metrics Table 405
SNP Genotype Report Setup 406
Stroke Selection 407
Text Penetrance Summary Setup 408
Track 409
UDF Import Summary 411
Universal Data Importer - Map Column Headers 412
User Preferences 414
View coordinates in UCSC browser 420
View Preferences 422

Reports 427
Text Aberration Summary Reports 427

SNP Reports 429

Configuration Settings 431
CGH Properties File 431

6 Statistical Algorithms 437

Overview of CGH Algorithms 439
Preprocessing algorithms 440
Aberration detection algorithms 441
Visualization algorithms 445
Algorithms for multi-array analysis 445
Algorithms for joint CGH-expression analysis 446
SNP analysis algorithms 447
Definitions Used in the Statistical Algorithms Sections 450

Preprocessing Algorithms 452
Centralization Algorithm 452
GC Correction Algorithm 455
Fuzzy Zero 458
How the data centering algorithms can affect aberration calls 461
Error Model and Combining Replicates 462

Aberration Detection Algorithms for CGH Analysis 465
Z-Scoring for Aberrant Regions 466
ADM-1 468
ADM-2 471
HMM 472
CBS 483
Noise Estimation – the Derivative Log Ratio Spread 491
Interpretation and Visualization of the Detection Algorithms 493

Visualization Algorithms 502
Triangular Smoothing 502

Algorithms for Multi-array Analysis 506
Penetrance 506

Common Aberration Analysis 508
Differential Aberration Analysis 512
Cluster Analysis 515
Copy Number Variation Region 518
Interpretation and visualization of the algorithms for the comparison of commonly aberrant regions 520

Algorithms for Joint CGH-Expression Analysis 526
Enrichment of extreme expression values in a single sample 527
Regional correlation scores with multiple-matched CGH and Expression samples 528
Enrichment of CGH/Expression correlation 530
Enrichment of external gene annotation 531

Algorithms for CGH+SNP Analysis 533
ASCN (SNP CN) – Allele-specific copy number detection algorithm 534
Assignment of SNP genotypes 541
LOH (Loss or lack of heterozygosity) algorithm 544

Appendix 555
Enrichment Analysis - the Hypergeometric distribution 555
Cluster Analysis - Distance-based methods 557
Cluster Analysis - Similarity-based methods 559
Cluster Analysis - Linkage methods 561
Multiple testing correction – False Discovery Rate in the Discrete Wavelet Transform 562
Segmentation- Discrete Haar Wavelet Transform and Gaussian mixtures for HMM 564

References 566

1 Getting Started

What is the CGH Application? 19
Using Agilent Genomic Workbench on a Mac 21
Entering a License and Starting the CGH Application 22
Using Main Window Components to Display/Analyze Data 25
Switching Applications 29
Using Tabs and Command Ribbons 30
Using the Navigator to Search for Data 33
Using the Genomic Viewer to Display Data and Results 36
General Instructions for Displaying Microarray Data/Results 40
General Instructions for Setting CGH Interactive Analysis Options 41
Getting Help 43

This chapter gives an overview of how to use Agilent Genomic Workbench Lite Edition to import, display, and analyze Comparative Genomic Hybridization (CGH) and Comparative Genomic Hybridization with Single Nucleotide Polymorphism (CGH+SNP) data.

Before or after you import extracted data into the program, you can assign identification information and attributes to the samples through the Sample Manager tab. See the *Sample Manager User Guide*.

To display or analyze imported CGH data, you organize the data files into logical units called *experiments*. Experiments are used to define the data you want to display or analyze using Agilent Genomic Workbench Lite Edition. You can then use the Preprocessing, Analysis, Discovery and Reports tabs of the program to interactively analyze the data in the experiment for aberrations.

Experiments can also serve as the basis for automated, unattended CGH or CGH+SNP analyses in the Workflow tab. The commands under this tab also let you set up image files for automated, unattended Feature Extraction before array analysis. See the *Workflow User Guide* for more information.

What is the CGH Application?

The FE lets you display and analyze aberration patterns from multiple Comparative Genomic Hybridization (CGH) or Comparative Genomic Hybridization with Single Nucleotide Polymorphism (CGH+SNP) microarray profiles, either interactively or by setting up an analysis method to use with a workflow.

With CGH interactive analysis, you set up an experiment and apply analysis algorithms to your data one experiment at a time. You also set parameters for report templates that are used for automated analysis using a Workflow.

You do most of your operations with the Preprocessing, Analysis, Discovery and Reports tabs. See Table 2 on page 30. With the interactive portion of the FE, you can:

- Import Agilent Feature Extraction data, Axon data or UDF files and use the Genomic Viewer to display this data along the chromosome.
- Preprocess the data
  - Apply feature and array filters. Combine designs and/or inter- or intra-array replicates.
  - Apply GC Content correction.
  - Apply a centralization calculation to the data.
  - Display QC metrics on the original data.
- Analyze the data
  - Calculate and show a moving average of log ratios.
  - Use robust statistical aberration detection algorithms to detect and map aberration regions with high confidence.
  - Perform SNP analysis, including SNP Copy Number (with or without manual reassignment of peaks) and LOH (Loss/Lack of Heterozygosity). Use Discovery options to display the analysis results.
  - Apply aberration filters to the results.
  - Merge adjacent CNVs into CNVRs (Copy Number Variation Regions).
  - Find common aberrations between several CGH samples, as well as do a differential aberration analysis.

  - Calculate a correlation analysis or enrichment analysis of gene expression and CGH data.
  - Do a cluster analysis and display data with a “heatmap”.
  - Generate a Genotype Reference file.
- Generate reports
  - Make text or graphical aberration reports
  - Make SNP genotype and aberration & LOH reports
  - Make penetrance summary reports
  - Make customizable Cytogenetic Reports for individual CGH and CGH+SNP samples.
- Change the way you view the data
  - Display chromosomal deletions and amplifications on different scales simultaneously.
  - Display allele-specific copy numbers and regions of LOH for CGH+SNP arrays.
  - Customize the way data is displayed in the genomic viewer.

With workflow analysis, you configure an analysis method ahead of time, and then run it in a workflow. This method is useful for unattended operation for consistent analyses of multiple data sets. To review workflow results, you use the Genomic Viewer. For more details and quick–start instructions to set up and run a CGH workflow, see the *Workflow User Guide*.

Using Agilent Genomic Workbench on a Mac

The content of this User Guide applies to both the Windows and Mac versions of Agilent Genomic Workbench. Both of these versions have the same features. However, when you use the Mac version of the program, please note the following:

Windows command	Equivalent Mac command
Right-click	• Command-click (⌘-click) • On Macs with trackpads, other options are available. On certain machines, you place two fingers on the trackpad while you press the button below the trackpad. See the user guide for your specific machine. • If you have a third-party mouse that has more than one button, you may be able to use one of the buttons as a right mouse button.
Control-click	Control-click (Same as the Windows command)
Shift-click	Shift-click (Same as the Windows command)
 (Close button)	 (Close button)

Entering a License and Starting the CGH Application

This section describes how to open the CGH application in Agilent Genomic Workbench, and enter your CGH license.

When you start Agilent Genomic Workbench for the first time, the program opens in the **Home** tab, with the **Open Application** tab displayed. From this tab, you can click any of the application areas, or click **Help** to open the User Guide for that application.

NOTE

If a message appears about performance when analyzing 1M feature arrays, read the message, mark the **Please do not show this warning again** box, and click **OK**. Instructions for how to increase heap memory are located in the *Agilent Genomic Workbench Lite Edition 6.5 Installation Guide*.

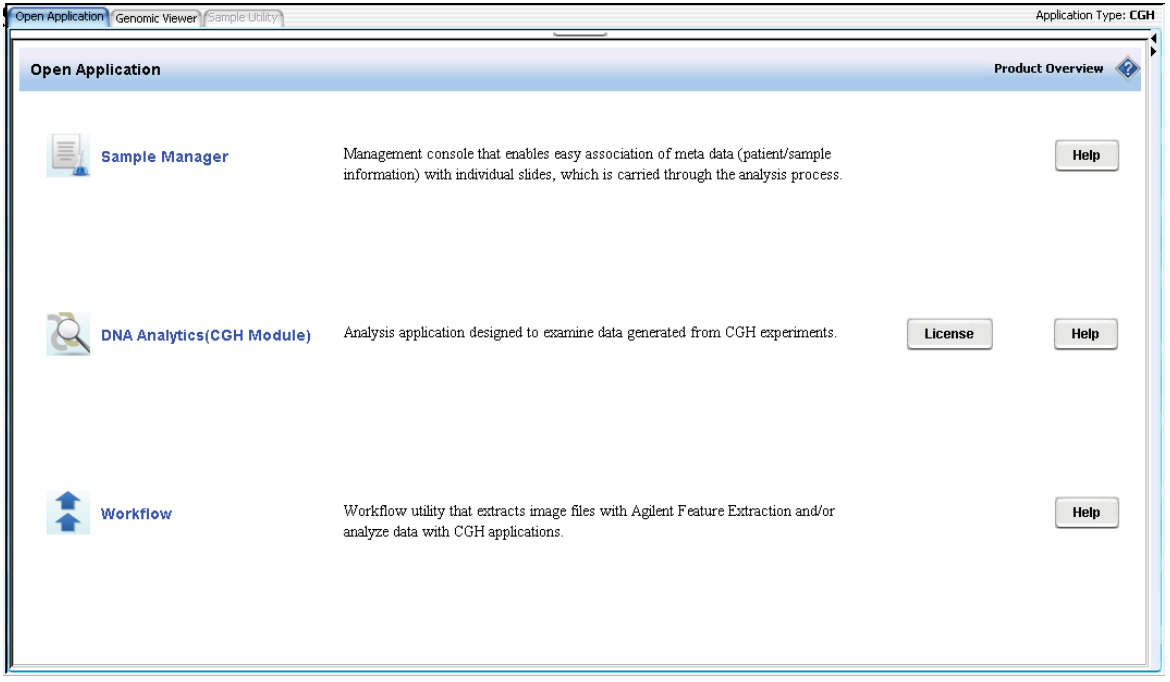


Figure 1 Open Application tab of Agilent Genomic Workbench Lite Edition for CGH

To enter a license for analyzing CGH data interactively

1 Click the **Open Application** tab if it is not already displayed.

2 Click **License** next to the description of DNA Analytics in the Open Application pane.

The License tab of the User Preferences dialog box appears.

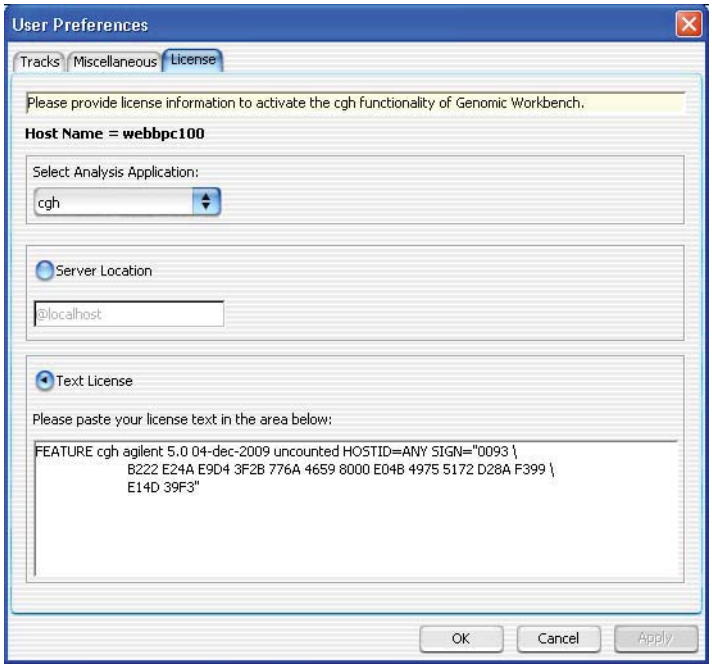


Figure 2 License screen for CGH module

There are two ways to provide the license information:

Use a Server Location

1 Unzip the license .txt file into a folder on your computer, to which the program has access.

2 Copy the path for that folder to the Clipboard.

3 In the User Preferences License tab, click **Server Location**.

4 Paste the license folder path into the box below Server Location. (To paste the license for both Windows and Mac computers, hold down the **ctrl** key and press **V**.)

5 Click **Apply**, or click **OK** to apply the license and close the dialog box.

Enter a Text License

1 Find the folder that contains the CGH module license .txt file.

2 Double-click the license name to open the file in Notepad, and copy the text in the Notepad window.

3 In the User Preferences License tab, click **Text License**.

4 Paste the license information into the License text box. (To paste the license for both Windows and Mac computers, hold down the **ctrl** key and press **V**.)

5 Click **Apply**.

6 If you have no other licenses, click **OK**.

OR

If you have another license, click the arrow from the Select Analysis Application list, select the DNA Analytics module and repeat steps 1-5.

To start the CGH module

- In the Open Application tab, click the **DNA Analytics (CGH Module)** icon  .

  The CGH module starts and the Genomic Viewer is displayed.

Using Main Window Components to Display/Analyze Data

You can use the data *viewing* capability in Agilent Genomic Workbench with or without a license to view data for many types of arrays, including CGH, CGH+SNP, ChIP, and Methylation (CH3). You can use the data *analysis* capability in Agilent Genomic Workbench only if you have a license for one or more of the DNA Analytics programs (CGH, ChIP, or Methylation).

What are the main window components?

You use four primary components of the Agilent Genomic Workbench main window to import, manage, export, display and analyze extracted data.

- Home tab commands – import, manage and export data
- Navigator – create and fill new experiments with array data

  When you make the experiment active, the data appear in the display, called Genomic Viewer.
- Genomic Viewer – display data and content in four Views: Genomic View, Chromosome View, Gene View, and Tab View
- View tab commands – change appearance of Genomic Viewer display

Figure 3 shows the main window of Agilent Genomic Workbench Lite Edition when the Genomic Viewer tab is selected, and identifies the names of its components.

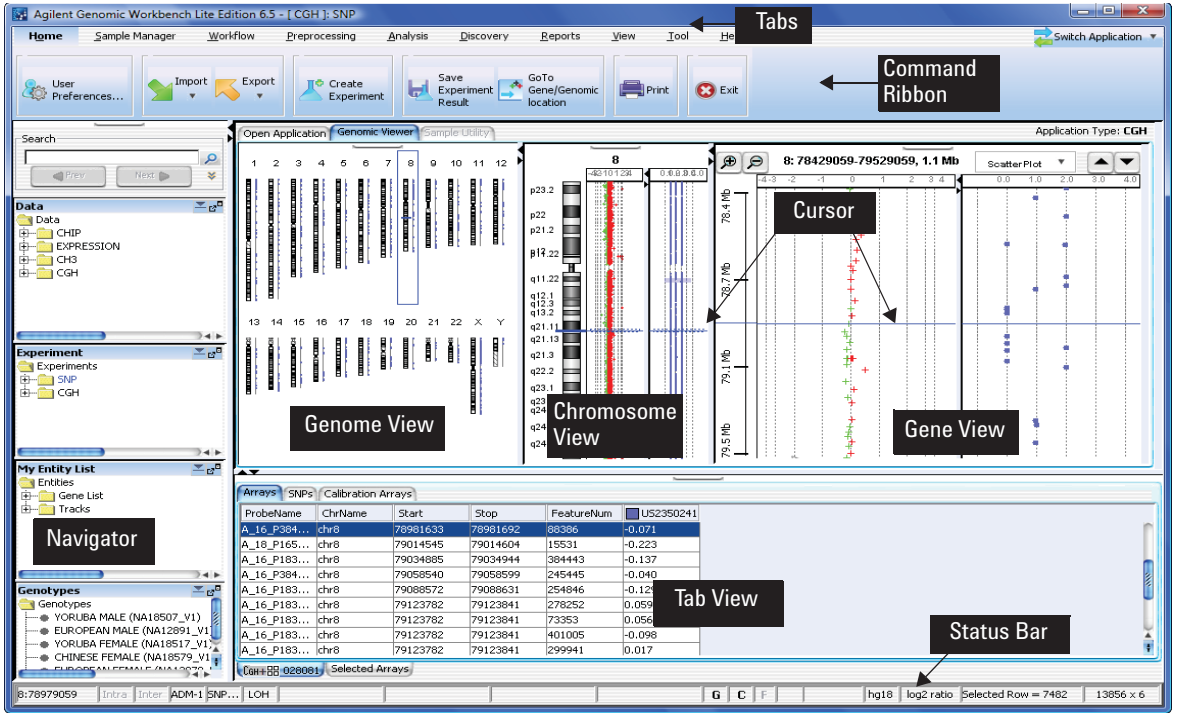


Figure 3 Agilent Genomic Workbench Lite Edition main window, showing major components for CGH

What can you do with the main components for display of data and results?

See the table below for the parts of the main window you use to display log ratio data and results.

Table 1 Components of Agilent Genomic Workbench Lite Edition main window for display of data and results

To do this	Use this part of the main window
Change program among CGH, ChIP, or Methylation (CH3)	**Switch Application button**: Click the button and click the program you want to open. The scatter plot options are different for the different program types
Show properties, rename, or delete designs and data in the program	**Data pane of the Navigator:** Double-click to expand data node, then right-click to select option.
Import or export data	**Home tab**: Click the **Import** or **Export** button to select the data you want to import or export. See Chapter 2, "Importing, Managing, and Exporting CGH and CGH+SNP Data and Other Content" for more information.
Select array data to display in the three graphical views or in the Tab View as a table	**Experiment pane of the Navigator**: Create an experiment with the imported data, select the experiment, and then select the data within the experiment to display or analyze. See Chapter 3, "Displaying CGH and CGH+SNP Data and Other Content" for more information.
Display array data/results for only a certain portion of a chromosome	**Genome View**: Select a chromosome to display in Chromosome View. You cannot view log ratio data points here. **Chromosome View**: Select a gene region to display in Gene View. You can display log ratio data points here if you select **Scatter Plot** in the View Preferences dialog box. **Gene View**: See the log ratio data next to a selected region of a chromosome, with associated genes and track-based annotation. See Chapter 5, "CGH Interactive Analysis Reference" for details about these Views.

Table 1 Components of Agilent Genomic Workbench Lite Edition main window for display of data and results (continued)

To do this	Use this part of the main window
Show/Hide or customize the data points for the scatter plots	**Gene View**: Move the mouse pointer over **Scatter Plot** to display the options. Or, right-click and then click **View Preferences**. **Chromosome View**: Right-click and then click **View Preferences**. **View tab**: Click **View Preferences**. See Chapter 3, "Displaying CGH and CGH+SNP Data and Other Content" for information on how to do this.
Display array data next to tracks or gene lists	**My Entity List pane of Navigator**: Add or select a track or gene list to have it appear in Gene View. See Chapter 3, "Displaying CGH and CGH+SNP Data and Other Content" for information on how to do this.
Change the appearance of the display	**View Tab**: Click **View Preferences**. From the View Preferences dialog box, you can change the orientation, select what type of data to view, and configure scatter plot options. **Genomic Viewer**: Right-click any View except the Tab View and select **View Preferences**. In the View Preferences dialog box, you can select to show or hide the scatter plots and how to display them, including results. See Chapter 3, "Displaying CGH and CGH+SNP Data and Other Content" for more information.
Analyze or reanalyze displayed data	**Preprocessing Tab**: Click this tab to display commands you use to manipulate the data before you apply the algorithms. **Analysis Tab:** Click this tab to display commands you use to analyze the data. **Discovery Tab**: Click this tab to display commands you use to display, filter, compare, and further analyze the results. **Reports Tab:** Click this tab top display commands you use to generate and manage reports. For more information on what you can do in these tabs, see "Tabs" on page 30.

Switching Applications

You can use the Agilent Genomic Workbench to work with a variety of different data types. Because the requirements for the display of data (and calculation of results, if using a license) are different for different data types, you must switch the application module for the type of data you want to display.

The Switch Applications menu, located at the upper right corner of the Agilent Genomic Workbench window, is used to change the DNA Analytics module. The selected module is marked . The selected module is also displayed in the title bar of the Genomic Workbench main window.

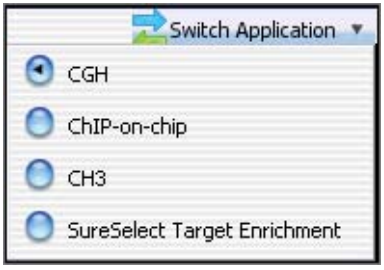


Figure 4 Switch Application menu

Using Tabs and Command Ribbons

Tabs

When you click a *tab*, groups of commands or single commands appear that are specific for that tab. The tabs that are displayed change depending on what licenses you have, and what DNA Analytics module is selected (CGH, ChIP, CH3).



Figure 5 Tabs for CGH interactive analysis

The following table summarizes what you can do from the CGH interactive analysis tabs of Agilent Genomic Workbench Lite Edition.

Table 2 Capabilities in CGH Interactive Tabs

Tabs	CGH Capabilities
Preprocessing	Edit and turn on or off data filters
	Turn on and configure GC Correction (recommended for SNP Copy Number and LOH calculation) and display plot of corrected data
	Turn on and configure centralization (recommended for SNP Copy Number and LOH calculation)
	Combine array designs and replicates
	Display QC metrics
Analysis	Calculate a moving average on log ratio data
	Select an aberration algorithm and set up its parameters (required for SNP Copy Number and LOH calculation)
	Turn on and configure SNP Copy Number (required for LOH calculation)
	Turn on LOH calculation (requires SNP Copy Number)

Table 2 Capabilities in CGH Interactive Tabs (continued)

Tabs	CGH Capabilities
Discovery	Add aberration filters
	Compare arrays with common aberrations
	Make graphical penetrance diagrams
	Set up to display CNVR (copy number variant regions)
	Compare CGH data with expression data
	Compare arrays with different aberrations
	Do a cluster analysis
	Do a heatmap analysis
	Generate a genotype reference file
Reports	Aberration report
	SNP Genotype report
	SNP Aberration & LOH report
	Penetrance report
	Cyto report
View	Set up preferences for display of data
	Copy displayed data to the Clipboard
	Turn on or off display of Views and Navigator
	Turn on or off tabular display of signal intensity and annotations
	Turn on or off display of Cytoband information in Gene View
	Turn on or off highlight of nonunique probes
	Turn on or off display of custom data
Tool	Set up and run plug-in programs

Commands

The area where commands appear is called a *command ribbon*. The command ribbon that appears when you click the Home tab for CGH is shown below. The commands that appear in the command ribbon change depending on what application module is selected, and which tab in that application module is selected.



Figure 6 Home command ribbon and tabs for CGH interactive analysis

For a complete description of all of the command ribbons and commands you see in Agilent Genomic Workbench, see "Command Ribbons" on page 176.

Using the Navigator to Search for Data

This section gives you instructions on how to search for design files, extracted FE data, experiments and other information in the Navigator of Agilent Genomic Workbench Lite Edition. The Navigator contains different panes when you select the Sample Manager or Workflow tabs. See the User Guides for those modules for information on the Navigator contents.

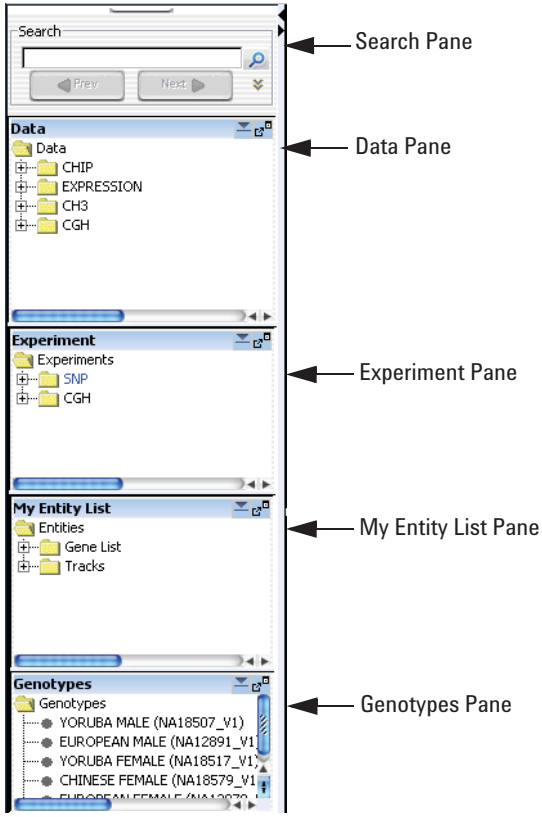


Figure 7 Navigator panes for CGH

The Navigator shows the array data, experiments, and other content stored in Agilent Genomic Workbench that is available in the program. It contains the following panes:

Pane	Comments
Search	Lets you search within any pane of the Navigator for a specific item (array or build, for example). You must type the entire array name or term; otherwise, use asterisks (*) as wildcards for unspecified strings. For example, type *1234* to find any item that contains "1234".
Data	Contains microarray data files, organized by application type, then by design and genome build. Shows all probe groups and microarray designs that are available to you, organized by folders. For the SureSelect Target Enrichment application type, the program shows all bait groups and libraries. In general, you can: • Expand or collapse folders to show or hide content. • Look at the icon that appears with an item to monitor its status. • Right-click the name of a folder or item to open a shortcut menu that lets you take action on the item. See "Data pane – icons, special text, and buttons" on page 206 and "Data pane – actions and shortcut menus" on page 207.
Experiment	Contains Agilent Genomic Workbench experiments. Experiments are organizational units that contain links to microarray data and design files. In data analysis modules, experiments also contain saved results. See "Experiment pane – icons, special text, and buttons" on page 209 and "Experiment pane – actions and shortcut menus" on page 210.
My Entity List	Contains gene lists and tracks: • **Gene Lists** are collections of genes of interest. You can create them within the program, import and export them, and apply them to Gene View and Chromosome View. • **Tracks** are collections of annotation or other information that map to specific genomic locations. You can import, export, and combine tracks, and display them in Gene View with your array data and analysis results. See "My Entity List pane – Icons, buttons, and special text" on page 220 and "My Entity List pane – actions and shortcut menus" on page 220.
Genotypes	Shows SNP genotype reference samples in the database. You can import, display details, rename, or delete genotype references from this pane. See "To import a genotype reference file" on page 58.

To search the Navigator

You can search one or all of the panes of the Navigator for items that match a specific search term. Figure 8 shows the search pane of the Navigator, and identifies a couple of its elements.

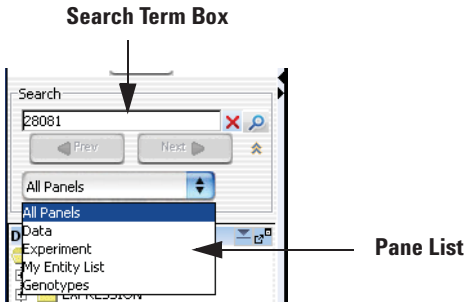


Figure 8 Search pane of the Navigator

1. At the top of the Navigator, in the Pane list, select the pane to be searched. To search in all panes, select **All Panels.** If the pane list is not visible, click to show it.
2. In the search term box, type the desired search term. The search term is not case sensitive, but it must contain the complete entry that you want to find. You can use asterisks (*) to represent one or more unspecified characters. For example, type *12345* to find any item that contains "12345".
3. Click .

   The program searches the selected pane(s) for items that match your search term. If it finds matching items, the program expands the appropriate folders, and displays the names of the matching items in red. The first matching item is highlighted in yellow.
4. Do any of the following:
   - To highlight the next matching item, if one is available, click Next.
   - To highlight the previous matching item, click Prev.
5. After you complete the search, click to clear the results of the search, as well as your search term.

Using the Genomic Viewer to Display Data and Results

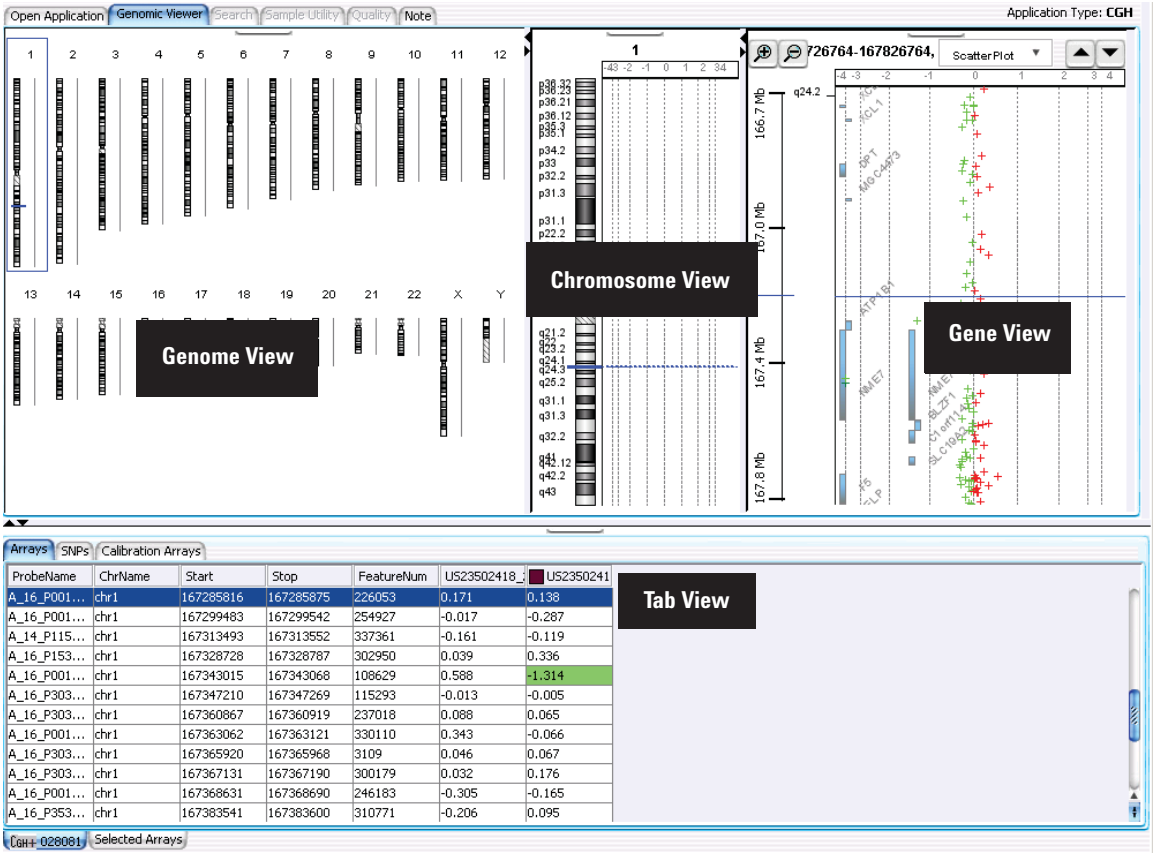
What is the Genomic Viewer?

Genomic Viewer is the graphics and tabular display section of the Agilent Genomic Workbench main window. In the Genomic Viewer, extracted data and analysis results can be tabulated and displayed next to depictions of the genome, selected chromosome, and selected genes of the species whose array data you are analyzing.

There are four main views in the Genomic Viewer, as shown in Figure 9.

- **Genome View** – A graphical representation of the entire genome for the selected species. Use this view to select the chromosome to show in the other views.
- **Chromosome View** – A graphical representation of the selected chromosome, displayed with cytobands and a plot area. Click or drag the mouse to select a region to display in the Gene View.
- **Gene View** – A more detailed view of the chromosomal region selected in the Chromosome View.
- **Tab View** – Displays CGH design annotation and log ratio data, and CGH+SNP genotype data related to the chromosome you select in Chromosome View.

For more information on the Genomic Viewer and its views, see Chapter 5, “CGH Interactive Analysis Reference”.



ProbeName	ChrName	Start	Stop	FeatureNum	US23502418_...	US2350241
A_16_P001...	chr1	167285816	167285875	226053	0.171	0.138
A_16_P001...	chr1	167299483	167299542	254927	-0.017	-0.287
A_14_P115...	chr1	167313493	167313552	337361	-0.161	-0.119
A_16_P153...	chr1	167328728	167328787	302950	0.039	0.336
A_16_P001...	chr1	167343015	167343068	108629	0.588	-1.314
A_16_P303...	chr1	167347210	167347269	115293	-0.013	-0.005
A_16_P303...	chr1	167360867	167360919	237018	0.088	0.065
A_16_P001...	chr1	167363062	167363121	330110	0.343	-0.066
A_16_P303...	chr1	167365920	167365968	3109	0.046	0.067
A_16_P303...	chr1	167367131	167367190	300179	0.032	0.176
A_16_P001...	chr1	167368631	167368690	246183	-0.305	-0.165
A_16_P353...	chr1	167383541	167383600	310771	-0.206	0.095

Figure 9 Genomic Viewer in vertical orientation

To change the size of and detach panes from the Agilent Genomic Workbench main window

- To change the size of a pane in the main window, drag one of its inside borders.
- To detach a pane from the main window and open it in a separate window, click its **Detach** button .

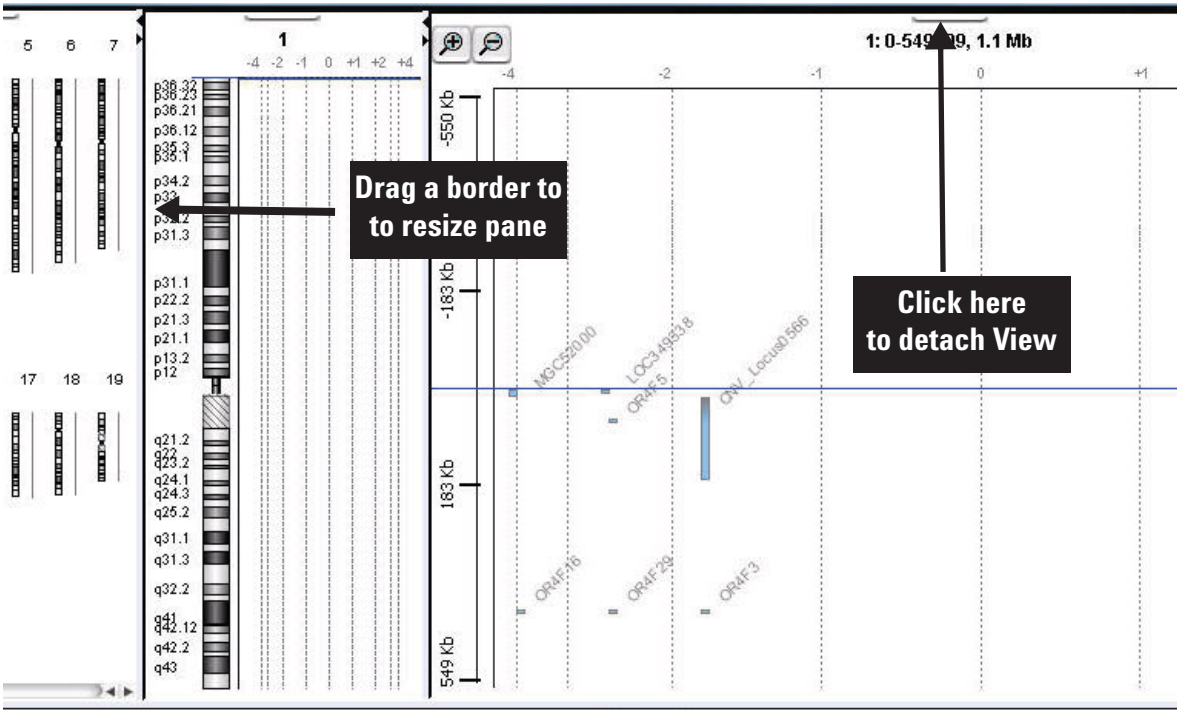


Figure 10 Changing the size of and detaching panes

To maximize and reattach panes to the Agilent Genomic Workbench main window

- To display a view full-screen in a separate window, click its **Maximize** button.
- To reattach a view in a separate window to the main window, click its **Close** button.

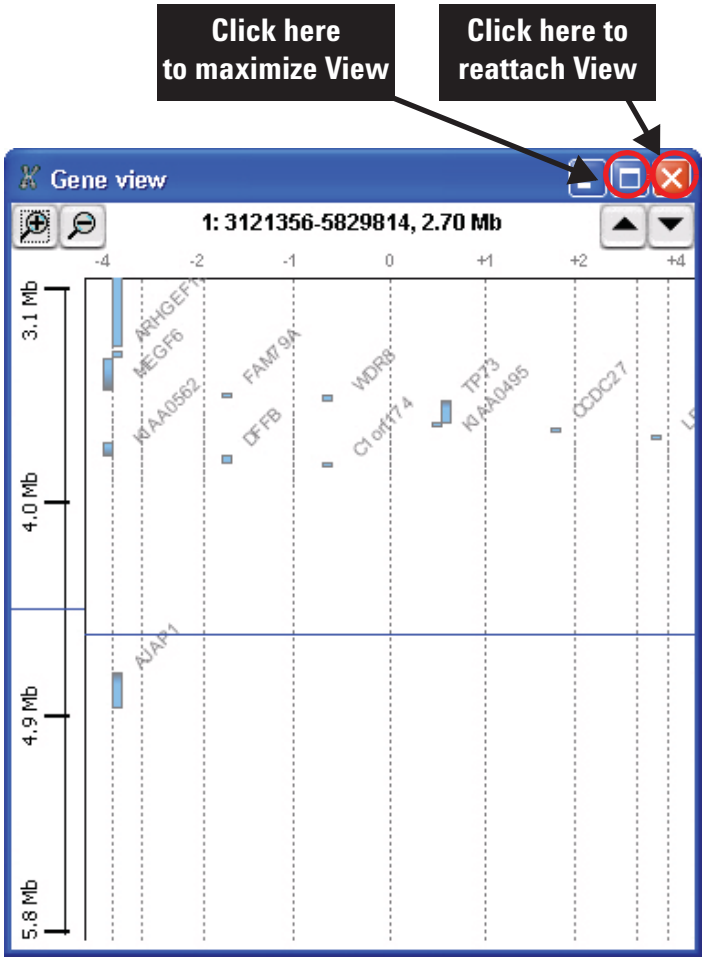


Figure 11 Maximizing and reattaching panes

General Instructions for Displaying Microarray Data/Results

An *experiment* is the folder that holds data from any array set you select for the experiment. The folder also holds analysis results.

You set up experiments to display all data and results in the Genomic Viewer. To set up an experiment you:

- Import data
- Create a new experiment
- Add the imported data to the experiment
- Select the experiment to display data

For step-by-step instructions on how to display data, see the *Data Viewing User Guide*.

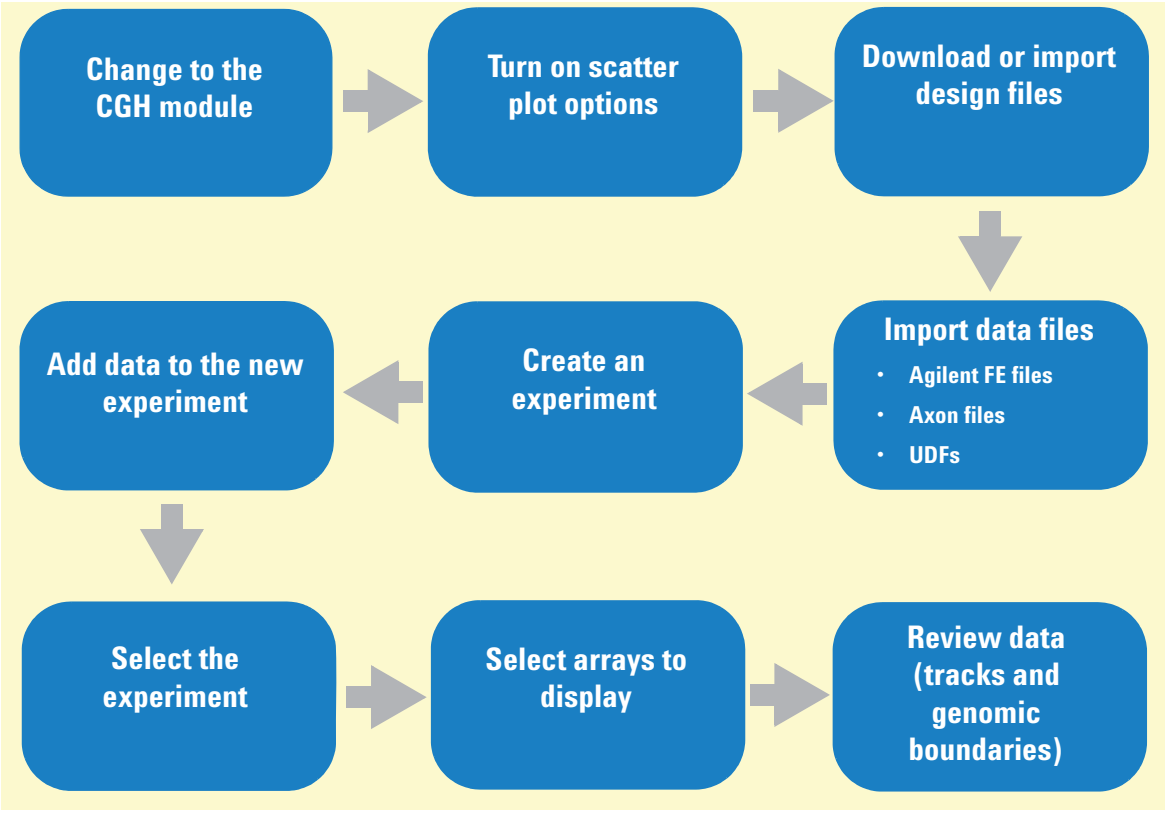


Figure 12 Typical pathway for displaying microarray data/results

General Instructions for Setting CGH Interactive Analysis Options

After you import data and set up experiments, you can set up preprocessing and analysis calculations. This can be done before or after you select the experiment. You can create and apply filters, combine designs and replicate probes, apply various algorithms to show aberration calls and create Cyto Report templates. Once an experiment is selected, the program recalculates the results immediately after you change each setting.

After analysis, you can apply aberration filters and select Discovery options.

For more information on how to change analysis settings interactively, see Chapter 4, “Setting Up CGH Interactive Analysis”.

For information on how to analyze CGH data as part of a Workflow, see the *Workflow User Guide*.

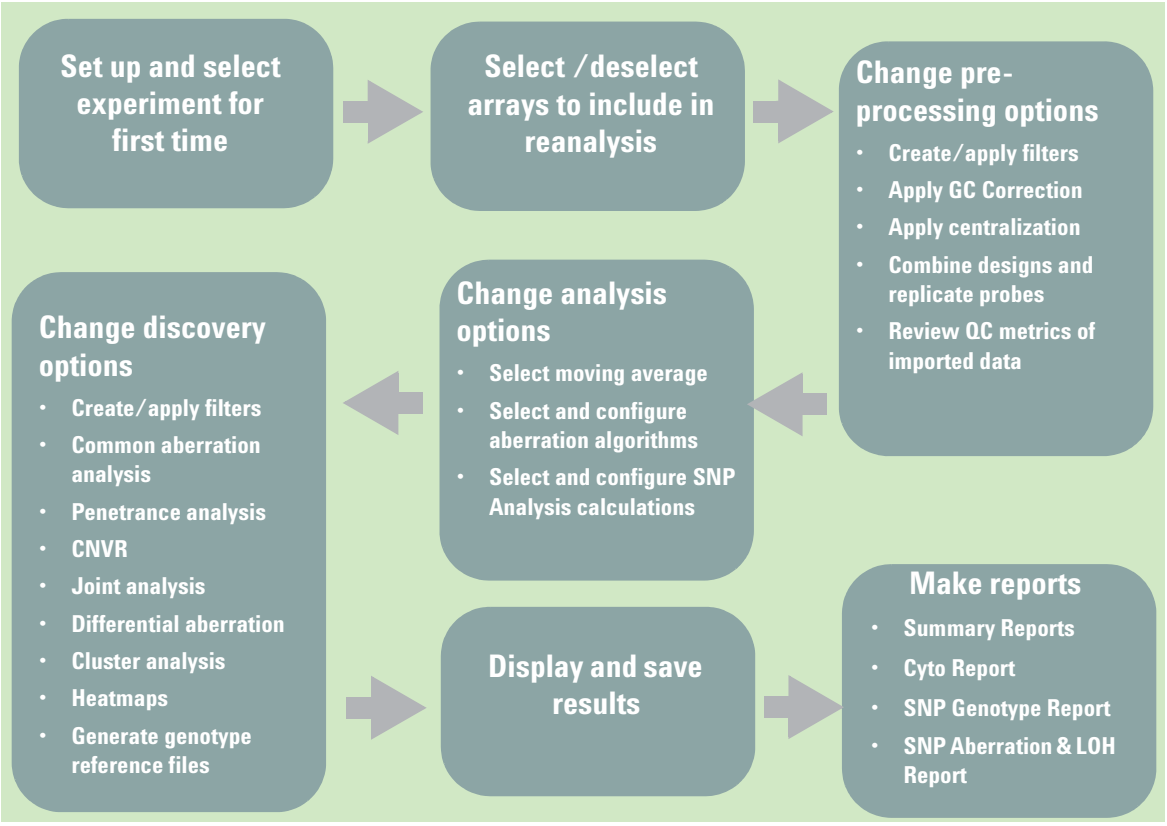


Figure 13 Typical CGH interactive analysis pathway

Getting Help

To get help within Agilent Genomic Workbench

Agilent Genomic Workbench has several help resources. Help guides open with Adobe® Reader®.

Help Resource	Description/Instructions
CGH Interactive Analysis User Guide	This user guide, which you are now reading, supplies comprehensive help on all available CGH tasks. You can access it easily from anywhere within the program. **1** In any tab of Agilent Genomic Workbench, click the **Help** tab. **2** On the Help Ribbon, click **Application Guide.** The CGH Interactive Analysis User Guide opens.
Other User Guides	The Help tab in Agilent Genomic Workbench lets you view any of the available user guides that apply to the currently selected application type. **1** Set the desired application module from the Switch Application menu. **2** In the Agilent Genomic Workbench tab bar, click **Help.** The names of the available user guides appear in the command ribbon. **3** Click the desired help guide. The selected guide opens with.
Product Overview Guide	An additional guide gives an overview of the capabilities within Agilent Genomic Workbench and describes how to start and find help for all of the programs. In addition, it helps you with system administration and troubleshooting. **1** In any interactive analysis tab of Agilent Genomic Workbench, click the **Open Application** tab. **2** At the upper right corner of the Open Application tab, click **Product Overview**.

To contact Agilent Technical Support

Technical support is available by phone and/or e-mail. A variety of useful information is also available on the Agilent Technical Support Web site.

Resource	To find technical support contact information
Agilent Technical Support Web site	**1** Go to http://chem.agilent.com. **2** Select a country or area. **3** Under Quick Links, select **Technical Support**. **4** Select from the available links to display support information.
Contact Agilent Technical Support by telephone or e-mail (United States and Canada)	Telephone: (800-227-9770) E-mail: informatics_support@agilent.com
Contact Agilent Technical Support by telephone or e-mail (for your country)	**1** Go to http://chem.agilent.com. **2** Select **Contact Us**. **3** Under Worldwide Sales and Support Phone Assistance, click to select a country, and then click **Go**. Complete e-mail and telephone contact information for your country is displayed.

To learn about Agilent products and services

To view information about the Life Sciences and Chemical Analysis products and services that are available from Agilent, go to www.chem.agilent.com.

2 Importing, Managing, and Exporting CGH and CGH+SNP Data and Other Content

Importing Files 46
Working with Experiments to Organize Imported Data 59
Managing Content 70
Exporting and Saving Content 82

This chapter describes how to import, organize, manage, and export CGH and CGH+SNP data and other content within the user interface of Agilent Genomic Workbench.

Importing Files

You can use the Home tab to import many kinds of files into Agilent Genomic Workbench. The table below summarizes the kinds of files you can import, and the topics in this section that describe how to import them.

The Data pane of the Navigator displays all of the content available in the program. See "Navigator" on page 202 for more information on the Navigator panes and how to use them.

Type of file	Comments	See these topics
Microarray data files	• Agilent Feature Extraction (*.txt) data files • Axon (*.gpr) data files • Universal Data Files (UDFs) (*.txt files)	"To import Agilent FE or Axon data files" on page 51 "To import a UDF file" on page 52
Microarray design files	• Agilent GEML (*.xml) design files • Axon (*.gal) design files	"To import Agilent GEML design files" on page 48 "To import Axon design files" on page 49
Genome builds	Agilent-supplied genome information for human, mouse and rat genomes or user-supplied genome information for other species	"To import a genome build" on page 50
Tracks	BED format annotation track files	"To import tracks" on page 55
Array attributes	.txt files that you have created yourself or previously exported from Agilent Genomic Workbench	"To import array attributes" on page 56
Experiments	ZIP format file of exported experiments	"To import an experiment file" on page 57

Type of file	Comments	See these topics
Filters	.xml file that contains filters created using Agilent Genomic Workbench	"To import filters" on page 58
Genotype Reference	Text or .xls file that contains reference genotype and expected number of cuts for each SNP probe in the sample.	"To import a genotype reference file" on page 58 "To generate a genotype reference file" on page 164

To select a different location for data files

By default, the program stores microarray and experimental data files in **C:\Program Files\Agilent\Agilent Genomic Workbench Lite Edition <version>\data**. If you want, you can select a different location.

CAUTION Do not select a location that contains a backup data folder; the data in the folder you select will be overwritten.

1 In the Home tab, click **User Preferences.**

The User Preferences dialog box appears. See "User Preferences" on page 414.

2 In the Miscellaneous tab, under Data Location, click **Browse.**

An Open dialog box appears.

3 Select a location, then click **Open.**

NOTE Make sure you have full permissions in the data location.

The selected location appears in the User Preferences dialog box, in Data Location.

4 Click **OK.**

To import Agilent GEML design files

The Agilent Genomic Workbench database must contain designs that match the Agilent Feature Extraction data files you want to import. The design file must be present before any extraction data files can be imported. Your imported GEML files contain array-specific information such as probe names, annotations, and chromosomal locations, and are associated with a specific genome build.

To import an Agilent GEML file, use the following procedure:

1 In the Home tab, click **Import > Design Files > GEML File.**

The Import Design Files dialog box appears. See "Import" on page 354. The dialog box shows only *.xml files.

2 To select a file for import, click its name. To select additional files, hold down the **ctrl** key while you click their names.

3 Click **Open.**

The program validates the selected file(s), and the Import GEML Design Files dialog box appears. See "Import GEML design files" on page 358.

- If a design file passes validation, the Status column shows **Valid** in green.
- If the design is an Agilent Catalog design, and is not yet downloaded from the eArray Web site, the Status shows **Not Allowed** in red. You must download the file from the eArray Web site.
- If a design file already exists in the database, the Status shows **Overwrite** in yellow. If you continue, the existing design will be replaced with the imported design.
- If a design file fails validation, **Corrupt** appears in the Status column beside it, and the program will not import the file. To remove the corrupt design from the list, click its **Remove** button  .

4 Click **Start Import.**

The program imports the file(s). The files appear as designs under the appropriate data type in the Data pane of the Navigator, with the genome build as a node within the folder.

You can import two design files with the same name that are associated with different genome builds. If you do, the program creates a single design folder with two nodes, one for each genome build.

To import Axon design files

You can import Axon (*.gal) microarray design files into Agilent Genomic Workbench. The program requires the Axon design files that match all Axon array data files you import.

1 In the Home tab, click **Import > Design Files > Axon File.**

The Import Axon Design Files dialog box appears. See "Import" on page 354. The dialog box shows only *.gal files.

2 To select a file to import, click its name. To select additional files, hold down the **ctrl** key while you click their names.

3 Click **Import.**

The program validates the selected file(s), and the Set genome build and species for Axon design files dialog box appears. See "Set genome build and species for Axon design files" on page 398.

- If a design file passes validation, the Status column will show **Update** in green.
- If a design file fails validation, **Corrupt** appears in the Status column beside it, and the program will not import the file. To remove the corrupt design from the list, click its **Remove** button  .

4 For each design file, select the appropriate Species and Genome Build.

5 Click **Start Import.**

The program imports the file(s). The files appear as new design folders in the Data pane, organized by application (CGH, ChIP, or methylation, for example).

To use eArray to update design files

Agilent regularly makes updates to probe annotations on its eArray Web portal. If you have imported Agilent array designs into Agilent Genomic Workbench, and you are a registered eArray user, you can download the updated design files from within Agilent Genomic Workbench. For more information about eArray, go to https://earray.chem.agilent.com and click **Help**.

1 In the Home tab, click **User Preferences**.

The User Preferences dialog box appears.

2 In the Miscellaneous tab, under eArray User Details, type your eArray Username and Password. For more information on User Preferences, see "User Preferences" on page 414.

3 Click OK.

To import a genome build

In general, the program uses the genome build specified in the array design file, and protects it from changes. If a genome build is not available in the program, you can import one.

NOTE Use arrays from a single genome build in an experiment.

1 In the Home tab, click **Import > Genome Build.**

The Import Genome Build dialog box appears. See "Import Genome Build" on page 359.

2 Set the following. All are required.

Setting	Instructions
Species	• Type the genome's species of origin, as you would like it to appear within the program.
Build Name	• Type the name of the genome build you want to import, as you would like it to appear within the program.
Refseq File	This file contains information on gene locations for Gene View. **a** Click **Browse.** A dialog box appears. **b** Select the file, then click **Open.**
Cyto-band File	This file contains the graphic information on the cytobands for Genome and Chromosome Views. **a** Click **Browse.** A dialog box appears. **b** Select the file, then click **Open.**

3 Click **OK.**

To import Agilent FE or Axon data files

You can import several types of microarray data files:

- Agilent Feature Extraction (FE) *.txt data files
- Axon (*.gpr) data files
- Universal Data Files (UDFs) (*.txt files) See "To import a UDF file" on page 52 for instructions on how to import this file type.

To import Agilent Feature Extraction files, the representative GEML array design files must imported first. In order to import Axon data files, the representative Axon.gal design files must be imported first. See "To import Agilent GEML design files" on page 48 or "To import Axon design files" on page 49.

1 In the Home tab, do one of the following:
   - To import Agilent FE data files, click **Import > Array Files > FE File.**
   - To import Axon data files, click **Import > Array Files > Axon File.**

   A dialog box appears. Only data files of the appropriate type appear. See "Import" on page 354.

2 To select a file for import, click its name. To select additional files, hold down the **ctrl** key while you click their names.

3 Do one of the following:
   - For Agilent FE files, click **Open.**
   - For Axon files, click **Import.**

   The Agilent Feature Extraction Importer dialog box appears. See "Agilent Feature Extraction Importer" on page 245.

4 Set the following:

Setting	Comments
Name	The names of imported arrays are often cryptic. You can give any array a more meaningful label. **a** Double-click the name of the array. **b** Type the name. **c** Press **Enter**.

Setting	Comments
Dye Flip	For each array: • Select **Normal** if: ◦ The test samples were labeled with cyanine-5 (red). ◦ The control samples were labeled with cyanine-3 (green). ◦ The imported ratio (test/control) should be reported directly. • Select **Flipped** if: ◦ The test samples were labeled with cyanine-3 (green). ◦ The control samples were labeled with cyanine-5 (red). ◦ The imported ratio (control/test) should be reported with the ratio inverted (test/control). The program does not combine dye-flip pairs.
Overwrite arrays with duplicate names	If you mark this option, the program deletes an existing array data file if it has the same name as one you import.

5 Do one of the following:

- To import the file(s) while you wait, click **OK.**
- To import the file(s) in the background, click **Run in Background.** This lets you work while the program imports the files.

To import a UDF file

UDF files are plain text files that contain array data in tab-delimited format. Files must contain the following six columns of information, in any order. Each column must contain the following column names, as column headers, or you must "map" the names from the file to these columns in Agilent Genomic Workbench:

- Probe name
- Chromosome name
- Start position
- Stop position
- Description
- Signal intensity or (log) ratio data (The file can contain additional columns, each with data from an additional array.)

When you import a UDF file, the program creates a new design based on the information you enter during import, and the information in the file itself. The program also creates a new experiment that contains the arrays.

1 In the Home tab, click **Import > Array Files > UDF File.**

The UDF Files dialog box appears. See "Import" on page 354. Only *.txt files appear in the dialog box.

2 Select the UDF file, then click **Open.**

The Select data type for experiments dialog box appears. "Select data type for experiments" on page 397.

3 For each array, set the following:

Setting	Comments
Experiment Name	By default, the program creates an experiment with the same name as the imported file. To change the name: **a** Double-click the name. **b** Edit the name. **c** Press **Enter.**
Data type	• Select the mathematical form of the signal intensity data for the array. The options are **ratio, $\log_2$ ratio, $\log_{10}$ ratio,** and **ln ratio.**
Design type	• Select **cgh, expression,** or **CH3.**

4 Click **Continue.**

NOTE When you "map" a column, you assign the column heading (in an external file) to a column heading in Agilent Genomic Workbench.

The Universal Data Importer – Map column headers dialog box appears. The main table in the dialog box contains the first few rows of data from the file. Column headings derived from the first line of the file appear at the top of the table as a guide, but the program does not interpret these headings. See "Universal Data Importer - Map Column Headers" on page 412.

5 Below each column heading, select the label that identifies the content of the column. Use each label exactly once, except for LogRatio, which you can use many times. Or, in Select Mapping, select a saved column map.

These options are available:

Column Label	This column contains:
ProbeName	Names of probes.
ChrName	Names of chromosomes.
Start	First chromosomal location to which each probe is designed.
Stop	Last chromosomal location to which each probe is designed.
Description	Text annotation related to the probe.
LogRatio	Array data values that correspond to each probe. You can use this label more than once.

6 Under Species Info, select the species and Genome Build appropriate to the data in the file.

7 If you expect to import many similar UDFs in the future, follow these steps to save the column map:

a Under **Mapping Info,** click **Save Mapping As.**
An Input dialog box appears.

b Type a name for the column map, then click **OK.**
The name of the saved map appears in Select Mapping.

In the future, you can select this mapping and apply it to any UDF file that you import.

8 By default, the program creates a "Virtual Array ID" that becomes the ArrayID attribute for the array(s) in the UDF. To create your own virtual Array ID, follow these steps:

a Under ArrayID Info, clear **Use System Generated Array ID.**

b Double-click the number in **Virtual Array ID,** then type your own Array ID.

For more information on Array IDs, see the *Sample Manager User Guide*.

9 Click **Import**.

The program validates your column mapping. A dialog box appears. If you need to fix the column map, the dialog box has a list of the missing column label(s). If the column map is complete, a message asks if you want to import additional files with the same mapping.

10 Do one of the following:

- If you want to import additional files with the same column mapping, follow these steps to include these files in the import:
  - **a** Click **Yes.**
    The UDF Files dialog box appears.
  - **b** Click the name of a file to select it for import. Hold down the **ctrl** key while you click the names of additional files.
  - **c** Click **Open.**
- If you do not want to include additional file(s) in the import, click **No.**

The Program imports all requested files, and the UDF Import Summary dialog box appears. This dialog box shows the imported files, the number of lines of data that were imported for each file, and the number of lines that were skipped, if any. If a file name appears in red, the program may not have imported the file. See "UDF Import Summary" on page 411.

11 Click **OK.**

In the Data pane, a new design folder appears in the appropriate design type folder. The design folder contains the imported array data.

A new experiment appears in the Experiments folder in the Experiment pane, that contains the array data. This experiment has the name of the imported UDF file, unless you changed it during import.

To import tracks

You can import BED format track files into Agilent Genomic Workbench. Track files contain specific features correlated with chromosomal locations, and apply to a specific genome build of a given species.

1 In the Home tab, click **Import > Track.**

The Import Track dialog box appears. See "Import Track" on page 360.

2 Set the following. All are required.

Setting	Instructions
Species	• Select the species to which the track applies.
Build Name	• Select the specific genome build of the species to which the track applies.
Track Name	• Type a name for the track. This name identifies the track within the program, including the name that appears if you include the track in Gene View.
Track File	**a** Click **Browse.** A dialog box appears. **b** Select the name of the track (*.bed) file to import. **c** Click **Open.** The location of the file appears in Track File.

3 Click **OK.**

The program imports the track. To display the track in Gene View, see "To show tracks in Gene View" on page 111. To manage tracks, see "Managing Content" on page 70.

To import array attributes

An array attributes file is a tab-delimited *.txt file that contains a list of arrays by Array ID, and values for array attributes. Attributes are pieces of array-specific information, such as the hybridization temperature or the name of an array set that contains the array.

Although you can import array attributes with this function, the Sample Manager module lets you import and assign array attributes more easily. See the *Sample Manager User Guide* for more information. From the Home tab, click **Import** and then select **ArrayAttributes**.

The Import Attribute Files dialog box appears. See "Import" on page 354.

4 Select the microarray attributes file, then click **Import.**

The program imports the file. If the ArrayIDs in the file do not match the ArrayIDs of arrays in the program, a dialog box appears. The dialog box has a list of the ArrayIDs in the file that do not match existing ArrayIDs. Click **No** to stop the import process, or click **Yes** to continue anyway.

To import an experiment file

In Agilent Genomic Workbench, an experiment is a set of links to microarray data and design files, and any associated results. An Agilent Genomic Workbench experiment file is a single ZIP file that contains the design and data files for one or more experiments. You can import

- Experiment files created in Agilent Genomic Workbench on another computer
- Agilent Genomic Workbench 5.0 and 6.X experiment files

1 In the Home tab, click **Import > Experiments.**

The Import Experiments dialog box appears. See "Import" on page 354.

2 Select the ZIP file that contains the experiment(s) you want to import, then click **OK.**

The program imports the experiment file. Designs appear as new folders in the Data pane, in the applicable design type folder. Array data appears within the applicable design folder, organized by genome build. In addition, the experiment(s) appear in the Experiment pane, with the appropriate arrays.

NOTE

Agilent Genomic Workbench exported experiment files contain all of the design and array data files for an experiment, but do not include any analysis parameter settings, array selections, or analysis results. To export the data and design files from one or more experiments, see "To export experiments" on page 83.

To import filters

Filters are used in Agilent Genomic Workbench to include or exclude data from an analysis, based on filter criteria. Filters are created in the interactive CGH and ChIP modules, or in workflow setup.

1 In the Home tab, on the Command Ribbon, click **Import > Filters**.

The Import dialog box appears. See "Import (filters)" on page 356 for more information.

2 Select the file that contains the exported filter(s) for import. and then click **Import**.

3 In the filters Import dialog box, mark the **Import** box next to each filter you want to import, and then click **OK.**

To import a genotype reference file

A genotype reference sample is required in order to analyze a CGH+SNP microarray. A genotype reference file contains reference genotypes for one or more genotype reference samples. See also "To generate a genotype reference file" on page 164.

1 From the Home tab, click **Import > Genotype References**.

The Import Genotype Reference Files dialog box appears.

2 Browse to a location and select the genotype reference file to import.

3 Click **Open**.

The Genotype Reference Importer dialog box appears. See "Genotype Reference Importer" on page 332.

4 Click **OK**.

The imported genotype references appear in the Navigator, in the Genotypes pane.

Working with Experiments to Organize Imported Data

This section describes how to organize imported array data and designs into *experiments*. Experiments, shown in the Experiment pane of the Navigator, contain links to specific array data and design files in the Data pane. After you set up an experiment, you can then analyze selected array data within the experiment.

Because experiments only contain *links* to the actual data and design files, any number of experiments can use a given set of files. In the data analysis modules (CGH, ChIP, or methylation, for example), experiments also can contain saved experiment results.

To display the array designs and data in the program

- To display the directory of data in the program, use the Data pane (Figure 14). Double-click a folder to expand or collapse it, or click the ⊞ and ⊟ buttons.

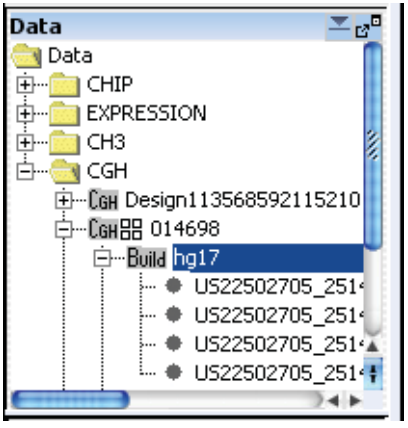


Figure 14 Data pane of the Navigator

In the Data pane, the program organizes design files by the module (CGH, ChIP, or methylation, for example) to which they apply. It organizes array data files by genome build under the design with which they are associated.

You can right-click many elements of the Data pane to open shortcut menus. For more information, see "Data pane – actions and shortcut menus" on page 207.

Many icons can appear in the Data pane. See "Data pane – icons, special text, and buttons" on page 206 for a complete list.

The Search pane can help you find specific data files or other content. See "To find specific items in the Navigator" on page 71.

To create a new experiment

In Agilent Genomic Workbench, *experiments* are organizational units that contain links to data and design files. To display or analyze data, you must first create an experiment and associate the data files with it. Because experiments only contain *links* to the actual data and design files, any number of experiments can use a given set of files. In data analysis modules (CGH, ChIP, or methylation, for example), experiments can also contain saved experiment results.

1 In the Home tab, click **Create Experiment.**

  The Create Experiment dialog box appears. See "Create Experiment" on page 278.

2 Type a **Name** and an optional **Description** for the experiment.

3 Do one of the following:

  - To create an empty experiment, and add data to it later, click **OK.** The program creates the experiment. To add arrays to the experiment later, see "To add arrays to an experiment" on page 62.

- To create an experiment and add data to it, follow these steps: (You can add or remove data from the experiment later.)

  a Click **Properties.**
  The Experiment Properties dialog box appears. See "Experiment Properties" on page 313.

  b Under Select Design, select the design and genome build associated with the desired array data.
  The applicable arrays appear in Array List.

  c In Array List, click the name of an array that you want in your experiment. Hold down the **ctrl** key while you click the names of additional arrays.

  d Click  .
  The program transfers the selected arrays to the Selected Array List.
  The dialog box also has other options for adding arrays. See "Experiment Properties" on page 313 for more information.

  e Click **OK.**
  The program creates the new experiment, and adds data to it from the selected arrays.

- To create an experiment and add data to it using the "drag and drop" method, follow these steps:

  a To create an empty experiment, click **OK.**
  The program creates the experiment.

  b From the Data pane, expand a design to see the build and array data.

  c Drag an array from the Data pane and drop it onto the experiment folder in the Experiment pane.

In all cases, a folder with the name of the new experiment appears in the Experiment pane of the Navigator.

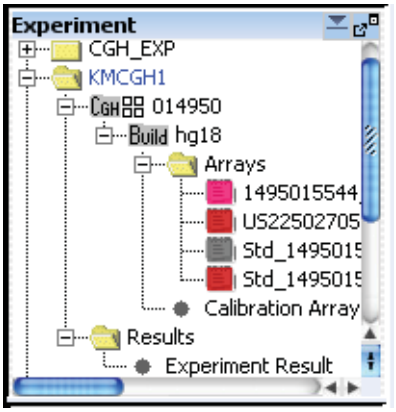


Figure 15 Experiment pane of the Navigator

To add arrays to an experiment

After you create an experiment, or import one, you can add arrays to it. When you add arrays to an experiment, you create links between the experiment and the array data and design files. Because the program does not move the actual files, multiple experiments can share the same arrays.

1. In the Experiment pane, double-click the **Experiments** folder to expand it.
2. Right-click the name of the experiment, then click **Show Properties.**

   The Experiment Properties dialog box appears. See "Experiment Properties" on page 313.
3. Under Select Design, select the design file and genome build for the arrays to add.

   The arrays for the selected design file and genome build appear in Array List.
4. In Array List, select the arrays to add to the experiment. To select a single array, click its name. To select additional arrays, hold down the **ctrl** key while you click their names.
5. Click > .

The program transfers the selected arrays to the Selected Array List.

The dialog box also has other options for adding arrays. See "Experiment Properties" on page 313 for more information.

6 Click **OK.**

To add array data to an experiment using the "drag and drop" method,

1 From the Data pane, expand a design to see the build and array data.

2 Drag an array from the Data pane and drop it onto the experiment folder in the Experiment pane.

If needed, the program adds appropriate design and genome build folders to your experiment folder in the Experiment pane. It places the arrays you selected in the appropriate genome build folder.

To associate a genotype reference with a CGH+SNP array

In order to perform CGH+SNP analysis on a microarray, you must first select a genotype reference from those available in the database.

1 In the Home tab, in the Data pane of the Navigator, expand a design until you see the individual microarrays.

2 Right-click the CGH+SNP microarray, and select **Show Properties**.

The Microarray Properties dialog box appears.

3 In the Attributes tab of the Microarray Properties dialog box, click the arrow next to Green Sample (or for dye-flipped arrays, Red Sample), and select a genotype reference for the microarray.

4 Click **Close**.

OR

1 In the Home tab, in the Experiment pane of the Navigator, expand an experiment until you see the individual microarrays.

2 Right-click the CGH+SNP microarray, and select **Show Properties**.

The Microarray Properties dialog box appears.

3 In the Attributes tab of the Microarray Properties dialog box, click the arrow next to Green Sample (or for dye-flipped arrays, Red Sample), and select a genotype reference for the microarray.

4 Click **Close**.

OR

- In the Sample Manager module, click the Green Sample box (or, for dye-flipped arrays, Red Sample) for the array, and select the genotype reference from the list. For more information, see the *Sample Manager User Guide*.

NOTE In order to select a genotype reference for an array, you must first import the genotype reference to the database. See "To import a genotype reference file" on page 58.

To change the order of arrays in an experiment

When you select an experiment, a table appears in the Tab View of Genomic Viewer that contains log ratio values and, if selected, signal intensities for arrays in the experiment. See "Tab View" on page 233. You can change the order in which the arrays appear in the table. If you display separate (stacked) scatter plots in Gene View and Chromosome View for each array, the array order also determines the order in which these plots appear. You can use this feature to organize your arrays more logically, or to make it more convenient to display certain arrays. It is especially useful if you have many arrays.

1. In the Experiment pane, right-click the name of the experiment, then click **Edit Array Order.**

   The Edit Array Order dialog box appears. See "Edit Array Order" on page 309.

2. In Design, select the design that contains the arrays whose order you want to change.

   The arrays from the selected design appear in Array Name.

3. Do any of the following:
   - To move an array up in the list, click its name, then click  .
   - To move an array down in the list, click its name, then click  .
   - To sort the list based on a specific microarray attribute, Order by, select the attribute.

4. Click **OK.**

To change the display names for arrays in an experiment

You can change the name displayed for arrays in an experiment, based on array attributes. When you change the display names for arrays in an experiment, the array names are changed only for the selected experiment. The display names are unchanged in the Data pane and in the other experiments.

1. Expand the folders in the Experiment pane until you see the experiment you want to change.
2. Right-click the experiment name, and select **Show Properties**.
3. In the Experiment Properties dialog box, click **Display Name by** and select an attribute to use for display of array names.
4. Click **OK**. The names of the arrays in the experiment are changed to the selected attribute. If the attribute does not exist for an array, the Global Display Name is displayed.

NOTE To change the name of an array throughout Agilent Genomic Workbench, change its Global Display Name using Sample Manager. For more information, see the *Sample Manager User Guide*.

To rename an array in an experiment

When you rename an array in an experiment, you change the array's name only within the context of a selected experiment. The name of the array is unchanged in the Data pane, and in other experiments.

1. Expand the folders in the **Experiment** pane until you can see the array you want to rename.
2. Right-click the name of the array, then click **Rename.**

   An Input dialog box appears.
3. Type the new name for the array, then click **OK.**

   The name of the array in the tab view of the selected experiment is renamed. The global display name of the array is not changed.

NOTE To see the original name of the array, move the mouse pointer over the array name. A ToolTip appears that displays the original array name.

To remove arrays from an experiment

When you remove arrays from an experiment, you only remove the links between the experiment and the data files. The files are still available in the program for use in other experiments. To completely remove files from the program, see "To remove data or design files from the program" on page 73.

1. In the Experiment pane, expand folders until you can see the experiment, and the array(s) to remove from it.
2. In the Arrays or Calibration Arrays folder of the experiment, click the name of an array to select it for removal. Hold down the **ctrl** key while you click the names of additional arrays.
3. Right-click one of the selected array names, then click **Delete.**

   A Confirm dialog box appears.
4. Click **Yes.**

   The program removes the links between the experiment and the selected array data files. If the removal of arrays leaves a design folder in the experiment empty, the program removes this folder as well.

To select or remove calibration array(s)

After you add an array to an experiment, you can select it as a calibration array. The program shows calibration arrays within the Calibration Arrays folder of the experiment with a special icon  . You can also remove the calibration designation from an array.

To select an array as a calibration array

1 Expand the folders of the Experiment pane until you can see the array to select as a calibration array.

2 Right-click the name of the array, then click Select for Calibration. To select all of the arrays of a given design in the experiment as calibration arrays, right-click the genome build folder of the design, then click **Set for Calibration**.

The program selects the array as a calibration array. In the Calibration Arrays folder of the applicable genome build and design within the experiment, the array appears with a special icon .

To deselect an array from calibration

1 Expand the folders of the Experiment pane until you can see the array. The program lists calibration arrays in the Calibration Arrays folder(s) of the applicable genome build(s) and design(s) within each experiment.

2 Right-click the name of the array, then click **Deselect from Calibration.**

To deselect multiple calibration arrays at once, select all of the arrays. Right-click one of the arrays, then click **Deselect from Calibration.** (To select multiple arrays, click the name of one array, then hold down the **ctrl** key and click the names of additional arrays. To select a contiguous block of arrays, click the name of the first array, then hold down the **shift** key and click the name of the last one.)

The program removes the array(s) from calibration, and moves the arrays to the Arrays folder of the applicable genome build and design within the experiment. The icons of the arrays change to the standard (non-calibration) array icon .

To show or hide array attributes in an experiment

Sample attributes are pieces of information specific to an array, such as Array ID or hybridization temperature. You can show or hide attributes for the arrays in the experiment with the Sample Attributes dialog box. See "Sample Attributes" on page 390.

NOTE You cannot hide the *required* attributes. These include Array ID, Global Display Name, Green Sample, Red Sample (for 2-color arrays), and Polarity.

1. Right-click the experiment whose attributes you want to show or hide, or to change.
2. Click **Sample Attributes**.

   You see the array attributes and their values that were set up in the Sample Manager table. See the *Sample Manager Guide*.
3. Click **Show/Hide Attributes**.

   The Show/Hide Columns dialog box appears. See "Show/Hide Columns" on page 403.
4. Mark the check boxes for the attributes you want to show, or clear the check boxes for the attributes you want to hide. These changes are applied globally for the arrays.
5. Click **Save**.
6. In the Show/Hide Columns dialog box, click **Close**.
7. Click **Close**.

NOTE You cannot create new attributes using this dialog box. To do this, you must use the Sample Manager tab. See the *Sample Manager User Guide*.

To display or edit array attributes in an experiment

1. Right-click the experiment whose attributes you want to display or edit.
2. Click **Sample Attributes**.

   You see the array attributes and their values that were set up in the Sample Manager table. See the *Sample Manager Guide*. See "Sample Attributes" on page 390.
3. Double-click the cell whose array attribute value you want to change.

NOTE You cannot change Array ID, Polarity, Extraction Status, or IsMultiPack attributes for extracted or UDF arrays.

4 Click **Save Changes**.

5 Click **Close**.

To display or edit the attribute values of a specific array

Array attributes are pieces of information specific to an array, such as array type or hybridization temperature. Sample attributes are usually set using the Sample Manager tab. For more information, see the *Sample Manager User Guide*. In the Navigator of the CGH module, you can display or change attributes for each array. You can also select a Genotype Reference to use for a selected CGH+SNP microarray.

1 Expand the folders of the Data pane or the Experiment pane until you can see the array of interest.

2 Right-click the name of the array, then click **Show Properties.**

The Microarray Properties dialog box appears, with a list of array attributes. See "Microarray Properties" on page 367. You can also edit the attributes of a specific array from this dialog box. In addition, if the array is an Agilent array, you can see header and feature information sent from the Agilent Feature Extraction program.

3 When you are finished, click **Close.**

NOTE You use the Sample Manager tab to organize, create, import, and export array attributes. See the *Sample Manager User Guide*.

Managing Content

This section describes how to create, find, rename, update, combine, and/or remove content such as data, gene lists, and tracks, stored in Agilent Genomic Workbench. To display the data, gene list and track content, see Chapter 3, "Displaying CGH and CGH+SNP Data and Other Content".

To display a list of the content stored in the program

The Data, Experiment, My Entity List, and Genotypes panes of the Navigator show the content stored in Agilent Genomic Workbench.

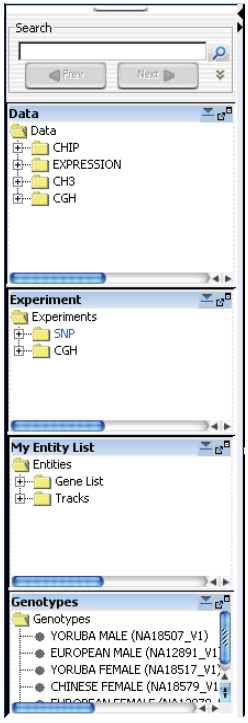


Figure 16 Navigator

Data pane – Shows all of the design and array data files stored in the database. For more information, see "To display the array designs and data in the program" on page 59.

Experiment pane – Shows the experiments that were created or imported to the program. To select an experiment, double-click its name. To display the contents of an experiment, right-click the experiment name and then select **Expand Node**.

My Entity List pane – Shows the gene lists and tracks stored in the program. To display the names of gene lists or tracks available in the program, double-click the names of folders to expand or collapse them, or click the ⊞ or ⊟ buttons.

Genotypes pane – Shows the genotype reference samples in the program.

To find specific items in the Navigator

At the top of the Navigator is a search pane that can help you find specific items. See "Search pane" on page 204.

1. Type a search term in the box at the top of the Navigator. The search term is not case-sensitive, but it must reflect the entire name of the item to find. You can use asterisks (*) as wildcards to represent a group of unspecified characters. For example, if you type *1234*, the search will find all items that contain "1234" in the name.
2. By default, the program searches all panes of the Navigator. To limit your search to a specific pane, click  . In the list that appears, select the desired pane.
3. Click  .

   The program searches the selected pane(s). If it finds item(s) that match your search term, it expands folders so that the items are visible, and highlights them in red. You may need to scroll down to see all the search results.
4. To clear the results of a search, click  .

To display the properties of a specific design

Design properties include general information about a design, such as its name, application type, and associated species. They also include a list of the names and chromosomal locations of probes.

1 Expand the folders of the Data pane until you can see the genome build folder(s) in the selected design folder.

2 Right-click the genome build folder, then click **Show Properties.**

The Design Properties dialog box appears. See "Design Properties" on page 299.

To update probe annotation in design files

Agilent regularly makes updates to probe annotations on its eArray Web portal. If you have imported Agilent array designs into Agilent Genomic Workbench, and you are a registered eArray user, you can download the updated design files from within Agilent Genomic Workbench. For more information about eArray, go to https://earray.chem.agilent.com and click **Help.**

1 In the Home tab, click **User Preferences.**

The User Preferences dialog box appears.

2 In the Miscellaneous tab, under **eArray User Details,** type your eArray **Username** and **Password.** See "User Preferences" on page 414.

3 Click **OK.**

4 Expand the folders of the Data pane until you can see the design to update.

5 Right-click the design, then click **Update from eArray.** This option appears only for Agilent designs.

A confirmation dialog box appears.

6 Click **Yes.**

The program transfers an updated design, if one is available.

To rename an array in the Data pane

This topic describes how to rename an array in the Data pane, which changes the Global Display Name for the array. If you rename an array in this way, and subsequently add the array to an experiment, the array appears in the experiment with the new name. It also changes the array name in any experiment to which it is already linked. To rename an array only within the context of a specific experiment, see "To rename an array in an experiment" on page 65.

1 Expand the folders of the Data pane until you can see the array you want to rename.

2 Right-click the name of the array, then click **Rename.**

An Input dialog box appears.

3 Type a new name for the array, then click **OK.**

The program renames the array.

NOTE To see the original name of the array, move the mouse pointer over the array name. A ToolTip appears that displays the original array name.

To remove data or design files from the program

You can delete array design and data files from the program when you are finished with them.

1 If an array to delete is associated with an experiment, first delete it from the experiment. See "To remove arrays from an experiment" on page 66.

2 In the Data pane, expand folders until you can see the design folder or array to delete.

3 Do one of the following:

- For array data files, click the name of the first array, then hold down the **ctrl** key while you click the names of additional arrays within the same design.

- For array design folders, click the name of the first design folder, then hold down the **ctrl** key while you click the names of additional ones. This selects the designs and all array data files within them for deletion.

4 Right-click the name of a selected design folder or array data file, then click **Delete.**

A confirmation dialog box appears.

5 Click **Yes.**

The program deletes the selected files.

CAUTION When you delete files from the Data pane, you permanently remove them from Agilent Genomic Workbench. To restore deleted files, you must import them again.

To create a gene list

When you create a gene list, you create a list of the genes in a contiguous chromosomal region that you select. To create a list of genes in multiple regions, create multiple gene lists and combine them. See "To add one gene list to another" on page 76.

1 Follow these steps to select a chromosomal region for your gene list. If you know the exact start and end locations of the chromosomal region, skip to step 2.

a In Genome View, select the chromosome.
The selected chromosome appears in Chromosome View. See "Chromosome View" on page 226.

b In Chromosome View, in the plotting area to the right of the chromosome, drag the pointer over the chromosomal region of interest.
The program draws a blue box around the region, and displays the region in greater detail in Gene View.

c In Gene View, adjust the view so only the genes of interest appear. For a description of the adjustment commands available in Gene View, see "Gene View" on page 228.

2 Right-click anywhere within the log ratio plotting area in Gene View, then click **Create Gene List**.

The Create Gene List dialog box appears. See "Create Gene List" on page 279.

3 In the dialog box, set the Name, Description and Color.

4 In the dialog box, select the chromosomal region for the new gene list.

5 Click **OK.**

The new gene list appears in the Gene List folder in the My Entity List pane of the Navigator.

To import a gene list

A gene list file is a plain text (*.txt) file that contains one gene name per line. When you import a gene list into Agilent Genomic Workbench, it appears in the Gene List folder in the My Entities List pane. You can use the gene list to highlight specific genes, or to show or hide the appearance of genes and data, in Gene and Chromosome Views. See "To show gene lists in Gene View" on page 109.

1 In the **My Entities List** pane, double click the **Entities** folder to expand it.

2 Right-click the **Gene List** folder, then click **Import Gene List.**

An Import dialog box appears. See "Import" on page 354.

3 Select the desired gene list file. To select additional gene list files, hold down the **ctrl** key and click their names.

Click **OK.**

To display the genes in a gene list

You can display the genes in a gene list as a table.

1 Expand the folders in the My Entity List pane until you can see the gene list.

2 Right-click the gene list, then click **View In Table.**

The Gene List dialog box appears, with a table that contains the names of the genes in the gene list. You can also use this dialog box to edit the description of the gene list and its display color. See "Gene List" on page 328.

You can also create gene lists. For more information, see "To create a gene list" on page 74.

To add one gene list to another

You can add one gene list (a source gene list) to another (the target gene list). The program appends the source gene list to the end of the target gene list, and leaves the source gene list unchanged.

1. Expand the folders in the My Entity List pane until you can see the gene lists to combine.
2. Right-click the source gene list, then click **Add to Gene List.**

   A dialog box appears. For more information, see "Add Gene List <name> to" on page 243.
3. In Select target gene list, select the target gene list.
4. Click **OK.**

To rename a gene list

The name of a gene list identifies it within the Gene List folder of the My Entity List pane. You can rename gene lists.

1. Expand the folders of the My Entity List pane until you can see the gene list to rename.
2. Right-click the gene list, then click **Rename.**

   An Input dialog box appears.
3. Type a new name for the gene list, then click **OK.**

To delete gene list(s)

1 In the My Entity List pane of the Navigator, click to expand the **Gene List** folder.

2 Click the name of a gene list to delete. Hold down the **ctrl** key while you click the names of additional gene lists.

3 Right-click one of the selected gene lists, then click **Delete.**

A confirmation dialog box appears.

4 Click **Yes.**

To create a track

When you create a track, you create a list of the genes in a contiguous chromosomal region that you select. To create a list of genes or other annotations, such as CNV or miRNA, in multiple regions, create additional tracks, and combine them. See "To combine tracks" on page 78.

1 Follow these steps to select a chromosomal region for your track. If you know the exact start and end locations of the chromosomal region, skip to step 2.

   a In Genome View, select the chromosome.
   The selected chromosome appears in Chromosome View.

   b In Chromosome View, in the plot area to the right of the chromosome, drag the pointer over the approximate chromosomal region of interest.
   The program draws a blue box around the region, and displays the region in greater detail in Gene View.

   c In Gene View, adjust the view so only the genes of interest appear. For a description of the adjustment commands available in Gene View, see "Gene View" on page 228.

2 Right-click anywhere within the log ratio plot area in Gene View, then click **Create Track.**

The Create Track dialog box appears. See "Create Track" on page 284.

3 In the dialog box set the Name, Description and Color.

4 In the dialog box select the chromosomal region for the new track.

5 Click **OK.**

The new track appear in the My Entity List of the Navigator in the Tracks folder.

To display the details of a track

You can display a table that contains the values for a list of track attributes.

1 In My Entity List pane, expand the Tracks folder to see the track.

2 Right-click the name of the track, then click **View Details.**

Data describing the track appears in a Track table. See "Track" on page 409.

To combine tracks

You can create a track that contains elements from two or more existing tracks. The existing tracks must be available in Agilent Genomic Workbench, and they must be associated with the same genome build.

1 In the My Entities List pane, double-click the **Entities** folder to expand it, if necessary.

2 Right-click the **Tracks** folder, then click **Combine Tracks.**

The Combine Tracks dialog box appears. See "Combine Tracks" on page 261.

3 In Name, type a name for the combined track. The program uses this name to identify the track in the Tracks folder, and to identify the track if it appears in Gene View.

4 Click **New Condition.**

A new row appears in the Track/Operator list.

5 Under Track, select the first track to combine.

6 Click New Condition, then select another Track/Operator pair. You can set up as many Track/Operator pairs as you like, but you must set up at least two. When you add a track, the program automatically assigns the AND operator to the previous track.

To remove the bottom row from the list, click **Delete Condition.** To delete all rows from the list, and erase any entry in Name, click **Reset.**

7 Under Operator, select one of the following:

Operator	Comments
AND	Creates a combined track out of 2 tracks that will contain elements that appear in both tracks
OR	Creates a combined track out of 2 tracks that will contain elements that appear in either of the tracks
MINUS	Removes the elements of the second track from the first track.

8 Click **Save.**

Your combined track appears in the Tracks folder of the My Entity List pane. The Combine Tracks dialog box remains open so you can combine additional tracks.

9 Click **Close** when you are done combining tracks.

To rename a track

The name of a track identifies it both within the Tracks folder of the My Entity List pane, and in Gene View when you select **Show In UI** for the track. You can rename tracks.

1 Expand the folders of the My Entity List pane until you can see the track to rename.

2 Right-click the track, then click **Rename.**

An Input dialog box appears.

3 Type a new name for the track, then click **OK.**

To delete tracks

1. In the My Entity List pane of the Navigator, expand the Tracks folder.
2. Click the name of a track to delete. Hold down the **ctrl** key while you click the names of additional tracks.
3. Right-click one of the selected tracks, then click **Delete.**

   A confirmation dialog box appears.
4. Click **Yes.**

To display genotype reference details

1. In the Genotypes pane of the Navigator, right-click the name of the genotype reference you want to display.
2. Click **Show Properties**.

   The Genotype Reference Details dialog box appears. See "Genotype Reference Details" on page 331.

To rename a genotype reference

1. In the Genotypes pane of the Navigator, right-click the name of the genotype reference you want to rename.

   The Input dialog box appears.
2. Type the new name for the genotype reference, and then click **OK**.

To delete a genotype reference

1. In the Genotypes pane of the Navigator, right-click the name of the genotype reference you want to delete.
2. Click **Delete**.

   A confirmation dialog appears.
3. Click **Yes**.

NOTE When you delete a genotype reference, the green and red sample attributes for any microarray associated with this genotype reference are reset.

Exporting and Saving Content

This section describes how to export different kinds of files from the program.

To export array attributes

You can export selected array attributes for any imported arrays. You first select the arrays and then the attributes for the selected arrays. You can export array attributes from the Home tab or from the shortcut menu for an experiment.

1 Click **Home > Export > Array Attributes.**

OR

In the Experiment pane of the Navigator, right-click an experiment of interest, and click **Export Attributes**.

The Export Array Attributes dialog box appears with the Array tab displayed. See "Export Array Attributes" on page 316.

If you opened this dialog box by right-clicking an experiment, only those arrays selected for the experiment appear in the Selected Array List. You can add or subtract from the list.

2 Under Select Design, select the design file and genome build for the arrays to add.

The arrays for the selected design file and genome build appear in Array List.

3 In Array List, select the arrays whose attributes you intend to export. To select a single array, click its name. To select additional arrays, hold down the **ctrl** key while you click their names.

4 Click  .

The program moves the selected arrays to the Selected Array List.

5 Click **Next** to select attributes for the selected arrays.

The Export Array Attributes dialog box appears with the Attribute tab displayed. See "Export Array Attributes" on page 316.

All of the attributes for the arrays are already located in the Selected Attribute List.

6 Move any attributes you don't want to export to the Attribute List.

   a In the Selected Attribute List, highlight those attributes you do not intend to export. To select additional attributes, hold down the **ctrl** key while you click their names.

   b Click  .

7 Click **OK**.

   The Export dialog box appears. See "Export" on page 315.

8 Select the folder in which to save the attributes, and click **Export**.

   The attributes will be saved to the selected folder as a .txt file.

To export experiments

You can export experiments as a ZIP file to transfer them to another computer. Exported experiments contain only the associated design and array data files. The program does not export information about array selections, or any analysis parameters or results.

1 In the Home tab, click **Export > Experiments.**

   The Export Experiments dialog box appears. See "Export Experiments" on page 320.

2 Mark the experiments to export. To export all experiments, click **Select All.** Click **OK.**

   An Export dialog box appears. See "Export" on page 315.

3 Select a location and type a name for the exported ZIP file.

4 Click **Export.**

   The program exports all selected experiment(s) together as a single ZIP file.

To export a gene list

You can export a gene list as a text file that contains one gene per line.

1 In the My Entity List pane, in the Gene List folder, right-click the gene list to export, then click **Save As.**

   A Save As dialog box appears.

2 Select a location and type a name for the file.

3 Click **Save.**

   A message appears when the operation is complete.

4 Click **OK.**

To export tracks

You can export selected tracks as a BED format track file. You can then import this file into Agilent Genomic Workbench on another computer, or into a genome browser that accepts BED format files.

1 In the Home tab, click **Export > Tracks**.

   The Export Tracks dialog box appears. See "Export Tracks" on page 322.

2 Mark the tracks to export. To select all tracks for export, click **Select All.**

3 Click **OK.**

   An Export dialog box appears.

4 Select a location and type a name for the exported track file, then click **Export.**

   The program exports the track(s) as a single BED format track file.

To export filters

You can export selected array, feature, design, metric, and aberration filters that are available in some data analysis modules in Agilent Genomic Workbench. The program exports all selected filters as a single *.xml file that you can import at a later time.

1 In the Home tab, click **Export > Filters.**

The Export Filters dialog box appears. See "Export Filters" on page 321.

2 Under Export, mark the check boxes beside the filter(s) to export. To select all filters for export, click **Select All.**

3 Click **OK.**

An Export dialog box appears.

4 Select a location and type a name for the exported file, then click **Export.**

The program exports all selected filters as a single *.xml file.

To copy what you see in the main window

You can copy panes of the main window to the Clipboard as images, and then paste them into a new document in another program (such as Microsoft® Word, or PowerPoint). The images contain only what actually appears on your screen; regions to which you must scroll are not included.

1 In the View tab, click **Copy.**

2 In the shortcut menu that appears, click the name of the pane to copy. You can copy any view, or the Navigator. To copy all of the panes, click **All.**

The program copies the selected pane(s) to the Clipboard.

NOTE To adjust how data is displayed in the panes, use the View Preferences dialog box. See "View Preferences" on page 422 for more information.

3 Open a document in a program that accepts images. In that program, click **Edit > Paste,** or the appropriate paste command.

To copy the list of array colors for an experiment

You can copy the list of arrays in an experiment, and the colors assigned to them, to the Clipboard as an image. You then paste the image into a document in another program such as Microsoft® Word or PowerPoint.

1 In the Experiment pane, expand the **Experiments** folder.

2 Right-click the name of the experiment, then click **Edit Array Color.**

The Edit Array Color dialog box appears. See "Edit Array Color" on page 308.

3 In the dialog box, click **Edit > Copy.**

The program copies the names of the arrays and their colors to the Clipboard as an image.

4 Open a program that accepts images. Click **Edit > Paste,** or the appropriate paste command for the specific program.

To save data and design information from an experiment

You can save the data and design information from a single design in an experiment as a tab-delimited text file.

1 In the Experiment pane, expand the **Experiments** folder until you see the genome build(s) for the design you want to export.

2 Right-click the name of the genome build, then click **Save As Text File**.

A dialog box appears.

3 Select a location and type a name for the file, then click **Save.**

To save SNP data and design information from an experiment

You can save the SNP data and design information from a single design in a CGH+SNP experiment as a tab-delimited text file.

1 In the Experiment pane, expand the **Experiments** folder until you see the genome build(s) for the CGH+SNP design you want to export.

2 Right-click the name of the genome build, then click **Save As SNP Text File**.

The Save SNP Design dialog box appears.

3 Select a location and type a name for the file, then click **Save.**

3 Displaying CGH and CGH+SNP Data and Other Content

Selecting an Experiment for Analysis 90
Displaying Array Data 96
Displaying Gene Lists and Tracks 109
Displaying CGH+SNP Content and Results 117
Searching for Probe and Gene Information 119

This chapter shows you how to display calculated data from imported feature extraction CGH and CGH+SNP data files, as well as gene and track content, in the Genomic Viewer. It also gives you instructions on how to customize the display of data and content to meet your needs.

Selecting an Experiment for Analysis

An experiment is a set of links to microarray data and design files, and any associated results. You can see a list of the experiments in the Experiments pane of the Navigator. See "Navigator" on page 202 for more information.

When you select an experiment and the Preprocessing and Analysis options have not been turned on or set to apply, the program shows the log ratio data of selected arrays in the active experiment. See "To locate and display data (or results) within the Views" on page 103 for more information.

When you select an experiment and Preprocessing and Analysis options have been turned on or set to apply, the program automatically begins the analysis of the selected array data with current settings and displays its results, if certain options have been set.

This section describes how to select an experiment, select or deselect arrays for further analysis, and analyze arrays one at a time.

To select an experiment

When you select an experiment, the program begins the analysis with the current settings. You can either set the Preprocessing and Analysis parameters before you select the experiment, or change the settings one at a time after the first analysis and reanalyze. Every time you change a Preprocessing or Analysis setting for a selected experiment, the program recalculates results.

1 If necessary, do one of the following to add the desired experiment to the Experiment Pane in the Navigator:

- Create a new experiment and add data to it. See "To create a new experiment" on page 60.
- Import a saved DNA Analytics 5.0 or 6.X CGH experiment. See "To import an experiment file" on page 57.

2 In the Navigator, double-click the name of the experiment.

The Experiment Selection dialog box appears.

3 Click **Yes.**

In the Experiment pane of the Navigator, the name of the experiment turns blue. The name also appears in the title bar of the main window. Tables of data and design information appear in Tab View. For more information on the tabs, see "Command Ribbons" on page 176.

If you have selected to show the moving average or results of an aberration calculation, then the moving average or aberration results appear for the first array when you select the experiment, if you have not selected any other arrays.

You can select or deselect arrays in the experiment both before and after you select it. Every time you select or deselect an array in an *active* experiment or change a setting, the program reanalyzes the new data set with the changed settings. See "To select or deselect arrays in the experiment" on page 91.

When you select the experiment after deselecting it or selecting another one, the experiment is simply restored if the settings haven't changed. If they have changed, the program reanalyzes all of the arrays assigned when the experiment was last selected.

To select or deselect arrays in the experiment

To include arrays for display and analysis, you select them from the arrays available, either in an inactive experiment or the selected one. When you first create an experiment, the program automatically sets the first array in the experiment for analysis. If you do not select additional arrays for analysis, only the first one will be analyzed when the experiment is selected.

To select the arrays for analysis before experiment selection:

1 Hold down the **shift** key to highlight contiguous arrays or hold down the **ctrl** key to highlight noncontiguous arrays.

2 Right-click the highlighted arrays, and click **Select**.

Even though the selected arrays do not change color, they will change color after activation.

In the Navigator, the color or an array's icon has the following meaning after experiment selection:

 Array is not selected.

 Array is selected. The specific color matches the color of the column headings for the array in Tab View. In addition, the program displays aberration results and moving averages related to this array in this color. To configure a custom color for the array, see "To change the display color of an array" on page 93.

To select or deselect arrays in a *selected* experiment:

1. In the Navigator, expand the folders of the selected experiment.
2. Click the name of an array you want to include in the display.

   To include additional arrays, hold down the **ctrl** key while you click their names. To include a contiguous block of arrays, click the name of the first array in the block, then hold down the **shift** key while you click the name of the last one.
3. Right-click the name of one of the highlighted arrays, then click **Select.**

   After you select the arrays, the program reanalyzes the data set within the experiment and displays the data in Genome, Chromosome, and Gene Views. You can see the data and results for just the selected arrays in the Selected Arrays tab in Tab View.

   To show analysis results if they do not appear, see "To show results of analysis" on page 142.

   To customize the appearance of the results in Genome, Chromosome, and Gene Views, see "To change scatter plot appearance" on page 101.

You can also use the headings of columns in Tab View that contain array data to select and deselect arrays.

- Click a column heading to select that array only.
- Hold down the **ctrl** key while you click a column heading to select or deselect an array without changing the status of other arrays.
- Right-click a column heading to open a shortcut menu with options that let you select or deselect that array, or all arrays.

For more information on Tab View, see "Tab View" on page 233.

To change the display color of an array

The color assigned to an array sets the color of its icon when you select the array within an experiment. It also changes the colored square in the array's column heading in Tab View. Agilent Genomic Workbench displays the aberrations for each selected array in its assigned color.

1 In the Experiment pane of the Navigator, in the Experiments folder, expand the folder of an experiment until you can see the array of interest.

2 Right-click the desired array, then click **Edit Array Color.**

The Select Color dialog box appears. The dialog box gives three different ways to select the desired color. See "Select Color" on page 394.

3 Select the desired color in one of the following ways:

Dialog box tab	Instructions
Swatches	• Click the desired color swatch.
HSB (Hue/Saturation/Brightness)	Type or adjust the values in H (Hue), S (Saturation), and B (Brightness), or alternately, follow these steps: **a** Select **H**, then drag the slider to select a hue based on the color strip to its right. **b** Click an appropriate location in the large color box to the left of the slider to set the saturation and brightness levels of the color. Both the HSB and equivalent RGB values of the color appear in the dialog box. Note these values; they will be useful if you need to duplicate this color in the future.
RGB (Red/Green/Blue)	Do one of the following. Note the final RGB Values; they will be useful if you need to duplicate this color in the future. • Drag the Red, Green, and Blue sliders. • Type or adjust values in the boxes to the right of the sliders.

Samples of the color in different contexts appear under Preview. The upper half of the color sample on the right shows the original color for comparison.

4 Adjust the color as desired, then click **OK.**

You can also manage all of the colors for all of the arrays in an experiment. Right-click the desired experiment, then click **Edit Array Color.** For more information, see "Edit Array Color" on page 308.

To analyze an experiment one sample at a time

You can use Sample-By-Sample View to analyze or reanalyze the *active* experiment one array at a time. When you work in this special mode, the program displays only one array at a time, and it also adds a new Sample-By-Sample View to the main window, where you can create or edit the Karyotype and Research Notes attributes of each array as you go.

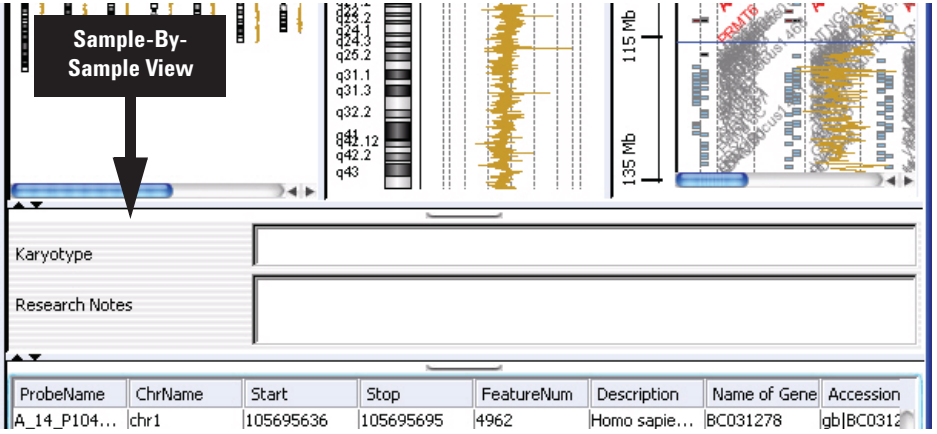


Figure 17 Sample-By-Sample View

1 Select an experiment. See "To select an experiment" on page 90.

2 Click **View > Show,** then mark **SampleBySample View.**

Sample-By-Sample View appears. The program also displays the first array in the experiment in Genome, Chromosome, Gene, and Tab Views.

3 Analyze the first array, as desired. You can also type comments in **Karyotype** and **Research Notes.**

The program saves what you type in the Karyotype and Research Notes attributes of the array.

4 Do any of the following:

- To show and analyze another array, right-click its name within the selected experiment in the Navigator, and click **Select**.
- To leave Sample-By-Sample mode, click **View > Show,** then clear **SampleBySample View.**

Displaying Array Data

After you select an experiment, you can change how data appear within the Views or change the appearance of the Views that contain the data (or results).

To display the scatter plots

Within the Chromosome and Gene views, there are up to three possible scatter plot display panels that you can turn on and off. These panels are used to let you examine different types of data in more detail. You can select to show Log Ratios, Signal Intensities, and (for CGH+SNP arrays) Copy Number scatter plots.

NOTE At least one scatter plot panel must be selected. The SNP data panel is only used for CGH+SNP arrays.

By default, display of log ratio scatter plots in Gene View is turned On. If you do not see the scatter plot(s), or if you want to turn on scatter plots for all views, do one of the following:

1 From the View tab, click **View Preferences**. See "View Preferences" on page 422 for more information.

2 In the View Preferences dialog box, under Data Visibility, select **All views** and then mark the box next to **Scatter Plot**.

OR

1 Right-click in any of the views, and select **View Preferences**. See "View Preferences" on page 422 for more information.

2 In the View Preferences dialog box, under Data Visibility, select **All views** and then mark the box next to **Scatter Plot**.

NOTE You use the View Preferences to select the array design type and select what data you want to display in the scatter plots. See "To turn on or off data visibility in scatter plots" on page 99.

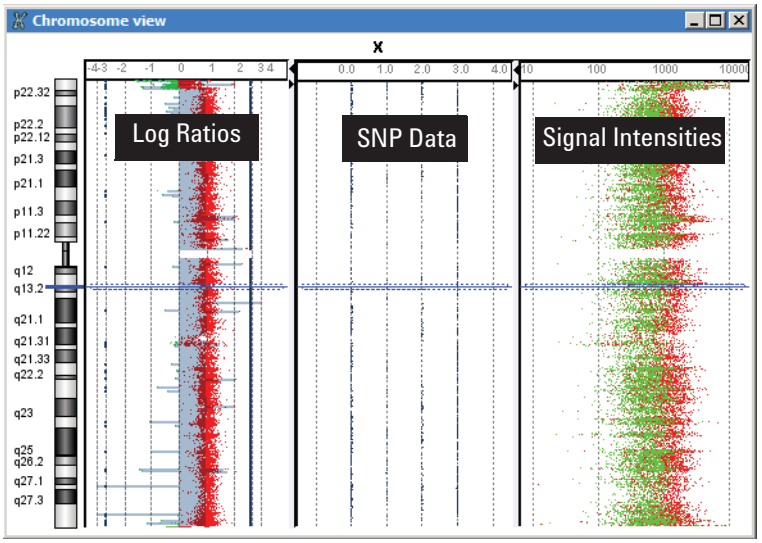


Figure 18 Scatter plot panels in Chromosome view

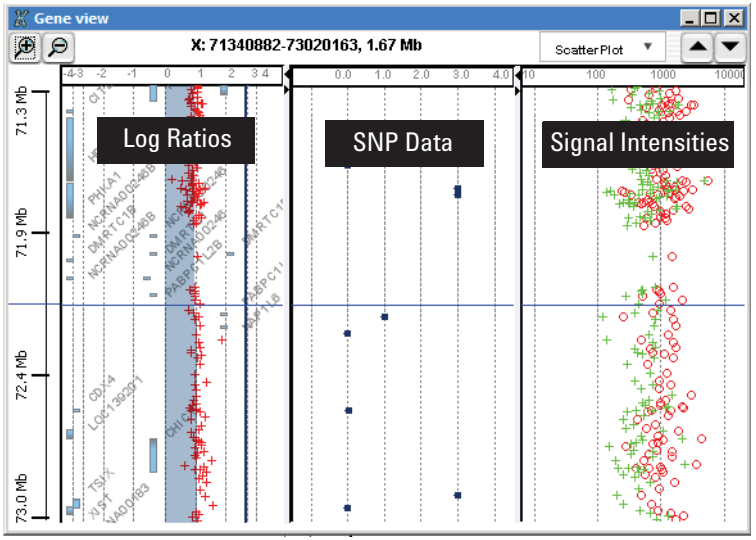


Figure 19 Scatter plot panels in Gene view

To show or hide data in the scatter plots

1 In the Gene View, move the mouse pointer over the down arrow next to Scatter Plot until the Scatter Plot box appears, and do any of the following:

To do this	Follow these steps
Show or hide the log ratio values in the Log Ratios plot	• To show the data points - Mark the **Log Ratios** check box and select **Log Ratio Values** from the list. • To hide all data points - Clear the **Log Ratios** check box.
Show or hide LogRatios color-coded by Probe Score Values in the Log Ratios plot	• To show the data points - Mark the **Log Ratios** check box and select **Probe Score Values** from the list. • To hide the data points - Clear the **Log Ratios** check box.
Show or hide Intensity values in the Signal Intensities plot	• To show the data points - Mark the **Signal Intensities** check box and select **Intensity Values** from the list. • To hide all data points - Clear the **Signal Intensities** check box.
Show or hide Signal Intensities color-coded by Channels in the Signal Intensities plot	• To show the data points - Mark the **Signal Intensities** check box and select **Channels** from the list. • To hide the data points- Clear the **Signal Intensities** check box.
Show or hide Signal Intensities color-coded by Probe Score values in the Signal Intensities plot	• To show the data points - Mark the **Signal Intensities** check box and select **Probe Score Values** from the list. • To hide the data points- Clear the **Signal Intensities** check box.
Show or hide SNP data panel	• To show the SNP data panel - Mark the **Show SNP Data Panel** check box. • To hide the SNP data panel - Clear the **Show SNP Data Panel** check box.
Change the ranges and colors for scatter plot and signal intensities panels	• Click **Configure Color and Ranges** to enter ranges and change colors. See "Configure Coloring Ranges and Shades" on page 264 for more information.

2 Click **X** to close the Scatter Plot window.

To turn on or off data visibility in scatter plots

You can select to show or hide different types of data that appear in the scatter plots. You can also select in what views (Genome, Chromosome, Gene) the data will appear.

NOTE Only data that is selected for viewing appears in the scatter plots.

1. Click the **View** tab, and then click **View Preferences.**

   OR

   Right-click in the Chromosome or Gene view, and select **View Preferences.**
2. Under Data Visibility, click the arrow next to **View** and select the view you want to display the data (Genome View, Chromosome View, Gene View, or All views).

   Data choices for the selected view are displayed.
3. Mark the check boxes for the data you want to show in the scatter plots.
4. Clear the check boxes for the data you do not want to show in the scatter plots.
5. Click **Apply** to apply the changes and leave the dialog box open, OR click **OK** to apply the changes and close the dialog box.

To customize scatter plot ranges and colors

To make it easier to see significant results, you can customize the display of scatter plot data. For each data type (log ratio, signal intensity) you can set custom ranges and colors for the display. For channels, you can set custom colors only. Copy numbers are always displayed using the selected array color. See "To change the display color of an array" on page 93.

Add and customize a plot

1 In Gene View, move the mouse pointer over **Scatter Plot** to display the options.

OR

Right-click in any of the views, and select **View Preferences**. See "View Preferences" on page 422.

2 Mark the **Log Ratios** check box under Configure Coloring schemes.

3 Select a data type from the "color by" list.

4 Click **Configure Color and Ranges**.

The Configure Coloring Ranges and Shades dialog box appears where you set ranges and colors for any of the data types. For more information, see "Configure Coloring Ranges and Shades" on page 264.

5 In the Configure Coloring Ranges and Shades dialog box, click the **Log Ratios** or **Signal Intensities** tab and then select the "color by" data type to configure.

6 Type minimum and maximum numbers to define a range for the data type.

7 Click **Color** to open the Select Color dialog box. Use the tabs to select a color for the range. See "Select Color" on page 394 for more information.

8 Click **OK** to close the Select Color dialog box and return to the Configure Coloring Ranges and Shades dialog box.

9 Click **Add Range** to add the custom range to the range list.

10 When you are done, click **OK** to close the dialog box.

Edit or remove a range

1 In the Configure Coloring Ranges and Shades dialog box, click the **Log Ratios** or **Signal Intensities** tab and then select the "color by" data type to configure.

2 In the range list, mark the **Edit/Delete** box to select the range. You can mark more than one range.

3 Click **Edit Range** to change the minimum and maximum values, or to change the color for the selected range.

4 Click **Delete Range** to delete the selected range.

5 Click **OK** to close the dialog box.

To change scatter plot appearance

You use the View Preferences dialog box to change the appearance of the scatter plots in Chromosome and Gene Views.

1 In the Genomic Viewer, right-click in the Gene View or Chromosome View, and then click **View Preferences**.

Or, click the View tab, and then click **View Preferences**.

The View Preferences dialog box appears. See "View Preferences" on page 422.

2 Do any of the following:

To do this	Follow these steps
Show or hide the scatter plot	**a** Under Data Visibility, in View, select **All Views.** **b** Do one of the following: To show the scatter plot, mark **Scatter Plot.** To hide the scatter plot, clear **Scatter Plot.** **c** Click **OK.**
Change the symbol that appears for data points	You can select the symbol separately for each design type. **a** Under Rendering patterns, select the desired **Design type.** **b** Under Styles, for each data type, select the desired symbol. **c** Click **Apply.**
Show a separate scatter plot in Gene and Chromosome Views for each selected array	**a** Under View Alignment, under Rendering Style, select **Stacked.** **b** Click **Apply.**
Show one scatter plot that contains data for all selected arrays	**a** Under View Alignment, under Rendering Style, select **Overlaid.** **b** Click **Apply.**
Enable ToolTips for the scatter plot in Gene View	ToolTips show information about an individual data point when you place the pointer over it. **a** Under Data Visibility, in View, select **Gene View.** **b** Mark **Scatter Tool Tip.** **c** Click **Apply.**

3 You can also configure ranges and colors for the scatter plot from the **View Preferences** dialog box. For more information, see "To customize scatter plot ranges and colors" on page 99 and "View Preferences" on page 422.

4 Click **OK.**

To print the scatter plot

You can print the scatter plot as it appears in Genome, Chromosome, and Gene views. Each view selected in the analysis is printed on a separate page. Chromosomes and genes appear on the printed pages, but tracks do not.

1 In the Home tab, click **Print**.

The Print dialog box appears.

2 Set print options, as desired, then click **OK.**

To create custom scales for Views

You can customize the scale used for display in the Chromosome View and Gene View. Custom scales are applied to both views.

1 Click the View tab and then click **View Preferences.**

2 In the View Preferences dialog box, under Configure Scales, mark the box next to **Apply** for the Log Ratios, Signal Intensities, and/or Copy Number.

3 In Range, type a value to use for the range. The range you type changes the scale for the display of the selected data.

NOTE The default range for the SNP data panel is -1 to the maximum copy number value for that chromosome, equally divided. When you apply a custom scale, the SNP data panel range is set from -1 to the specified range. This will create plots with integer scale values.

To locate and display data (or results) within the Views

To look through the data of the selected arrays, do any of the following. The views are synchronized; if you select a location or region in one view, the other views move there as well.

To do this	Follow these steps
Select a specific chromosome to display	• In Genome View, click the desired chromosome. All other views switch to the selected chromosome.
Display data in a region of the selected chromosome	• In Chromosome View, drag the pointer over the desired region. Gene View expands (or shrinks) to show only the selected region. Tab View scrolls to the new cursor location.
Zoom in and out in Gene View	• Click  to zoom in. • Click  to zoom out.
Scroll through the selected chromosome	• Click  to scroll up. • Click  to scroll down. **Note:** These arrows will appear side by side for horizontal orientation.
Recenter Gene View or Chromosome view	• Click anywhere in Chromosome View, or anywhere within the scatter plot in Gene View. The location you click becomes the new cursor location.
Move all views to a specific genomic location	**a** Click **Home > Go To Gene/Genomic location.** A dialog box appears. **b** Under Genomic Location, select a Chromosome, and type a **Base Position.** **c** Click **Go.** All views move to the selected location.
Display the location of a specific gene in the center of all Views	**a** Click **Home > Go To Gene/Genomic location.** A dialog box appears. **b** Under RefSeq by Symbol, either select the desired gene (if available) or type the name of the gene. **c** Click **Go.** All views move to the location of the selected gene.

To do this	Follow these steps
Display the data selected in Tab View in the center of Chromosome and Gene Views	• In Tab View, click any entry in any table, except a column heading. The genetic location of the selected data appears in the center of the Gene View, and the cursor moves to the position of the selected data in Chromosome View.
Scroll to a specific column in Tab View (for the selected chromosome)	**a** In Tab View, right-click any column heading, then click **Scroll To Column.** The Scroll to Column dialog box appears. See "Scroll to Column" on page 391. **b** In Select Column, select the desired column. **c** Click **OK.**
Search for a specific column entry in Tab View, and move the cursor there	**a** In Tab View, right-click any entry except a column heading, then click **Find in column.** The Find in column dialog box appears. See "Find in column" on page 325. **Note:** The Find in column function works within the chromosome that is selected in the Genome View. **b** Set the desired search parameters, then click **Find Next.** The program searches the column within the selected chromosome, using your search parameters, and highlights the row of the first entry that matches. The cursor moves to the location selected in the highlighted row.
Display the exact chromosomal location of the cursor	At the bottom of the main window, look at the first cell of the Status bar. The location appears as the chromosome followed by the base position. For more information on the status bar, See "Status Bar" on page 238.

To create a Signal Intensity Bar Chart

You can create a histogram that displays the median signal intensities in a genomic region of selected Agilent arrays.

1 In the Genomic Viewer, right-click the Gene View and select **Show Intensity BarCharts**.

The Create Signal Bar Chart dialog box appears, where you can select what region to include in the bar chart. See "Create Signal Bar Chart" on page 283 for more information.

NOTE Certain options are available only if an aberration calculation is selected in the Analysis tab.

2 Click **OK** to create the bar chart.

The bar chart appears.

3 To copy or save the image in a file, right-click in the chart and select from the menu options. See "Show Intensity Bar Charts" on page 404 for more information.

To smooth and plot CGH log ratio data

You use a plug-in to create separate, stacked plots of smoothed log ratio data for each of the selected CGH arrays in the selected experiment. It can perform one of several varieties of moving-average-like smoothing, and plots the data associated with the selected chromosome.

The Plugin Settings command lets you change the parameters when you have selected to produce the plot immediately after you click Plugin.

1 Select an experiment.

2 Select the arrays whose log ratio data you want to smooth and plot.

3 Select the chromosome whose log ratio data you want plotted.

4 Click **Tool > Plugin > CGHSmooth**.

The CGHSmooth Parameters dialog box appears. See "CGHSmooth Parameters" on page 251.

5 Type the parameter values for the selected arrays in the active experiment and the selected chromosome.

6 Click **OK**.

The CGHSmooth Plot appears, showing a plot of the log ratio vs. chromosomal position for each of the selected arrays. See "CGHSmooth Plot" on page 253.

7 (optional) To add a title, legend and change the appearance of the plot(s), right-click the plot, and click **Properties**.

See "Chart Properties" on page 255. You can also right-click the plot to save, print, zoom in or out, or change the auto range.

To enter parameters from the Plugin Settings command and display the plot directly from the Plugin command

1 After step 5 above, mark **Don't Show Again**, then click **OK**.

The plot appears.

2 To change the parameter values, exit the plot, and click **Plugin Settings**.

3 Change the values you want to change, and click **OK**.

4 Click **Plugin**.

The plot appears.

To show the CGHSmooth Parameters dialog box again when you click Plugin, click **Plugin Settings**, clear **Don't Show Again**, and click **OK**.

To produce an echo example plot

The Echo Example Plot is the output of the Echo Example plug-in. It displays the log ratio data for the selected chromosome in the active experiment. Data from all of the selected arrays in the experiment appear as a series of stacked plots, one for each array.

1 Select an experiment.

2 Select the arrays that contain the log ratio data to smooth and plot.

3 Select the chromosome that contains the log ratio data of interest.

4 Click **Tool > Plugin > EchoExample**.

5 Type the parameter values for the selected arrays in the active experiment and the selected chromosome.

6 Click **OK**.

The Echo Example Plot appears, showing a plot of the log ratio vs. chromosomal position for each of the selected arrays. See "Echo Example Plot" on page 305.

7 (optional) To add a title or legend and change the appearance of the plot(s), right-click the plot, and click **Properties**.

See "Chart Properties" on page 255.

8 (optional) To see options to save, print, zoom in or out, or change the auto range of the plot, right-click the plot.

To produce a MovAvg Example plot

The MovAvgExample plug-in computes a 10-point moving average of the data from the selected chromosome in the selected arrays of the active experiment. It displays stacked plots of moving averages for all the arrays, one plot per array.

The plug-in program itself (**MovAvg Example.pl**, located in the Plugins folder of the Agilent Genomic Workbench installation folder on your computer) is a short Perl program. It is a good example of how computed columns are processed. You must have Perl installed on your computer to use this plug-in.

1 Select an experiment.

2 Select the arrays that contain the log ratio data to smooth and plot.

3 Select the chromosome that contains the log ratio data of interest.

4 Click **Tool > Plugin > MovAvg Example**.

5 Type the parameter values for the selected arrays in the active experiment and the selected chromosome.

6 Click **OK**.

The MovAvg Example Plot appears, showing a plot of the log ratio vs. chromosomal position for each of the selected arrays. See "MovAvg Example Plot" on page 373.

7 (optional) To add a title and legend and change the appearance of the plot(s), right-click the plot, and click **Properties**.

See "Chart Properties" on page 255.

8 (optional) To see options to save, print, zoom in or out, or change the auto range of the plot, right-click the plot.

To enter parameters from the Plugin Settings command and display the plot directly from the Plugin command

1 After step 5 above, mark **Don't Show Again**, then click **OK**.

The plot appears.

2 To change the parameter values, exit the plot, and click **Plugin Settings**.

3 Change the values you want to change, and click **OK**.

4 Click **Plugin**.

The plot appears.

To show the MovAvg Example Parameters dialog box again when you click **Plugin**, click **Plugin Settings**, clear **Don't Show Again**, and click **OK**.

Displaying Gene Lists and Tracks

To show gene lists in Gene View

A gene list is a set of genes of interest. Within the program, you can highlight the genes in the gene list in Gene View, or limit the display of data, genes, and tracks to the regions selected by a gene list.

You can import gene lists into Agilent Genomic Workbench, and you can also create them in the program and export them. See "To import a gene list" on page 75, and "To export a gene list" on page 84.

In Gene View, the names of all genes normally appear in gray. When you apply a gene list, the program highlights the listed genes in their defined display color. You can limit the genes and/or data that appear in Gene View and Chromosome View to only the listed genes.

1 In the My Entity List of the Navigator, expand the Gene List folder. If the gene list does not appear, create or import it. See "To create a gene list" on page 74, or "To import a gene list" on page 75.

2 Right-click the gene list, then do one of the following to apply it:

- To show all genes and all data, and highlight the listed genes in their display color, click **Highlight.**
- To show only the listed genes and only the data for those genes, click **Show only.**

Gene and Chromosome views change accordingly. In the My Entity List of the Navigator, the name of the gene list appears in italics.

To remove the effects of a gene list, right-click the active gene list in the Navigator, then click **Show All.**

To select gene list display color

In Gene View, the names of all genes normally appear in gray. When you apply a gene list, the program highlights the listed genes in their defined display color. You can customize this color.

1 In the My Entity List of the Navigator, expand the Gene List folder.

2 Right-click the name of the gene list whose color you want to change, then click **View in Table.**

The Gene List dialog box appears.

3 Under Color, click **Color.**

A dialog box appears.

4 Select the desired color.

The dialog box offers three different ways to select the desired color. See "Select Color" on page 394.

5 Adjust the color as desired, then click **OK.**

6 In the Gene List dialog box, click **OK.**

To display a gene list as a table

You can display the description of a gene list and the names of the genes in it.

1 In the My Entity List of the Navigator, in the Gene List folder, right-click the desired gene list, then click **View in Table.**

The Gene List dialog box appears. See "Gene List" on page 328. The names of the genes appear in Gene Names. You can also use this dialog box to edit the description of the gene list, or to change its display color. To change the display color, see "To select gene list display color" on page 110.

2 When you are finished displaying the list, click **OK.**

You can also export a gene list. See "To export a gene list" on page 84.

To change the appearance of genes in Gene View

You use the User Preferences dialog box to change the appearance of the genes in Chromosome and Gene views.

1 Right-click any part of the Gene View, then click **User Preferences.**

The User Preferences dialog box appears.

2 Click **Tracks**.

See "Tracks tab" on page 415.

3 Do any of the following:

To do this	Follow these steps
Show or hide genes in Gene View	**a** Under Visualization Parameters: To show genes – Under Genes, mark **Show Gene Symbols.** To hide genes – Under Genes, clear **Show Gene Symbols.** **b** Click **Apply.**
Change the display font for genes (and track annotations) in Gene View	**a** Under Font, select a new **Font, Font Style,** and **Font Size.** **b** Click **Apply**
Change the display angle for genes (and track annotations) in Gene View	**a** Under Visualization Parameters, under Genes, in Orientation (Degrees), type a new orientation, in degrees. 0° is horizontal. **b** Click **Apply.**

4 Click **OK.**

To show tracks in Gene View

Tracks contain information for specific genomic locations. A multitude of tracks from diverse sources is available for many species. You can display tracks next to genes and microarray data in Gene View.

1 Select and show microarray data. See "To select an experiment" on page 90.

2 In the My Entity List pane, open the Tracks folder.

3 Right-click the track you want to display, and click **Show In UI**.

Or, you can do this:

1 In Gene View, right-click anywhere within the scatter plot, then click **User Preferences.**

The User Preferences dialog box appears. See "User Preferences" on page 414.

2 Click **Tracks**.

3 Mark the **Show In UI** check box of each desired track.

4 Click **OK.**

The program displays the selected tracks in Gene View.

To change the appearance of tracks

Within the Tracks tab of the User Preferences dialog box, you can change the appearance of tracks, as described in the table below.

To do this	Follow these steps
Include track information in reports	**a** In the list of tracks, in the Show in Report column, mark the check boxes of the desired tracks. **b** Click **Apply.** Doing this adds a column with the hits from the track file. For each aberrant interval, it reports the entries from the track file for that interval in that separate column.
Show or hide annotations in all tracks	• To show annotations in all tracks: under Tracks, mark **Show Annotations.** • To hide annotations in all tracks: under Tracks, clear **Show Annotations.**
Display all selected tracks as a single track	• Under Tracks, mark **Show Overlaid.** The program combines the annotations of all selected tracks into a single track named Overlaid Track. • To show tracks individually again, clear **Show Overlaid.**
Display the parameters and the list of annotations of a track	• In the list of tracks, for the desired track, click **Details.**

To do this	Follow these steps
Change the display font for track annotations (and genes)	**a** Under Font, select a new Font, Font Style, and Font Size for track annotations. **b** Click **Apply.** The program changes the display font of track annotations and genes in Gene View.
Change the order in which tracks appear in Gene View.	The order of tracks in the Gene Symbols tab controls the left-to-right order of tracks in Gene View. **a** Click the name of the track you want to move. **b** Do any of the following: • To move the track up in the list of tracks (and farther left in Gene View), click its name, then click **Up.** • To move the track down in the list of tracks (and farther right in Gene View), click its name, then click **Down.** **c** Click **Apply.**
Change the display angle of track annotations (and genes)	• Under Genes, in Orientation, type a new orientation (in degrees). 0° is horizontal. The program changes the display angle of track annotations and genes in Gene View.

To show track information in reports

1 In the list of tracks, in the Show in Report column, mark the check boxes of the desired tracks.

2 Click **Apply.**

Doing this adds a column with the hits from the track file. For each aberrant interval, it reports the entries from the track file for that interval in that separate column.

To limit aberration calls to the genomic boundaries of the track

1 In the list of tracks in My Entity List, right-click the track whose boundaries you want to use to limit the display of the aberration calls.

2 Mark **Genomic Boundaries**.

You can remove the boundaries by clearing the check box.

To display tracks in UCSC Browser

1 Right-click Gene View, and click **Show in UCSC**.

The View coordinates in UCSC browser dialog box appears. See "View coordinates in UCSC browser" on page 420.

2 Complete the dialog box with the track parameters, and click **OK**.

The UCSC Browser appears, if you are connected to the Internet. You may need to enable pop-ups or set other preferences on the UCSC Web site; see the browser help for information.

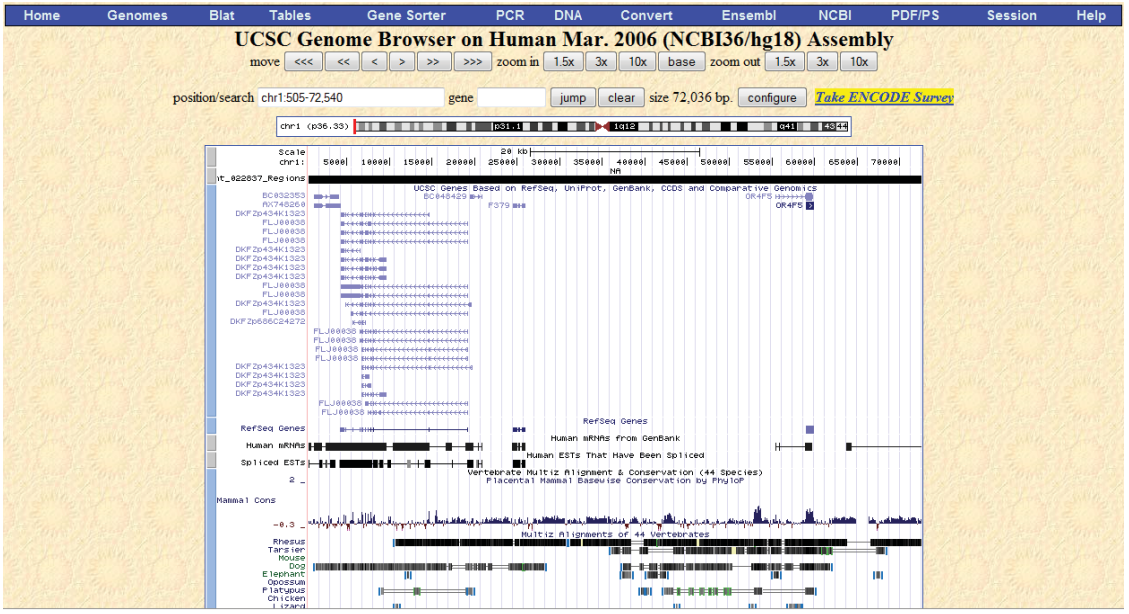


Figure 20 Track displayed in UCSC browser

3 Follow the instructions on the Web site for what you want to do.

To change the graphical display to a different genome build

The default graphical display for Genome, Chromosome and Gene Views represents human genome build 18.

- To change the graphical display to a different genome build, select an experiment whose data are based on a design file of a different genome build.

The display automatically changes when you select an experiment that contains a design file with a different genome build, such as human genome build 17, or a mouse or rat genome build.

If a genome build is not available for the design file you import, you must import the genome build first. See "To import a genome build" on page 50.

The program will not let you add arrays that belong to one genome build to an experiment that contains arrays of a different genome build.

See also "To create a new experiment" on page 60, "To add arrays to an experiment" on page 62, and "To select an experiment" on page 90.

Displaying CGH+SNP Content and Results

To show SNP data panel

SNP copy numbers are displayed graphically in the SNP data panel of the Chromosome and Gene views. See Figure 227 on page 553. You must turn on the SNP data panel in order to see the SNP data.

NOTE In order for copy number data to appear in the panel, the required SNP algorithms and calculations must be selected in the Preprocessing and Analysis tabs.

1. In the Chromosome or Gene view, right-click and then select **View Preferences**.

   The View Preferences dialog box appears. See "View Preferences" on page 422.
2. In the View Preferences dialog box, under Configure Coloring schemes, mark the box next to **Show SNP data Panel**.
3. Under Rendering Patterns, click the arrow next to Design type, and select **CGH+SNP.**
4. Under Data Visibility, click the arrow next to View, and select **Gene View** or **Chromosome View**. The copy number data panel is automatically turned on for both views.
5. Under Data Visibility, mark **Scatter Plot** and **SNP Copy Number**.
6. Click **OK**.

To show LOH regions in genomic viewer

Regions that show a loss or lack of heterozygosity (LOH) are displayed as colored bars. (See Figure 225 on page 551.)

1. In the Chromosome or Gene view, right-click and then select **View Preferences**.

   The View Preferences dialog box appears. See "View Preferences" on page 422.

2. In the View Preferences dialog box, under Configure Coloring schemes, mark the box next to **Show SNP Data Panel**.
3. Under Rendering Patterns, click the arrow next to Design type, and select **CGH+SNP.**
4. Under Data Visibility, click the arrow next to View, and select **Gene View** or **Chromosome View**, or **All Views**. The copy number data panel is automatically turned on for Gene and Chromosome views.
5. Under Data Visibility, mark **Scatter Plot** and **LOH**.
6. Click **OK**.

Searching for Probe and Gene Information

To search Tab View for specific probe information

You can find a specific entry in a column of a data table in Tab View. For more information on Tab View, see "Tab View" on page 233.

1 In Tab View, right-click anywhere in the column you want to search, then click **Find in column.** See "Find in column" on page 325.

The Find in column dialog box appears. The column to be searched also appears in the title bar of the dialog box.

NOTE The Find in column function works within the selected chromosome.

2 Set the search parameters, as described below.

Parameter	Comments/Instructions
Find in column	• Type the text you want to find (the *search term*). This can be an entire entry, or part of one.
Direction	• Select one of these options: ◦ Up – Search the column upwards from the current cursor location (the highlighted row of the table). ◦ Down – Search the column downwards from the current cursor location (the highlighted row of the table). Tip: Click a row in Tab View to highlight it.
Conditions	• Mark any of these, as desired: ◦ Match Case – Return entries that match upper and lower case characters in the search term. ◦ Match whole word – Return an entry only if the entire entry matches the search term.

3 Click **Find Next.**

If the program finds a match, it highlights the row that contains the matching entry, and resets the cursor to the corresponding position. You can click **Find Next** as many times as you like, and the program

continues to search for additional matching entries in the column for the selected chromosome. If it finds no match, the message: **String not found** appears in black in the lower part of the dialog box.

4 When you finish your search, click **Cancel** to close the dialog box.

To search Agilent eArray for probe information

You can use the chromosomal region that appears in Gene View, or another chromosomal region as the basis for a probe search on the Agilent eArray Web site. eArray is a powerful microarray design system for CGH, ChIP and gene expression applications. It contains a massive database of validated, annotated probes, and a full complement of tools for custom microarray design.

Before you can search for probes in eArray, you must be a registered eArray user. For more information, go to eArray.chem.agilent.com. You must also provide your eArray user name and password in the Miscellaneous tab of the User Preferences dialog box. See "User Preferences" on page 414.

1 In Gene View, right-click anywhere in the plotting area, then click **Search probes in eArray**.

The Search probes in eArray dialog box appears. See "Search Probes in eArray" on page 392.

2 Do one of the following to define the chromosomal region for your search:

- To set the region to the one that currently appears in Gene View, select **For complete gene view.**
- To set the region numerically, select **User Defined,** then select a **Chromosome** and type **Start** and **Stop** locations for the desired region.

3 Click **OK.**

The eArray Web portal opens in your internet browser.

To search the Web for information on probes in Tab View

You can use any entry in a table in Tab View as the basis for a Web search.

1 In Tab View, right-click any data table entry other than a column heading.

2 Click one of the available sites.

If the site of interest does not appear in the shortcut menu, you can create a custom search link. See "To create a custom Web search link" below.

The selected site opens in your Internet browser. The program sends the table entry to the site as a search string.

To create a custom Web search link

If you need to search a different database or site based on data table entries, you can create your own custom search link. When you right-click a table entry in Tab View, a shortcut menu opens, and your custom link appears in it. If you select this link, Agilent Genomic Workbench opens the site in your Web browser and sends the table entry to the site as a search string.

1 Right-click any data table entry in Tab View, except a column heading, then click **Customize Link.**

The Customize Search link dialog box appears. See "Customize Search Link" on page 286.

2 Click **New.**

3 In the Input dialog box, in **URL name,** type a name for the link.

This name will appear in the shortcut menu that opens when you right-click a data table entry.

4 Click **OK.**

5 In URL, type the complete URL needed to send a search string to the site. Use `<target>` as the query string value.

For example, this URL sends selected table entries to Google.com:
`http://www.google.com/search?hl=eng&q=<target>`

6 Click **Update**, then click **Yes.**

To update or delete a custom Web search link

1 Right-click any data table entry in Tab View other than a column heading, then click **Customize Link.**

The Customize Search link dialog box appears.

2 In URL Name, select the custom search link to update or delete.

3 Do one of the following:

To do this	Follow these steps
Update a Web search link	**a** Edit the **URL name** and the **URL** as needed. **b** Click **Update.** A Confirm dialog box appears. **c** Click **Yes.**
Delete a Web search link	• Click **Delete.**

4 Click **Close.**

4 Setting Up CGH Interactive Analysis

Working with Interactive Analysis Options 124
Changing Preprocessing Options 126
Changing Analysis Options 136
Displaying Results Before Selecting Discovery Options 142
Displaying Results with Discovery Options 150
Creating Reports 166

This chapter gives instructions on how to set up the interactive analysis functions for CGH experiments. These include the Preprocessing, Analysis, Discovery and Reports tabs. To prepare the CGH or CGH+SNP data to be analyzed, see Chapter 2, “Importing, Managing, and Exporting CGH and CGH+SNP Data and Other Content” and Chapter 3, “Displaying CGH and CGH+SNP Data and Other Content” of this guide.

Working with Interactive Analysis Options

For a detailed description of the interactive analysis tabs – Preprocessing, Analysis, Discovery, Reports – and their commands, see "Command Ribbons" on page 176.



Figure 21 Interactive Analysis tabs

Table 3 identifies the sections in this chapter where tasks for the interactive analysis options are described.

Table 3 Interactive analysis topics

Subject	See these topics
"Changing Preprocessing Options"	"To create or modify an array filter" on page 126 "To create or modify a feature filter" on page 127 "To create or modify a design filter" on page 127 "To apply a filter" on page 128 "To apply GC Correction" on page 129 "To apply centralization" on page 130 "To combine (fuse) arrays" on page 130 "To combine intra-array replicates" on page 132 "To combine interarray replicates" on page 133
"Changing Analysis Options"	"To set up a moving average calculation to smooth the data" on page 136 "To apply the aberration algorithm" on page 137 "To apply fuzzy zero correction" on page 138 "To apply the nesting filter" on page 139 "To calculate SNP copy number" on page 139 "To manually reassign peaks" on page 140 "To calculate LOH" on page 141

Table 3 Interactive analysis topics (continued)

Subject	See these topics
"Displaying Results Before Selecting Discovery Options"	"To show results of analysis" on page 142 "To move through View displays to observe results" on page 143 "To save a result" on page 145 "To restore a saved result" on page 146 "To print moving average and aberration calculation results" on page 146 "To customize the appearance of analysis results" on page 146
"Displaying Results with Discovery Options"	"To create or modify an aberration filter" on page 150 "To create common aberration summaries" on page 151 "To display common aberrations" on page 152 "To display graphical probe penetrance summaries" on page 154 "To create interval penetrance summaries" on page 154 "To visualize a CNVR (copy number variant region)" on page 156 "To set up and perform a joint analysis on CGH and expression data" on page 156 "To create, modify, or apply a differential aberration filter" on page 160 "To set up a clustering analysis" on page 161 "To display results as a heatmap" on page 162 "To generate a genotype reference file" on page 164
"Creating Reports"	"To create and export common aberration text reports" on page 166 "To create CGH aberration text reports" on page 166 "To generate graphical CGH aberration reports" on page 167 "To create a new Cyto Report template" on page 169 "To open an existing Cyto Report template" on page 170 "To generate a cytogenetic report" on page 170 "To generate a SNP Genotype Report" on page 167 "To generate a SNP Aberration & LOH Report" on page 168

Changing Preprocessing Options

The preprocessing options let you manipulate the data in the active experiment before you apply (or reapply) moving average or aberration detection algorithms. You can also filter, centralize, and combine data. In some cases, you must first select the experiment, and in others, this is not required.

To create or modify an array filter

An array filter includes or excludes an array from the active experiment based on the compliance of the array with the conditions of the filter. You can create a new array filter, or modify an existing one.

1 In the Preprocessing tab, under Filters, click **Array Filters > Edit Filter.**

The Array Level Filters dialog box appears. See "Array Level Filters" on page 247.

2 Do one of the following:

- To create a new filter, click **New.** In the dialog box that appears, type a name for your new filter, then click **OK.**
- To edit an existing filter, select its name.

The name of the filter and its conditions appear in the dialog box.

3 Create or change the conditions of the filter, as desired.

To restore changed values to their original values, click **Reset.**

4 Click **Update** to save the filter.

5 Click **Close.**

NOTE To restore the previous values after you update them, you must retype the values manually.

To create or modify a feature filter

A feature filter includes or excludes feature data from the active experiment based on conditions specified in the filter. The conditions of feature filters are based on feature attributes that are derived from the FE files. For example, a feature filter can be used to exclude saturated or nonuniform features. You can create a new feature filter, or modify an existing one.

1 In the Preprocessing tab, under Filters, click **Feature Filters > Edit Filter.**

   The Feature Level Filters dialog box appears. See "Feature Level Filters" on page 323.

2 Do one of the following:

   - To create a new filter, click **New**. In the dialog box that appears, type a name for your new filter, then click **OK.**
   - To edit an existing filter, select its name.

     The name of the filter and its conditions appear in the dialog box.

3 Change the conditions of the filter, as desired.

   To restore changed values to their original values, click **Reset.**

4 Click **Update** to save the filter.

5 Click **Close.**

NOTE To restore the previous values after you update them, you must retype the values manually.

To create or modify a design filter

A design filter includes or excludes data from the selected experiment based on conditions specified in the filter. The conditions of design filters are based on attributes of design files. For example, a design filter can exclude nonunique features that have multiple perfect matches in the

genome (NumPMHits), fail the homology filter, are in PseudoAutosomal regions of the genome, or have a low probe score. You can create a new design filter, or modify an existing one.

1 In the Preprocessing tab, under Filters, click **Design Filters> Edit Filter.**

The Design Level Filter dialog box appears. See "Design Level Filters" on page 297.

2 Do one of the following:

- To create a new filter, click **New**. In the dialog box that appears, type a name for your new filter, then click **OK.**
- To edit an existing filter, select its name.

  The name of the filter and its conditions appear in the dialog box.

3 Change the conditions of the filter.

To restore changed values to their original values, click **Reset.**

4 Click **Update** to save the filter.

5 Click **Close.**

To apply a filter

You can apply one of each type of filter. The program applies the filter(s) to the selected experiment, if any. Otherwise, the program applies the filter(s) to experiments when you select them.

1 In the Preprocessing tab, under Filters, do either of the following:

- To apply an array filter, click **Array Filters > Apply.**
- To apply a feature filter, click **Feature Filters > Apply.**
- To apply a design filter, click **Design Filters > Apply.**

2 Mark the check box next to the desired filter.

To remove a filter

You can "turn off" (remove) a filter. When you remove a filter, the program turns off the filter for the selected experiment, as well as for any other experiments that you select.

1 In the Preprocessing tab, under Filters, do either of the following:

- To remove an array filter, click **Array Filters > Apply.**
- To remove a feature filter, click **Feature Filters > Apply.**
- To remove a design filter, click **Design Filters > Apply.**

2 Mark **No Filter.**

To apply GC Correction

The GC Correction algorithm corrects for "wavy" artifacts that are associated with the GC content of genomic regions. These artifacts can interfere with detection algorithms and lead to inaccurate aberration, copy number, and LOH calls.

1 Click the **Preprocessing** tab.

2 Under Normalization, in GC Correction, mark **Apply.**

NOTE GC Correction is recommended for SNP Copy Number and LOH analyses.

3 Click the arrow under Window Size, and select the size of the window to be used for the calculation.

NOTE You can quickly check the status of GC Correction. If a bold, black **G** appears in the Status Bar at the bottom of the window, GC Correction is turned on. The G is gray if GC Correction is not turned on. See "Status Bar" on page 238.

To display a plot of GC corrected data

1 Click the **Preprocessing** tab.

2 Under Normalization, in GC Correction, make sure Apply is selected.

3 Click **Plot Distribution**.

A plot that shows the normalized and unnormalized log ratios for the GC correction is displayed. See Figure 191 on page 457.

To apply centralization

The centralization algorithm adds or subtracts a constant value from each log ratio measurement. This recenters the log ratio values, and make sure that the zero-point reflects the most-common-ploidy state. If centralization is turned on, the program applies it to all array data, even if the experiment has not been selected.

1 Click the **Preprocessing** tab.

2 Under Centralization, mark **Apply.**

3 Change the Threshold and Bin Size parameters, if necessary.

For most cases, do not change the Threshold or Bin Size parameters. For information about these parameters, see "Centralization Algorithm" on page 452.

NOTE

You can quickly check the status of centralization. If a bold, black **C** appears in the Status Bar at the bottom of the window, centralization is turned on. The C is gray if centralization is not turned on. See "Status Bar" on page 238.

To combine (fuse) arrays

If you have two arrays that use different design files, you can combine (fuse) them into one larger virtual array. You do this to increase the coverage of the genome in your design. If for example, you have a catalog array, you can design another array to add probes between the catalog probes to increase the density of coverage. With the Fuse function, you

combine the array data to see all the probe data in the display at once. The program cannot combine arrays from more than two different design files.

Requirements for fusing arrays:

- Each array must be associated with a different design file.
- All of the arrays must be of the same analysis type (for example, CGH).
- None of the arrays can be fused arrays.
- The samples you hybridize to the arrays must all be aliquots from the same preparation.
- (Preferred) Hybridization and labeling occur for all samples together under the same conditions.

1 Create and select a new experiment. See "To create a new experiment" on page 60 and "To select an experiment" on page 90.

2 Add the arrays to be combined to the experiment. See "To add arrays to an experiment" on page 62.

3 Assign the same value to the ArraySet attribute of every array you want to combine. Follow these steps for each array:

   a In the Experiment pane of the Navigator, right-click the name of the array, then click **Show Properties.**
   The Microarray Properties dialog box appears.

   b Next to the ArraySet attribute, under **Value,** click  .
   The field becomes active.

   c Type a value in the text box. Type the same value for the ArraySet attribute of every array.

   d Click **Close.**

4 In the Preprocessing tab, under Combine, click **Fuse.**

   The Array Set dialog box lists the arrays to be fused.

NOTE Double-check the values in the ArraySet Attribute column of the dialog box. Agilent Genomic Workbench combines all of the array pairs that have the same value for this attribute.

5 Set any of these options, as desired:

- **Select Normalization** – Select **None** or **Centralization.** Centralization adds or subtracts a constant value from each log ratio measurement. This re-centers the log ratio values, and makes sure that the zero-point reflects the most common ploidy state.
- **Remove arrays from experiment after fuse** – To delete the initial individual arrays from the experiment, mark this option so that the experiment will not contain duplicate data.

6 Click **Continue.**

The program combines the arrays. The new array appears in the Experiment pane of the Navigator in a new design folder within the active experiment. The folder name contains the names of both designs.

NOTE

- If the arrays you combine have probes in common, these probes appear as replicates in the fused array. You can combine these replicates. See "To combine intra-array replicates" on page 132.
- If the original arrays you combined change, you can manually update the combined array. Fuse the same set of arrays again.
- If you want to combine many arrays, consider importing an array attributes file. See "To import array attributes" on page 56.

To combine intra-array replicates

Intra-array replicates are features within the same array that contain the same probe. The probe information is contained in the design file for each type of array. You can combine these replicates to increase the statistical power of your analyses. For a discussion of the statistical model Agilent Genomic Workbench uses to combine replicates, see "Error Model and Combining Replicates" on page 462.

To combine intra-array replicates for the selected arrays of the active experiment:

1 Click **Preprocessing.**

2 Under Combine, in Replicates, mark **Intra Array.**

Agilent Genomic Workbench combines the replicates in the selected arrays within the experiment.

If you set up this option before activating an experiment, it will be applied when you select the experiment and select the arrays.

To restore the replicates to their uncombined state, clear the check box.

To combine interarray replicates

When you combine interarray replicates, you select an array attribute. Agilent Genomic Workbench combines replicates from the arrays with the same value for the selected attribute. For a discussion of the statistical model Agilent Genomic Workbench uses to combine replicates, see "Error Model and Combining Replicates" on page 462.

To combine inter array replicates for all arrays of the active experiment:

1. Click **Preprocessing.**
2. Under Combine, in Replicates, mark the check box for **Inter Array.**
3. In **Group By,** select the desired array attribute.
4. Click **Go.**

   Agilent Genomic Workbench combines replicates from the arrays with the same value for the selected attribute in an active experiment.

   If you set this option before you select an experiment, the program applies it when you select the experiment and select the arrays.

   To restore the replicates to their uncombined state, clear **Inter Array,** then click **Go.**

Microarray experiments can include biological or technical replicate arrays, which are often combined to create one virtual array for downstream analysis. Biological or technical replicates of the physical arrays may be used to mitigate systematic variation arising from sample preparation or array processing.

Because the replicates are hybridized to microarrays with the same design, inter-array replicates are simply features on different arrays that contain the same probe. The combined signal values give an increase in statistical power for the measurement of log ratios.

To display QC metrics of arrays and set array QC status

QC Metrics are indicators of the quality of data from an array. QC Metrics, which are available only for Agilent arrays, are either passed through from the FE program, or are calculated by the CGH module itself. You can view QC Metrics for an individual array, the arrays in an experiment, or the arrays associated with a specific design and genome build.

1 In the Preprocessing tab, do one of the following:

- To display the QC metrics of Agilent arrays in the active experiment – In the ribbon, click **QC Metric.**
- To display the QC metrics of a specific Agilent array – In the **Data** or **Experiment** panes of the Navigator, right-click the name of the array. In the shortcut menu that appears, click **QC Metrics.**
- To display the QC metrics of all Agilent arrays associated with a genome build within a specific design folder – In the **Data** or **Experiment** panes of the Navigator, right-click the name of the genome build folder within the desired design folder. In the shortcut menu that appears, click **QC Metrics.**

In each case, the QC Metrics Table appears. For each array, the QC Metrics Table color codes each metric as Evaluate (pink), Good (turquoise), Excellent (yellow) or NA (white). This dialog box also offers several other options for display of the QC metrics. See "QC Metrics Table" on page 378.

2 In the QCMetricStatus column, select the desired overall status for each array.

This sets the QCMetricStatus attribute of each array. You can use the value of this attribute as the basis for an array filter.

3 Click **Close.**

To display SNP QC Metrics

SNP QC Metrics are indicators of the quality of data from an array that contains SNP content. SNP QC Metrics are viewed on a per-experiment basis.

1 In the Navigator, double-click an experiment that contains CGH+SNP arrays.

2 Right-click on the experiment folder, and then select **SNP QC Metrics**.

The SNP QC Metrics table appears. The table shows calculated metrics for the selected arrays in the experiment. See "SNP CN QC Metrics Table" on page 405.

Changing Analysis Options

Before you analyze or reanalyze the data, you can change these options. Doing the tasks in the section may or may not trigger analysis or reanalysis of the experiment. If you select an experiment before you change the options, the data will be reanalyzed immediately after you make a change. If you make the changes before you select an experiment, the changes will be applied when you select the experiment.

To set up a moving average calculation to smooth the data

With the Moving Average tool, a moving average is computed for each point in the data set using a window size centered on that point. See "Visualization Algorithms" on page 502.

1 In the Analysis tab, under Moving Average, mark or clear **Show** to show or hide the Moving Average Line Plot.

2 In Algorithm, select **Linear** or **Triangular.**

- **Linear** – The linear algorithm calculates a standard, unweighted average using every probe Log Ratio score within a specified window size (specified by either a number of adjacent measurements or a positional range). This average is applied sequentially to every point by moving the window along the chromosome.
- **Triangular –** The triangular algorithm calculates a weighted average using every probe Log Ratio within a specified window size. The triangular algorithm is more sensitive to localized variations in the data.

3 In Line width, select the thickness (in pixels) of the Moving Average Line Plot.

The range is 1-5 pixels.

4 In Window, select the size of the moving average window.

Moving averages can be computed with windows of sizes based on either a specific length of base-pairs (5 Kb to 50 Kb or .1 Mb to 50 Mb) or a fixed number of data points (1 pt. to 60 pt.).

To apply the aberration algorithm

Detection of contiguous aberrant regions is available in Agilent Genomic Workbench in a variety of approaches. See "Aberration" on page 185 and "Aberration Detection Algorithms for CGH Analysis" on page 465 for more information about the aberration detection algorithms.

1 In the Analysis tab, under Aberration, in **Algorithm,** select the desired aberration detection algorithm. Use the following as a guide:

Algorithm	Description/Instructions
Z-score	The *Z*-score algorithm is used to determine enrichment of over- or under-abundance probes in a region. • In Threshold, type the desired threshold value, or use the slider to select a value. • In Moving Average (Log Ratio) select the **Window** size to be used in the calculation.
ADM-1 ADM-2	The aberration detection method algorithms (ADM-1 and ADM-2) look for regions that are statistically different from a log ratio score of 0, or no change, between channels. **a** In Threshold, type the desired threshold value, or use the slider to select a value. A recommended threshold starting point for both the ADM-1 and ADM-2 algorithms is a value of 6. **b** If needed, turn on Fuzzy Zero correction, and apply the nesting filter. See "To apply fuzzy zero correction" on page 138 and "To apply the nesting filter" on page 139.
CBS	The circular binary segmentation, or CBS, algorithm identifies regions with constant copy numbers.
HMM	The Hidden Markov Model, or HMM algorithm partitions a large number of probe scores into a smaller number of discrete states. The sequence of states (gain, loss, or no change) forms the basis to identify aberrant regions. • Under State Parameters, select the number of states, and type the desired value in FDRQ. See "Aberration" on page 185.

2 Under Aberration, mark **Show.**

The program applies the selected aberration detection algorithm to the selected data in the active experiment. An aberration plot appears in Chromosome and Gene views.

NOTE

- Moving Averages and the Z-score algorithm both use the window size specified in the Moving Average as input. See "Moving Average" on page 184.
- The optimal threshold values for the Z-score algorithm, ADM-1, and ADM-2 are different, even for the same array. The meaning of threshold also differs for each algorithm.
- For CGH+SNP arrays, you must select an aberration algorithm in order to calculate SNP Copy Number and LOH. Z-score and HMM aberration algorithms are not allowed for SNP analysis.

To apply fuzzy zero correction

Fuzzy Zero is turned off by default in Agilent Genomic Workbench. See "Fuzzy Zero" on page 458 for more information.

1 Click **Analysis.**

2 Under Fuzzy Zero, mark **Turn On.**

If an experiment is active, this algorithm is immediately applied to the data.

You can quickly check the status of Fuzzy Zero in the **Status Bar**. If the sixth cell displays a bold '**F**', Fuzzy Zero is applied. The 'F' is gray if Fuzzy Zero is not applied. See "Status Bar" on page 238.

The ADM-1 and ADM-2 algorithms may identify extended aberrant segments with average probe log ratios that are low. Often such aberrations represent noise, and are detected because a large number of successive probes in that segment that have correlated errors, but the ADM algorithms treat the errors as independent. If long, low aberrations are detected in your analysis, you can apply the fuzzy zero algorithm to correct for the assumption of independence between probes in long genomic intervals.

To apply the nesting filter

1 Click **Analysis.**

2 In Nesting Level, mark **Apply Nesting Filter** and type an integer value for nested aberrations allowed.

ADM-1 and ADM-2 scores iteratively identify all aberrations that differ significantly from log ratios that show no change between sample channels. In the iterations, the algorithms identify nested, or “child”, aberrations that are contained within other “parent” aberrations but show a significant difference from the parent aberration log ratio. The nesting level is the number of “child” aberrations to be identified.

By default, the nesting level is set to a maximum integer value (2147483647) to find as many discrete aberrations as possible. When set to this maximum value, no filter is applied to the number of possible “child” aberrations.

To filter the number of aberrations identified, you can set this integer value to any desired value between 0 and 2147483647. For example, if this value is set to 0, you will filter out all nested (child) aberrations, retaining only the parent aberrations. If you set it to a value of 1, only the first nested level is retained.

To calculate SNP copy number

For CGH+SNP microarrays that combine CGH and SNP probes on the same microarray, this calculates an expected copy number for uncut alleles for each SNP site on the array. For details on how the algorithm works, see “ASCN (SNP CN) – Allele-specific copy number detection algorithm” on page 534.

1 Click the **Analysis** tab.

2 Under SNP Copy Number, mark **Calculate**.

3 Under SNP CN Conf. Level, type a confidence level to use.

NOTE In order to calculate SNP copy number, you must first select an aberration calculation (ADM-1, ADM-2, or CBS) in the Analysis tab. GC Correction and Centralization are recommended.

To manually reassign peaks

Copy numbers assigned by the software can be displayed graphically and changed, if desired.

1 In the Analysis tab, under SNP Copy Number, mark **Calculate.**

2 Mark **Manually Reassign Peaks**.

The Manually Reassign Peaks dialog box appears. See "Manually Reassign Peaks" on page 364.

3 If the experiment contains multiple microarrays, click the arrow next to **Select Microarray** and select the microarray for which you want to change copy numbers.

4 In the table at the bottom of the dialog box, locate the peak of interest, and double-click the Copy Number you want to change.

The data points in the graphs for that peak turn green.

5 Type the new copy number for the peak.

6 Click **ReCalculate**.

The graphs are updated to display the changes.

7 To accept the new copy number, click **Accept**.

OR

To reset the copy number to the original value and reset the plots, click **Reset**.

NOTE The Manually Reassign Peaks dialog box appears whenever you analyze data, as long as the Manually Reassign Peaks box in the Analysis tab is marked.

To calculate LOH

For SNP analysis, the LOH algorithm identifies copy-neutral genomic regions with a statistically significant scarcity of heterozygous SNP calls, and reports the regions where the LOH score exceeds a user-defined threshold. For more information on how the LOH algorithm works, see "LOH (Loss or lack of heterozygosity) algorithm" on page 544.

1 In the Analysis tab, under LOH, mark **Calculate**.

2 If needed, type a new Threshold for the analysis.

NOTE In order to calculate LOH, you must first select to calculate SNP Copy Number.

Displaying Results Before Selecting Discovery Options

The section shows you how to make sure the aberration analysis results are displayed automatically after you select an experiment.

To show results of analysis

After you set up an experiment and analyze it, the program displays the results automatically in Genome, Chromosome, and Gene Views, if the display is turned on. Figure 22 shows an example of an aberration result displayed in all views. To customize the way the results appear, see "To customize the appearance of analysis results" on page 146.

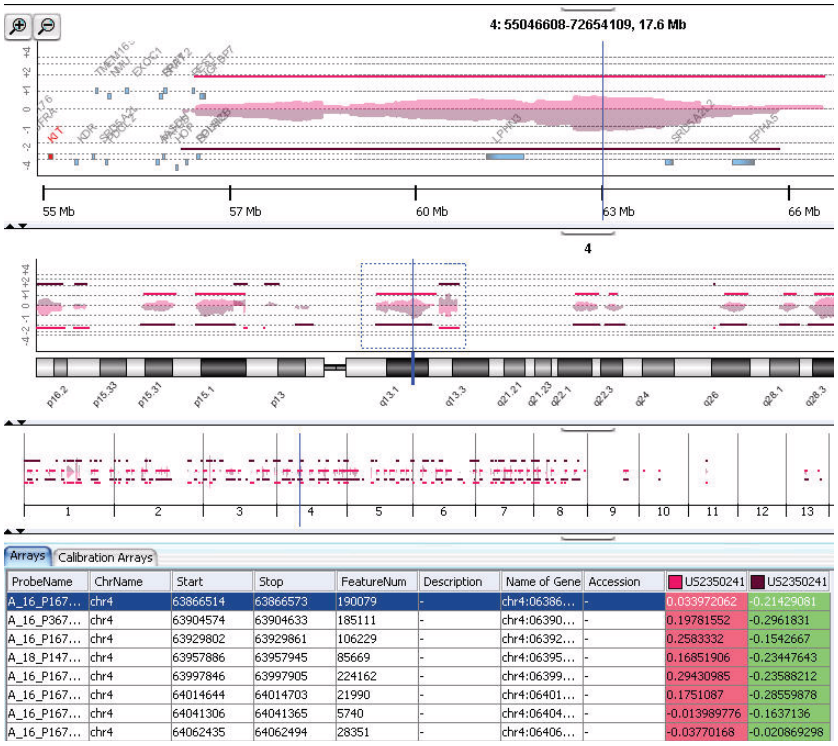


Figure 22 Aberration results of a dye-flip experiment

If the analysis results do not appear in any of the Views

Check these potential problem areas in this order:

- In the Analysis tab, make sure you have marked the Show check box. Mark **Show** for Moving Average or for Aberration or for both.
- Make sure you have selected the experiment or selected the arrays of interest. See "To select an experiment" on page 90.
- Make sure you have turned on the display for the moving average or the aberration algorithm. Do the following:
  1. Right-click any of the Views, and click **View Preferences**.
  2. Under Data Visibility, in View, select **All views,** then mark any of the check boxes for information to display.

     You can also select a single view.
  3. Click **OK.**
- Make sure you are looking in the right region of Chromosome View.

To move through View displays to observe results

- To move through the results of the selected arrays, do any of the following. In general, all views are synchronized; if you select a location or region in one view, the other views move there as well.

To do this	Follow these steps
Select a specific chromosome to display	• In Genome View, click the desired chromosome. All other views switch to the selected chromosome.
View data in a region of the selected chromosome	• In Chromosome View, in the scatter plot, drag the pointer over the desired region. Gene View expands (or shrinks) to show only the selected region. Tab View scrolls to the new cursor location.
Zoom in and out in Gene View	• Click  to zoom in. • Click  to zoom out.
Scroll through the selected chromosome	• Click  to scroll up. • Click  to scroll down.

To do this	Follow these steps
Re-center Gene View or Chromosome view	Click anywhere in Chromosome View, or anywhere within the scatter plot in Gene View. The location you click becomes the new cursor location.
Move all views to a specific genomic location	**a** Click **Home > Go To Gene/Genomic location.** A dialog box appears. **b** Under Genomic Location, select a Chromosome, and type a **Base Position.** **c** Click **Go.** All views move to the selected location.
Center all views on the location of a specific gene	**a** Click **Home > Go To Gene/Genomic location.** A dialog box appears. **b** Under RefSeq by Symbol, either select the desired gene (if available) or type the name of the gene. **c** Click **Go.** All views move to the location of the selected gene.
Center Chromosome and Gene views based on data in Tab View	• In Tab View, click any entry in any table, except a column heading. Chromosome and Gene views become centered on the genomic location corresponding to the selected entry.
Scroll to a specific column in Tab View	**a** In Tab View, right-click any column heading, then click **Scroll To Column.** The Scroll to Column dialog box appears. See "Scroll to Column" on page 391. **b** In Select Column, select the desired column. **c** Click **OK.**
Search for a specific column entry in Tab View, and move the cursor there	**a** In Tab View, right-click any entry except a column heading, then click **Find in column.** The Find in column dialog box appears. See "Find in column" on page 325. **b** Set the desired search parameters, then click **Find Next.** The program searches the column based on your search parameters, and highlights the row of the first entry that matches. The cursor moves to the location defined in the highlighted row.
View the exact chromosomal location of the cursor	At the bottom of the main window, look at the first cell of the Status bar. The location appears as the chromosome followed by the base position. For more information on the status bar, see "Status Bar" on page 238.

To save a result

The program lets you save the current result of the active experiment. You can run many different analyses in the same experiment, and save each one. Later, you can restore any of your saved results.

If you are saving a result for the first time for the experiment:

1 In the Experiment pane of the Navigator, right-click the experiment, and click **Save Experiment Result**, or

Click **Home > Save Experiment Result.**

A dialog box asks if you want to save the results of the current experiment.

2 Click **Yes**.

The Save experiment result dialog box appears.

3 Type a name for the result, then click **OK.**

If you have already saved at least one result for the experiment:

1 In the Experiment pane of the Navigator, expand the folders of the current experiment.

The selected result, if any, appears in blue in the Results folder.

2 Click **Home > Save Experiment Result** or

Right-click the experiment, and click **Save Experiment Result**.

A dialog box asks if you want to select one of the following actions:

- To replace the current result with another saved result, click **Overwrite Current Result.**
- To add the current results to the list of experimental results, click **Create New Result.**
- To change views to another result without changing the current result, click **Continue Without Saving.**

To restore a saved result

1 If necessary, select the experiment that contains the result to see. See "To select an experiment" on page 90.

2 In the Experiment pane of the Navigator, expand the folder of the active experiment, then expand its **Results** folder.

3 Right-click the desired result, then click **Restore result.**

The restored result appears in Genome, Chromosome, and Gene Views.

To print moving average and aberration calculation results

You can print the moving average and aberrations as they appear in Genome, Chromosome, and Gene Views. Each view selected in the analysis is printed on a separate page. Chromosomes and genes appear on the printed pages, but tracks do not.

1 Click **Home > Print.**

The Print dialog box appears.

2 Set print options, as desired, then click **OK.**

To customize the appearance of analysis results

You use the View Preferences dialog box to customize the appearance of the analysis results in all Views.

1 Right-click any part of Chromosome or Gene View, then click **View Preferences.**

The View Preferences dialog box appears. See "View Preferences" on page 422.

2 Do any of the following:

To do this	Follow these steps
Change the orientation of the views	**a** In View Preferences, under Orientation, select **Horizontal** or **Vertical**.

To do this	Follow these steps
Show or hide the analysis results	**a** In View Preferences, under Data Visibility, in View, select **All Views.** **b** Do one of the following: To show any of the analysis results, mark the check boxes next to any available choice. To hide any of the analysis results, clear the check mark next to any available choice. **c** Click **OK.**
Show a separate result plot in Gene and Chromosome Views for each selected array	**a** In **View Preferences, under View Alignment,** under Rendering Style, select **Stacked.** **b** Click **Apply.**
Show one result plot that contains data for selected arrays	**a** In View Preferences, under View Alignment, under Rendering Style, select **Overlaid.** **b** Click **Apply.**
Select a Style for the Scatter Plot	**a** In View Preferences, under Styles, click the arrow next to each data type, and select a style for display of the data points.
Select to display and select a Style for Green or Red Intensity	**a** In View Preferences, under Data Visibility, mark the check box next to Green Intensity or Red Intensity to turn on the display. **b** Under Styles click the arrow and select a style for display of the data points.
Select a Style for the Moving Average.	**a** In View Preferences, under Styles, click the arrow next to Moving Average and select **Continuous, Dashed, Dotted,** or **Do not show area.**
Select a Style for the Aberrations.	**a** In View Preferences, under Styles, in Aberration, select **Semi transparent filled**, **Hatched** or **Do not show area.**
Select a Style for SNP Copy Number	**a** In View Preferences, under Styles, under Design Type, select **CGH+SNP.** **b** Under Styles, click the arrow next to SNP Copy Number and select **Colored filled circle, + sign, x sign, circle, rectangle, or filled rectangle.**
Select a Style for LOH	**a** LOH is displayed as a continuous line in the color of the selected array. No other choices are allowed.

To do this	Follow these steps
Show or hide the analysis results	**a** In View Preferences, under Data Visibility, in View, select **All Views.** **b** Do one of the following: To show any of the analysis results, mark the check boxes next to any available choice. To hide any of the analysis results, clear the check mark next to any available choice. **c** Click **OK.**
Show a separate result plot in Gene and Chromosome Views for each selected array	**a** In **View Preferences, under View Alignment,** under Rendering Style, select **Stacked.** **b** Click **Apply.**
Show one result plot that contains data for selected arrays	**a** In View Preferences, under View Alignment, under Rendering Style, select **Overlaid.** **b** Click **Apply.**
Select a Style for the Scatter Plot	**a** In View Preferences, under Styles, click the arrow next to each data type, and select a style for display of the data points.
Select to display and select a Style for Green or Red Intensity	**a** In View Preferences, under Data Visibility, mark the check box next to Green Intensity or Red Intensity to turn on the display. **b** Under Styles click the arrow and select a style for display of the data points.
Select a Style for the Moving Average.	**a** In View Preferences, under Styles, click the arrow next to Moving Average and select **Continuous, Dashed, Dotted,** or **Do not show area.**
Select a Style for the Aberrations.	**a** In View Preferences, under Styles, in Aberration, select **Semi transparent filled**, **Hatched** or **Do not show area.**
Select a Style for SNP Copy Number	**a** In View Preferences, under Styles, under Design Type, select **CGH+SNP.** **b** Under Styles, click the arrow next to SNP Copy Number and select **Colored filled circle, + sign, x sign, circle, rectangle, or filled rectangle.**
Select a Style for LOH	**a** LOH is displayed as a continuous line in the color of the selected array. No other choices are allowed.

To do this	Follow these steps
Change the size of the scatter plot data points	**a** In View Preferences, under Scatter Plot (Chr View), select a point size to use for display of scatter plot data in Chromosome View.
Create a custom scale for display of plots	**a** Under Configure Scales, mark **Apply** for the plot, and type a scale to use for the selected plot.
To change coloring schemes for scatter plots	**a** Under Configure Coloring schemes, click **Configure Colors and Ranges** and complete the Configure Coloring Ranges and Shades dialog box. See "Configure Coloring Ranges and Shades" on page 264.

3 Click **OK.**

Displaying Results with Discovery Options

This section explains how to create, modify, and apply aberration filters and display results through the Discovery options.

To create or modify an aberration filter

You can create and modify aberration conditions after the analysis for a given experiment. For example, you can filter the aberration results.

1 Click **Discovery.**

The ribbon displays an option to create, edit and apply an aberration filter.

2 Click **Aberration Filters > Edit Filter.**

The Aberration Filters dialog box appears. See "Aberration Filters" on page 240.

3 Do one of the following:

- To create a new filter, click **New**. In the dialog box that appears, type a name for your new filter, then click **OK.**
- To edit an existing filter, select its name.

  The name of the filter and its conditions appear in the dialog box.

4 Change the condition values of the filter, as desired.

To restore changed values to their original values, click **Reset.**

5 Click **Update** to save the filter.

6 Click **Close.**

To apply or remove a filter, see "To apply a filter" on page 128 or "To remove a filter" on page 129.

NOTE To restore the previous values after you update them, you must retype the values manually.

To create common aberration summaries

Within two or more arrays, you can test for statistically significant aberrations that have common genomic intervals. The results are called common aberration summaries. Once created, common aberration summaries are added to the Experiment pane of the Navigator. The resultant common aberration summary node is used as the basis for common aberration report summaries.

NOTE Common aberration analysis can be applied to two or more arrays. Using any of the available aberration algorithms, you must first analyze each of the arrays you will use to create a common aberration summary.

1. In the folder of the active experiment, select the arrays to be used in the common aberration analysis, and make sure an algorithm is turned on. See "To select or deselect arrays in the experiment" on page 91.
2. In the Discovery tab, under Common Aberration, click **Text.**

   The Create Common Aberration dialog box appears. See "Create Common Aberration" on page 276.
3. From the list of available algorithms, select one of the methods for calling common aberrations. See "Common Aberration Analysis" on page 508 for more information on the available common aberration algorithms.
4. If you select Context Corrected Common Aberration as the algorithm, select either **Genome Scope** or **Chromosome Scope** to define the scope of the common aberration analysis.

   Genome Scope is selected by default and includes all chromosomes.
5. Change the input parameters if necessary.
6. Type a name for the new common aberration node to be created in the Navigator.
7. Choose to create either a new experimental result or a common aberration node under an existing result.
8. Depending on which option you chose in step 7, type the name of the new experimental result name or select the existing result from the list.
9. Click **OK.**

   The Navigator updates the node list with Common Aberrations.

To display common aberrations

You can display statistically significant aberrations that are shared in common genomic intervals between two or more samples. You can create differential subsets of the samples for comparative analysis, and you can also filter common aberrations by attribute(s).

1 In the Discovery tab, under Common Aberration, click **Graphical.**

The Graphical Common Aberration Setup dialog box appears. See "Graphical Common Aberration Setup" on page 339.

2 From the Data or Experiment pane of the Navigator, drag the arrays to be used in the common aberration analysis to the top pane of the dialog box.

3 From the list of available algorithms, select the method to be used for common aberration calling. See "Graphical Common Aberration Setup" on page 339 for more information on available common aberration algorithms.

4 If you select the Context Corrected Common Aberration algorithm, click either **Genome Scope** or **Interval Scope** to select the scope of the common aberration analysis.

Genome Scope is selected by default and includes all chromosomes.

5 Change the input parameters, and click **Calculate Aberration.**

The Graphical Aberration Summary dialog box appears. See "Graphical Aberration Summary (for common aberration analysis)" on page 337.

6 To explore the aberrations found in each array, click the < and > buttons to step through the genome one chromosome at a time.

The aberrations are sorted by chromosomal position and displayed as an array aberration color map, where red corresponds to aberration amplifications and green corresponds to deletions. By marking the appropriate check box, you can limit the color map display to Amplification aberrations, Deletion aberrations, or both.

7 To define up to two sets of arrays for comparison, select either **Set 1** or **Set 2** beneath each array aberration color map. To exclude an array, select **Ignore** below it.

8 Click **Common Aberration** to continue.

To create graphical common aberration filters

To focus on aberrations with specific attributes, you can filter graphical common aberration summaries by genomic position, aberration size, number of probes, aberration *p*-value, or by direction (amplification or deletion).

1 From the Graphical Common Aberration Summary pane, click **Create Filter**. The Interval Filter dialog box appears. See "Interval Filter" on page 361.

2 Click **New**. An Input text box appears.

3 Type a name for your new filter, then click **OK**.

4 Complete the Interval Filter dialog box by assigning attributes, operators, values, and logical expressions to match intervals.

5 Click **New Condition** to add another set of criteria.

Application of an interval filter reflects intervals that pass a boolean conditional across all criteria.

6 Select to include or exclude such intervals.

7 Click **Update** to apply the conditions to the named filter.

8 Click **Close** to return to the Graphical Common Aberration Summary pane.

9 Check **Apply Filter** to visualize results.

To generate gene lists for common aberrations

You can generate a list of genes that share a common aberration across samples within a genomic interval.

1 In the Graphical Common Aberration Summary dialog box, choose the genomic interval to display the Common Aberrations for a given chromosome.

2 Click **Create Gene List** to display the Common Aberration Create Gene List dialog box.

3 Type a name and description for the new gene list.

4 If necessary to distinguish from other gene lists, change the color code for the gene list node.

5 Click **OK**, and then **Close** to proceed.

A new gene list node appears in the Gene List folder of the My Entity List pane.

To display graphical probe penetrance summaries

Probe penetrance plots display the percentage of selected arrays that have an aberration at each probe position on the array for the selected chromosomes.

1 In the Discovery tab, under Graphical Penetrance, click **Probe.**

The Graphical Penetrance Summary dialog box appears. See "Graphical Penetrance Summary" on page 349.

2 Select the chromosome or chromosomes whose plots you want to display or export.

3 To display a summary in the Genomic Viewer display, double-click in the penetrance plot.

The cursor moves to the selected location.

4 To put the plots into a report, click **Edit > Copy Summary to Clipboard**. You can then paste the summary information in the Clipboard into a document of your choice.

5 Exit the window.

To create interval penetrance summaries

Interval penetrance plots display the percentage of selected samples that have an identified aberrant region on the array. You can create and export the plots and tab-delimited tables of the interval penetrance summaries.

To produce a penetrance plot:

1 In the Discovery tab, under Graphical Penetrance, click **Interval.**

The Graphical Interval Penetrance Summary dialog box opens. See "Graphical Interval Penetrance Summary" on page 351.

2 In the table, click a chromosome of interest.

The display for the selected chromosome appears. The blue cursor also appears on the graphical display to outline the region of the aberration you clicked in the table.

To create/edit and apply a filter to the plot:

1 Select a Filter from the drop-down list of existing filters.

To create a new filter, click **New Filter.** See "Interval Filter" on page 361.

2 Click **Apply Filter** to apply the selected interval filter to the penetrance plot.

To export the penetrance plots:

1 Adjust the zoom level to focus on aberrations of interest across the samples.

2 Click **Edit > Copy Summary to Clipboard** to place the plots into a report.

To export the penetrance tables:

1 In the Graphical Interval Penetrance Summary, click **Export Table.**

2 Move to the folder where you want to save the probe penetrance summary report and type a file name.

3 Click **Save.**

To create a gene list from the penetrance summary:

1 In the Graphical Interval Penetrance Summary, click **Create Gene List.**

The Create Gene List dialog box opens. See "Create Gene List" on page 279.

2 Type the attributes for the genes to be included in the gene list.

3 Click **OK**, then click **Close.**

To visualize a CNVR (copy number variant region)

A copy number variation (CNV) is a form of structural variation in which a segment of DNA, at least 1 kb in length, is found at different copy numbers in different people. CNVs are a result of deletion or duplication events. The copy number variant region (CNVR) algorithm uses an aberration algorithm output as a list of potential adjacent intervals to be joined together into a putative CNV region.

1 Click **Discovery > CNVR.**

The Create CNVR dialog box appears. See "Create CNVR" on page 274.

2 Type the **CNVR Node Name.**

3 Mark or clear the **Show CNVR(s)** check box.

4 Choose to create either a new experimental result or a CNVR node under an existing result.

5 Depending on which option you chose in step 4, type the name of the new experimental result, or select the existing result from the list.

6 Click **OK**.

The CNVR node appears under the active experiment result or new experiment.

If you selected to show CNVRs, they appear as green regions in the Genomic Viewer.

To set up and perform a joint analysis on CGH and expression data

The CGH module of Agilent Genomic Workbench lets you perform two types of analyses to compare CGH data with expression data: an enrichment analysis and a correlation analysis.

To set up and perform a correlation analysis

Correlation analysis is a process that examines the relationship between gene expression and CGH results for given labeled samples.

1 Click **Discovery > Joint- Analysis > Correlation Analysis > Setup.**

The Correlation Analysis Setup dialog box appears.

2 Select from the options available.

See "Correlation Analysis Setup" on page 270.

3 Click **Perform Analysis.**

This opens the Matched Sample dialog box, where you can view the paired CGH-expression arrays, and start the correlation analysis.

See "Matched Sample" on page 365.

4 Click **Continue** if the pairs are matched.

The Correlation Results dialog box appears. See "Correlation Results" on page 272.

5 Click **OK** to accept the results.

For details on this kind of analysis, see "Algorithms for Joint CGH-Expression Analysis" on page 526.

To set up and perform an enrichment analysis

Enrichment analysis computes an enrichment score for each genomic interval. The intervals can represent the entire genome, or only detected aberrant regions. An enrichment score is a measure of the increase in the number of probes with high values of a selected type as compared to what can be expected at random for a given interval. You can calculate an enrichment score for any sample for which you have both CGH and expression data.

1 Click **Discovery > Joint-Analysis > Enrichment Analysis > Setup.**

The Enrichment Analysis Setup dialog box appears.

2 Select from the options available.

See "Enrichment Analysis Setup" on page 310.

3 Click **Perform Analysis**.

This opens the Matched Sample dialog box, where you can view the paired CGH-expression arrays, and start the enrichment analysis.

See "Matched Sample" on page 365.

4 Click **Continue** if the pairs are matched.

The Correlation Results dialog box appears. See "Correlation Results" on page 272.

5 Click **OK** to accept the results.

For details on this kind of analysis, see "Algorithms for Joint CGH-Expression Analysis" on page 526.

To set up a differential aberration analysis

Differential aberration analysis reveals aberrations that are significantly different in different samples. After you assign one or more arrays to each of two groups, the program compares the groups and evaluates the differences in the detected aberrations between the groups. For a description of the statistical algorithms used in this analysis, see "Differential Aberration Analysis" on page 512.

1 Set up an experiment with at least two arrays. See "To create a new experiment" on page 60.

2 Select the experiment, select the desired arrays, then apply an aberration detection algorithm. See "To apply the aberration algorithm" on page 137.

3 Select the chromosome of interest.

4 In the Discovery tab, click **Differential Aberration.**

The Differential Aberration Setup dialog box appears. See "Differential Aberration Setup" on page 303. A graphical summary of detected aberrations in the chromosome appears for each array.

5 Below the diagram for each array, select one of these options. You must assign at least one array to Set 1 and one array to Set 2.

- **Set 1** – Assigns the array to the first group of arrays. The program will compare the aberrations in this group of arrays to the ones you assign to Set 2.
- **Set 2** – Assigns the array to the second group of arrays. The program will compare the aberrations in this group of arrays to the ones you assign to Set 1.
- **Ignore** – The program assigns the array to neither of the comparison groups, and does not consider the array in the differential aberration analysis.

If you have many arrays, it can be easier to use the Define Sets dialog box to assign the arrays to comparison groups. To open this dialog box, click **Define Sets.** See "Define Sets" on page 287.

6 In Select Algorithm, the HyperGeometric algorithm is selected. It is used to evaluate the significance of differences in the detected aberrations between the two comparison groups.

- **HyperGeometric Tail** – The HyperGeometric distribution tests for enrichment, or the likelihood that there is an increase in the number of aberrations present in one group given the number of aberrations present in both groups. The HyperGeometric Tail is therefore a measure of the enrichment of the gains or losses in a group. For more information, see "Differential Aberration Analysis" on page 512 and "Enrichment Analysis - the Hypergeometric distribution" on page 555.

7 Click **Run.**

The program evaluates the two sets of arrays, and summarizes the differences in the Graphical Differential Aberration Summary dialog box. "Graphical Differential Aberration Summary" on page 344.

You can customize and export the information in this dialog box in several ways:

Task	Details/Instructions
Filter the differential aberration results	You can include or exclude differentially aberrant regions based on filter criteria that you select. See "To create, modify, or apply a differential aberration filter" on page 160.
Export the table of differential aberration results	You can export the summary of analysis parameters and the table of differentially aberrant regions as a *.xls format file that you can open and work with in Microsoft Excel. If you have applied a filter to the results, the exported file contains only the regions that pass the filter. **a** Click **Export Table.** A dialog box appears. **b** Type a name and location for the file, then click **Open.** A dialog box asks if you would like to display the exported file. **c** To save the file, and open it in Microsoft Excel, click **Yes.** Otherwise, to save the file, only, click **No.**
Export the contents of the dialog box to another program as a graphic	**a** At the top of the dialog box, click **Edit > Copy summary to clipboard.** The program copies the visible portions of the dialog box to the clipboard as an image. **b** Paste the image into a document in another program that accepts images. **Note:** The copied image does not contain the regions of the dialog box to which you must scroll.

Task	Details/Instructions
Create a list of genes found in a differentially aberrant region	**a** At the bottom of the dialog box, in the table of differentially aberrant regions, click anywhere within a row to select it. **b** Click **Create Gene List.** The Create Gene List dialog box appears. See "Create Gene List (from common or differential aberration results)" on page 282. The program automatically sets the genomic coordinates of the region and the genome build. **c** Type a name for the gene list, then click **OK.** The gene list appears in the Gene List folder in the My Entities pane of the Navigator. You can now apply this gene list to data and results as desired. For more information on gene lists, see the *Data Viewing Guide*.
To resize all or part of the dialog box	• Drag the side or bottom borders of the dialog box to resize the entire dialog box. To resize individual panes, drag any of the borders between panes up or down.

To create, modify, or apply a differential aberration filter

A differential aberration filter excludes or includes differentially aberrant regions from the list of detected regions in the Graphical Differential Aberration Summary dialog box. The filter also removes or includes these regions in the file that you can download from this dialog box.

To create or modify a differential aberration filter

1 Set up a differential aberration analysis. See "To set up a differential aberration analysis" on page 158.

The results of the analysis appear in the Graphical Differential Aberration Summary dialog box.

2 At the top of the dialog box, click **Create Filter.**

The Interval Filter dialog box appears. See "Interval Filter" on page 361.

3 Do one of the following:

- To create a new filter, click **New.**
- To modify an existing filter, select its name.

4 Add, change, or remove filter conditions as desired.

NOTE If you update an existing filter, the program overwrites it. To restore the original filter parameters, you must recreate them manually.

5 Click **Update.**

The program saves the filter.

6 Click **Close.**

To apply an existing differential aberration filter

1 Set up a differential aberration analysis. See "To set up a differential aberration analysis" on page 158.

The results of the analysis appear in the Graphical Differential Aberration Summary dialog box.

2 In **Filter,** select the desired filter.

3 Mark **Apply Filter.**

The program applies the filter to the differential aberration results.

To set up a clustering analysis

Cluster analysis in Agilent Genomic Workbench reveals related groups of aberrant regions based on CGH aberration calls or aberration scores. A variety of distance metrics and linkage methods are available.

1 To effectively use cluster analysis, set up an experiment with at least three arrays.

2 Select the experiment, select the desired arrays for analysis, and apply an aberration detection algorithm. See "To apply the aberration algorithm" on page 137.

3 Click **Discovery > Clustering.**

The Setting Clustering Method and Parameters dialog box appears.

4 Select the settings for the clustering algorithm.

See "Setting Clustering Method and Parameters" on page 400.

5 Click **OK.**

The program performs the analysis, and displays the results as a hierarchical tree.

To display results as a heatmap

You can display the detected aberrations from the selected arrays in the active experiment as a heatmap. For each array, this format displays detected aberrations as a row of false color columns below a diagram of each chromosome in the genome. The false color spectrum ranges from luminous green for deletions to luminous red for amplifications. Areas with low luminosity (tending towards black) represent regions of no detectable change. When you click on one or more heatmap rows, you toggle the display of line plots for the arrays you selected.

1 In the active experiment, select the desired arrays, then apply an aberration detection algorithm. See "To apply the aberration algorithm" on page 137.

2 Click **Discovery > Heatmap.**

The Aberration Heatmap dialog box appears. See "Aberration Heatmap" on page 242. Amplifications appear as red bars, and deletions as green bars. The brightness of each bar reflects the magnitude of each aberration. Figure 23 shows an example heatmap for two arrays.

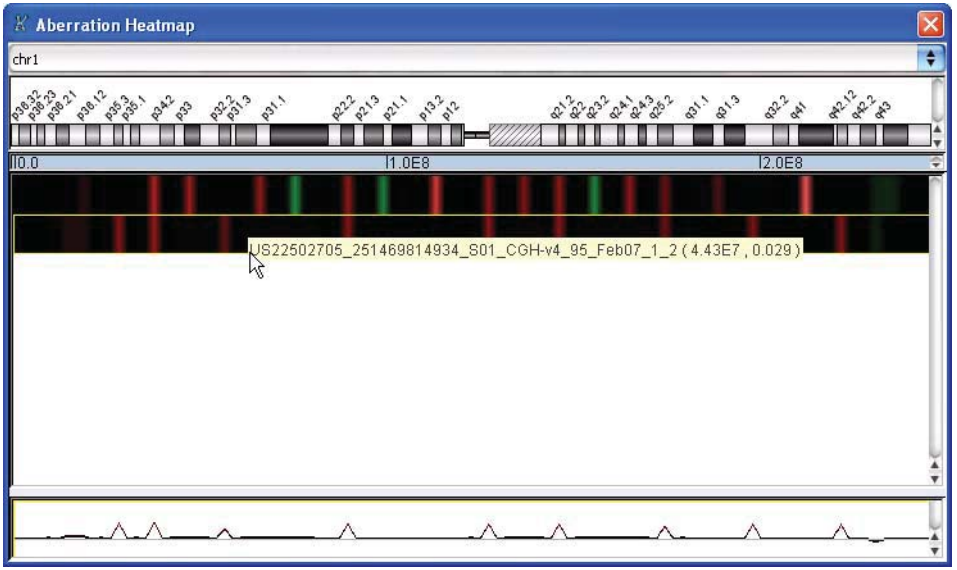


Figure 23 Aberration Heatmap dialog box, showing aberration heatmaps for two arrays. The aberrations for chromosome 1 are visible.

3 Do any of these tasks, as desired:

Task	Instructions
View the aberration heatmap for a different chromosome	• At the top of the dialog box, select the name of a different chromosome.
Display the name of the array associated with a heatmap	• Place the pointer over the desired heatmap. A ToolTip displays the name of the array that is associated with the heatmap. The genomic coordinates and data value that correspond to the location of the pointer also appear in the ToolTip.
Display the genomic coordinates and data value of a specific point on a heatmap	• Place the pointer over the desired point. A ToolTip displays the genomic coordinates and data value that correspond to the location of the pointer. The name of the array associated with the heatmap also appears in the ToolTip.

Task	Instructions
Display data for a specific array as a line plot	• Click the heatmap that corresponds to the desired array. A line plot of the array data appears at the bottom of the dialog box.
Hide line plots	• Click the heatmap that corresponds to the array whose line plot you want to hide.
Resize panes	Do any of the following: • To change the size of the entire dialog box, drag the side or bottom borders of the dialog box. • To change the relative sizes of the heatmap and line plot panes, drag the border between these panes up or down.

To generate a genotype reference file

If you have a reference sample that is not genotyped, you can create a genotype reference for that sample using the Agilent-provided or known genotype references available in the database. The genotype calls generated using the known genotype reference are used, along with a user-supplied confidence threshold and level, to generate a genotype reference file. This "custom" genotype reference file can then be imported to the database, and used to analyze additional CGH+SNP microarrays.

1 In the Home tab, import the reference microarray(s). See "To import Agilent FE or Axon data files" on page 51. The design file for the microarray must already exist in the database. Otherwise, import the design from the eArray Web site first. See "To import Agilent GEML design files" on page 48.

2 Create an experiment, and add microarrays that are associated with the known genotype references you want to use to generate the new genotype reference file. See "To create a new experiment" on page 60 and "To associate a genotype reference with a CGH+SNP array" on page 63.

3 In Sample Manager, in the Red Sample field for each microarray (Green for dye-flipped,) type the unknown reference sample name.

NOTE A validated genotype reference must be selected in the other sample channel.

4 In the Preprocessing and Analysis tabs, select the parameters required to perform a SNP analysis, then select the experiment you created in step 2, above. Analyze the microarray data for SNP Copy Number. See "To calculate SNP copy number" on page 139.

5 Select the Discovery tab, then click **Generate Genotype Reference**.

The Generating Genotype Reference File dialog box appears. See "Generating Genotype Reference File" on page 329.

6 In the Generating Genotype Reference Files dialog box, select the sample(s) to include in the genotype reference file.

7 Under Input Parameters, type a Confidence Threshold.

8 Under Input Parameters, select a Level of confidence

9 Click **Browse** and select a location to save the new genotype reference file.

10 Click **OK** to generate the new genotype reference file.

The file is saved in the specified location, with the name GenotypeReference_<Date>_<Design>_<Build>.txt.

Creating Reports

In Interactive mode, you can create and view many kinds of reports that help you understand and display the aberration results.

To create and export common aberration text reports

You can export tabulated reports by interval for shared aberrations in common genomic regions.

1 In the Experiment pane of the Navigator, right-click a Common Aberration node, then click **Generate Text Summary** in the shortcut menu.

2 Type a name and select a location for the common aberration output file, then click **Save.**

The program creates the common aberration text report(s) as *.xls format files, and saves them to the selected location. Multiple reports appear in a folder.

3 Click **Yes** to open the report.

NOTE You must first create a common aberration node within the experiment. See "To create common aberration summaries" on page 151.

To create CGH aberration text reports

You can export tabular aberration reports for any number of arrays.

1 In the Navigator, select array(s) for aberration summaries.

2 In the Reports tab, under CGH Aberration, click **Text.**

The Text Aberration Summary Setup dialog box appears. See "CGH Aberration Report Setup" on page 250.

3 Select one of these available Report Types:

- **Probe Based** – Reports by each probe on specified array(s).
- **Interval Based** – Reports by genomic intervals.
- **Probe & Interval Based** – Reports by probe and interval.

4 Select the scope of the aberration report. Per-Chromosome is selected by default, which produces one report for each chromosome in the genome.

5 Click **Browse.**

A Save dialog box appears.

6 Select a location and type a name for the report, then click **Save.**

The location for the report appears in Select File Location.

7 Click **Save.**

The program creates the aberration report(s) as *.xls format files, and saves them to the selected location. Multiple reports appear in a folder.

NOTE Interval based CGH aberration reports may be nested and therefore differ from what is seen in the display panel.

To generate graphical CGH aberration reports

The graphical CGH aberration report is a graphic display that summarizes aberrations for arrays that are selected in the experiment.

1 In the Reports tab, under CGH Aberration, click **Graphical**.

The Graphical Aberration Summary appears.

2 Click **Edit > Copy summary to clipboard** to copy the displayed summary to the Clipboard as an image. You can then paste the image into a document from another program that supports the Clipboard function.

For details, see "Graphical Aberration Summary (report)" on page 335.

To generate a SNP Genotype Report

1 Click the **Reports** tab.

2 Under SNP Genotype, click **Text**.

The SNP Genotype Report Summary Setup dialog box appears.

3 Select to generate the report for the Complete Genome, or Per-Chromosome.

For Complete Genome, one report is generated that contains summary information for the entire genome. For Per-Chromosome, individual reports are generated for each chromosome.

4 Click **Browse** and find the location where you want to save the report file, and type the name you want to use for the file.

5 Click **Save**.

The selected report(s) are created and saved in the file location. See "SNP Genotype Report" on page 429.

NOTE In order to generate a SNP Genotype report, you must first select to calculate SNP copy number in the Analysis tab.

To generate a SNP Aberration & LOH Report

You must select to calculate LOH in order to generate this report.

1 Click the **Reports** tab.

2 Under Aberration & LOH, click **Text**.

3 Click **Browse** and find the location where you want to save the report file. and type the name you want to use for the file.

4 Click **Save**.

The report is created and saved in the file location. See "SNP Aberration & LOH Report" on page 429.

NOTE In order to generate a SNP Aberration & LOH report, you must first select to calculate SNP copy number and LOH in the Analysis tab.

To generate probe penetrance text reports

Probe penetrance reports show the percentage of selected arrays that have an aberration at each probe position for the selected chromosomes. You can create tab-delimited report files of these summaries.

1 In the Reports tab, under Penetrance, click **Probe.**

The Text Penetrance Summary Setup dialog box appears. See "Text Penetrance Summary Setup" on page 408.

2 Click **Browse** to move to the folder where you want to save the probe penetrance summary report and type a file name.

3 Click **Save.**

The program creates the probe penetrance report(s) as *.xls format files, and saves them to the selected location. Multiple reports appear in a folder.

To create a new Cyto Report template

You can create cytogenetic reports for individual CGH samples in your Experiment or Data folders. First, you must set up a template to organize and define the report.

1 From the Reports tab, click **Cyto Reports > Design/Edit Template.**

The Design/Edit Cytogenetic Report Templates dialog box appears.

2 Click **New.**

3 Type the name of the report template, then click **OK.**

4 Accept or edit the default settings, then click **Save.**

See "Design/Edit Cytogenetic Report Templates" on page 288.

5 Click **OK.**

To open an existing Cyto Report template

After you have created a template, you can edit it at a later time by selecting it from a list.

1 From the Reports tab, click **Cyto Reports > Design/Edit Template.**

The Design/Edit CytoGenetic Report Templates dialog box appears.

2 In the Name list, click the arrow and select the template to edit.

3 Edit the settings, then click **Save.**

See "Design/Edit Cytogenetic Report Templates" on page 288 to learn how to change the settings for the report.

4 Click **OK.**

To generate a cytogenetic report

Generate individual sample reports one at a time

1 Select an individual sample in an Experiment folder, then from the Reports tab, click **Cyto Reports > Cyto Reports > "template name."**

Or right-click an individual sample under an Experiment or Data folder, and click **Cyto-Report > "template name."**

The "template name" is the name of the template you created in the Design/Edit CytoGenetic Report dialog box. Select the one you want to use to select content and format for the report.

If no template name appears, you must create a template. See "To create a new Cyto Report template" on page 169.

The Run Cytogenetic Report dialog box appears. See "Run Cytogenetic Report" on page 388.

2 Click **Browse.**

A Select report name dialog box appears.

3 Select a location and type a name for the report file, then click **Open.**

4 Select either **Run on selected experiment** or **Run by importing FE files.**

These two options are available only if you are generating a report through the Reports tab.

5 If you selected Run on selected experiment, go to step 6.

If you selected Run by importing FE files, follow these steps:

a Select the **Attribute File** (optional).

b Click **Add.**

c Select an FE file or files, and click **Open**.

6 Click **OK**.

The report is created using the template you selected. A message appears that displays the location of the report and asks if you want to open it.

7 Click **Yes** to open the pdf report. Click **No** to finish without a report display.

Generate individual sample reports simultaneously for all the samples in a folder

1 In the Data or Experiment navigator pane, right-click a build. The build can be located in a Data folder or an Experiment folder.

Or, right-click a named experiment.

The experiment does not have to be selected.

2 Click **Cyto-Report > "template name"**.

The Run Cytogenetic Report dialog box appears.

See "Run Cytogenetic Report" on page 388.

3 Type a folder path where you want to save your report files.

Or, click **Browse** to find the folder path where you want to save your report files.

4 Type a file name to be used for the report, and then click **Open**.

If you type the name of the folder into the Report File text box, do not add a backslash. Report generation will work with or without a .pdf extension on the Report File/Directory name.

Note that the options Run on selected experiment and Run by importing FE files are unavailable. These are available only if you generate a report through the Reports tab.

5 Click **OK**.

The program generates the reports using the template you selected, and places them in a folder generated from the report file name.

5 CGH Interactive Analysis Reference

Agilent Genomic Workbench Main Window 174
Switch Application Menu 175
Command Ribbons 176
Navigator 202
Genomic Viewer 224
Status Bar 238
Dialog Boxes 239
Reports 427
Configuration Settings 431

This chapter describes the parts of the Agilent Genomic Workbench main window that you use to import, organize, manage, export and display CGH and CGH+SNP array data and other content. It also gives details of the relevant tab commands, shortcut menus, and dialog boxes that appear.

Agilent Genomic Workbench Main Window

The sections that follow describe the main components of the Agilent Genomic Workbench main window – Switch Application Menu, the command ribbons, the Navigator and the Views. You use these to perform Preprocessing, Analysis and Discovery tasks on CGH and CGH+SNP data. For descriptions of the dialog boxes for these elements, see "Dialog Boxes" on page 239.

Figure 24 shows the main window of Agilent Genomic Workbench Lite Edition, and identifies its main parts.

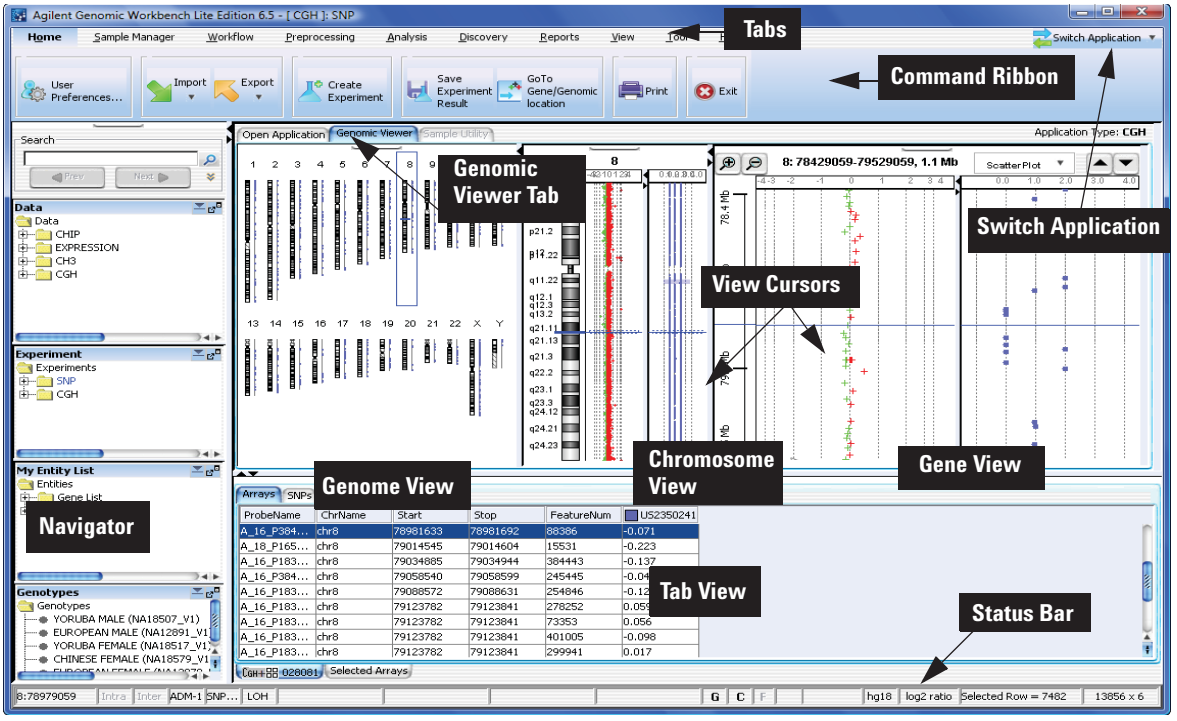


Figure 24 Agilent Genomic Workbench Lite Edition Main Window

Switch Application Menu

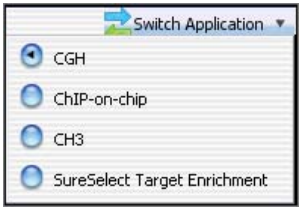


Figure 25 Switch Application menu

The Switch Application menu lets you to change to the other data display and analysis module types in Agilent Genomic Workbench. Select the desired application type.

CGH (Separate license required) Import, display, and analyze array-based comparative genomics hybridization (aCGH) data in both an interactive "analyze as you go" mode, and an automated workflow mode.

ChIP (Separate license required) Import, display, and analyze ChIP-on-Chip microarray data in both an interactive "analyze as you go" mode, and an automated workflow mode.

CH3 (Separate license required) Import and display data from microarray-based studies of genomic methylation patterns.

SureSelect Target Enrichment Use the Quality Analyzer function for SureSelect Target Enrichment. See the *SureSelect Target Enrichment User Guide* for more information.

Command Ribbons

When you click a tab at the top of the Agilent Genomic Workbench main window, groups of commands appear below the tab bar. This group of commands is called a command ribbon, and the commands that appear are available only for the selected tab. The tabs that are displayed change depending on what module is selected (such as CGH, ChIP, CH3). This section describes the ribbon commands used to import, manage, export and display data and results for CGH interactive analysis.



Figure 26 Tab bar with command ribbon for CGH module

Home command ribbon

The Home command ribbon shows the functions that let you import, manage, export and display CGH data and content (gene lists, tracks) for further CGH analysis.



Figure 27 Home command ribbon

User Preferences Opens the User Preferences dialog box with the following tabs:

Tab	Description
Tracks	Opens a dialog box that lets you manage the tracks to display in Genomic Viewer and how they appear. See "Tracks tab" on page 415.
Miscellaneous	Opens a dialog box where you can select a new location for your data files and set up access to the eArray web site. See "Miscellaneous tab" on page 417.
License	Opens a dialog box where you can enter an Agilent Genomic Workbench DNA Analytics module license, if you purchase one after using the unlicensed version. See "License tab" on page 418.

Import Opens a menu of file types that you can import:

Option	Description
Array Files	Opens a menu with these options: • **FE File** – Opens the Import FE Files dialog box, where you can select an Agilent Feature Extraction array data file to import. See "Import" on page 354 and "To import Agilent FE or Axon data files" on page 51. • **Axon File** – Opens the Import Axon Files dialog box, where you can select Axon (*.gpr) files for import. See "Import" on page 354 and "To import Agilent FE or Axon data files" on page 51. • **UDF File** – Opens the UDF Files dialog box, where you can select a Universal Data File (UDF) to import. See "Import" on page 354 and "To import a UDF file" on page 52.
Design Files	Opens a menu with these options: • **GEML File** – Opens the Import Design Files dialog box, where you can select Agilent GEML-based (*.xml) array design files for import. See "Import" on page 354 and "To import Agilent GEML design files" on page 48. • **Axon Design File** – Opens the Import Axon Design Files dialog box, where you can select Axon (*.gal) array design files for import. See "Import" on page 354 and "To import Axon design files" on page 49.

Option	Description
Genome Build	Opens the Import Genome Build dialog box, where you can import Agilent-supplied genome build files. See "Import Genome Build" on page 359 and "To import a genome build" on page 50.
Track	Opens the Import Track dialog box, where you can select a BED format track file for import, and create a display name for the track. See "Import Track" on page 360 and "To import tracks" on page 55.
Array Attributes	Opens the Import microarray attributes dialog box, where you can select a microarray attributes file. See "Import" on page 354.
Experiments	Opens the Import Experiments dialog box, where you can select a ZIP format experiment file for import. See "Import" on page 354 and "To import an experiment file" on page 57.
Filters	Opens the Import dialog box, where you select a filter file to import. For more information, see "Import" on page 354 and "To import filters" on page 58.
Genotype References	Opens the Import Genotype Reference Files dialog box, where you select a .txt or .xls file that contains one or more genotype references to use for SNP analysis. For more information see "To import a genotype reference file" on page 58.

Export Opens a menu that lets you export several kinds of files.

Option	Description
Experiments	Opens the Export Experiments dialog box, where you can select one or more experiments for export as a single ZIP file. See "Export Experiments" on page 320 and "To export experiments" on page 83.
Filters	Opens the Export Filters dialog box, where you can select one or more filters for export as a single *.xml file. See "Export Filters" on page 321 and "To export filters" on page 85.
Tracks	Opens the Export Tracks dialog box, where you can select one or more tracks to export as a single BED format file. See "Export Tracks" on page 322 and "To export tracks" on page 84.
Array Attributes	Opens the Export Array Attributes dialog box, where you can select arrays and their attributes for export. See "Export Array Attributes" on page 316.

Create Experiment Opens the Create Experiment dialog box, where you can create a new, empty experiment and add data to it. See "Create Experiment" on page 278 and "To create a new experiment" on page 60.

Save Experiment Result Opens a confirm dialog box. Click **Yes** and the Save experiment result dialog box opens where you enter a name and description for the results to save.

Go to Gene/Genomic Location Moves the cursor to the location in Chromosome and Gene Views that you select. See "Go To Gene/Genomic Location" on page 334.

Print Opens the Print window to print the display.

Exit Closes the program.

Preprocessing command ribbon

The Preprocessing tab opens a ribbon that contains four option groups: Filters, Centralization, Combine, and QC Metrics. When you apply these options to a selected experiment, they prepare the data before subjecting it to the formal analysis for aberrations.

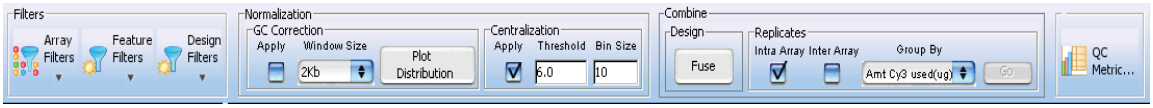


Figure 28 Preprocessing command ribbon

Filters

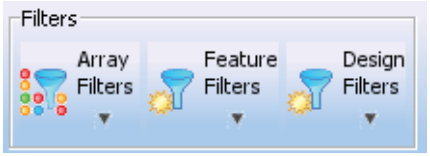


Figure 29 Preprocessing commands – Filters

Array Filters

Array filters let you include or exclude arrays in the current experiment based on their attributes. If you select this option, another menu opens with these options:

- **Edit Filter** – Opens the Array Level Filters dialog box, where you can create, change, or delete array-level filters. See "Array Level Filters" on page 247.
- **Apply** – Opens a menu that shows all available array-level filters. Mark the desired filter. To remove the applied filter, or to select no filter, mark **No Filter.**

Feature Filters

Feature-level filters let you include or exclude data from specific microarray features based on information from the imported Feature Extraction output files. For example, the DefaultFeatureFilter removes data for features that have a non-uniformity or saturated flag set for either channel (for example, if one or more of glsFeatNonUnifOL, rlsFeatNonUnifOL, glsSaturated, or rlsSaturated is set to **True.**)

Select one of these options:

- **Edit Filter** – Opens the Feature Level Filters dialog box, where you can create, change, or delete feature-level filters. See "Feature Level Filters" on page 323.
- **Apply Filter** – Opens a menu that shows all available feature-level filters. Mark the desired filter. To remove the applied filter, or to select no filter, mark **No Filter.**

Design Filters

Design filters let you include or exclude microarray data based on design information from imported design files. A design filter can be used, for example, to filter out features that fail the homology filter or have a low probe score.

Select one of these options:

- **Edit Filter** – Opens the Design Level Filters dialog box, where you can create, change, or delete design level filters. See "Design Level Filters" on page 297.
- **Apply Filter** – Opens a menu that shows all available design-level filters. Mark the desired filter. To remove the applied filter, or to select no filter, mark No Filter.

GC Correction

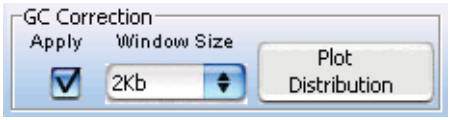


Figure 30 Preprocessing commands – GC Correction

When GC Correction is applied, data is corrected for "wavy" artifacts that are associated with the GC content of genomic regions. These artifacts can interfere with detection algorithms and lead to inaccurate aberration, copy number, and LOH calls. GC Correction is recommended for SNP Copy Number and LOH analyses.

Apply Mark this option to apply GC Correction to the log ratio data.

Window Size Select a window for the calculation. For GC Correction to work properly, the array design file must contain GC content values for the genomic regions that flank each CGH probe. The design files for Agilent Catalog arrays currently contain GC content values for window sizes of 2 kb, 10 kb, and 40 kb surrounding each CGH probe.

Plot Distribution Click this to open a plot that shows corrected and uncorrected log ratio data.

Centralization



Figure 31 Preprocessing commands – Centralization

Turning on centralization turns on and configures the centralization normalization algorithm. Centralization re-centers log ratio values. It finds a constant value to subtract from or add to all values, and ensures that the zero-point reflects the most-common-ploidy state. For a description of the centralization algorithm, see "Centralization Algorithm" on page 452.

Turn On Mark this option to apply the centralization algorithm to all interactive analyses.

Threshold This value is the ADM-1 or ADM-2 threshold used to call aberrations for the centralization algorithm. Ideally, set this value to the ADM-1 or ADM-2 threshold defined in the current analysis. However, because the centralization value is fairly robust to a wide range of threshold values, Agilent does not recommend that you change the default settings for this algorithm.

Bin Size This value is used to average contiguous probes across the genome. This makes the calculation faster. The default value is recommended.

Combine

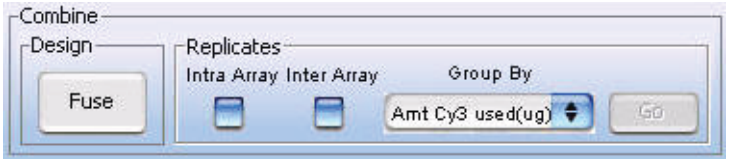


Figure 32 Preprocessing commands – Combine

The settings and commands in Combine enable you to merge multiple array designs into a larger virtual combined design. They also let you combine both intra- and inter-array replicate probes. For a discussion of the error model used for combining replicates, see "Error Model and Combining Replicates" on page 462.

Fuse Click to open the Array Set dialog box, where you can view the names of arrays to be fused, set options, and fuse two designs together to form a larger virtual design. See "Array Set" on page 249, and "To combine (fuse) arrays" on page 130.

Intra Array Mark this check box to combine intra-array replicates. The program uses probe names to identify replicates in arrays. This option is available only if Agilent microarrays are selected. The program combines the replicates when you click **Go.** See "To combine intra-array replicates" on page 132.

Inter Array Mark this check box to combine replicates across multiple arrays. The program uses the array attribute you select in **Group By** and combines replicates from arrays that have the same value defined for that attribute. The program combines the replicates when you click **Go.** See "To combine interarray replicates" on page 133.

Group By Select an array attribute. The program combines replicates from arrays that have the same value assigned to the attribute you select.

Go Click **Go** to combine replicates in the selected arrays in the current experiment according to the other settings under Replicates. To select arrays, see "To select or deselect arrays in the experiment" on page 91. Combining replicates in very large arrays, or from many arrays, can require extended computation time.

QC Metric

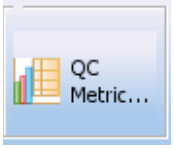


Figure 33 QC Metric command

The QC Metric command opens the QC Metrics Table, which displays the available QC metrics for Agilent Arrays in the selected experiment. You can also use this dialog box to set the QCMetricStatus array attribute for each array. See "QC Metrics Table" on page 378 and "To display QC metrics of arrays and set array QC status" on page 134.

Analysis command ribbon

The Analysis command ribbon lets you select parameters for the analysis algorithms.



Figure 34 Analysis command ribbon

Moving Average



Figure 35 Analysis commands – Moving Average

The settings under Moving Average control the calculation and display of line plots that represent smoothed data for each selected array. These plots can appear in Genome, Chromosome, and Gene views.

Show Mark this check box to show moving average line plots, or clear the check box to hide them. You turn the display of moving average line plots for specific views on or off from the View Preferences dialog box.There are two ways to open the View Preferences dialog box:

- In View Preferences, under Data Visibility, select the desired view, then mark or clear **Moving Average.** See "View Preferences" on page 422.
- In the Gene View, Genome View, or Chromosome View, right-click and then click **View Preferences**.

Algorithm Select one of these options:

- **Linear** – The linear algorithm calculates a standard, unweighted average using every probe log ratio score within a selected window size (defined by either a number of adjacent measurements or a positional range). The program applies this average sequentially to every point by moving the window along the chromosome.
- **Triangular** – The triangular algorithm calculates a weighted average using every probe log ratio within a selected window size. The triangular algorithm is more sensitive to localized variations in the data. See "Triangular Smoothing" on page 502.

Line width Select the thickness (in pixels) for the moving average line plots. You can select a thickness from 1 to 5 pixels.

Window Select the desired size of the moving average window. You can select either a number of base pairs (5 Kb to 50 Mb), or a number of data points (1 pt to 60 pt). You can also type a value. The program calculates a moving average for each selected array based on a window of the given size, centered on each point in the array.

NOTE Both moving averages and Z-scores use window size as input, and you can plot them individually or together. See "Z-Scoring for Aberrant Regions" on page 466. Changing this window size does not affect aberration calls using the ADM, CBS, or HMM algorithms, but does change the window size used by the Z-score detection algorithm.

Aberration

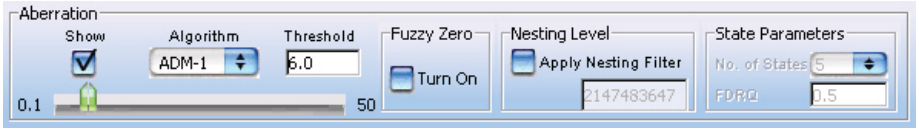


Figure 36 Analysis commands – Aberration

You use the settings under Aberration to select, configure, and apply an aberration algorithm to the selected array data.

Show Mark this check box to apply the chosen aberration algorithm to the selected data, and to display the aberration plot. The aberration plot can appear in Genome, Chromosome, and/or Gene views. You turn on or off the display of aberration plots for specific views from the View tab, under View Preferences. In View Preferences, under Data Visibility, select the desired view, then mark or clear **Aberration**. See "View Preferences" on page 422.

Algorithm Select one of these aberration detection algorithms:

- **Z-score**
- **ADM-1** (Aberration Detection Method 1)
- **ADM-2** (Aberration Detection Method 2)

- **CBS** (Circular Binary Segmentation)
- **HMM** (Hidden Markov Model)

For more information on these algorithms, see "Aberration Detection Algorithms for CGH Analysis" on page 465.

Threshold/Slider Drag the slider, or type a value between 0.1 and 50.0 in Threshold.

The optimal threshold value, and the meaning of threshold, differ among algorithms, even for the same array. For example, for the ADM-1 and ADM-2 algorithms, the threshold is applied to each interval score identified by the algorithm (see "ADM-1" on page 468), while in the *Z*-score algorithm, the threshold is used as a cutoff to classify data as significantly above or below the mean (see "Z-Scoring for Aberrant Regions" on page 466).

NOTE

- This threshold has no meaning when either the CBS or HMM algorithms are applied.
- A recommended starting point for both the ADM-1 and ADM-2 algorithms is a threshold value of 6 (the default in the UI).

Fuzzy Zero (ADM-1 and ADM-2 algorithms only) To apply fuzzy zero correction, mark **Turn On.**

ADM-1 and ADM-2 scores may identify extended aberrant segments with low absolute mean ratios. Often such aberrations represent noise, and are detected because of a high number of probes in the region. If the program detects long, low aberrations in your analysis, you can apply the fuzzy zero algorithm to correct for the reliance on segment probe number. For more information on this algorithm, see "Fuzzy Zero" on page 458.

Nesting Level (ADM-1 and ADM-2 algorithms only) To apply the nesting filter, mark **Apply Nesting Filter,** and type an integer in the box.

ADM-1 and ADM-2 scores iteratively identify all aberrations that differ significantly from log ratios that show no change between sample channels. In the iterations, the algorithms identify nested, or "child," aberrations that are contained within other "parent" aberrations, but show a significant difference from the log ratio of the parent aberration.

The nesting level is the number of "child" aberrations to be identified. By default, the nesting level is set to a maximum integer value (2147483647), to find as many discrete aberrations as possible. When set to this maximum value, no filter is applied to the number of possible "child" aberrations.

To filter the number of identified aberrations, you can set this integer value to any desired value between 0 and 2147483647. For example, if you set this value to 0, you will filter out all nested (child) aberrations, retaining only the parent aberrations. If you set it to a value of 1, you retain only the first nested level.

State Parameters

(HMM algorithm only) See "HMM" on page 472. Set these parameters:

- **No. of States** – This is the number of distinct aberration states into which the observed data is to be partitioned. For example, if the number of states equals 3, then this would correspond to amplification, deletion and "no change" aberration status. You can set the number of states to either 3, 5, or 7.
- **FDRQ** – This is a False Discovery Rate threshold used in the calculation of the Discrete Haar Wavelet Transform. It is used to keep only high-amplitude coefficients, which mark occurrences of true breakpoints (rather than noise) in the data. Stringent thresholding will reduce the number of spurious transitions between states in the HMM, but may result in loss of sensitivity to local changes in the data. For details, see "Segmentation- Discrete Haar Wavelet Transform and Gaussian mixtures for HMM" on page 564.

SNP Copy Number

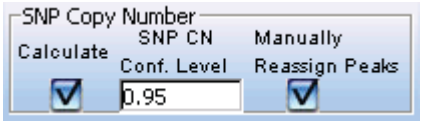


Figure 37 Analysis commands – SNP Copy Number

For CGH+SNP microarrays that combine CGH and SNP probes on the same microarray, this calculates an expected copy number for uncut alleles for each SNP site on the array. For details on how the algorithm works, see "SNP analysis algorithms" on page 447.

Calculate Mark this box to turn on calculation of SNP copy number calculation.

SNP CN Conf. Level This number represents a threshold, or confidence value that is used to decide whether or not to include a copy number in the SNP analysis. A higher value makes the test more stringent.

Manually Reassign Peaks When you mark this box, the Manually Reassign Peaks dialog box opens, where you can review the copy number assignments and change them. See "To manually reassign peaks" on page 140.

NOTE In order to calculate SNP Copy Number, you must first select an Aberration in the Analysis tab. GC Correction and Centralization are recommended.

LOH

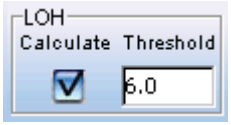


Figure 38 Analysis commands – LOH

For SNP analysis, LOH calculates statistically significant regions of consecutive homozygous SNPs, and reports the regions where the LOH score exceeds a defined threshold. For details on how the algorithm works, see "LOH (Loss or lack of heterozygosity) algorithm" on page 544.

Calculate Mark this to enable calculation of LOH regions.

Threshold Type a threshold value to use in determining whether or not to include an LOH call. A higher value makes the test more stringent.

NOTE In order to calculate LOH, you must first select to calculate SNP Copy Number.

Discovery command ribbon

The commands on the Discovery command ribbon let you display, filter, compare, and further analyze the results of aberration analysis.

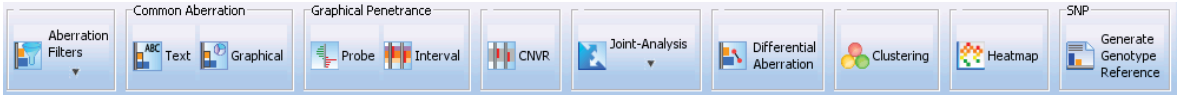


Figure 39 Discovery command ribbon

Aberration Filters



Figure 40 Discovery commands – Aberration Filters

Aberration filters reduce the number or extent of aberrations that the program displays, based on parameters defined in the filter. If you select this option, another menu opens with these options:

Apply

- Opens a menu that shows all available aberration filters. Mark the desired filter. To remove the applied filter, or to apply no filter, mark **No Filter.**

Edit Filter

- Opens the Aberration Filters dialog box, where you can create, change, or delete aberration filters. See "Aberration Filters" on page 240.

NOTE

The program always applies all four filter conditions in the order in which they appear in the Aberration Filters dialog box. To create a filter based on a single condition, set the other conditions to nonrestrictive, or "safe," values. For example, to create an aberration filter based only on absolute average log ratio, set **Minimum number of probes in region** to 1, set **Maximum number of aberrant regions** to 2000, and **Percent penetrance per feature** to 1.

Common Aberration

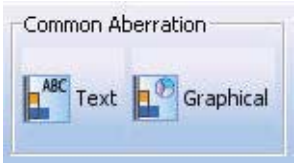


Figure 41 Discovery commands – Common Aberration

Text Opens the Create Common Aberration dialog box. See "Create Common Aberration" on page 276. You use this dialog box to create a new Common Aberration node within an experiment result in the Navigator. The program detects statistically significant aberrations among multiple samples, and lets you display these regions in Genome, Chromosome, and Gene views.

Graphical Opens the Graphical Common Aberration Setup dialog box, where you can select specific arrays and configure the parameters for common aberration analysis. See "Graphical Common Aberration Setup" on page 339. When you run the analysis, the program first analyzes each array individually. You then click **Common Aberration** to generate a graphical summary of common aberrations.

Graphical Penetrance

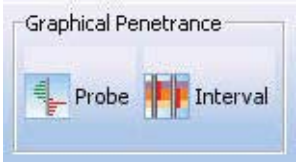


Figure 42 Discovery commands – Graphical Penetrance

Probe Opens the Graphical Penetrance Summary in a new window. See "Graphical Penetrance Summary" on page 349. This report displays graphically the percentage of selected arrays that have a significant

aberration at each probe position. You can select any number of chromosomes in the genome. The program creates a separate penetrance plot for each one.

Interval Opens the Graphical Interval Penetrance Summary in a new window. See "Graphical Interval Penetrance Summary" on page 351. This report summarizes all of the genomic intervals associated with significant aberrations, and gives the percentage of selected arrays that show a significant deletion or amplification in each interval. You can view both a table of all intervals with aberrations, and graphical maps of aberrations detected in each array. You can also export the table and view it in Microsoft Excel, and create a gene list from the report.

CNVR



Figure 43 Discovery commands – CNVR

(Copy Number Variant Region) Opens the Create CNVR dialog box. See "Create CNVR" on page 274. CNVRs are the result of joining any overlapping aberrant regions on a chromosome. You use this dialog box to create a new CNVR node within an experiment result in the Navigator. The program determines CNVRs based on the selected array data, then displays them next to log ratio data and detected aberrations in Genome, Chromosome, and Gene views. See "Copy Number Variation Region" on page 518.

Joint-Analysis



Figure 44 Discovery commands – Joint Analysis

Opens another menu with the options listed below. These options let you correlate CGH data with gene expression data.

Correlation Analysis

Opens another menu with the options listed below. A correlation analysis determines if correlations exist between selected CGH and gene expression arrays.

- **Setup** – Opens the Correlation Analysis Setup dialog box, where you can configure and then perform the analysis. See "Correlation Analysis Setup" on page 270.
- **Perform** – (If available) Runs the correlation analysis you previously configured.

Enrichment Analysis

Opens another menu with the options listed below. You can run a CGH-Expression enrichment analysis for any sample in which both CGH and expression were measured. The analysis produces a list of genomic intervals for each sample, with a statistical analysis that shows the enrichment of over- and under-expressed genes in each interval.

- **Setup** – Opens the Enrichment Analysis Setup dialog box, where you can configure and then perform the analysis. See "Enrichment Analysis Setup" on page 310.
- **Perform** – (If available) Runs the enrichment analysis you previously configured.

Differential Aberration



Figure 45 Discovery commands – Differential Aberration

Opens the Differential Aberration Setup dialog box, where you can assign samples to different comparison groups, select an algorithm, and start the differential aberration analysis. See "Differential Aberration Setup" on page 303. After running the analysis, the Graphical Differential Aberration Summary dialog box appears. See "Graphical Differential Aberration Summary" on page 344. Differential aberration analysis reveals aberrations that are significantly different in different samples.

Clustering



Figure 46 Discovery commands – Clustering

Opens the Setting Clustering Method and Parameters dialog box, where you can set the parameters for cluster analysis. See "Setting Clustering Method and Parameters" on page 400. Cluster analysis reveals related groups of aberrant regions based on CGH aberration calls or aberration scores. For a description of the clustering algorithm, see "Cluster Analysis" on page 515.

Heatmap



Figure 47 Discovery commands – Heatmap

Opens the Aberration Heatmap dialog box in a new window, where you can view significant aberrations as heat maps aligned with the ideogram of each chromosome in the genome. See "Aberration Heatmap" on page 242.

Generate Genotype Reference

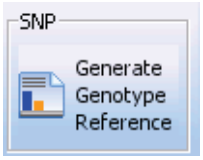


Figure 48 Discovery commands – Generate Genotype Reference

Opens the Generating Genotype Reference Files dialog box, where you set up and create a genotype reference file from selected CGH+SNP reference microarray(s). This genotype reference can be imported later into the Agilent Genomic Workbench database, and used to analyze CGH+SNP data. See "To generate a genotype reference file" on page 164 and "Generating Genotype Reference File" on page 329.

Reports command ribbon

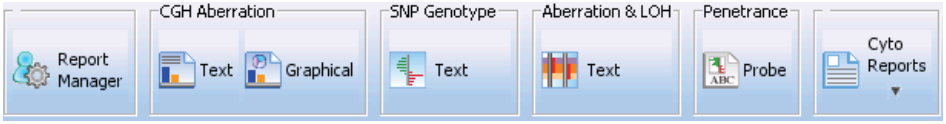


Figure 49 Reports command ribbon

The Reports tab gives you access to several different kinds of reports. Some reports provide details about the analysis of the current experiment, while others let you perform additional analyses. The program saves report files to a location that you select. You use other programs, such as a spreadsheet program or Adobe Acrobat Reader, to display saved reports.

Report Manager Opens the Report Management dialog box where you can view and manage reports created with Agilent Genomic Workbench. See "Report Management" on page 386.

CGH Aberration

Text Opens the Text Aberration Summary Setup dialog box, where you can configure the report and select a location for it. See "CGH Aberration Report Setup" on page 250.

The report details the regions that have high aberration scores, according to the selected algorithm for the active experiment. You can report these regions by genomic interval, by probe, or both. The program reports aberrations separately for each selected array in the active experiment. When you save the report, the program creates one or more *.xls format files that you can work with in Microsoft Excel.

Graphical Opens the graphical display of the Aberration Summary for Chromosome *, where the * represents the number or letter of the chromosome, and displays the location of the amplifications and deletions for the selected arrays.

SNP Genotype

Genotype Opens the SNP Genotype Report Summary Setup dialog box, where you select to generate a SNP genotype report for the complete genome, or separate reports for each chromosome. You also type or select a location and name for the report file(s). The report is saved in .xls format. See "SNP Genotype Report Setup" on page 406 and "SNP Genotype Report" on page 429.

Aberration & LOH

Text Opens the Aberration & LOH Report dialog box, where you select a location for the Aberration & LOH report, and type a name for the report. The report is saved in .xls format, See "SNP Aberration & LOH Report" on page 429.

Penetrance

Probe Opens the Text Penetrance Summary Setup dialog box, where you can configure the Probe Penetrance Summary report, and select a location for it. See "Text Penetrance Summary Setup" on page 408. The report shows each probe associated with a significant aberration, and gives the percentage of selected arrays that show a significant deletion or amplification in the region associated with each probe. Amplifications and deletions are considered separately. When you save the report, the program creates one or more *.xls format files that you can work with in Microsoft Excel.

Cyto Reports

Opens a menu with the options listed below. Cyto reports summarize analysis settings and detected aberrations by array.

Design/Edit Template Opens the Design/Edit Cytogenetic Report Templates dialog box, where you can create, change, or delete the cyto report templates available in the program. See "Design/Edit Cytogenetic Report Templates" on page 288. See "To create a new Cyto Report template" on page 169, and the topic following it. You can use these templates to generate cyto reports for individual samples or microarrays.

"Name of template" (The names of any saved templates appear in the Cyto Reports menu. If Empty appears, no cyto report templates are available.) Opens the Run Cytogenetic Report dialog box, where you can configure the input for the

report, and select a location for it. See "Run Cytogenetic Report" on page 388. The program generates a separate PDF format file for each array. See "To generate a cytogenetic report" on page 170.

View command ribbon

The View command ribbon lets you change the display of data and results in Genomic Viewer.

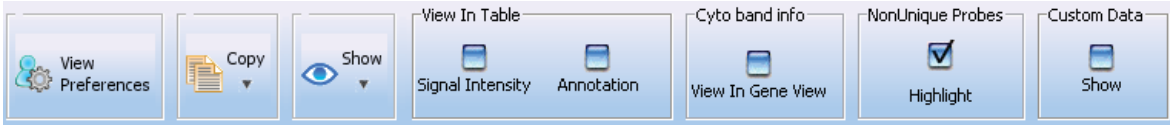


Figure 50 Command ribbon of the View tab

View Preferences Opens the View Preferences dialog box, where you can customize the display of data and results in the Genomic Viewer. For more information, see "View Preferences" on page 422.

Copy This command opens a menu with the options listed below. In general, the Copy command copies pane(s) of the main window to the Clipboard as an image. You can then paste the image into a document in another program. See "To copy what you see in the main window" on page 85.

Option	Description
All	Copies all panes of the main window to the Clipboard as an image.
Navigator	Copies only the Navigator to the Clipboard as an image.
Tab View	Copies only Tab View to the Clipboard as an image.
SampleBySample view	(Available only in data analysis modules, when selected) Copies only Sample-by-sample View to the Clipboard as an image.
Genome view	Copies only Genome View to the Clipboard as an image.
Chromosome view	Copies only Chromosome View to the Clipboard as an image.
Gene view	Copies only Gene View to the Clipboard as an image.

Show Opens a menu with all available elements of the main window. Mark the check box for the ones you want to display in Genomic Viewer.

View In Table

Signal Intensity Mark the check box to see the red and green raw signal intensities of the log ratio data in the Tab View.

Annotation Mark the check box to show annotations in the Tab View.

Cyto band info

View In Gene View Mark the check box to display cytobands in the Gene View.

NonUnique Probes

Highlight Nonunique probes in a microarray design are probes that map to more than one genomic location. Because the probes represent the same sequence, the probe log ratio reflects a combination of log ratios from the redundant locations. Mark the check box to display nonunique probes in a different color.

Custom Data

Show Mark the check box to display custom data in the Genomic Viewer.

Tool command ribbon

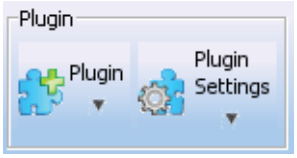


Figure 51 Tool command ribbon

Plugin

Plugins are ancillary programs that process the selected array data in the active experiment in specific ways.

Opens a menu with the options listed below. If you or another user has created custom plug-ins, they also appear in this menu.

CGHSmooth Opens the CGHSmooth Parameters dialog box. See "CGHSmooth Parameters" on page 251. You can set the parameters of the CGHSmooth plug-in, and create separate, stacked plots of smoothed log ratio data for each of the selected arrays in the current experiment. The plug-in plots the data associated with the selected chromosome.

When you open this dialog box, you see the parameters you select under Plugin Settings.

Echo Example Creates separate, stacked plots of log ratio data for each of the selected arrays in the current experiment. The plug-in plots the data associated with the selected chromosome. The plot appears in a new window. Although simple, this plug-in gives you a convenient way to display the log ratio data for selected arrays as separate plots. See "Echo Example Plot" on page 305.

MovAvg Example Opens the MovAvg Example Parameters dialog box. See "MovAvg Example Parameters" on page 371. You can set the parameters of the MovAvg Example plug-in. The plug-in calculates a 10-point moving average of each column of selected microarray data, and produces stacked plots of all of the input data and moving averages. To use this plug-in, you must have Perl installed on your computer.

When you open this dialog box, you see the parameters you select under Plugin Settings.

Plugin Settings

Opens another menu with these options:

CGHSmooth Opens the CGHSmooth Parameters dialog box, where you can set the parameters for the plug-in when you have selected to not show the parameters dialog box when you click Plugin. See "CGHSmooth Parameters" on page 251.

MovAvg Example Opens the MovAvg Example Parameters dialog box, where you can set the parameters for the plug-in when you have selected to not show the parameters dialog box when you click Plugin. See "MovAvg Example Parameters" on page 371.

Help command ribbon

The Help command ribbon lets you display the available Agilent Genomic Workbench help guides, and get information about software version, installation history, and check for software updates. Help guides are opened in Adobe Reader.



Figure 52 Help command ribbon for the CGH module

Table 4 Table of Agilent Genomic Workbench Help

Help Command	Action
Application Guide	Opens the Agilent Genomic Workbench user guide for the selected module.
Sample Manager	Opens the *Sample Manager User Guide,* that shows how to use the Sample Manager module of Agilent Genomic Workbench to organize microarrays and edit their attributes.
Workflow	Opens the *Workflow User Guide*, that describes how to use the Workflow module of Agilent Genomic Workbench to extract image files with Agilent Feature Extraction software and/or analyze data using CGH and ChIP analysis software.
Data Viewing	Opens the *Data Viewing User Guide* that describes how to import, organize, manage, export and display data and other content (experiments, gene lists, tracks) within Agilent Genomic Workbench. It is targeted for users who have no DNA Analytics module license(s).
About	Opens a message with information about the version number and copyright of the program.

An additional guide is available in the Open Application tab of the program. The *Agilent Genomic Workbench Product Overview Guide* gives an overview of the capabilities within the Standard Edition of Agilent Genomic Workbench. It also describes how to start each of the component

programs and find help, and how to enter your license information. In addition, it helps you with system administration and troubleshooting. To open this guide, click the **Open Application** tab, then click **Product Overview**.

Navigator

This section describes the parts of the Navigator, and the shortcut menus and other functionality available within it.

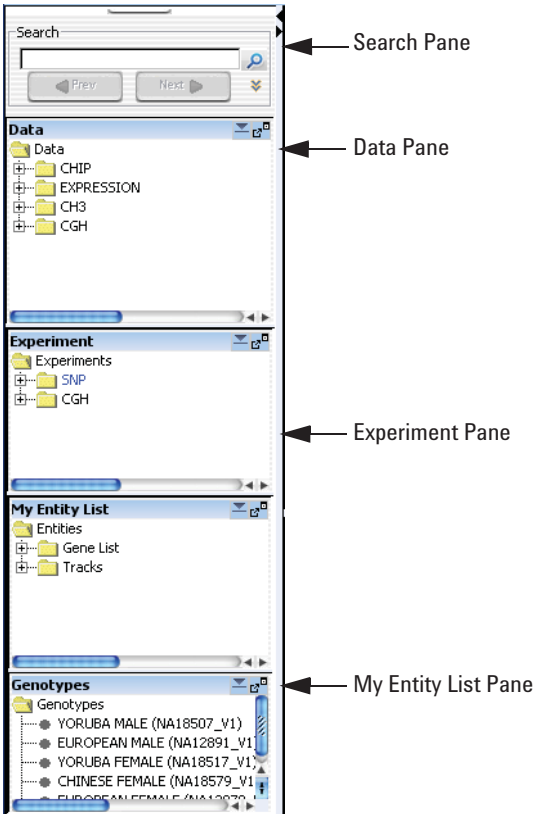


Figure 53 Navigator panes

The Navigator shows the array data, experiments, and other content stored in Agilent Genomic Workbench. It contains the following panes:

Pane	Comments
Search	Lets you search within any pane of the Navigator for a specific item (array or build, for example). You must type the entire array name or term; otherwise, use asterisks (*) as wildcards for unspecified strings. For example, type *1234* to find any item that contains "1234".
Data	Contains microarray data files, organized by application type, then by design and genome build. Shows all probe groups and microarray designs that are available to you, organized by folders. For the SureSelect Target Enrichment application type, the program shows all bait groups and libraries. In general, you can: • Expand or collapse folders to show or hide content. • Look at the icon that appears with an item to monitor its status. • Right-click the name of a folder or item to open a shortcut menu that lets you take action on the item. See "Data pane – icons, special text, and buttons" on page 206 and "Data pane – actions and shortcut menus" on page 207.
Experiment	Contains Agilent Genomic Workbench experiments. Experiments are organizational units that contain links to microarray data and design files. In data analysis modules, experiments also contain saved results. See "Experiment pane – icons, special text, and buttons" on page 209.
My Entity List	Contains gene lists and tracks: • **Gene Lists** are collections of genes of interest. You can create them within the program, import and export them, and apply them to Gene View and Chromosome View. • **Tracks** are collections of annotation or other information that map to specific genomic locations. You can import, export, and combine tracks, and display them in Gene View with your array data and analysis results.
My Networks	Contains the biological networks/pathways that you found using Network Search or that you create using a literature search in eArrayXD. For more information, see the *eArray$_{XD}$ Users Guide*
Genotypes	Shows the genotype references in the database. From this pane, you can import genotype reference files, and show properties, rename, or delete genotype references.

Search pane

The Search pane lets you find all occurrences of a specific search term in the Data, Experiment, and/or My Entity List panes. See "To find specific items in the Navigator" on page 71. It also contains several buttons that you can use to move, hide, show or resize the Navigator.

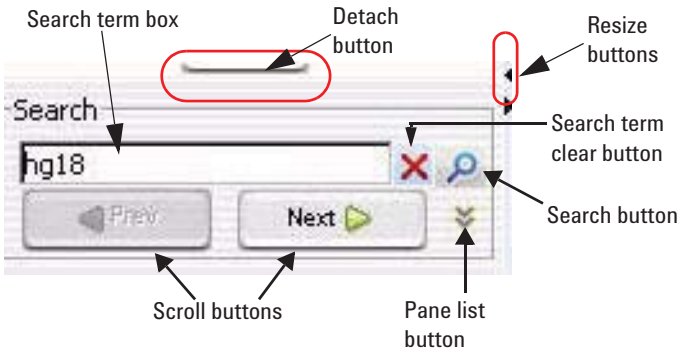


Figure 54 Navigator – Search pane

Detach button Click to move the Navigator from the main window of the program and open it in a new, separate window.

Resize buttons Click to hide, show, or expand the Navigator.

Search term box The place for you to type your desired search term. Search terms are not case-sensitive, but they must reflect the entire name of an array or other content item to find. You can use asterisks (*) as wildcard to represent groups of unspecified characters. For example, type *12345* to search for any content that contains the string "12345".

Pane list Lets you limit a search to a specific pane. Select the name of the desired pane from the list. To select all panes, select **All Panels.** By default, the program searches all panes.

(Show Pane List button, available only if the Pane list is not visible) Makes the Pane list visible.

(Hide Pane List button, available only if the Pane list is visible) Hides the Pane list.



(Search button) Searches the pane(s) selected in the Pane list for all occurrences of the term you typed in the Search term box. If the program finds a matching item, it expands the folder structure to make the matching item(s) visible, makes the lettering of each item red and highlights the current item in yellow. Note: The search term is not case-sensitive, but it must contain the entire name of the desired items. You can use asterisks (*) as wildcards to represent groups of unspecified characters.

Scroll buttons (Available only after a search) Lets you scroll up and down the lists of highlighted search items after a search.



(Clear button, available only after a search) Clears the search term from the Search term box, and resets the color of any matching item to its original color.

Data pane – icons, special text, and buttons

Item	Comments
	An unexpanded folder (domain) that contains subfolders or other items.
	An expanded folder. The items that it contains are visible in the Navigator.
	Expands a folder to show its contents.
	Collapses a folder to hide its contents.
	A methylation array design. This folder contains array data for the design, organized by genome build.
	A CGH array design. This folder contains array data for the design, organized by genome build.
	A CGH+SNP array design. This folder contains array data for the design, organized by genome build.
	A gene expression array design. This folder contains array data for the design, organized by genome build.
	A ChIP array design. This folder contains array data for the design, organized by genome build.
	A genome build folder within a specific design folder. This folder contains arrays for the specific genome build and design.
	A single array data file.
	Data created from a multi-pack array.
red text	An item that matches the search term in a search.
	The current result when you search the Navigator. Click **Next** to highlight the next result, and **Previous** to highlight the previous one.
	(Dock out button) Moves the Data pane from the Navigator, and opens it in a, separate window.
	(Collapse button, available only if the Data pane is not collapsed) Collapses the Data pane, and shows its title bar at the bottom of the Navigator.
	(Expand button, available only if the Data pane is collapsed) Expands the Data pane.

Data pane – actions and shortcut menus

The Data pane of the Navigator shows available content items that are stored in the program for the selected application type, and any external content that you imported.

Data Folder

- Double-click any folder to expand or collapse it.
- Double-click an imported designs folder (ChIP, Expression, CGH, CH3) to display the imported designs for that data type.
- Double-click the name of a genome build folder to display imported arrays for that build.

Design Folder

- Right-click the name of a design folder to display these options:

Option	Description
Update from eArray	Updates the annotations for your array design from the eArray Web site. Agilent regularly updates annotations in eArray as new ones become available. See "To update probe annotation in design files" on page 72. **Note:** In order to use this function, you must enter your eArray Username and Password in the Miscellaneous tab of the User Preferences dialog box. See "User Preferences" on page 414.
Delete	Opens a Confirm dialog box. If you click **Yes,** the program permanently deletes the design and all arrays associated with it.

Genome Build Folder

- Right-click the name of a genome build folder to display these options:

Option	Description
Show Properties	Opens the Design Properties dialog box. See "Design Properties" on page 299.
Cyto-Report	Opens another menu with a list of cytogenetic report templates available in the program. Select a template, if available, to open the Run Cytogenetic Report dialog box. See "Run Cytogenetic Report" on page 388. After you select a location, the program generates PDF format reports for all arrays associated with the genome build. "To generate a cytogenetic report" on page 170.
QC Metrics	Opens the QC Metrics table for all arrays in the genome build. The QC Metrics Table is available only for Agilent arrays. See"QC Metrics Table" on page 378, and "To display QC metrics of arrays and set array QC status" on page 134.
Delete	Opens a Confirm dialog box. If you click **Yes,** the program permanently deletes all of the arrays in this genome build folder.

Individual Arrays

- Right-click the name of an array to display these options:

Option	Description
Show Properties	Opens the Microarray Properties dialog box. See"Microarray Properties" on page 367 and "To display or edit array attributes in an experiment" on page 68.
Cyto-Report	Opens another menu with a list of cytogenetic report templates available in the program. Select a template, if available, to open the Run Cytogenetic Report dialog box. See "Run Cytogenetic Report" on page 388. After you select a location, the program generates PDF format reports for the array(s) you have selected. "To generate a cytogenetic report" on page 170.
QC Metrics	Opens the QC Metrics table for the array(s) you have selected. The QC Metrics Table is available only for Agilent arrays. See"QC Metrics Table" on page 378, and "To display QC metrics of arrays and set array QC status" on page 134.

Option	Description
Rename	Opens an Input dialog box, where you can type a new name for the array. Click **OK** to rename the array.
Delete	Opens a Confirm dialog box. If you click **Yes,** the program permanently deletes the array.

- Drag an array from the Data pane to an experiment folder in the Experiment pane to associate it with an experiment. You can drag multiple arrays at once from one genome build in a design. Hold down the **ctrl** key and click the additional arrays to select them. You can also select a contiguous block of arrays; click the first array in the block, then hold down the **shift** key and click the last one.

Experiment pane – icons, special text, and buttons

Item	Comments
⊞	Click to expand a folder and display its contents.
⊟	Click to collapse a folder and hide its contents.
	A folder that contains files or other folders. A folder with a W means that it contains workflow results. See the *Workflow User Guide*.
CH3	A methylation array design. This folder contains array data for the design, organized by genome build.
CGH	A CGH array design. This folder contains array data for the design, organized by genome build.
CGH+	A CGH+SNP array design. This folder contains array data for the design, organized by genome build.
EXP	A gene expression array design. This folder contains array data for the design, organized by genome build.
CHIP	A ChIP array design. This folder contains array data for the design, organized by genome build.
Build	A genome build folder within a specific design folder. This folder contains arrays for the specific genome build and design.

Item	Comments
	An array that is not selected for view and analysis.
	An array that is selected for view and analysis. The specific color of this icon can vary.
	A calibration array.
	An empty folder.
	Data created from a multi-pack array.
blue text	The active experiment. All data and results that appear in Chromosome, Gene, and Tab Views are from this experiment.
red text	An item that matches the search term in a search.
	(Dock out button) Moves the Experiment pane from the main window, and opens it in a, separate window.
	(Collapse button, available only if the Experiment pane is not collapsed) Collapses the Experiment pane, and shows its title bar at the bottom of the Navigator.
	(Expand button, available only if the Experiment pane is collapsed) Expands the Experiment pane.

Experiment pane – actions and shortcut menus

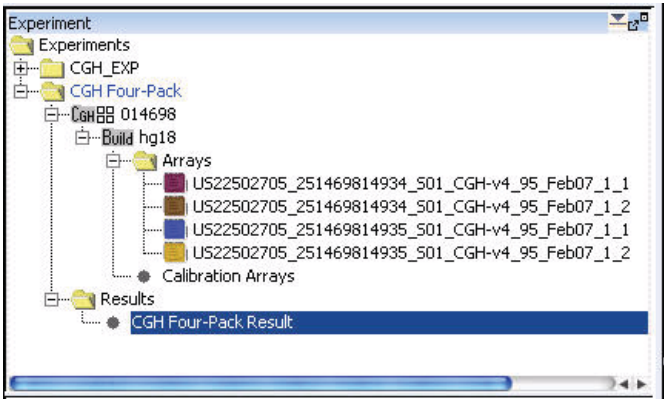


Figure 55 Expanded Experiment Pane

You can use many items in the Experiment Pane of the Navigator to open shortcut menus or take other actions.

- In general, double-click the Experiments folder within the Experiment Pane, and the folders within it, to expand and collapse them. Exception: Double-click the name of an unselected experiment to select it for analysis. This action opens the Experiment Selection dialog box. To select the experiment for analysis, click **Yes.**

NOTE The options that are displayed change depending on the user and status of the designs, builds, and arrays. You may not see all of the options that are described below.

Experiments Folder

- Right-click the **Experiments** folder to display these options:

Option	Description
New Experiment	Opens the Create Experiment dialog box (see "Create Experiment" on page 278), where you can name the new experiment, and open another dialog box that lets you add microarray data to the experiment. See "To create a new experiment" on page 60.
Export	Opens the Export Experiments dialog box, where you can export one or more experiments as a single ZIP file. See "Export Experiments" on page 320 and "To export experiments" on page 83.

Specific Experiment Folder

- In the Experiments folder, right-click the name of a specific experiment folder to display these options:

Option	Description
Select Experiment	(Appears only if the experiment is not selected.) Opens the Experiment Selection dialog box, which asks if you want to select the experiment. Click **Yes** to select the experiment for display and analysis. Or, in the Experiments folder, double-click the name of an experiment that is not selected to open the Experiment Selection dialog box. To select the experiment for analysis, click **Yes.** If you switch experiments, a Confirm dialog box asks if you want to save the current result. Select one of these options: • **Overwrite Current Result** – Replaces the selected experiment result in the Navigator with the result that appears on your screen. • **Create New Result** – Opens the Save experiment result dialog box, where you can save the result that appears in the main window as a new experiment result. See "To save a result" on page 145. • **Continue Without Saving** – The program does not save the result that appears on your screen. In some cases, the Confirm dialog box offers only Yes and No choices. If you click **Yes,** the Save experiment result dialog box appears, where you can save the result that appears on your screen with the name of your choice.
Deselect Experiment	(Appears only if the experiment is selected.) If the results are unsaved, a Confirm dialog box opens with these options: • **Overwrite Current Result** – Replaces the selected experiment result in the Navigator with the result that appears on your screen. • **Create New Result** – Opens the Save Experiment result dialog box, where you can save the result that appears on your screen as a new experiment result. • **Continue Without Saving** – The program does not save the result that appears on your screen. In all three cases, the program then removes the experiment data and results from all views.

Option	Description
Save Experiment Result	Opens a dialog box that asks if you want to save the results of the current experiment. When you click OK, one of three dialog boxes open: • If you have not yet saved a result for the experiment, the Save experiment result dialog box opens, where you can type a name for the experiment result. • If you have previously saved the result, the confirm dialog box appears, which lets you overwrite the current result, create a new result, or discard any changes. See "Confirm" on page 267. • If you have previously saved the result, and no changes have occurred, a message informs you that the current experimental condition is the same as an existing one.
Show Properties	Opens the Experiment Properties dialog box. Use this dialog box to see the names of the arrays in the experiment, and to add or remove arrays from the experiment. See "Experiment Properties" on page 313.
Sample Attributes	Opens the Sample Attributes dialog box, where you can display and change the values for the attributes assigned to the arrays in the experiment. Changes made are applied globally. To add new attributes, you must use the Sample Manager. See the *Sample Manager User Guide* for information. See "Sample Attributes" on page 390.
Export	Opens the Export Experiments dialog box, where you can export this and other experiments as a single ZIP file. See "Export Experiments" on page 320, and "To export experiments" on page 83.
Export Attributes	Opens the Export Array Attributes dialog box, where you can save a file that contains selected attributes of the arrays in your experiment. See "Export Array Attributes" on page 316.
Cyto-Report	Opens another menu with a list of cytogenetic report templates available in the program. Select a template, if available, to open the Run Cytogenetic Report dialog box. See "Run Cytogenetic Report" on page 388. After you select a location, the program generates PDF format reports for all arrays associated with the experiment. "To generate a cytogenetic report" on page 170.
QC Metrics	Opens the QC Metrics table for all arrays in the experiment. The QC Metrics Table is available only for Agilent arrays. See"QC Metrics Table" on page 378, and "To display QC metrics of arrays and set array QC status" on page 134.

Option	Description
SNP QC Metrics	Available only for CGH+SNP experiments, opens the SNP QC Metrics Table, that displays QC Metrics for the CGH+SNP arrays in the experiment. See "SNP CN QC Metrics Table" on page 405.
Edit Array Color	Opens the Edit Array Color dialog box, where you can select a display color for each of the arrays in the experiment. See "Edit Array Color" on page 308.
Edit Array Order	Opens the Edit Array Order dialog box, where you can change the order of the arrays in the experiment pane of the Navigator, and in Chromosome, Gene, and Tab Views. See "Edit Array Order" on page 309.
Rename	Opens an Input dialog box, where you can type a new name for the experiment. Click **OK** to rename the experiment.
Delete	Opens a Confirm dialog box that asks if you want to delete the Experiment. Click **Yes** to delete it. Note: You can delete any experiment except the selected one.
Expand Node	Expands the selected node to display all folders and their contents.
Collapse Node	Closes all folders for the selected node.

Design Folder

- In the folder of a specific experiment, right-click the name of a design to display a menu with a Delete command. If you click **Delete,** a Confirm dialog box opens. Click **Yes** to disassociate all of the arrays under the design from the experiment.

Genome Build Folder

- In the folder of a specific experiment, in a specific design folder, right-click the name of a genome build to display these options:

Option	Description
Set for Calibration	Because it is not recommended for cgh to use another array to calculate noise for the sample array, this is not used.
Cyto-Report	Opens another menu with a list of cytogenetic report templates available in the program. Select a template, if available, to open the Run Cytogenetic Report dialog box. See "Run Cytogenetic Report" on page 388. After you select a location, the program generates PDF format reports for all arrays in the experiment that are associated with the specific genome build and design. "To generate a cytogenetic report" on page 170.
QC Metrics	Opens the QC Metrics Table for all arrays in the genome build. The QC Metrics Table is available only for Agilent arrays. See "QC Metrics Table" on page 378, and "To display QC metrics of arrays and set array QC status" on page 134.
Save As Text File	Opens the Save Design dialog box, where you can save all of the data for the genome build and design within the experiment as a tab-delimited text file.
Delete	Opens a Confirm dialog box that asks if you want to disassociate all arrays under the design from the experiment. Click **Yes** to remove the links between the arrays and the experiment. Note: • If you delete a design from an experiment, the program removes the links between the experiment and the design and its arrays. The actual design and array data stay in the Data folder. • Saved results become unavailable if they include arrays you delete with this command.

Individual Arrays

- In a specific experiment, right-click the name of an individual array to display these options:

Option	Description
Select	(Available only if the array is not already selected) Selects the array for display and analysis.
Deselect	(Available only if the array is selected) Removes the array data from Genome, Chromosome, and Gene views, and excludes it from the analysis. Also removes the array data from the Selected Arrays tab in Tab View.
Select for Calibration	(Available only in the Arrays folder.) Designates the array as a calibration array. Within the specific experiment, the program moves the name of the array to the Calibration Arrays folder. Within the specific design tab in Tab View, the program also moves the data for the array from the Arrays tab to the Calibration Arrays tab.
Deselect from Calibration	(Available only within the Calibration Arrays folder.) Designates the array as a "regular" non-calibration array. Within the specific experiment, the program moves the name of the array to the Arrays folder. Within the specific design tab in Tab View, the program also moves the data for the array from the Calibration Arrays tab to the Arrays tab.
Rename	Opens an Input dialog box, where you can type a new name for the array in this experiment. Click **OK** to accept the new name for the array. The array name is changed only for the selected experiment.
Delete	Opens a Confirm dialog box that asks if you want to disassociate the array from the experiment. Click **Yes** to remove the link between the array and the experiment. **Note:** • If you delete an array from an experiment, the program removes the link between the experiment and the array. The actual array data remains in the Data folder. • You cannot restore an experiment result that includes a deleted array.

Option	Description
Show Properties	Opens the Microarray Properties dialog box, where you can view and edit microarray attributes. See "Microarray Properties" on page 367. For array files from the Agilent Feature Extraction program, you can also view the headers and feature data from the file. See "To display or edit array attributes in an experiment" on page 68.
Cyto-Report	Opens another menu with a list of cytogenetic report templates available in the program. Select a template, if available, to open the Run Cytogenetic Report dialog box. See "Run Cytogenetic Report" on page 388. After you select a location, the program generates a PDF format report for the array. See "To generate a cytogenetic report" on page 170.
QC Metrics	Opens the QC Metrics Table for the array. The QC Metrics Table is available only for Agilent arrays. See "QC Metrics Table" on page 378, and "To display QC metrics of arrays and set array QC status" on page 134.
Edit Array Color	Opens the Select Color dialog box, where you can select a display color for the array. See "Select Color" on page 394.
Edit Array Order	Opens the Array Order dialog box, where you can change the order of the arrays in the Experiment pane of the Navigator, and in Chromosome, Gene, and Tab Views. See "Edit Array Order" on page 309.

Results Folder

- In the **Results** folder of an experiment, right-click the name of a saved result to display these options:

Option	Description
Restore result	Replaces the result that appears in Genome, Chromosome, Gene, and Tab Views with the saved result. See "To restore a saved result" on page 146. The experiment associated with the saved result must be the selected experiment.
Rename	Opens an Input dialog box. Type a new name for the result, then click **OK**.

Option	Description
Delete	Opens a Confirm dialog box that asks if you want to delete the result. Click **Yes** to delete the result from the experiment.
Show Properties	Opens a Properties dialog box that you can use to display or edit a description of the result, and to display other attributes of the result. See "Properties (of an experiment result)" on page 377.

- In the Results folder of an experiment, in the CNVR folder of a specific result, right-click the name of a CNVR to display these options:

Option	Description
View Details	Opens the Copy Number Variant Region dialog box, where you can view the attributes of the CNVR. See "Copy Number Variant Region" on page 269.
Show CNVR(s)	(Available only if the result that contains the CNVR(s) is selected, and the CNVR(s) do not already appear in Genome, Chromosome, and Gene Views.) Displays CNVR(s) in Genome, Chromosome, and Gene views. Depending on the view and magnification, a CNVR appears as a green line or box.
Hide CNVR(s)	(Available only if the result that contains the CNVR(s) is selected, and the CNVR(s) appear in Genome, Chromosome, and Gene views.) Removes the CNVR(s) from Genome, Chromosome and Gene views.
Compare CNVRs	Opens the Compare CNVRs dialog box, where you can select a track. The program compares the CNVR result with the selected track, See "Compare CNVRs" on page 263. The program creates a *.xls file that contains the comparison results. You can open this file in Microsoft Excel.
Generate Text Summary	Opens the Select Report File dialog box, where you can select a location for the CNVR text summary. The program creates a *.xls format file that you can open in Microsoft Excel.
Create Gene List	Opens the Create Gene List dialog box, where you can provide a name for the gene list, type a description, and choose a gene list color. See "Create Gene List (from CNVR results)" on page 281. The program creates a gene list that contains the genes found within the CNVR(s).
Rename	Opens an Input dialog box, where you can type a new name for the set of CNVR(s). Click **OK** to accept your edits. The program renames the CNVR.

Option	Description
Export	Opens an Export dialog box, where you can select a location for the exported CNVR. The program exports CNVRs as a BED format track file.
Delete	(Available only if the specific CNVR(s) do not appear in Genome, Chromosome, and Gene views.) Opens a Confirm dialog box that asks if you want to delete the nodes. To delete the CNVR(s), click **Yes.**

- In the **Results** folder of an experiment, in the **Common Aberrations** folder of a specific result, right-click the name of a common aberration node to display these options:

Option	Description
View Details	Opens the Common Aberration dialog box, where you can view the attributes of the specific common aberration result. See "Properties (for common aberration result)" on page 375.
Show Common Aberration(s)	(Available only if the result that contains the common aberrations is selected, and the common aberrations do not already appear in Genome, Chromosome, and Gene Views.) Displays the specific common aberrations in Genome, Chromosome, and Gene views. See "To display common aberrations" on page 152.
Hide Common Aberration(s)	(Available only if the result that contains the common aberrations is selected, and the common aberrations appear in Genome, Chromosome, and Gene views.) Removes common aberrations from Genome, Chromosome, and Gene views. See "To display common aberrations" on page 152.
Generate Text Summary	Opens the Select Report File dialog box, where you can select a location for the common aberration text summary file. The program generates the report in *.xls format. You can open the file in Microsoft Excel. See "To create CGH aberration text reports" on page 166.
Rename	Opens an Input dialog box, where you can type a new name for the specific common aberration node. Click **OK** to accept your edits. The program changes the name.
Delete	(Available only if the specific common aberrations do not appear in Genome, Chromosome, and Gene views.) Opens a Confirm dialog box, that asks if you want to delete the nodes. If you click **Yes,** the program removes the specific common aberration node from the Navigator.

My Entity List pane – Icons, buttons, and special text

Item	Comments
	Click to expand a folder and display its contents.
	Click to collapse a folder and hide its contents.
	A folder that contains files or other folders.
	An individual gene list or track.
red regular text	An item that is an exact match with the search term in a search, or a gene list that has not been applied or that has red chosen as its custom color.
colored italics	A gene list that has been applied.
red bold italics	A track that is selected for display in Gene View.
black bold italics	A "combined" track that is selected for display in Gene View. A combined track contains information from two or more individual tracks associated by logical criteria.
	(Dock out button) Moves the My Entity List pane from the main window, and opens it in a, separate window.
	(Collapse button, available only if the My Entity List pane is not collapsed) Collapses the My Entity List pane, and shows its title bar at the bottom of the Navigator.
	(Expand button, available only if the My Entity List pane is collapsed) Expands the My Entity List pane.

My Entity List pane – actions and shortcut menus

Gene List folder

- Right-click the **Gene List** folder to display an **Import Gene List** option. This command opens an Import dialog box that you can use to import a gene list into the program. See "To import a gene list" on page 75 and "Import" on page 354.
- Double-click the **Gene List** folder to show or hide its gene lists.

- In the **Gene List** folder, right-click the name of a gene list to display these options:

Option	Description
View In Table	Opens the Gene List dialog box, where you can display the list of genes. You can also edit the description of the gene list, and change the display color of the genes. See "Gene List" on page 328 and "To display the genes in a gene list" on page 75.
Rename	Opens an Input dialog box, where you can type a new name for the gene list. Click **OK** to accept the new name.
Delete	Opens a confirm dialog box that asks if you are sure you want to delete the gene list. Click **Yes** to confirm.
Save As	Opens a Save As dialog box, where you can save the gene list as a text (*.txt) file. See "To export a gene list" on page 84.
Add to Gene List	Opens the Add gene list dialog box, where you can add the gene list to any other one in the Gene List folder. See "Add Gene List <name> to" on page 243 and "To add one gene list to another" on page 76.
Highlight	(Available if the gene list is not selected.) Displays all genes in Gene View, and highlights the genes from the gene list in their display color. See "To show gene lists in Gene View" on page 109.
Show Only	(Available only if all genes appear in Gene View, or if the gene list is not selected) Limits the genes in Gene View to those on the gene list. No other genes appear. The program displays the genes in their display color. See "To show gene lists in Gene View" on page 109.
Show All	(Available only for the selected gene list.) In Gene View, displays all genes, without highlighting. See "To show gene lists in Gene View" on page 109.

Tracks folder

- Right-click the Tracks folder to display these options:

Option	Comments
Import Tracks	Opens the Import Track dialog box, where you can import a BED format track file into the program. See "Import Track" on page 360 and "To import tracks" on page 55.
Export Tracks	Opens the Export Tracks dialog box, where you can select tracks for export as a single BED format track file. See "Export Tracks" on page 322 and "To export tracks" on page 84.
Combine Tracks	Opens the Combine Tracks dialog box, where you can associate two or more individual tracks by logical criteria to create a new combined track. See "Combine Tracks" on page 261 and "To combine tracks" on page 78.

Track name

- Right-click the name of a track to display these options:

Option	Comments
Show In UI	Mark this option to display the track in Gene View next to the data and results of the selected experiment. See "To show tracks in Gene View" on page 111.
Show in Report	Mark this option to display the track in the reports.
Genomic Boundaries	Click to use the genome track to limit the regions that aberration detection algorithms will process. You can choose to do this for only one track.
Show In UCSC	Opens the UCSC Genome Browser in your Web browser and uploads the track. You can then display the track.
View Details	Opens a table that displays all the chromosome locations defined in the track. See "Track" on page 409.
Rename	Opens an Input dialog box, where you can type a new name for the track. Click **OK** to rename the track.
Delete	Opens a Delete Track dialog box that asks if you are sure you want to delete the track. Click **Yes** to delete the track.

Genotypes pane

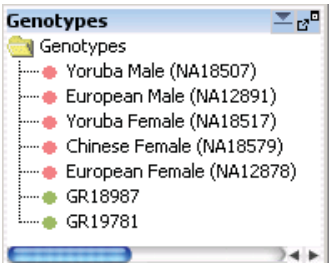


Figure 56 Genotypes pane of the Navigator

In order to perform SNP analysis, there must be genotype references in the database for the CGH+SNP microarrays you want to analyze. The pre-loaded and imported genotype references in the database are displayed in this pane, along with genotype references you imported.

Genotypes pane – Actions and shortcut menus

- Right-click the Genotypes folder and click **Import Genotype Reference** to select a genotype reference file to import to the database.
- Right-click on a genotype reference in the list, and select from the following options:

Genomic Viewer

Genomic Viewer is the display for the Agilent Genomic Workbench modules. It includes the three Views – Genome, Chromosome and Gene Views – the Tab View and the View Cursor.

Genome View

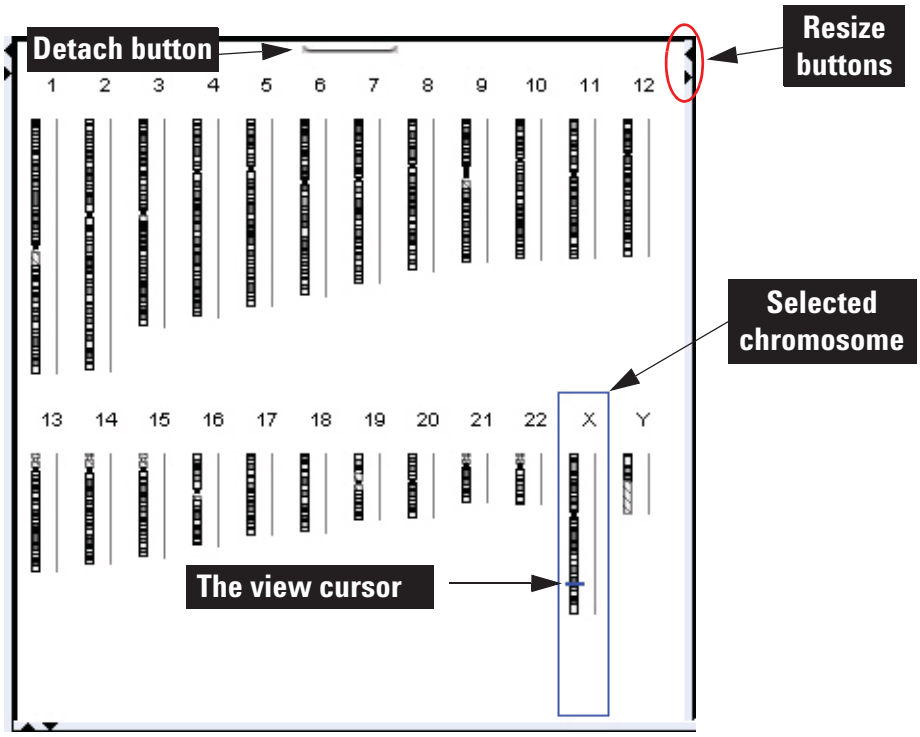


Figure 57 Genome View (vertical orientation), with human chromosomes. The X chromosome is selected.

Genome View shows pictures of each of the distinct types of chromosomes in the selected genome. A blue box encloses the selected chromosome, and the cursor appears as a blue line across the chromosome.

Genome View actions and shortcut menus

- Click a chromosome within the Genome View to select it. When you select a chromosome, Chromosome, Gene, and Tab Views show only genomic regions, genes, and data associated with it. The specific location in which you click the chromosome sets the position of the cursor. See "The View Cursor" on page 232.
- On the selected chromosome, click anywhere to reposition the cursor. See "The View Cursor" on page 232. This also repositions the cursor in Chromosome, Gene, and Tab Views.
- Right-click anywhere within Genome View to display a menu. If you click **View Preferences,** the View Preferences dialog box opens, where you can set preferences for the display. See "View Preferences" on page 422.
- Click the **Detach** button (located at the top center of the pane) to remove Genome View from the main window and open it in a separate window. To reattach the view, click its **Close** button  .
- Drag the side or bottom borders of the pane to resize it.
- On a border of the pane, click a resize button (for example,  or ) that points away from the pane to move that border all the way to the edge of the main window. To move the border back to its previous location, click the other resize button.

Chromosome View

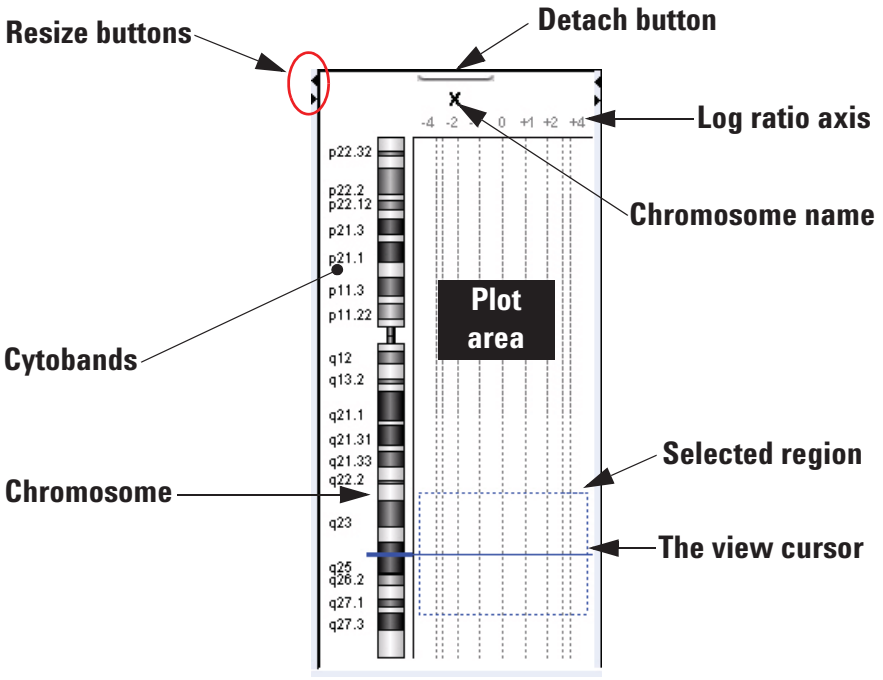


Figure 58 Chromosome View (vertical orientation), human X chromosome shown

Chromosome View shows a more detailed diagram of the chromosome you select in Genome View.

- Cytobands and a plot area appear next to the chromosome.
- When you select arrays for display, their data appear in the plot area.
- The cursor appears as a solid blue line across the chromosome and the plot area.
- The selected region of the chromosome (if any) appears as a dotted blue box in the plot area.

Chromosome View actions and shortcut menus

- Click a cytoband, any part of the chromosome, or anywhere in the plot area to reposition the cursor at that location. See "The View Cursor" on page 232.
- Drag the pointer over any part of the plot area to select a chromosomal region for display in Gene View. Drag parallel to the chromosome. This also repositions the cursor to the center of the selected region. See "The View Cursor" on page 232.
- Right-click anywhere within Chromosome View to display a menu. If you click **View Preferences,** the View Preferences dialog box opens, where you can set preferences for the display. See "View Preferences" on page 422.
- Click the **Detach** button (located at the top center of the pane) to remove Chromosome View from the main window and open it in a separate window. To reattach the view, click its **Close** button  .
- Drag the side or bottom borders of the pane to resize it.
- On a border of the pane, click a resize button (for example,  or ) that points away from the pane to move that border all the way to the edge of the main window. To move the border back to its previous location, click the other resize button.

Gene View

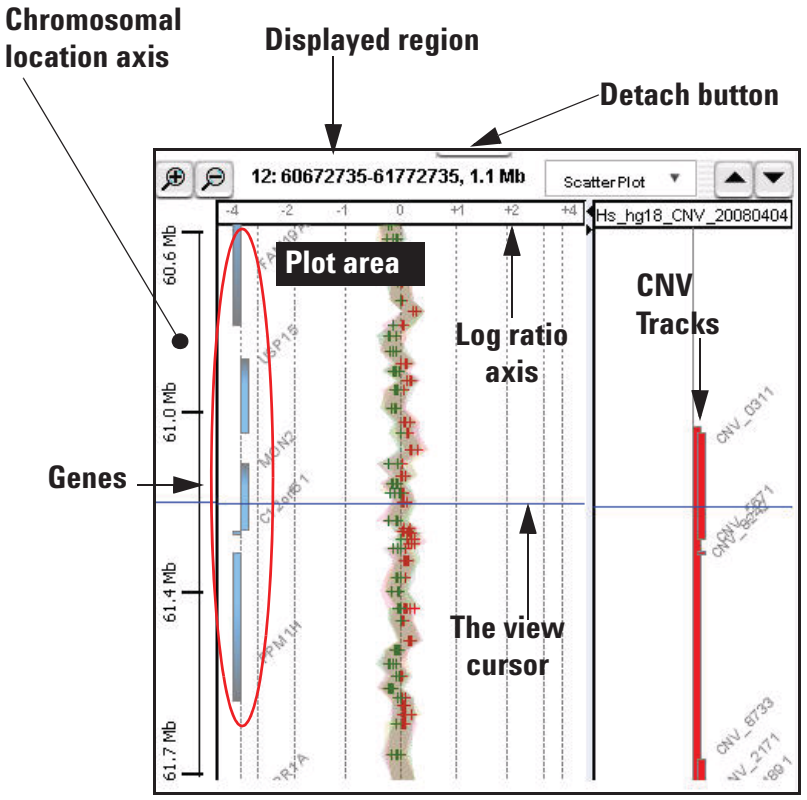


Figure 59 Gene View (vertical orientation), with log ratio data from an experiment and CNV tracks

Gene View shows a more detailed view of the chromosomal region you select in Chromosome View. See "Chromosome View" on page 226.

- Regions occupied by genes appear as small blue boxes. Gene names appear nearby. You can customize the appearance of gene names. Also, you can use a gene list to highlight genes of interest, or to limit the genes that appear to those in the list. See "To change the appearance of genes in Gene View" on page 111, and "To show gene lists in Gene View" on page 109.
- Log ratio data from selected arrays in the active experiment appear as a scatter plot. You can also customize the scatter plot. See "To change scatter plot appearance" on page 101.
- The location of the cursor matches the location of the cursors in other views. See "The View Cursor" on page 232.
- The name of the chromosome, and the coordinates and size of the displayed chromosomal region appear at the top of the pane.
- Cytobands can also be displayed in Gene View.
- Imported tracks can also be displayed in Gene View. See "To show tracks in Gene View" on page 111.

Scatter Plot

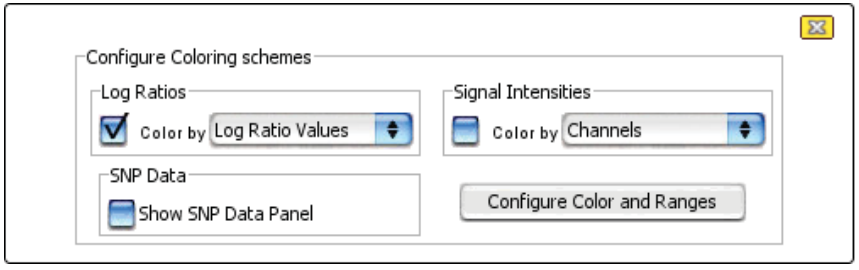


Figure 60 Scatter Plot command group in CGH Gene View

You access the scatter plot command group in Gene View or View Preferences from the View tab. The selections differ for each Agilent Genomic Workbench module. Scatter plot data appear in the Chromosome and Gene Views, but only if they have been selected under Data Visibility in the View Preferences dialog box. See "View Preferences" on page 422.

Log Ratios Mark the box to enable the Log Ratios scatter plot. Choices for the plot are Log Ratio Values or Probe Score Values.

Signal Intensities Mark the box to enable the Signal Intensities scatter plot. Selections for the plot are Channels, Probe Score Values, or Intensity Values.

SNP Data Mark the box to enable the SNP data panel that shows copy number and LOH data for CGH+SNP microarrays.

Configure Color and Ranges Opens the Configure Coloring Ranges and Shades dialog box, where you can set up the colors and ranges for Primary and Secondary scatter plots. For more information, see "Configure Coloring Ranges and Shades" on page 264.

Gene View buttons


Zooms in to see a smaller region in more detail.


Zooms out to see a larger region in less detail.


When in vertical orientation, scrolls up through the genes and data to lower-numbered chromosomal coordinates.


When in vertical orientation, scrolls down through the genes and data to higher-numbered chromosomal coordinates.


When in horizontal orientation, scrolls left through the genes and data to lower-numbered chromosomal coordinates.


When in horizontal orientation, scrolls right through the genes and data to higher-numbered chromosomal coordinates.


(**Resize** buttons) The button that points away from Gene View expands the view. The other button restores the view to its original size.(These buttons will appear horizontal if the view orientation is horizontal.)

(**Detach** button) Removes Gene View from the main window, and opens it in a separate window.

Gene View shortcut menu and other actions

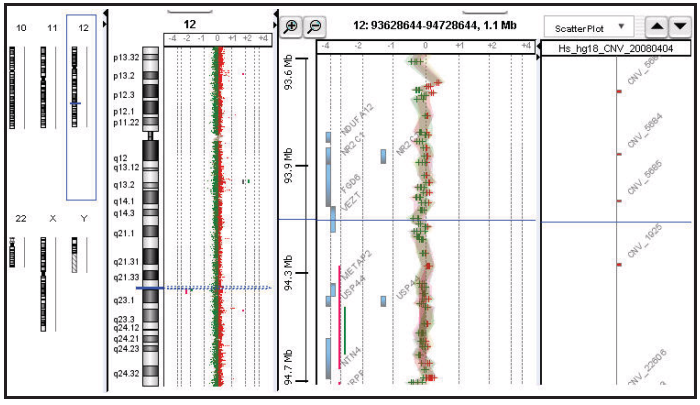
- Click anywhere in the plot area of Gene View to move the cursor to that location. See "The View Cursor" on page 232.
- Drag an inside border of Gene View to resize the view.
- Right-click anywhere in the plot area of Gene View to display these options:

Option	Description
Create Gene List	Opens the Create Gene List dialog box, where you can create a new gene list based on the selected (or another) chromosomal region. See "Create Gene List" on page 279 and "To create a gene list" on page 74.
Create Track	Opens the Create Track dialog box, where you select the chromosome locations for the track. See "To create a track" on page 77 and "Create Track" on page 284.
Show Intensity Bar Charts	Opens the Create Signal Bar Chart dialog box, where you select parameters to create a signal intensity chart for the data. See "Create Signal Bar Chart" on page 283.
Show in UCSC	Opens the View Coordinates in UCSC Browser dialog box where you select track information for display in the UCSC (University of California at Santa Cruz) Genome Browser. You can then view the track.
Search probes in eArray	Opens the Search probes in eArray dialog box, where you can search for probes based on their chromosomal location. See "Search Probes in eArray" on page 392.
User Preferences	Opens the User Preferences dialog box, where you can set user preferences on three separate tabs. See "User Preferences" on page 414 and the related pages that follow.
View Preferences	Opens the View Preferences dialog box, where you set preferences for the Genomic Viewer. See "View Preferences" on page 422.

The View Cursor

The View cursor reflects the center of the current chromosomal location of interest. It appears in several Views:

- In Genome View, it appears as a blue bar across the selected chromosome.
- In Chromosome View, it is a blue bar that appears across the chromosome and across the plot area of the view.
- In Gene View, it is a blue bar that appears across the plot area and tracks of the view.



The position of the cursor in one view is also the position of the cursor in all views. The exact chromosomal location of the cursor appears in the first cell of the Status bar. Several actions affect the position of the cursor:

- In Genome View, click anywhere on a chromosome to move the cursor to that location.
- In Chromosome View, click a cytoband name, part of the chromosome, or anywhere in the plot area to move the cursor to that location.
- In Gene View, click anywhere in the plot area to move the cursor to that location.

  The cursor used in Gene View is the same cursor used for the tracks.
- In Tab View, click a row of a data table to move the cursor to the chromosomal location associated with that row.

Tab View

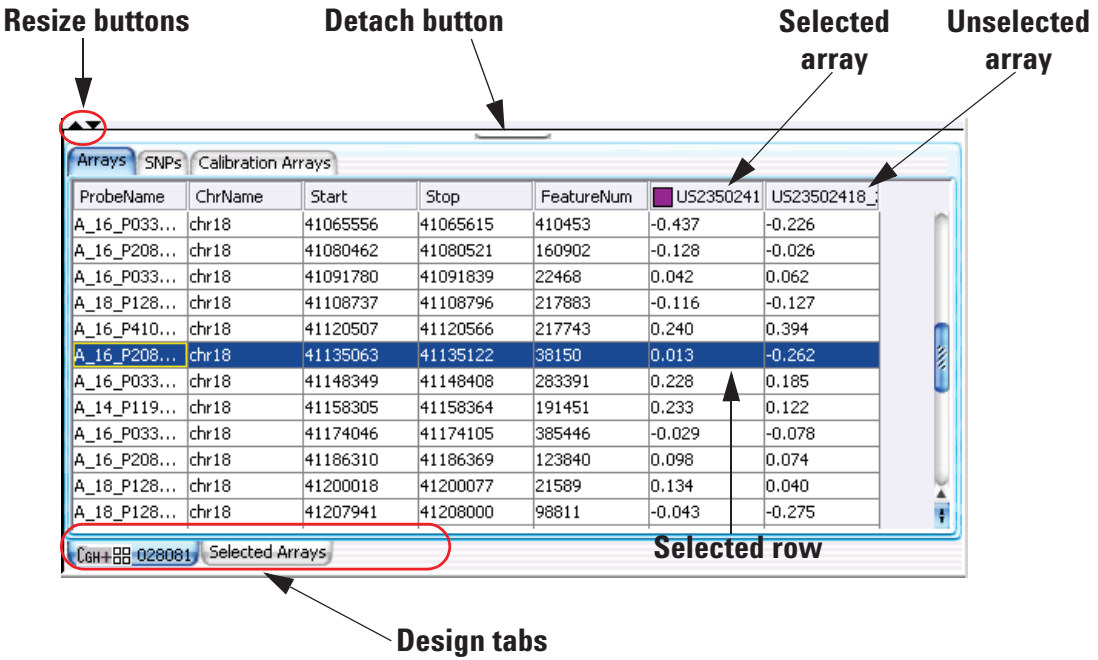


Figure 61 Tab View

Tab View displays design annotation and log ratio data related to the chromosome you select in Chromosome View. CGH probes are displayed in the Arrays tab. For CGH+SNP arrays, an additional SNPs tab is available with SNP data for the selected chromosome. This tab does not show results unless the SNP algorithms are selected and data is analyzed.

- The exact column content of the tables depends on the tab and design, but it always includes chromosomal locations of probes.
- The selected row of data is highlighted in blue. This row represents data that corresponds approximately with the location of the cursor.
- Columns of log ratio data appear below the names of the arrays to which they correspond. If an array is selected for display in Chromosome and Gene views, a colored square appears next to its name.

- Signal intensity (raw signals) and/or annotations appear if selected from the View command ribbon.
- SNPs display two probes, separated by a comma. This corresponds to a probe for each strand. Corresponding feature numbers are displayed, also separated by a comma.

Tab View tabs and buttons

You can see the following tabs and buttons in Tab View. See Figure 61 for a diagram that identifies some of these elements.

Design tabs A separate tab appears for each microarray design included in the active experiment. The name of the design appears on each tab, along with an icon:

- CH3 – A methylation array design
- CGH – An aCGH array design.
- CGH+ – A CGH+SNP array design.
- EXP – A gene expression array design.
- CHIP – A ChIP-on-Chip array design.

When you click a design tab, the data and annotation for the arrays in the design appear in Tab View. The program separates the arrays of the design into the Arrays tab and the Calibration Arrays tab (see below).

Arrays tab (Available when you click a specific design tab.) Contains a table of data and annotation for all arrays in a design that contain biological data.

SNPs tab (Available when CGH+SNP array is selected.) Contains a table of SNP data for the selected chromosome. Includes genotype information on a per-array basis.

Calibration Arrays tab Contains a table of arrays that are selected as calibration arrays. See "To select or remove calibration array(s)" on page 66.

Selected Arrays tab Contains a table of data and annotation for the selected arrays from all designs in the active experiment.

(**Resize** buttons) The button that points away from Tab View expands the view. The other button restores the view to its original size.

(**Detach** button) Removes Tab View from the main window, and opens it in a separate window.

Tab View actions and shortcut menus

- Click the *name of an array in a column heading* to select the array for display.
- Right-click the *name of an array in a column heading* to display these options:

Option	Description
Rename Array	Opens an Input dialog box, where you can type a new name for the array. This only changes the name of the array within the active experiment.
Remove Array From Experiment	Opens a confirmation dialog box. Click **Yes** to remove the link between the array and the active experiment. This command does not delete the data file from the program. To do this, see "To remove data or design files from the program" on page 73.
Select Array	(Available if the array is not selected.) Selects the array for display. A colored square appears next to the name of the array.
Deselect Array	(Available if the array is selected.) Removes the array data from scatter plots, and removes the column of the array from the Selected Arrays tab.
Select for Calibration	Selects the array for calibration. Moves the selected array to the Calibration Arrays tab and to the Calibration Arrays folder in the Experiment pane. Calibration arrays in the Experiment pane are marked with a "C". See "To select or remove calibration array(s)" on page 66.
Edit Array Color	Opens the Select Color dialog box, where you can change the display color of the array. See "Edit Array Color" on page 308 and "To change the display color of an array" on page 93.
Edit Array Order	Opens the Edit Array Order dialog box, where you can change the order in which the names of the arrays in a given design of the active experiment appear in Tab View and in the Navigator. In Gene View, when you view separate scatter plots for each array, the plots also appear in this order. See "Edit Array Order" on page 309 and "To change the order of arrays in an experiment" on page 64.
Select All Arrays	Selects all arrays in all designs in the active experiment for display. All arrays appear in the Selected Arrays tab.

Option	Description
Deselect All Arrays	Removes all arrays from display, and from the Selected Arrays tab.
Select All Arrays for Calibration	Selects all arrays in the table as calibration arrays. Moves the selected arrays to the Calibration Arrays tab and to the Calibration Arrays folder in the Experiment pane. Calibration arrays in the Experiment pane are marked with a "C".
Deselect All Arrays from Calibration	Removes all calibration arrays from the Calibration Arrays tab and Calibration Arrays folder in the Experiment pane.
Scroll To Column	Opens the Scroll to Column dialog box, where you can select a column in the current tab. The program then scrolls the data table in the tab so you can see the selected column. "Scroll to Column" on page 391.

- Right-click a *heading of a column other than an array data column* to display a Scroll To Column option. If you click this option, the Scroll to Column dialog box appears, where you can select a column in the current tab. The program then scrolls the data table in the tab so you can see the column.
- Click an *entry in a data table* to select the row in which it appears. This also moves the cursor to the location of the data point that corresponds to the selected row.
- Right-click a *data table entry* to display these options:

Option	Description
Find in column	Opens the Find in column dialog box, where you can search for a specific text string within the column you clicked. "Find in column" on page 325.

Option	Description
Google LocusLink PubMed UCSC HG15(April '03) UCSC HG16(July'03) UCSC HG17(May'04) UCSC HG18(March'06) UCSC mm8(Feb'06) UCSC mm9(July'07) DGV(hg18) GO KEGG(HUMAN)	Opens your Web browser, and passes the column entry you clicked as a search string to the selected site. The UCSC links search the indicated University of California, Santa Cruz database related to the indicated genome build. See "To search the Web for information on probes in Tab View" on page 121.
Customize Link	Opens the Customize Search link dialog box, where you can create or edit a custom Web link that appears in this shortcut menu. When you click a custom link, the program opens your Web browser, and passes the column entry you clicked as a search string to the site. See "Customize Search Link" on page 286 and "To create a custom Web search link" on page 121.
(other options)	If other options appear in this shortcut menu, they are custom Web search links. Click them to open your Web browser, and pass the column entry you clicked as a search string to the site.

Status Bar

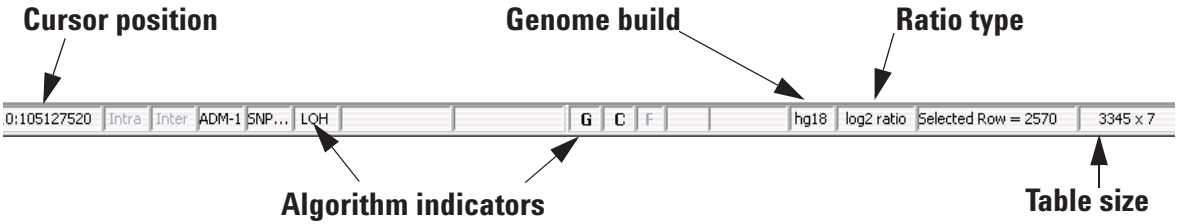


Figure 62 Status bar

The Status Bar displays information related to the displayed data.

Cursor position The chromosomal location of the cursor. See "The View Cursor" on page 232.

Algorithm indicators Calculations that are currently selected in the Preprocessing and Analysis tabs are shown in **bold**.

Genome build The genome build associated with the displayed data.

Ratio type The mathematical type of the array data. The possible types are:

- **ratio**
- **$\log_2$ ratio**
- **$\log_{10}$ ratio**
- **ln (natural log) ratio**

Selected Row The row in the displayed data table that is selected. The location of the cursor is approximately the chromosomal location associated with this row.

Table size The number of rows and columns in the displayed tab. The size appears as
<# of rows> x <# of columns>.

Dialog Boxes

This section describes the dialog boxes that can appear when you import, organize, manage, export and display array data/results and other content in Agilent Genomic Workbench, as well as those associated with CGH interactive analysis. The dialog boxes appear in alphabetical order by title.

Aberration & LOH Report Setup

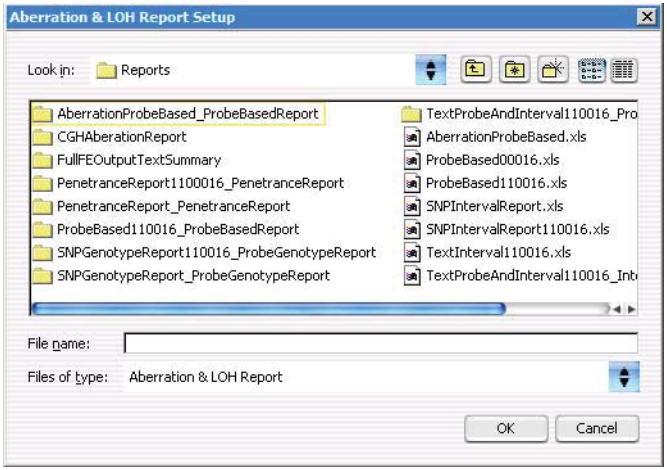


Figure 63 Aberration & LOH Report Setup dialog box

Purpose: Select a location and name for an Aberration & LOH Report and generate the report.

To open: In the Reports tab, under Aberration & LOH, click **Text**.

File name Type the name to use for the report file.

Files of type The program limits the list of files to the specific types expected for the report. To display all files, click ,then select **All Files.**

OK Creates the report, and saves it to the specified location.

Cancel Closes the dialog box without creating the report.

Aberration Filters

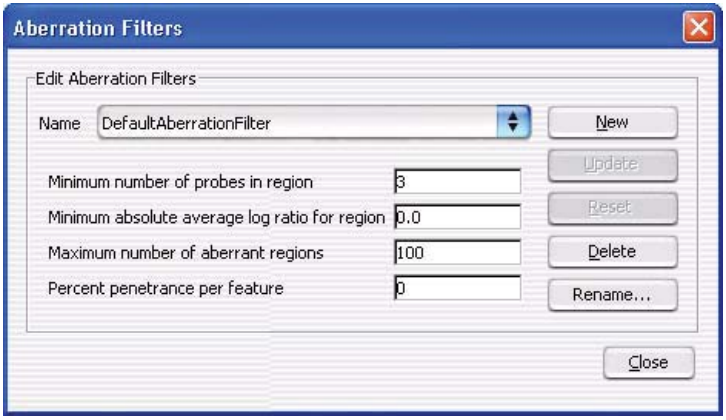


Figure 64 Aberration Filters dialog box

Purpose: This dialog box lets you create, change, or delete aberration filters. Aberration filters limit the aberrations that appear in the analysis results of the selected experiment, based on specific conditions. See "To create or modify an aberration filter" on page 150.

To open: Click **Discovery > Aberration Filters > Edit Filter.**

Name Select the name of the filter you want to change or delete. To create a new aberration filter, click **New.**

New Opens an Input dialog box, where you can type a name for the new aberration filter. To accept the name, click **OK.** The program creates the filter, and adds the new name to the Name list.

Update Saves any changes you make to the filter conditions.

Reset Restores the values of the filter conditions to what they were before you made any changes to them.

Delete Opens a Confirm dialog box that asks you if you want to delete the selected filter. To delete the filter, click **Yes.**

Rename Opens an Input dialog box where you can type a new name for the filter. To accept the name, click **OK.**

Minimum number of probes in region Type a whole number. The filter excludes putative aberrant regions that contain fewer probes than the number you type.

Minimum absolute average log ratio for region Type a value. The filter excludes putative aberrant regions if the average $\log_2$ ratio within the region is less than the value you type.

Maximum number of aberrant regions Type a whole number. For each microarray, the filter includes up to this number of aberrant regions in the genome that have the highest statistical significance.

Percent penetrance per feature Type a value. The filter excludes putative aberrations that have less than the specified minimum percent penetrance across the set of selected arrays.

Close Closes the dialog box. If you created or changed an aberration filter, but did not update it, a Confirm dialog box opens. Click **Yes** to accept the changes, **No** to reject the changes, or **Cancel** to return to the dialog box.

Aberration Heatmap

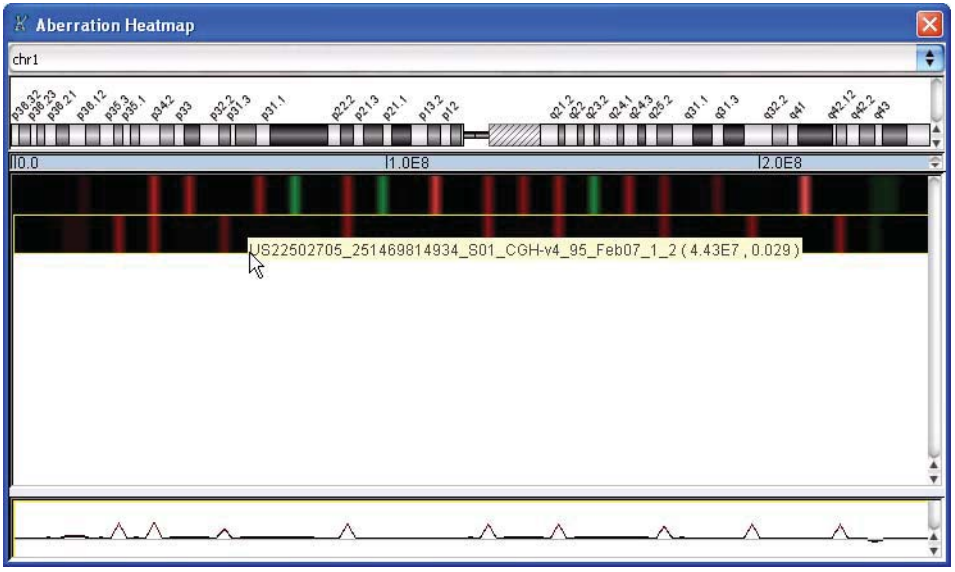


Figure 65 Aberration Heatmap dialog box

Purpose: The heatmaps correspond to the detected aberrations in the selected arrays of the active experiment. They are another way for you to visualize these aberrations. See "To display results as a heatmap" on page 162.

To open: click **Discovery > Heatmap.**

List of chromosomes Displays all of the chromosomes in the genome under study. When you select a chromosome, the ideogram for the chromosome appears in the dialog box, and the heatmaps associated with the selected chromosome appear.

Heatmaps Heatmaps appear for each selected array in the active experiment. Amplifications appear as red bars, and deletions as green bars. The magnitude of the aberration is reflected in the intensity of the bars.

ToolTip Place the pointer over a specific point in a heatmap to display a ToolTip that shows the name of the array associated with the heatmap, the chromosomal coordinates of the specific point, and the data value associated with that point.

- Click anywhere within a specific heatmap to display a line plot of the data for the array. To hide the line plot, click the heatmap again.

Line Plot(s) Plot(s) that represent the data from specific array(s). To display a line plot for an array, click its heatmap. To hide a line plot, click the specific heatmap again.

Drag the border between the heat map and line plot panes to change the relative size of the panes.

Add Gene List <name> to

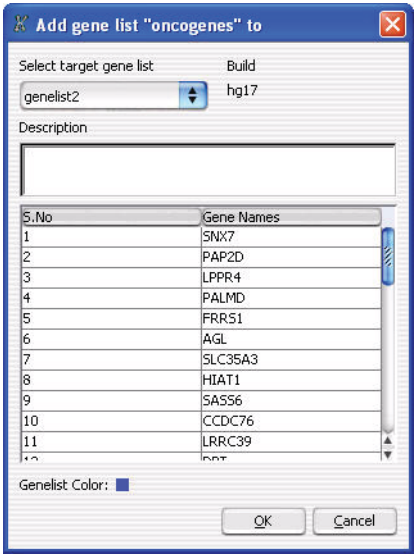


Figure 66 Add Gene List <name> to

Purpose: Adds genes from one gene list (the source gene list) to another (the target gene list).

To open: In the My Entity List pane, right-click the name of a gene list, then click **Add to Gene List.**

Select target gene list The gene list to which genes will be added. Select one from the list.

Build (Read-only) The genome build associated with the genes in the list. The builds of the two gene lists must match.

Description (Optional) Description of the combined gene list.

List of genes A list of the genes in the target gene list.

Gene List Color (Read-only) The current display color of the target gene list.

OK Adds the genes from the source gene list to the target gene list.

Cancel Closes the dialog box without adding any genes to the target gene list.

Agilent Feature Extraction Importer

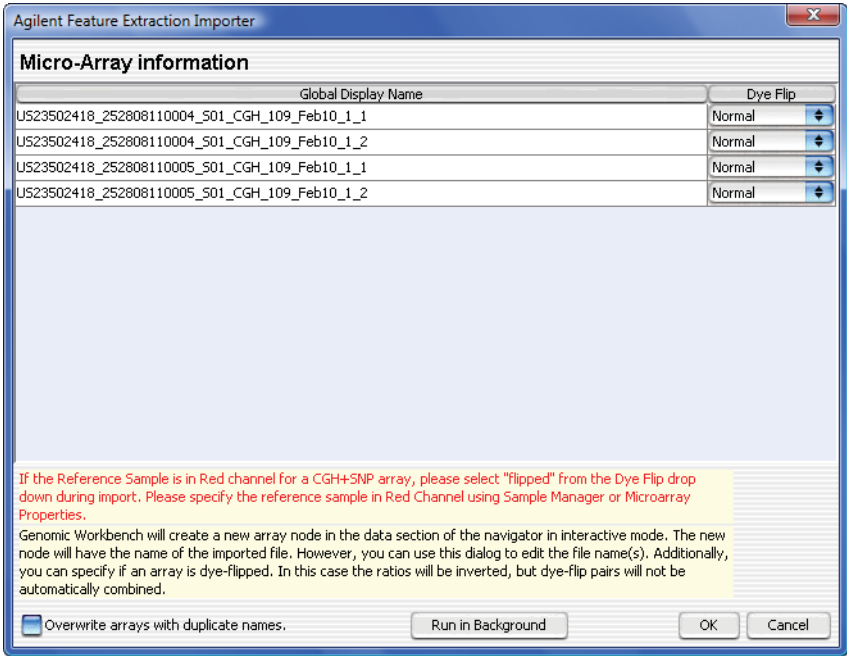


Figure 67 Agilent Feature Extraction Importer

Purpose: Lets you edit the name of the FE data file you intend to import, and to select if you want to flip the red/green ratio for the data.

To open: In the Home tab, click **Import > Array Files > FE File,** select the desired FE data file(s), then click **Open.**

Global Display Name Shows the global display name of each microarray to import. You can change the names of the files to names that you are more likely to recognize or remember.

Dye Flip For each array:

Select **Normal** if:

- The test samples were labeled with cyanine-5 (red).
- The control samples were labeled with cyanine-3 (green).

- The imported ratio (test/control) should be reported directly.

Select **Flipped** if:

- The test samples were labeled with cyanine-3 (green).
- The control samples were labeled with cyanine-5 (red).
- The imported ratio (control/test) should be reported with the ratio inverted (test/control).

The program does not combine dye-flip pairs.

Overwrite arrays with duplicate names Mark this option to replace existing file(s) in the program with the imported one(s), if they have the same name(s).

Run in Background Imports the files, and lets you use your computer for other purposes while the import occurs. This is especially useful if you have many files to import.

OK Imports the files in the foreground. You cannot use your computer for other purposes while the import occurs.

Cancel Cancels the entire import process without importing anything.

Array Level Filters

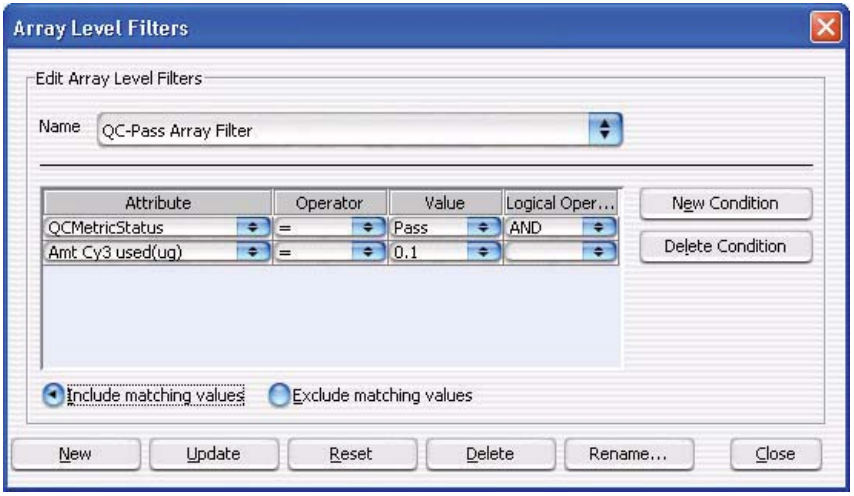


Figure 68 Array Level Filters dialog box

Purpose: This dialog box lets you create, change, or delete an array level filter. Array level filters include or exclude arrays from an analysis based on specific conditions. See "To create or modify an array filter" on page 126.

To open: Click **Preprocessing > Array Filters > Edit Filter.**

Name Select the name of the filter you want to change or delete. To create a new aberration filter, click **New.**

Filter conditions For each condition (row), select options from the lists. In **Value,** select an option from the list, if available. Otherwise, type a value, then click **Enter.** To add another row to the table, click **New Condition.**

Each condition has these elements:

- **Attribute** – The array attribute evaluated by the filter.
- **Operator** – How the filter uses the entry in Value to evaluate arrays. For example, the **>=** operator configures the filter to include or exclude arrays where the selected attribute is greater than or equal to the entry in Value.

- **Value** – The value the filter uses as the basis to evaluate an array. For example, if you select the **QCMetricStatus** attribute, the = operator, and a value of **Pass,** the filter includes or excludes an array if its QCMetricStatus is Pass.
- **Logical Operator** – (Available if you configure more than one condition) The relationship between the condition to the next one in the list. For example, if you select **AND** in **Logical Operator** for the first condition, the filter includes or excludes an array if it passes both the first condition and the next condition.

New Condition Adds a new, blank condition (row) to the table.

Delete Condition Removes a condition from the list. To remove a specific condition, click anywhere within the condition, then click **Delete Condition.**

Include/Exclude matching values Select one of these options:

- **Include matching values** – If an array passes the filter conditions, the program *includes* it in the analysis.
- **Exclude matching values** – If an array passes the filter conditions, the program *excludes* it from the analysis.

New Opens an Input dialog box, where you can type a name for the new filter. To accept the name, click **OK.** The program creates the filter, and adds the new name to the Name list.

Update Saves any changes you make to the filter conditions.

Reset Restores the values of the filter conditions to what they were before you made any changes to them.

Delete Opens a Confirm dialog box that asks you if you want to delete the selected filter. To delete the filter, click **Yes.**

Rename Opens an Input dialog box where you can type a new name for the filter. To accept the name, click **OK.**

Close Closes the dialog box. If you created or changed a filter, but did not update it, a Confirm dialog box opens. Click **Yes** to accept the changes, **No** to reject the changes, or **Cancel** to return to the dialog box.

Array Set

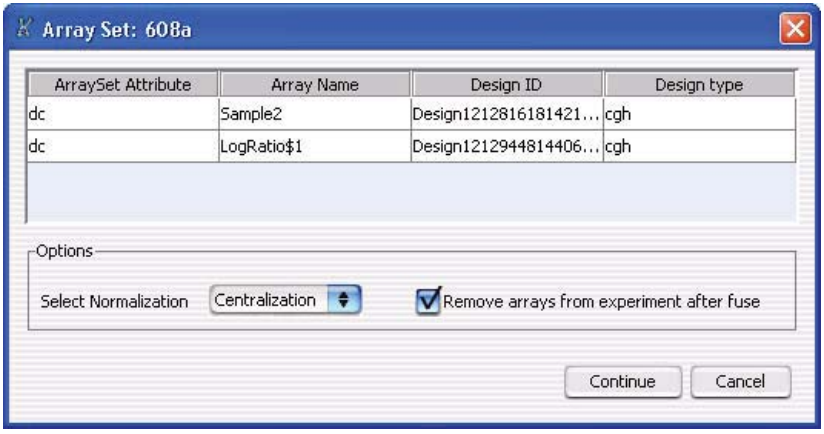


Figure 69 Array Set dialog box

Purpose: This dialog box lets you view the array designs to be fused, and to configure the details of the fusing process.

To open: Fuse two designs. See "To combine (fuse) arrays" on page 130.

Table (Read-only) Displays the arrays to be fused, arranged by their values for the ArraySet attribute. The program creates a separate fused array for each group of arrays with a given value for ArraySet. The table also shows the design ID associated with each array, and the design type.

Select Normalization Select Centralization normalization or none. See "Centralization Algorithm" on page 452.

Remove arrays from experiment after fuse Mark this option to delete the original unfused arrays after creating fused arrays. This reduces the duplication of data within the experiment.

Continue Click to create fused designs using the selected options.

Cancel Cancels any selections, and closes the dialog box.

CGH Aberration Report Setup

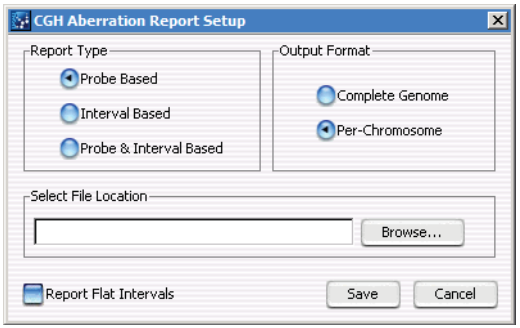


Figure 70 CGH Aberration Report Setup dialog box

Purpose: This dialog box allows you to configure the CGH Aberration Text Report, and select a location for it. This report details regions that have detected aberrations. You can report these regions by genomic interval, by probe, or both. The program reports aberrations separately for each selected array in the active experiment, and creates one or more *.xls format files that you can work with in Microsoft Excel.

To open: In the Reports tab, under CGH Aberration, click **Text.**

Report Type Configures the organization of reported aberrations. Select one of these options:

Option	Description
Probe Based	Creates a report that contains one line for each probe associated with an aberration.
Interval Based	Creates a report that contains one line for each aberrant genomic interval.
Probe & Interval Based	Creates both a probe-based report and an interval-based report.

Output Format Select one of these options:

Option	Description
Complete Genome	Creates a single report file for each requested report type.
Per-Chromosome	(Available only for probe based reports) Creates a separate report file for each chromosome.

Select File Location Displays the location to which the program saves the CGH Aberration Text report file(s). Click **Browse** in the Open dialog box that appears, type a name for the report(s) and select a location, then click **Open.**

Report Flat Intervals Reports aberration intervals without any nested structure.

Save Creates the CGH Aberration Text report, and saves it to the specified location.

Cancel Closes the dialog box without creating the CGH Aberration Text report.

CGHSmooth Parameters

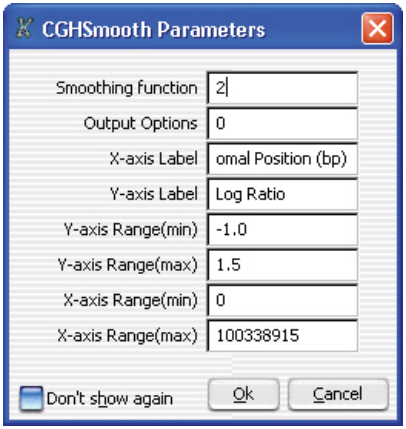


Figure 71 CGHSmooth Parameters dialog box

Purpose: This dialog box lets you configure the CGHSmooth plug-in. This plug-in creates separate, stacked plots of smoothed log ratio data for each of the selected arrays in the current experiment. It can perform one of several varieties of moving-average-like smoothing, and plots the data associated with the selected chromosome.

To open: The CGHSmooth Parameters dialog box opens when you click **Tool > Plugin > CGHSmooth.** It lets you configure the CGHSmooth plug-in.

Parameters Set any of these parameters:

Parameter	Description
Smoothing function	A number from 0 to 5. The number sets one of the following options as the weighting function used by the moving average algorithm. In general, the options weight measurements closer to the center position more heavily than those more distant from it. • 0 – **None.** The plug-in applies no smoothing, and returns the original data. In some cases, the plug-in averages data points with identical positions. This sets, in effect, a window size of 0. • 1 – **Rectangular.** The plug-in performs a standard moving average. All points within the rectangle (the window) receive the same weight. • 2 – **Gaussian.** Applies a Gaussian weighting function. • 3 – **Triangular.** Applies a triangular weighting function. • 4 – **Lorentzian.** Applies a Lorentzian weighting function. • 5 – **Biexponential.** Applies a biexponential weighting function.
Output Options	A number from 0 to 2. The number sets one of the following options: • **0** – Overlays the unsmoothed plot of each array on the smoothed plot. • **1** – Displays smoothed and unsmoothed plots for each array. • **2** – Displays smoothed, unsmoothed, and error plots for each array.
X-axis Label	The text that appears beneath the X-axis of the plot as a label.
Y-axis Label	The text that appears beside the Y-axis of the plot as a label.
Y-axis Range (min)	The minimum value on the Y-axis.
Y-axis Range (max)	The maximum value on the Y-axis.
X-axis Range (min)	The minimum value on the X-axis.
X-axis Range (max)	The maximum value on the X-axis.

Don't show again Mark this option to prevent the appearance of this dialog box in the future when you click **Tool > Plugins > CGHSmooth**. To restore the dialog box so it appears again, click **Tool > Plugin Settings > CGHSmooth,** then clear **Don't show again.**

OK Accepts the parameters and prepares the plot. You can further refine the appearance of the plot once the plug-in generates it.

Cancel Discards any changes you made, and closes the dialog box.

CGHSmooth Plot

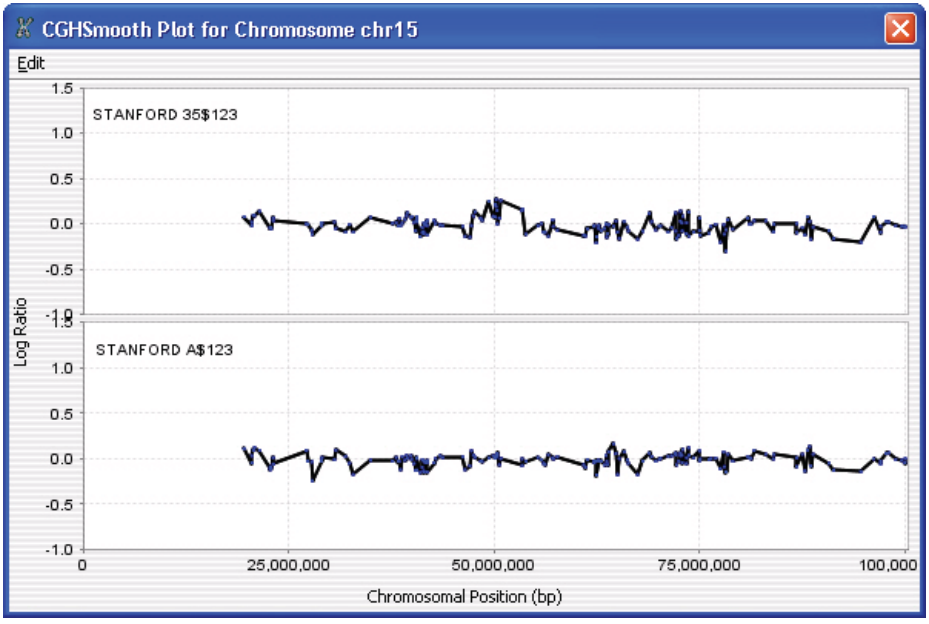


Figure 72 CGHSmooth Plot

Purpose: The CGHSmooth Plot is the output of the CGHSmooth plug-in. It contains separate, stacked plots of smoothed log ratio data for each of the selected arrays in the current experiment.

To open: Click **OK** in the CGHSmooth Parameters dialog box. See "CGHSmooth Parameters" on page 251.

Plot(s) Depending on the selected output option, the main plot area shows up to three plots for each array in the active experiment. The plots can include unsmoothed and smoothed log ratio plots, and an error plot.

Edit Opens a menu with a Copy plot(s) to Clipboard option. If you select this option, the program copies the plots to the clipboard as an image. You can then paste the image into a document in another program.

When you right-click anywhere within the plot area, a shortcut menu opens with these options:

Option	Description
Properties	Opens the Chart Properties dialog box, where you can change the appearance of many aspects of the plot. See "Chart Properties" on page 255.
Copy	Copies the chart to the Clipboard, where you can paste it into a word processing or other program.
Save as	Opens a Save dialog box, where you can select a location for the *.png image file of the plots.
Print	Opens the Windows Page Setup dialog box, where you can set page size, margins, and orientation, and select a printer. From there, you click **OK** to open the Windows Print dialog box, where you can set print options and print the plot.
Zoom In	Opens another menu that lets you zoom in the plot. You can zoom in several ways: • **Both Axes** – Zooms in the Domain (X) axis for all stacked plots, and zooms in the range Range (Y) axis for the specific plot in which you clicked. • **Domain Axis** – Zooms in the Domain (X) axis for all stacked plots. • **Range Axis** – Zooms in the Range (Y) axis for the specific plot in which you clicked. You can also drag across an area of one of the plots to select an area to zoom in on.

Option	Description
Zoom Out	Opens another menu that lets you zoom out the plot. You can zoom out several ways: • **Both Axes** – Zooms out the Domain (X) axis for all stacked plots, and zooms out the range Range (Y) axis for the specific plot in which you clicked. • **Domain Axis** – Zooms out the Domain (X) axis for all stacked plots. • **Range Axis** – Zooms out the Range (Y) axis for the specific plot in which you clicked.
Auto Range	Opens another menu that lets you zoom the plot to encompass the full range of the data. You can zoom in several ways: • **Both Axes** – Appropriately zooms both axes of the specific plot to show the full set of data. • **Domain Axis** – Zooms in the Domain (X) axis for all stacked plots so this axis encompasses the full range of X values of the data. • **Range Axis** – Zooms in the Range (Y) axis for the specific plot in which you clicked to encompass the full range of Y values of the data.

Chart Properties

Purpose: This dialog box is used to create or edit settings for plots.

To open: Use the CGHSmooth, Echo Example, or MovAvg Example plug-in to produce a plot. Right-click within the plot area, then click **Properties** in the shortcut menu. You can also open this dialog box when you right-click within the line plot in the Graphical Differential Aberration Summary dialog box. See "Graphical Differential Aberration Summary" on page 344.

This dialog box has three tabs. At any point, click **OK** to accept the settings in all three tabs, or click **Cancel** to close the dialog box without making any changes to the settings.

Title tab

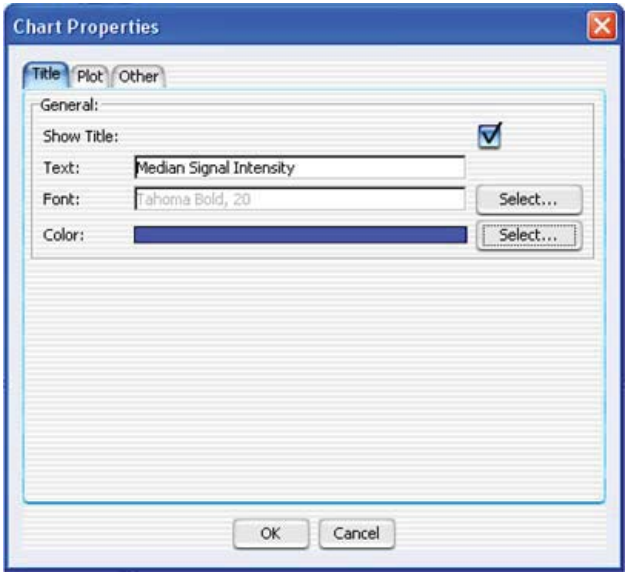


Figure 73 Chart Properties dialog box – Title tab

- **Show Title** – Mark this option to display a title across the top of the chart.
- **Text** – Type a title for the chart.
- **Font** – (Available if you mark **Show Title**) Click **Select** to open the Font Selection dialog box. Select the desired font attributes, then click **OK.**
- **Color** – (Available if you mark **Show Title**) Click **Select** to open the Title Color dialog box. Select or configure a color for the title, then click **OK.** This dialog box is identical to the Select Color dialog box. See "Select Color" on page 394.

Plot tab

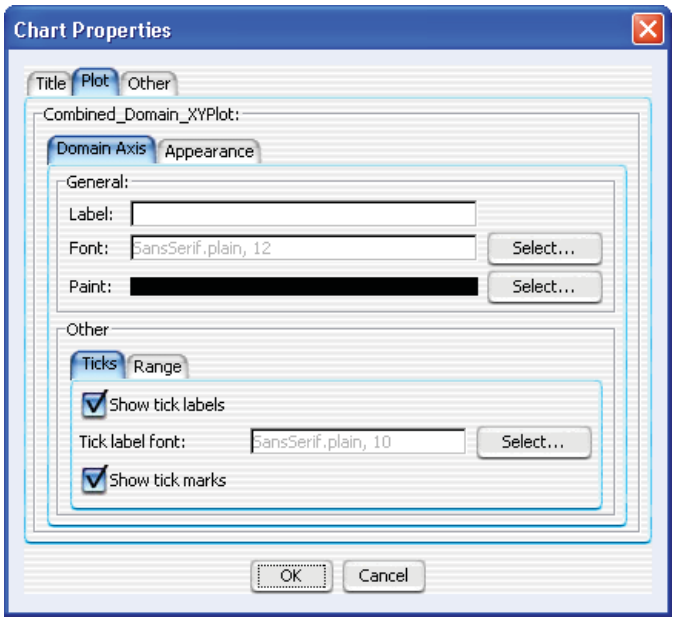


Figure 74 Chart Properties dialog box – Plot tab

- In the Plot tab, you can set these properties in the Domain Axis tab ("X" axis):

Property	Details
General	
Label	A custom label for the Domain (X) axis of the chart. Type the desired label.
Font	The font for the custom label on the Domain (X) axis. Click **Select** to open the Font Selection dialog box. Select the desired font attributes, then click **OK.**
Paint	The color of the custom label on the Domain (X) axis. Click **Select** to open the Label Color dialog box. Select the desired color, then click **OK.** This dialog box is identical to the Select Color dialog box. See "Select Color" on page 394.

Property	Details
Other – Ticks tab	
Show tick labels	Mark this option to show, or clear it to hide, the numerical values on the Domain (X) axis.
Tick label font	The font for the numerical values on the Domain (X) axis. Click **Select** to open the Font Selection dialog box. Select the desired font attributes, then click **OK.**
Show tick marks	Mark this option to show, or clear it to hide, tick marks on the Domain (X) axis.
Other – Range tab	
Auto-adjust range	Mark this option to automatically set the range of values on the X-axis to include all data.
Minimum range value	(Available if you do not mark Auto-adjust range) The lowest value represented on the X-axis.
Maximum range value	(Available if you do not mark Auto-adjust range) The highest value represented on the X-axis. The program automatically converts large numbers to scientific "E" notation – for example, **1.22E8.**

- In the Plot tab, you can set the following properties in the Appearance tab:

Property	Details
Outline stroke	The thickness of the lines that enclose each plot. Click **Select** to open the Stroke Selection dialog box. Select the desired line thickness, then click OK.
Outline paint	The color of the lines that enclose each plot. Click **Select** to open the Outline Color dialog box. Select the desired color, then click **OK.** This dialog box is identical to the Select Color dialog box. See "Select Color" on page 394.
Background paint	The color of the background within each plot area. Click **Select** to open the Background Color dialog box. Select the desired color, then click **OK.** This dialog box is identical to the Select Color dialog box. See "Select Color" on page 394.
Orientation	Select either Vertical (domain-axis on the bottom of the chart) or Horizontal (domain-axis on the left side of the chart).

Other tab

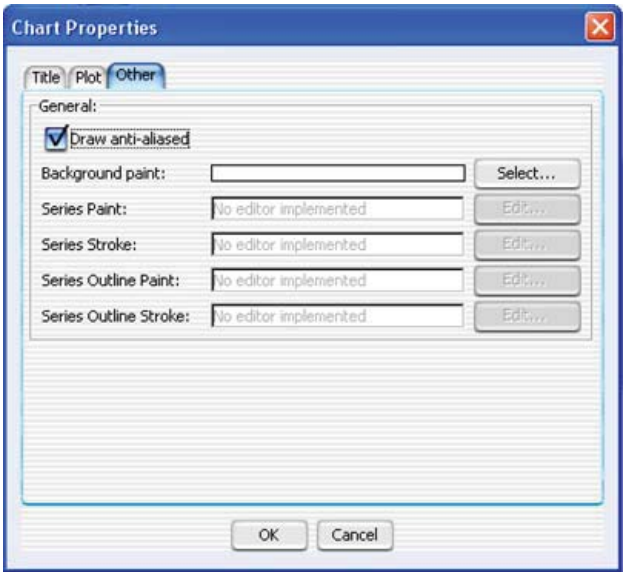


Figure 75 Chart Properties dialog box – Other tab

The Other tab offers these options:

- **Draw anti- aliased** – Mark this option to minimize distortion and visual artifacts in the plot image. This produces a smoother image, but it can be less sharp than the original one.
- **Background paint** – The color of the chart outside of the plot area and legend. Click **Select** to open the Background Color dialog box. Select the desired color, then click **OK.** This dialog box is identical to the Select Color dialog box. See "Select Color" on page 394.

The other options are for future expansion, and are not available in the current release of Agilent Genomic Workbench.

Chromosome Selection

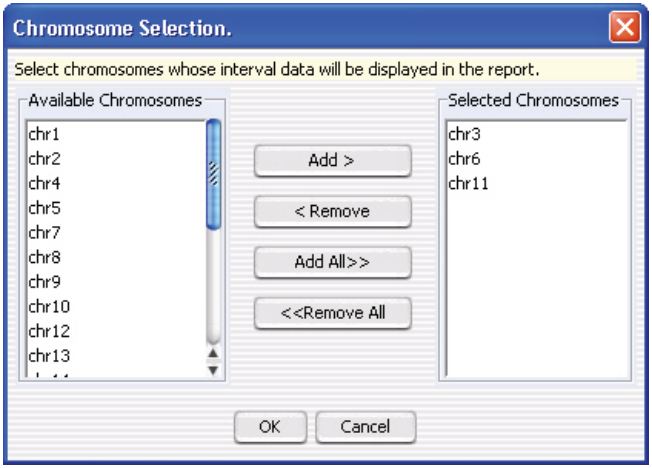


Figure 76 Chromosome Selection dialog box

Purpose: The Chromosome Selection dialog box lets you select the chromosomes to include in a cytogenetic report template.

To open: Click **Add** under **Chromosomes** in the Chromosome and Interval Selection tab of the Design/Edit Cytogenetic Report Templates dialog box See "Design/Edit Cytogenetic Report Templates" on page 288.

Available Chromosomes List of chromosomes that have not yet been selected for inclusion in the cyto report template.

Selected Chromosomes List of chromosomes that have been selected for inclusion in the cyto report template.

Add > Transfers chromosomes from the Available Chromosomes list to the Selected Chromosomes list. To transfer a chromosome, click its name under **Available Chromosomes** (you can hold down the **ctrl** key and click additional names), then click **Add >.**

< Remove Transfers chromosomes from the Selected Chromosomes list to the Available Chromosomes list. To transfer a chromosome, click its name under **Selected Chromosomes** (you can hold down the **ctrl** key and click additional names), then click **< Remove.**

Add All >> Transfers all of the chromosomes under Available Chromosomes to the Selected Chromosomes list.

<< Remove All Transfers all of the chromosomes under Selected Chromosomes to the Available Chromosomes list.

OK Closes the dialog box, and transfers the chromosomes under Selected Chromosomes to the list of chromosomes in the Design/Edit Cytogenetic Report Templates dialog box (Chromosome and Interval Selection tab).

Cancel Closes the dialog box without transferring any chromosomes.

Combine Tracks

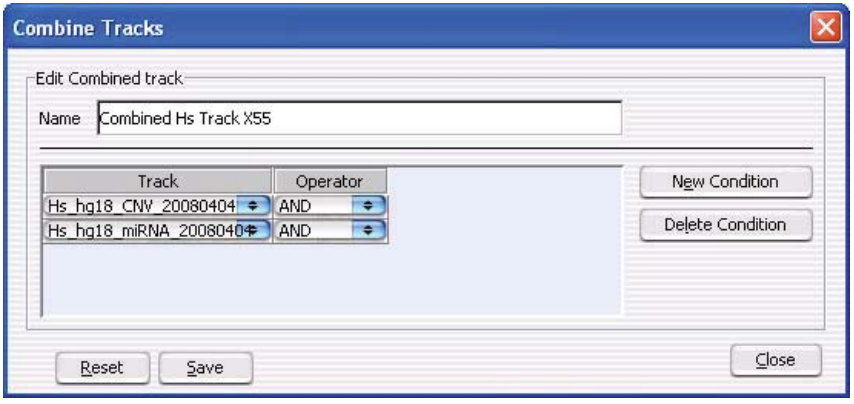


Figure 77 Combine Tracks dialog box

Purpose: Lets you create a combined track that contains elements of two or more source tracks, associated by logical conditions. See "To combine tracks" on page 78.

To open: In the **My Entity List** pane, right-click the **Tracks** folder, then click **Combine Tracks.**

Name The name of the combined track

New Condition Adds a new, empty row to the Track/Operator table in the dialog box.

Delete Condition Removes the bottom row from the Track/Operator table in the dialog box.

Track In each row, select a track to include in the combined track.

Operator In each row, select the desired logical operator. This operator controls the manner in which the program combines the track in this row with the others. Select one of these options:

Operator	Comments
AND	Places an element in the combined track if it appears in both this track and any of the others.
OR	Places an element in the combined track if it appears in either this track or any of the others. If you set this operator for all tracks in the list, the result is a nonredundant set of elements from all tracks.
MINUS	Removes the elements that appear in this track from the combined track, if they otherwise appear there.

Reset Removes all Track/Operator pairs from the table in the dialog box, and clears the Name of the combined track.

Save Creates the combined track, but does not close the dialog box.

Close Closes the dialog box. Opens the Confirm track save dialog box if you created a combined track, but did not save it.

Compare CNVRs

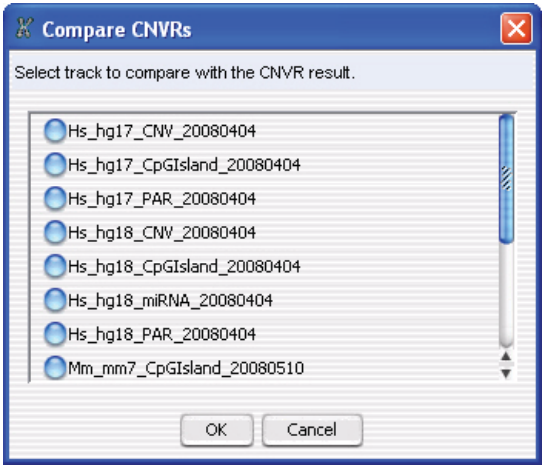


Figure 78 Compare CNVRs dialog box

Purpose: This dialog box lets you select a track to compare with the CNVR result.

To open: Right-click the name of a CNVR result in the Experiment pane of the Navigator, then select **Compare CNVRs.**

List of tracks Displays the available tracks in the program. Select the desired track to compare with the CNVR result.

OK Opens a Save dialog box, where you can select a location for the CNVR comparison result file. When you click **Save,** the program creates the file in *.xls format. You can open this file in Microsoft Excel.

Cancel Closes the dialog box without generating a CNVR comparison file.

Configure Coloring Ranges and Shades

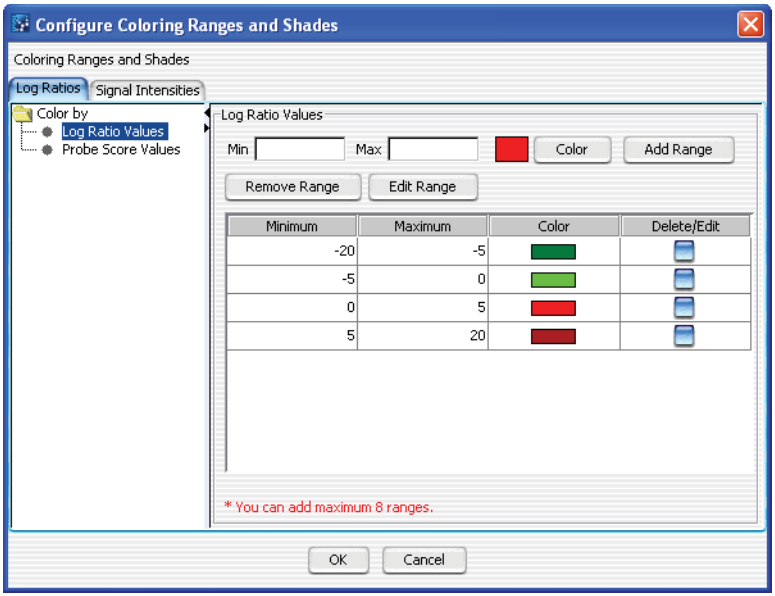


Figure 79 Configure Coloring Ranges and Shades dialog box

Purpose: This dialog box is used to enter ranges and select colors for scatter plot options. Tabs show selections for Log Ratios and Signal Intensities plots.

To open: In Gene View, move the mouse pointer over **Scatter Plot** to display the scatter plot options and then click **Configure Color and Ranges**. Or, click the **View** tab and click **View Preferences**. Then, under Configure Coloring schemes, click **Configure Colors and Ranges.**

Table 5 Log Ratios

Color by	Description
Log Ratio Values	
Min	Type a minimum value for the range.
Max	Type a maximum value for the range.
Color	Click to open the Select Color dialog box, where you can select the color you want to display for this range. See "Select Color" on page 394 for more information.
Add Range	Click to add a row to the range table, using the values displayed in Min and Max, and the selected Color.
Remove Range	Click to remove the ranges with Edit/Delete box marked.
Edit Range	Click to edit range(s) with Edit/Delete box marked.
Range table	This table displays the defined ranges, including minimum and maximum values, color for each range, and Edit/Delete selection.
Probe Score Values	
Min	Type a minimum value for the range.
Max	Type a maximum value for the range.
Color	Click to open the Select Color dialog box, where you can select the color you want to display for this range. See "Select Color" on page 394 for more information.
Add Range	Click to add a row to the range table, using the values displayed in Min and Max, and the selected Color.
Remove Range	Click to remove the ranges with Edit/Delete box marked.
Edit Range	Click to edit range(s) with Edit/Delete box marked.
Range table	This table displays the defined ranges, including minimum and maximum values, color for each range, and Edit/Delete selection.

Table 6 Signal Intensities

Color by	Description
Channels	
Green Intensity	Click to open the Select Color dialog box, where you can select the color you want to display for this channel. See "Select Color" on page 394 for more information.
Red Intensity	Click to open the Select Color dialog box, where you can select the color you want to display for this channel. See "Select Color" on page 394 for more information.
Probe Score Values	
Min	Type a minimum value for the range.
Max	Type a maximum value for the range.
Color	Click to open the Select Color dialog box, where you can select the color you want to display for this range. See "Select Color" on page 394 for more information.
Add Range	Click to add a row to the range table, using the values displayed in Min and Max, and the selected Color.
Remove Range	Click to remove the ranges with Edit/Delete box marked.
Edit Range	Click to edit range(s) with Edit/Delete box marked.
Range table	This table displays the defined ranges, including minimum and maximum values, color for each range, and Edit/Delete selection.
Intensity Values	
Min	Type a minimum value for the range.
Max	Type a maximum value for the range.
Color	Click to open the Select Color dialog box, where you can select the color you want to display for this range. See "Select Color" on page 394 for more information.
Add Range	Click to add a row to the range table, using the values displayed in Min and Max, and the selected Color.
Remove Range	Click to remove the ranges with Edit/Delete box marked.

Table 6 Signal Intensities

Color by	Description
Edit Range	Click to edit range(s) with Edit/Delete box marked.
Range table	This table displays the defined ranges, including minimum and maximum values, color for each range, and Edit/Delete selection.

Confirm



Figure 80 Confirm dialog box

Purpose: This dialog box lets you save the updated result or discard it.

To open: Switch results or experiments, when saved results already exist.

Overwrite current result Replaces the selected experiment result in the Navigator with the result that appears on your screen.

Create new result Opens the Save Experiment result dialog box, where you can save the result that appears on your screen as a new experiment result. See "To save a result" on page 145.

Continue without saving The program does not save the result that appears on your screen.

Confirm Overwrite

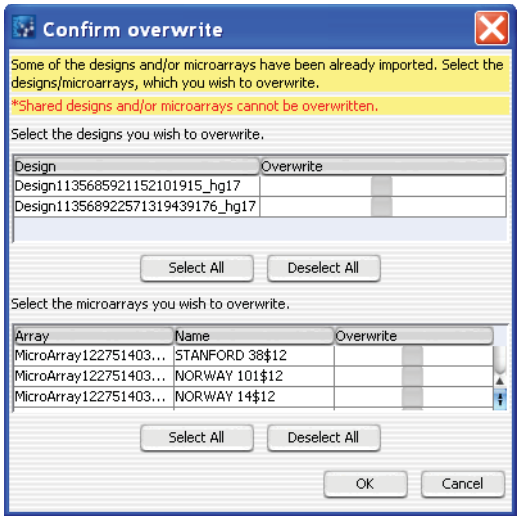


Figure 81 Confirm overwrite dialog box

Purpose: When you import an experiment, it can contain designs and/or arrays that have the same names as those already available in Agilent Genomic Workbench. This dialog box lets you select which designs and/or arrays to overwrite.

To open: This dialog box appears when you import a ZIP format experiment file, and it contains designs and/or arrays that are already available in Agilent Genomic Workbench. See "To import an experiment file" on page 57.

Select the designs you wish to overwrite

Design The names of the designs in the imported file that have the same names as designs that are already available in Agilent Genomic Workbench.

Overwrite Mark the check box next to each existing design to overwrite.

Select All Marks all of the check boxes under Overwrite.

Deselect All Clears all of the check boxes under Overwrite.

Select the microarrays you wish to overwrite

Array Identification number for the array

Name The name of the array in the imported file that has the same name as an array that is already available in Agilent Genomic Workbench.

Overwrite Mark the check box next to each existing array to overwrite.

Select All Marks all of the check boxes under Overwrite.

Deselect All Clears all of the check boxes under Overwrite.

OK Overwrites the selected files (both designs and arrays) and closes the dialog box.

Cancel Closes the dialog box, and returns you to the Import (experiments) dialog box. See "Import (experiments)" on page 357.

Copy Number Variant Region

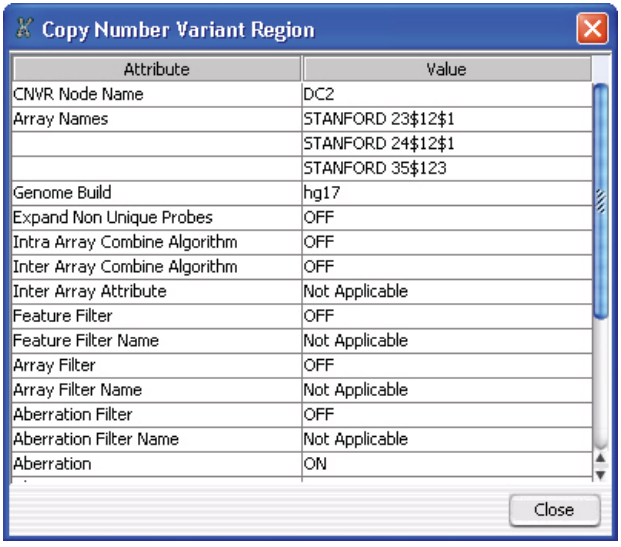


Figure 82 Copy Number Variant Region dialog box

Purpose: This dialog box lets you view the attributes of the CNVR result. You view the actual CNVRs in Genome, Chromosome, and Gene views.

To open: Right-click the name of a specific CNVR result in the Experiment pane of the Navigator, then select **View Details.**

Attribute list Displays specific attributes of the CNVR result, such as its name, the arrays included in the CNVR analysis, and the analysis settings.

Close Closes the dialog box.

Correlation Analysis Setup

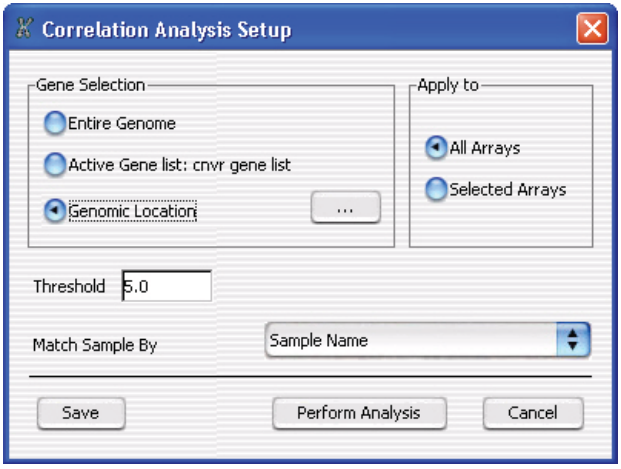


Figure 83 Correlation Analysis Setup dialog box

Purpose: This dialog box lets you configure the parameters of a correlation analysis, a process that examines the relationship between gene expression and CGH results for given labeled samples. You can also use the dialog box to run the analysis, and save the parameters for later use.

To open: Click **Discovery > Joint-Analysis > Correlation Analysis > Setup.**

Gene Selection Defines the scope of the analysis. Select one of these options:

- **Entire Genome** – Analyzes the whole genome
- **Active Gene list** – Analyzes regions defined by the active gene list.
- **Genomic Location** – Analyzes regions you define. Click ... to open the Select Chromosome Intervals dialog box, where you can define the genomic regions for the analysis. See "Select Chromosome Intervals" on page 393.

Apply to Specifies the arrays to be analyzed. Select one of these options:

- **All Arrays** – Applies the correlation algorithm to all arrays in the experiment.
- **Selected Arrays** – Applies the correlation algorithm only to the selected arrays in the experiment. To select arrays for analysis, see "To select or deselect arrays in the experiment" on page 91.

Threshold Sets the border between significant and insignificant correlations. In general, increase this number to make the correlation detection process more stringent.

Match Sample By The microarray attribute the program uses to match CGH arrays with expression arrays. The default selection is Green Sample.

Save Saves the correlation analysis parameters internally. Later, you can click **Discovery > Joint- Analysis > Correlation Analysis > Perform** to run a correlation analysis with these parameters. If you save the correlation analysis parameters, the program overwrites any existing set of saved parameters.

Perform Analysis Opens the Matched Sample dialog box, where you can view the pairs of arrays that the program will use for the correlation analysis. See "Matched Sample" on page 365.

Cancel Closes the dialog box and cancels any changes you made to the correlation analysis parameters.

Correlation Results



Figure 84 Correlation Results dialog box

Purpose: The Correlation results dialog box lets you view the correlation results in several formats.

To open: The Correlation Results dialog box opens after you set up a correlation analysis and click **Continue** in the Matched Sample dialog box See "Matched Sample" on page 365.

Show Entire Genome Mark this option to display the results for the entire genome. Otherwise, the dialog box displays results for one chromosome at a time.

Sort samples by the selected gene Mark this option to change the order of samples based on the expression data.

Show Plot Labels Mark this option to display axis names in the plots in the dialog box.

 Zooms in the view.

 Zooms out the view.

 Resets the zoom level to the initial setting.

 (Available only if you clear **Show Entire Genome**) Displays the results for the previous chromosome.

 (Available only if you clear **Show Entire Genome**) Displays the results for the next chromosome.

Enrichment Analysis Opens the Enrichment Analysis Setup dialog box, where you can configure the parameters for an enrichment analysis. See "Enrichment Analysis Setup" on page 310.

Results Panes These graphical results appear as the separate, detachable panes listed below. In addition, a tool tip with the current numerical coordinates of the pointer appears when you place the pointer within each plot.

- **Gene Expression and Aberration Values** – Gene expression and aberration data plotted together for one particular genomic position from all of the samples. The particular genomic position is selected by the blue vertical line (cursor) in the following three panes, and is synchronized with the highlighted row in the data table at the bottom of the dialog box. The left axis reflects expression level, and the right axis reflects aberration level.
- **Correlation Scores** – The correlation scores change depending on your selection in the column headings of the data table at the bottom of the dialog box. In addition, you can right-click within this plot to display an Input Threshold Value option. If you select this option, an Input dialog box opens. Type the desired threshold value, then click **OK.**
- **Heatmap for Expression Data** – Shows the gene expression values from the expression arrays as a heatmap. In addition, you can right-click within this plot to display an Input saturation value option. If you select this option, an input dialog box appears. Type the desired value, then click **OK.**

- **Aberration Summary** – Shows the CGH aberration calls from the CGH arrays. In addition, you can right-click within this plot to display an Input saturation value option. If you select this option, an input dialog box appears. Type the desired value, then click **OK.**

Data Table Shows the output scores produced by the correlation analysis algorithm, identified by chromosomal location and gene name. These results appear:

- **Amp-nonAmp t-test** – The student t-test results for amplification and non-amplification.
- **Del-nonDel t-test** – The student t-test results for deletion and non-deletion.
- **Pearson p-value test results** - Results using the Pearson correlation approach for correlation between expression and DNA copy number.
- **Max Score** – The maximum value from the above three test scores.

Create CNVR

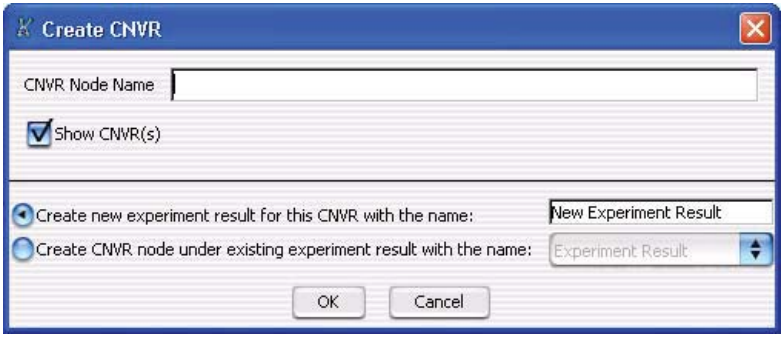


Figure 85 Create CNVR dialog box

Purpose: This dialog box lets you define a name and location for CNVRs detected by the program. See "To visualize a CNVR (copy number variant region)" on page 156.

To open: The create CNVR dialog box opens when you click **Discovery > CNVR.**

CNVR Node Name Type a name for the CNVR result as you would like it to appear within the Results folder of a specific experiment in the Navigator.

Show CNVRs Mark this option to show the detected CNVRs in Genome, Chromosome, and Gene views. CNVRs appear as green boxes that enclose the detected regions.

Location options Sets the location of the generated CNVR results within the Navigator. Select one of these options:

- **Create new experiment result for this CNVR with the name** – Creates a new folder within the Results folder of the selected experiment. Type a name for this new folder. This folder will contain a CNVR folder that contains the specific CNVR result.
- **Create CNVR node under existing experiment result with the name** – Saves the new CNVR result in an existing experiment folder. Select the name of the desired existing result folder from the list.

OK Generates the CNVRs, and saves them as specified in the dialog box.

Cancel Closes the dialog box without generating CNVRs.

Create Common Aberration

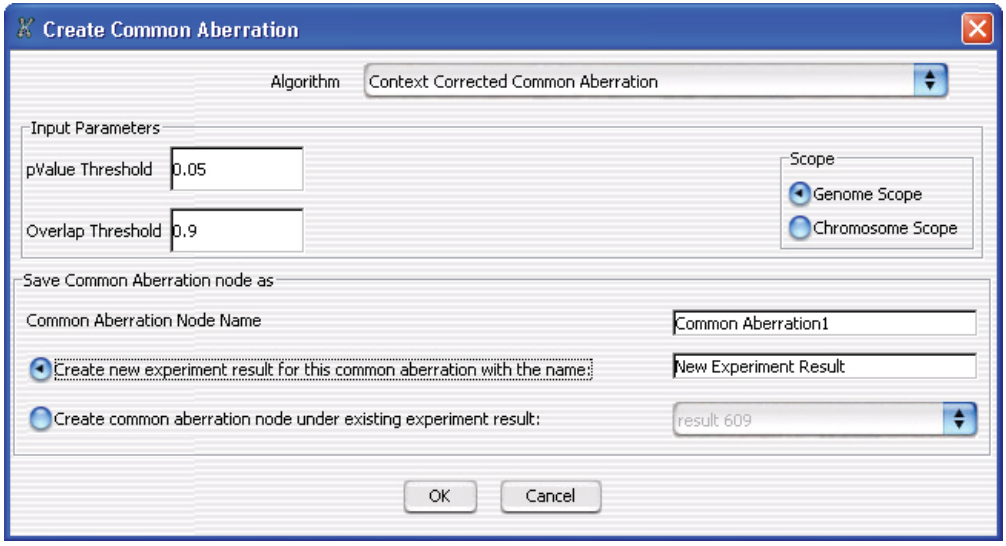


Figure 86 Create Common Aberration dialog box

Purpose: This dialog box lets you configure a common aberration analysis, and to select a location for the results. Common aberration analysis detects common aberrant regions across multiple samples. For instructions on how to do a common aberration analysis, see "To create common aberration summaries" on page 151.

To open: In the **Discovery** tab, under **Common Aberration,** click **Text.**

Algorithm The algorithm used to detect common aberrations among multiple samples. For a detailed description of the available algorithms, see "Algorithms for multi-array analysis" on page 445. Select one of these options:

- **Context Corrected Common Aberration** – This algorithm detects common aberration in a subset of samples to maximize the score. Also, it takes into account the probability of finding a common aberration by consideration of all of the other aberrations present in the genome.
- **T-Test Common Aberration** – Calculates results that may not be sensitive to large aberrations. The algorithm calculates the likelihood of detection of a given aberration.

Input Parameters Type values for these parameters for multi-sample experiments:

- **pValue Threshold** – Sets the *p*-value limits of the aberrations called.
- **Overlap Threshold** – Sets the limit of overlapping. A threshold of 0.9 covers the entire genome.

Scope (Available if you select the Context Corrected Common Aberration algorithm.) Select one of these options:

- **Genome Scope** – Displays results for the entire genome.
- **Chromosome Scope** – Displays results on a chromosome-by-chromosome basis.

The scope of the T-Test Common Aberration algorithm is always the entire genome.

Save Common Aberration node as Defines a name and location for the common aberration results.

In **Node Name,** type a name for the common aberration result as you want it to appear in the Navigator.

Select one of these locations:

- **Create new experiment result for this common aberration with the name** – Creates a new folder within the Results folder of the selected experiment. Type a name for this new folder. This folder will contain a Common Aberration folder that contains the specific common aberration result (node).
- **Create common aberration node under existing experiment result** – Saves the new common aberration result (node) in an existing experiment folder. Select the name of the desired existing result folder from the list.

OK Generates the common aberrations, and saves them to the specified location.

Cancel Closes the dialog box without generating common aberrations.

Create Experiment

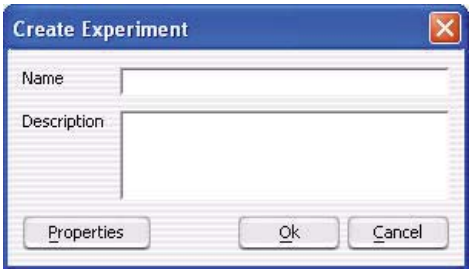


Figure 87 Create Experiment dialog box

Purpose: Creates an organizational unit (an *experiment*) that lets you view and analyze array data in Agilent Genomic Workbench. You add data to the experiment with links to array data files that are available in the program, a process that you can start from this dialog box. See "To create a new experiment" on page 60.

To open: In the Home tab of Agilent Genomic Workbench, click **Create Experiment.**

Name (Required) The name of the new experiment. This name identifies the experiment within the program and in exported reports and files.

Description (Optional) Brief information that will later help to identify the experiment.

Properties Opens the Experiment Properties dialog box, where you can select array data files to add data to the new experiment. See "Experiment Properties" on page 313.

OK Closes the dialog box and creates the new experiment.

Cancel Closes the dialog box without creating an experiment.

NOTE Click **Properties** to open the Experiment Properties dialog box to add data to your new experiment with array data. Otherwise, the program creates an empty experiment. You can also add arrays to the experiment later. See "To add arrays to an experiment" on page 62.

Create Gene List

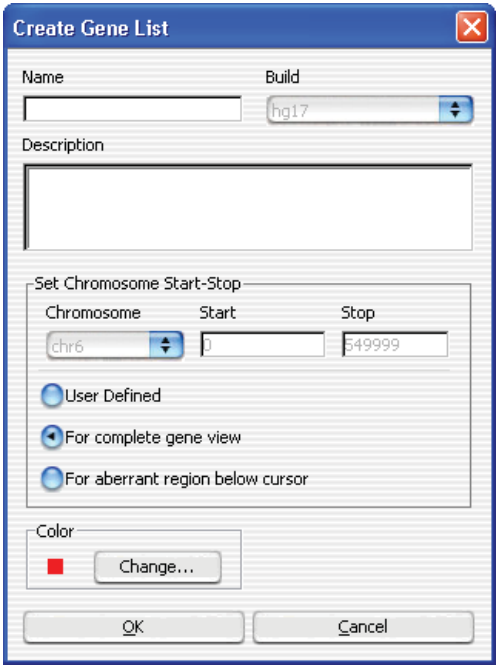


Figure 88 Create Gene List

Purpose: To limit the genes presented in Gene View to a preselected number valuable for interpreting data

To open: Right-click Gene View, and click **Create Gene List**.

Name Type in name of gene list.

Build Select the genome build for the genes to be selected for list.

Description Describe the type or nature of the genes in the list.

Set Chromosome Start-Stop

Select a chromosome and a region in Chromosome View for selecting the genes in the list before you open the Create Gene List dialog box.

User Defined Select to choose region from which the genes in Gene View will be selected. The chromosome selection list and the Start and Stop positions on the Y axis are enabled when this option is selected. With this option, you can override the selections you made before opening Create Gene List.

For complete gene view Select all the genes in Gene View.

For aberrant region below cursor Select those genes that appear in the aberrant region just below the cursor position in Gene View.

Chromosome If you select User Defined, you can select a different chromosome than had been selected before opening the Create Gene List dialog box.

Start If you select User Defined, you can type in a Start position to define the region that contains the genes to be in the list.

Stop If you select User Defined, you can type in a Stop position to define the region that contains the genes to be in the list.

Color

Change Click to change the color of the gene list name in Navigator.

Create Gene List (from CNVR results)

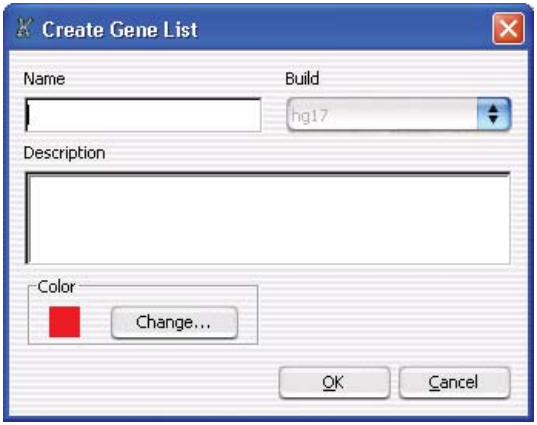


Figure 89 Create Gene List dialog box, specific for CNVR results

Purpose: This dialog box lets you create a gene list that contains the genes found within the regions defined in the CNVRs. For details on gene lists, see the first two chapters of this guide.

To open: Right-click the name of a specific CNVR in a specific experiment in the Navigator, then click **Create Gene List** in the shortcut menu. The result folder that contains the CNVR must be the selected result.

Name Type a name for the new gene list, as you want it to appear in the Gene List folder of the Navigator.

Build (If available) Select the applicable genome build.

Description Type descriptive text to be attached to the gene list for reference.

Color Shows the current gene list color. This is the highlight color for the genes in the list, as well as the color of the gene list name in the Gene List folder. To change the color, click **Change.** The Choose GeneList Color dialog box appears. This dialog box is identical to the Select Color dialog box. See "Select Color" on page 394.

OK Generates the CNVR gene list.

Cancel Closes the dialog box without generating a gene list.

Create Gene List (from common or differential aberration results)

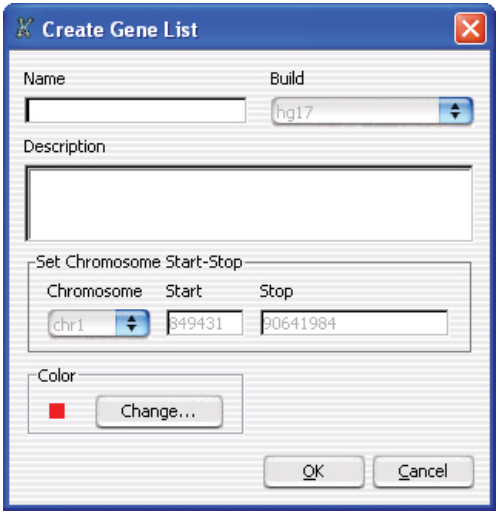


Figure 90 Create Gene List dialog box – Specific for results of common or differential aberration analyses.

Purpose: This lets you create a gene list that contains the genes found within one commonly or differentially aberrant region. See "Graphical Common Aberration Summary" on page 341 and "Graphical Differential Aberration Summary" on page 344.

To open: Click **Create Gene List** in either the Graphical Common Aberration Summary dialog box or the Graphical Differential Aberration Summary dialog box.

Name Type a name for the new gene list, as you want it to appear in the Gene List folder in the My Entity List pane of the Navigator.

Build Shows the genome build associated with the common or differential aberration results.

Description Type descriptive text to be attached to the gene list for reference.

Set Chromosome Start-Stop Shows the genomic interval that corresponds to the selected commonly or differentially aberrant region.

Color Shows the current gene list color. This is the highlight color for the genes in the list, as well as the color of the gene list name in the Gene List folder. To change the color, click **Change.** The Choose Gene List Color dialog box appears. This dialog box is identical to the Select Color dialog box. See "Select Color" on page 394.

OK Generates the gene list. The gene list appears in the My Entity List pane of the Navigator.

Cancel Closes the dialog box without generating a gene list.

Create Signal Bar Chart

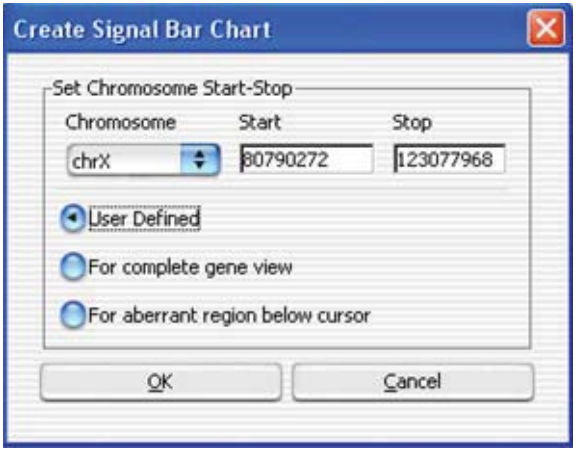


Figure 91 Create Signal Bar Chart dialog box

Purpose: This dialog box lets you set parameters to create a histogram of signal intensities. You can customize the region that you want to display by selections in Set Chromosome Start-Stop.

To open: Right-click in the Gene View and select **Show Intensity Bar Charts**.

Set Chromosome Start-Stop Defines the region of the chromosome for which the bar chart will be defined. Select one of these options:

- **User Defined** – Lets you define an arbitrary region of any chromosome. If you select this option, select the desired chromosome in **Chromosome,** then type the beginning **(Start)** and end **(Stop)** locations of the desired interval.
- **For complete gene view** – The chromosomal region that appears in Gene View.
- **For aberrant region below cursor** – All of the intervals that begin before the cursor position and end after the cursor position. (This option is available if an aberration algorithm is selected.)

OK Creates the histogram using the selected region.

Cancel Closes the dialog box without creating the histogram.

Create Track

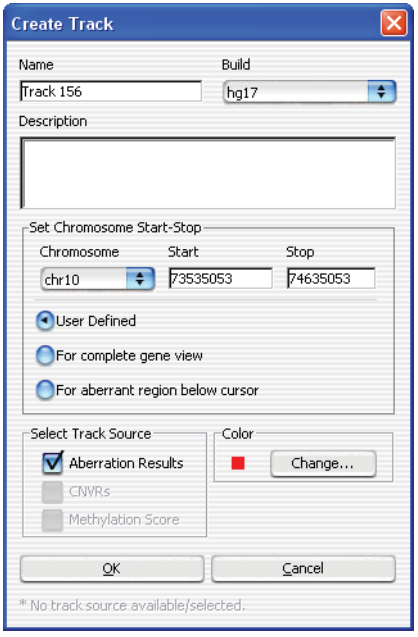


Figure 92 Create Track dialog box

Purpose: This dialog box lets you create a track for a chromosomal region, based on certain results or other parameters. You can display one or more tracks next to the genes, data, and results in Gene View. See "To show tracks in Gene View" on page 111.

To open: Right-click in the plot area of Gene View for the CGH module, then click **Create Track** in the shortcut menu.

Name Type a name for the track. This name identifies the track when it appears in lists and displays.

Build (Available if you select **User Defined** in **Set Chromosome Start-Stop.**) Select the genome build with which to associate the track.

Description Type descriptive text to attach to the track for reference.

Set Chromosome Start-Stop This parameter defines the region of the chromosome for which the track will be defined. Select one of these options:

- **User Defined** – Lets you define an arbitrary region of any chromosome. If you select this option, select the desired chromosome in **Chromosome,** then type the beginning **(Start)** and end **(Stop)** locations of the desired interval.
- **For complete gene view** – The chromosomal region that appears in Gene View.
- **For aberrant region below cursor** – All of the intervals that begin before the cursor position and end after the cursor position.

Select Track Source The type of analysis result the program uses to construct the regions defined in the track. Select one or both of these options:

- **Aberration Results** – The current aberration analysis results in the defined chromosomal region.
- **CNVRs** – The detected CNVRs in the defined chromosomal region.

Change Click to open the Choose Track Color dialog box to select the color to use for display of the track in the Tracks folder. See "Select Color" on page 394.

OK Creates the track. To display the track in Gene View, use the **Tracks** tab of the User Preferences dialog box to enable it. See "Tracks tab" on page 415. To export the track, see "Export" on page 315.

Cancel Closes the dialog box without creating a track.

Customize Search Link



Figure 93 Customize Search Link dialog box

Purpose: This dialog box lets you create a custom Web search link in the shortcut menu that appears when you right-click a table entry. The link opens the URL of your choice, and passes the table entry to it as a search string. See "To create a custom Web search link" on page 121.

To open: Right-click any entry in a tab in Tab View, other than a column heading, then click **Customize Link.**

URL Name The name of the custom Web search link that appears in the shortcut menu (see above). To edit an existing custom Web search link, select it from the list.

URL The full uniform resource locator (URL) of the desired search page. For the query string value, type `<target>`

For example, this URL passes the selected Tab View entry to google.com:

```
http://www.google.com/search?hl=eng&q=<target>
```

New Opens an Input dialog box, where you can type a name for a new custom Web search link. Click **OK** to accept the name and add it to the URL name list.

Update Saves the settings in the dialog box.

Delete Deletes the selected custom Web search link.

Close Closes the dialog box.

Define Sets

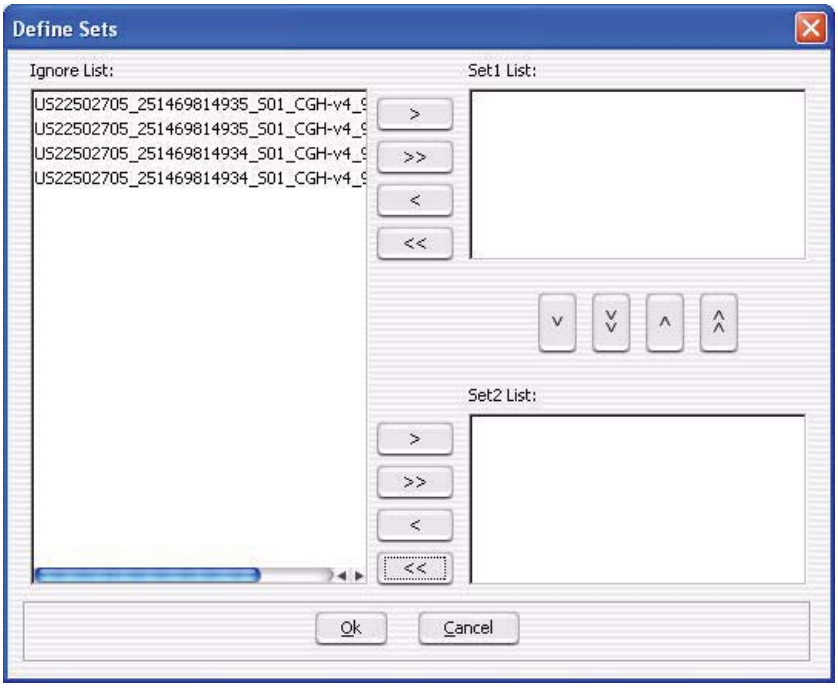


Figure 94 Define Sets dialog box

Purpose: This dialog box lets you assign arrays to comparison sets for the analysis. See "Differential Aberration Setup" on page 303 and "To set up a differential aberration analysis" on page 158.

To open: The Define Sets dialog box opens when you click **Define Sets** in the Differential Aberration Setup dialog box as you set up a differential aberration analysis.

The lists in this dialog box reflect the assignments you make with the Set1/Set2/Ignore options in the Differential Aberration Setup dialog box. Similarly, the selected Set1/Set2/Ignore options in the Differential Aberration Setup dialog box reflect the assignments you make in this Define Sets dialog box.

Set1 List Displays the arrays assigned to comparison set 1.

Set2 List Displays the arrays assigned to comparison set 2.

Ignore List Displays the arrays that are not assigned to a comparison set.



Moves selected arrays from one list to another, in the direction indicated by the arrowhead. To select arrays:

- To select an array, click its name.
- To select additional arrays, hold down the **ctrl** key and click their names.
- To select a contiguous block of arrays, click the name of the first array, then hold down the **shift** key and click the name of the last one.





Moves all of the arrays in one list to another, in the direction indicated by the arrowheads.

Ok Accepts your array assignments. These assignments will be reflected in the settings of the Set1/Set1/Ignore options for each array in the Differential Aberration Setup dialog box.

Cancel Discards any changes to array assignments, and closes the dialog box.

Design/Edit Cytogenetic Report Templates

Purpose: Used to create a cyto report, see "To generate a cytogenetic report" on page 170.

To open: Click **Reports > Cyto Reports > Design/Edit Template.** It lets you configure cyto report templates for later use.

The Design/Edit Cytogenetic Report Template dialog box contains the following tabs:

- **Report Settings** – Configures the general content and format of the report. See "Report Settings tab" on page 290.
- **Analysis Settings** – Configure the analysis the program applies when you run the report. See "Analysis Settings tab" on page 292.
- **SNP Analysis Setting** – Configures the SNP analysis settings that are applied when you run the report. See "SNP Analysis Setting tab" **on page 294**.
- **Chromosome and Interval Selection** – Define the regions of the genome to include in the report. See "Chromosome and Interval Selection tab" on page 295.

These buttons appear at the bottom of every tab:

Button	Details
New	Creates a new report template. When you click **New,** a Message dialog box opens. Type a name for the new report template, then click **OK.** The program creates a copy of the current template settings and enables it for editing.
Save	Saves any changes you make to the current template.
Reset	Resets the current tab to the default settings.
Delete	Deletes the selected report template.
Rename	Changes the name of the selected report template. When you click Rename, an Input dialog box appears. Type a new name for the report template, then click **OK.**
OK	Accepts your selections and closes the dialog box.
Cancel	Closes the dialog box, and discards any changes that you have not saved.

Report Settings tab

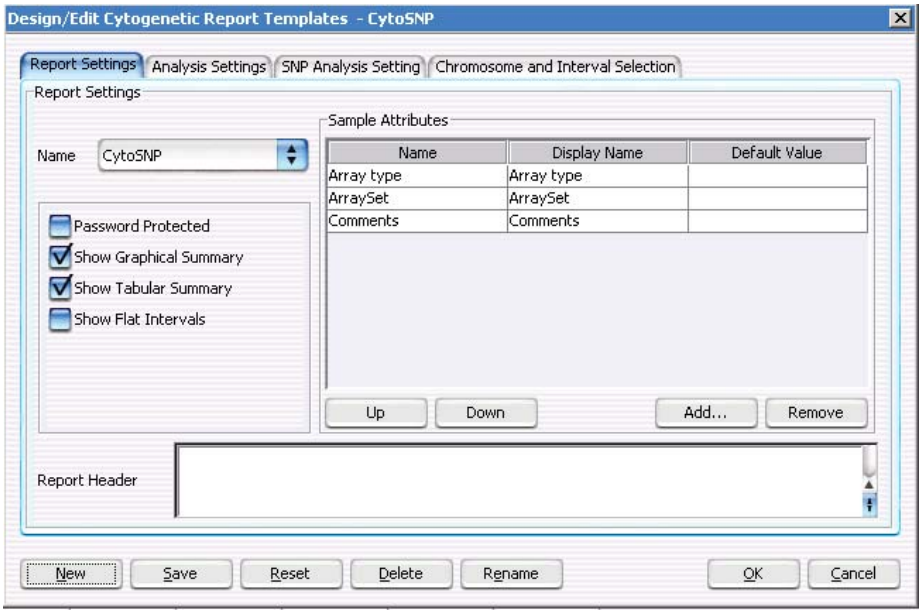


Figure 95 Design/Edit Cytogenetic Report Template – Report Settings tab

The report settings tab contains these options and commands:

Option/Command	Details
General Settings	
Name	Shows the Cyto Report templates available in the program. To edit an existing template, select its name. To create a new template, click **New** (see below).
Password Protected	Mark this option to limit editing of the template. When you mark this option, the Set Report Password dialog box opens. Type a password, type it again in **Confirm Password,** then click **OK.** A lock button appears.  Click this button to change the password. When you or another user tries to edit the template in the future, an Enter Password dialog box appears. You must type the password to display or edit the template.

Option/Command	Details
Show Graphical Summary	Mark this option to include a graph of the whole genome or selected chromosomes in the generated cyto report.
Show Tabular Summary	Mark this option to include a table of results in the generated cyto report.
Show Flat Intervals	Mark this to have aberration intervals reported without any nested structure.
Report Header	Type text to appear at the top of cyto reports that use this template.
Sample Attributes	
Name	(Read-only) The array attributes included in the report.
Display Name	The name that appears in the report for each array attribute. By default, the Display Name is the same as the name of the attribute. To change the Display Name, click it, edit it as desired, then press **Enter.**
Default Value	When the program generates the cyto report, the program automatically uses the actual value of each attribute. If the value is not defined in the properties of the sample, the program leaves it blank, unless you type a default value here.
Up	Moves a selected attribute up in the list. This moves the attribute up in the Sample Properties list in the generated report.
Down	Moves a selected attribute down in the list. This moves the attribute down in the Sample Properties list in the generated report.
Add	Opens the Select Sample Attributes dialog box. In this dialog box, mark the desired attributes, or mark **Select** to select all attributes, then click **OK.**
Remove	Removes a selected attribute from the list.

Analysis Settings tab

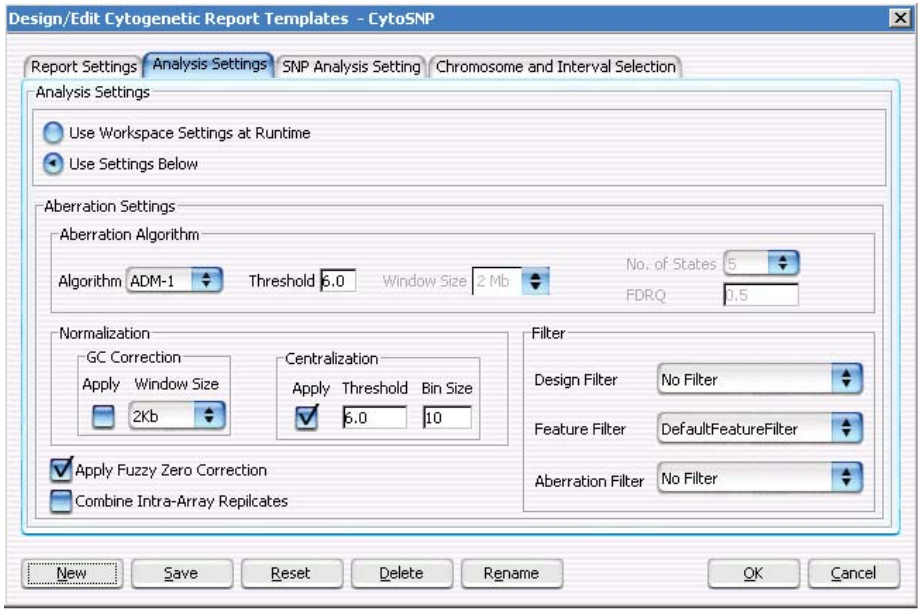


Figure 96 Design/Edit Cytogenetic Report Templates – Analysis Settings tab

The Analysis Settings tab contains these options and commands:

Option/Command	Details
Main options	
Use Workspace Settings at Runtime	Select this option to use the analysis settings from the Analysis tab for interactive analysis, or from the workflow analysis method at the time you run the report. See "Analysis command ribbon" on page 183.
Use Settings Below	Select this option to define analysis settings. If you select this option, the analysis parameters in the tab become available. These override the settings in the workspace when you run the report.

Option/Command	Details
Aberration Settings	
Aberration Algorithm	Select the desired aberration detection algorithm. Type a threshold value. For the Z-Score algorithm, select (or type) a window size in Mb, Kb, or data points (pt). For details on these algorithms, see "Aberration Detection Algorithms for CGH Analysis" on page 465.
GC Correction	Mark **Apply** to calculate GC Correction to the data, and select a **Window Size** for the calculation. GC Correction is recommended for SNP Copy Number and LOH analyses. For details on this algorithm, see "GC Correction Algorithm" on page 455.
Centralization	Mark **Apply** to apply this normalization algorithm. For details on this algorithm and its parameters, see "Centralization Algorithm" on page 452.
Filter	Select the desired feature, design, and aberration filters. You can also select **No Filter.** To create custom filters, see "To create or modify a design filter" on page 127, "To create or modify a feature filter" on page 127, and "To create or modify an array filter" on page 126.
Apply Fuzzy Zero Correction	Mark this option to apply this statistical correction to the data. For details on this algorithm and its parameters, see "Fuzzy Zero" on page 458.
Combine Intra-Array Replicates	Mark this option to combine data from replicates within arrays.

SNP Analysis Setting tab

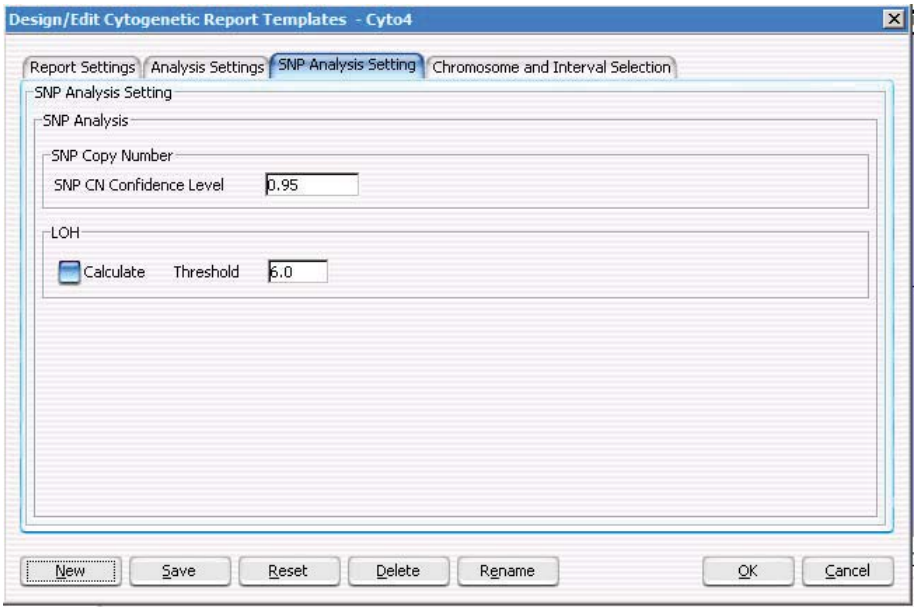


Figure 97 Design/Edit Cytogenetic Report Templates – SNP Analysis Setting tab

The SNP Analysis Setting tab contains these options:

Command/Option	Details
SNP CN Confidence Level	The threshold used to decide whether or not to include a copy number in the SNP analysis.
LOH	
Calculate	Mark to enable LOH calculations.
Threshold	A threshold value to use in determining whether or not to include an LOH call in the analysis.

Chromosome and Interval Selection tab

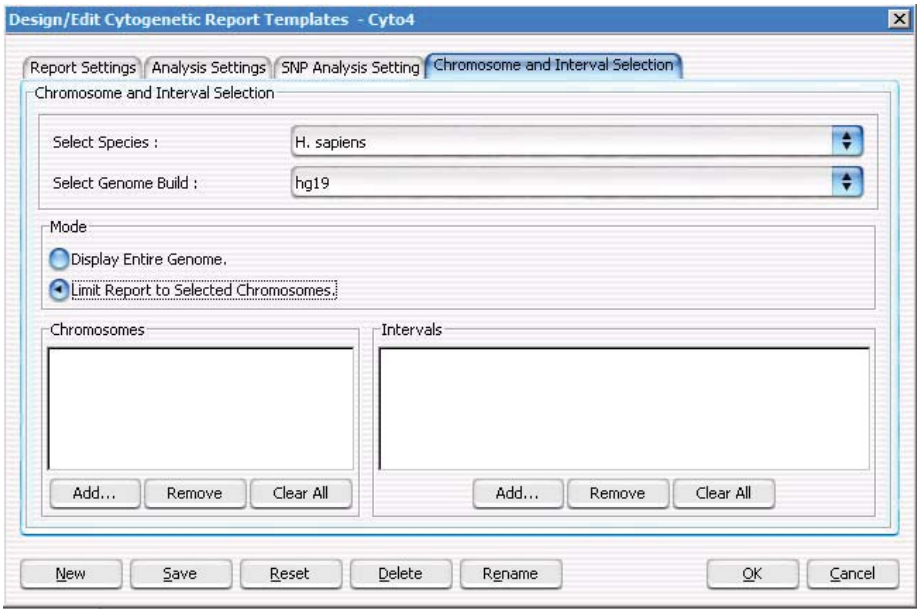


Figure 98 Design/Edit Cytogenetic Report Templates – Chromosome and Interval Selection tab

The Chromosome and Interval Selection tab contains these options and commands:

Option/Command	Details
Select Species	Select the genome's species of origin. The CGH module supports Human, Mouse, and Rat genomes.
Select Genome Build	Select a genome build from the list that matches the build used in your microarray design files.
Mode	
Display Entire Genome	Configures the report to include the entire genome

Option/Command	Details
Limit Report to Selected Chromosomes	Configures the report to include only selected chromosomal regions. If you select this option, the commands under Chromosomes and Intervals become available.
Chromosomes (Select these first)	
List of chromosomes	Displays the chromosomes to include in the report.
Add	Opens the Chromosome Selection dialog box, where you can select chromosomes to include in the report. The program displays the selected chromosomes in the list. See "Chromosome Selection" on page 260.
Remove	Removes chromosomes from the list. To remove a chromosome, click its name, then click **Remove.**
Clear All	Removes all chromosomes from the list.
Intervals	
List of intervals	Displays the chromosomal intervals to include in the report.
Add	Opens the Range for Intervals on <chromosome> dialog box, where you can define an interval to include in the report. To add an interval to the list, select the desired chromosome under **Chromosomes,** then under **Intervals,** click **Add.** In the dialog box that appears, type start and end locations for the region, then click **OK.**
Remove	Removes intervals from the list. To remove an interval, click its name, then click **Remove.**
Clear All	Removes all intervals from the list.

Design Level Filters

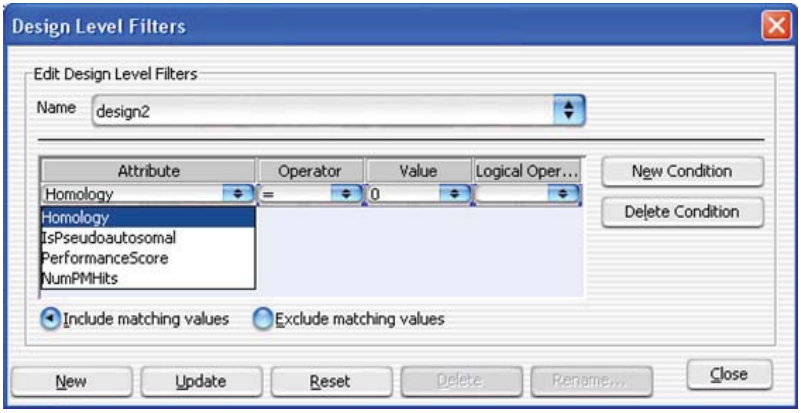


Figure 99 Design Level Filters dialog box

Purpose: This dialog box is used to create or edit design filters that let you include or exclude data from analysis based on selected design attributes.

To open: Click **Preprocessing > Design Filters > Edit Filter.**

Name Select the name of the design filter you want to edit. To create a new filter, and add its name to the list, click **New.**

Filter Conditions table For each condition (row), select options from the list or enter a value, then press **Enter.** To add another row to the table, click **New Condition.**

Each condition has these elements:

- **Attribute** – The design attribute evaluated by the filter.
- **Operator** – How the filter uses the entry in Value to evaluate arrays. For example, the **>=** operator configures the filter to include or exclude probes where the selected attribute is greater than or equal to the entry in Value.

- **Value** – The value the filter uses as the basis to evaluate a probe. You either select a value or, for some attributes, enter a value. For example, if you select the **Homology** attribute with an operator of =, the filter will include or exclude probes that exhibit homology if the value is set to **1.** (The homology attribute is assigned to probes that have more than one mapping in the genome or probes that have secondary hits that are not perfect matches.)
- **Logical Operator** – (Available only if you configure more than one condition) The relationship between the condition and the next one in the list. For example, if you select **AND** in **Logical Operator** for the first condition, the filter includes or excludes a probe if it passes both the first condition and the next condition.

New Condition Adds a new, blank condition (row) to the table.

Delete Condition Removes a condition from the list. To remove a specific condition, click anywhere within the condition, then click **Delete Condition.**

Include/Exclude matching values Select one of these options:

- **Include matching values** – If a probe passes the filter condition, the program *includes* it in the analysis.
- **Exclude matching values** – If a probe passes the filter condition, the program *excludes* it from the analysis.

New Opens an Input dialog box, where you can type a name for the new filter. To accept the name, click **OK.** The program creates the filter, and adds the new name to the Name list.

Update Saves any changes you make to the filter conditions without closing the dialog box.

Reset Restores the values of the filter conditions to what they were before you made any changes to them.

Delete Opens a Confirm dialog box that asks you if you want to delete the selected filter. To delete the filter, click **Yes.**

Rename Opens an Input dialog box where you can type a new name for the filter. To accept the name, click **OK.**

Close Closes the dialog box. If you created or changed a filter, but did not update it, a Confirm dialog box opens. Click **Yes** to accept the changes, **No** to reject the changes, or **Cancel** to return to the dialog box.

Design Properties

Purpose: Gives general and detailed information about a given microarray design. See "To display the properties of a specific design" on page 72.

To open: In the Data pane of the Navigator, right-click the name of a genome build within a design folder, then click **Show Properties.** Several tabs are available.

Attribute tab

Displays general identifying attributes of the array design, and statistics such as the total number of features in the design.

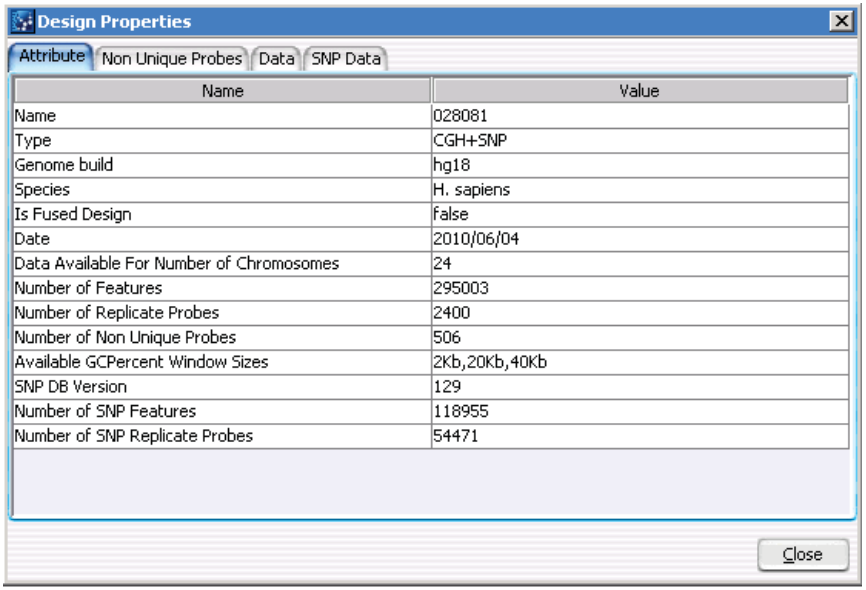


Figure 100 Design Properties dialog box – Attribute tab

Non Unique Probes tab

Displays the nonunique probes in the design. Nonunique probes have more than one mapping in the genome that is a perfect match.

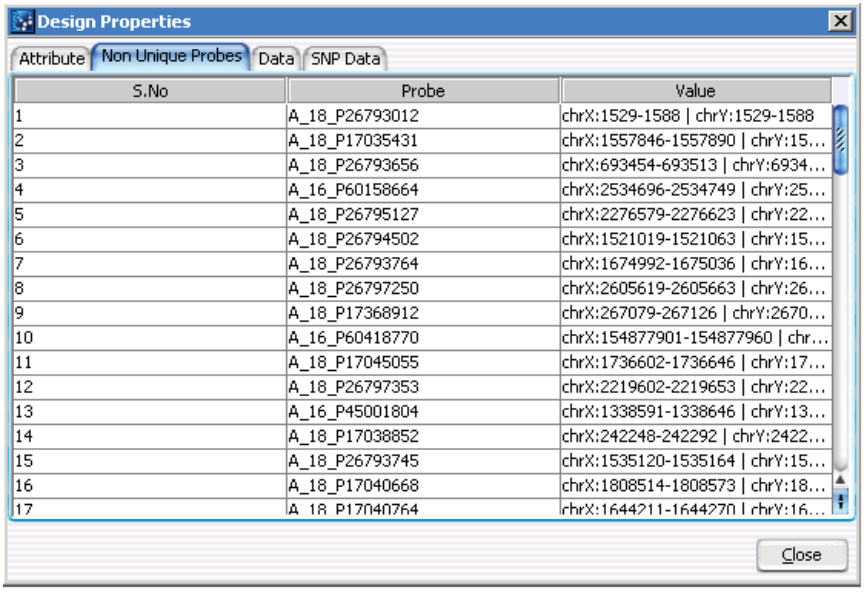


Figure 101 Design Properties dialog box – Non Unique Probes tab

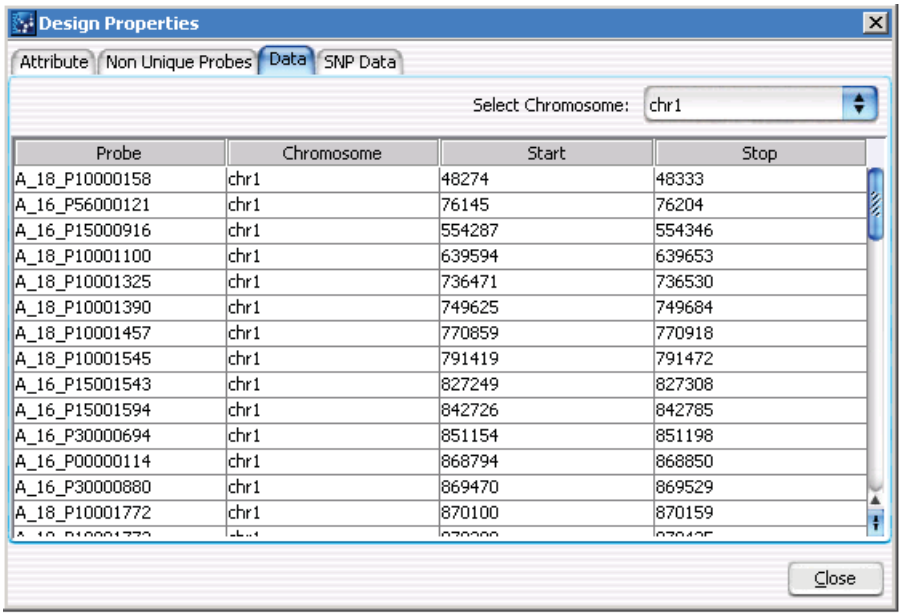
S. No The sequence order of the probes within the tab.

Probe The name of the each nonunique probe.

Value The chromosomal locations to which each of the probes maps. Because these are nonunique probes, two or more locations appear for each probe.

Data tab

Displays the names of the probes in the design and the genomic locations to which they are designed. The tab displays the probes for one chromosome at a time.



Probe	Chromosome	Start	Stop
A_18_P10000158	chr1	48274	48333
A_16_P56000121	chr1	76145	76204
A_16_P15000916	chr1	554287	554346
A_18_P10001100	chr1	639594	639653
A_18_P10001325	chr1	736471	736530
A_18_P10001390	chr1	749625	749684
A_18_P10001457	chr1	770859	770918
A_18_P10001545	chr1	791419	791472
A_16_P15001543	chr1	827249	827308
A_16_P15001594	chr1	842726	842785
A_16_P30000694	chr1	851154	851198
A_16_P00000114	chr1	868794	868850
A_16_P30000880	chr1	869470	869529
A_18_P10001772	chr1	870100	870159

Figure 102 Design Properties dialog box – Data tab

Select Chromosome The chromosome whose probes appear in the list. To display the probes designed to a different chromosome, select one from this list.

Probe The name (Probe ID) of each probe.

Chromosome The name of the chromosome to which the probe is designed.

Start The location on the selected chromosome of the first base pair to which each probe is designed.

Stop The location on the selected chromosome of the last base pair to which each probe is designed.

SNP Data tab

This tab shows design information for SNPs in the design.

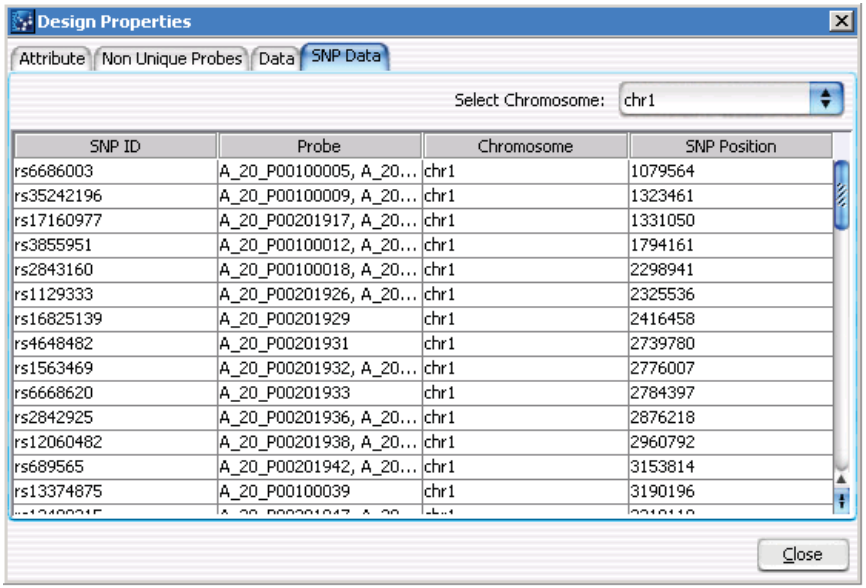


Figure 103 Design Properties dialog box – SNP Data tab

SNP ID The SNP identification.

Probe The name (Probe ID) of the probe. The probe names are separated with a comma.

Chromosome The chromosome on which the probe is located.

SNP Position The position of the SNP on the chromosome.

Differential Aberration Setup

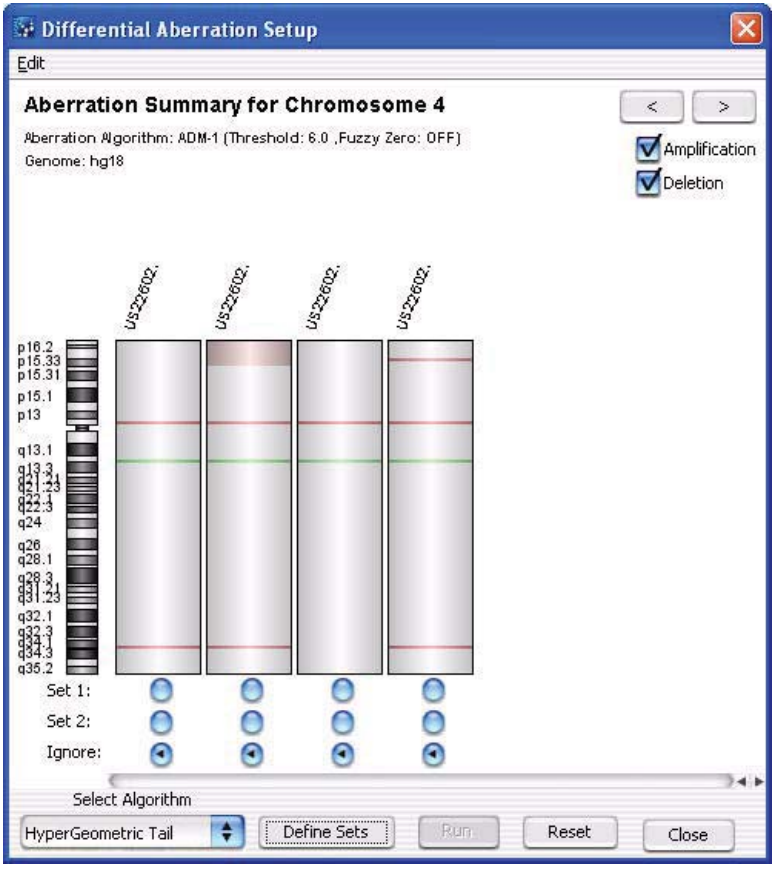


Figure 104 Graphical Aberration Summary dialog box – Special version used to set up a differential aberration analysis

Purpose: This dialog box lets you configure and start a differential aberration analysis.

To open: Click **Discovery > Differential Aberration.** You must first set up and select an experiment with the desired arrays, and apply an aberration detection algorithm. See "To set up a differential aberration analysis" on page 158.

Edit Opens a menu with a **Copy summary to clipboard** option. This option lets you move an image of the dialog box to a document in another program.



Displays aberrations for the previous chromosome in the genome.



Displays aberrations for the next chromosome in the genome.

Amplification Mark this option to display the aberrant regions that represent amplifications as red bars.

Deletion Mark this option to display the aberrant regions that represent deletions as green bars.

Graphical aberration plots Display the selected types of aberrant regions for the selected chromosome for all arrays.

Set 1 Assigns the aberrations in the given array to comparison set 1.

Set 2 Assigns the aberrations in the given array to comparison set 2.

Ignore Does not assign the aberrations in the given array to a comparison set.

Select Algorithm Sets the algorithm used to score differential aberrations. This algorithm is available:

- **HyperGeometric Tail** – The Hypergeometric distribution tests for enrichment, or the likelihood of an increase in the number of aberrations present in one group, given the number of aberrations present in both groups. The HyperGeometric Tail sums the likelihoods as extreme or greater than the original number of aberrations from one group, to calculate an enrichment *p*-value.

For further information on this algorithm, see "Differential Aberration Analysis" on page 512.

Define Sets Opens the Define sets dialog box, where you can assign arrays to comparison groups. This is an alternative to the use of the Set1/Set2/Ignore options that appear below each array. See "Define Sets" on page 287.

Run Runs the differential aberration analysis with the options you selected. The results appear in the Graphical Differential Aberration Summary dialog box. See "Graphical Differential Aberration Summary" on page 344.

Reset Resets the assignment of all arrays to comparison set 1.

Close Closes the dialog box.

Echo Example Plot

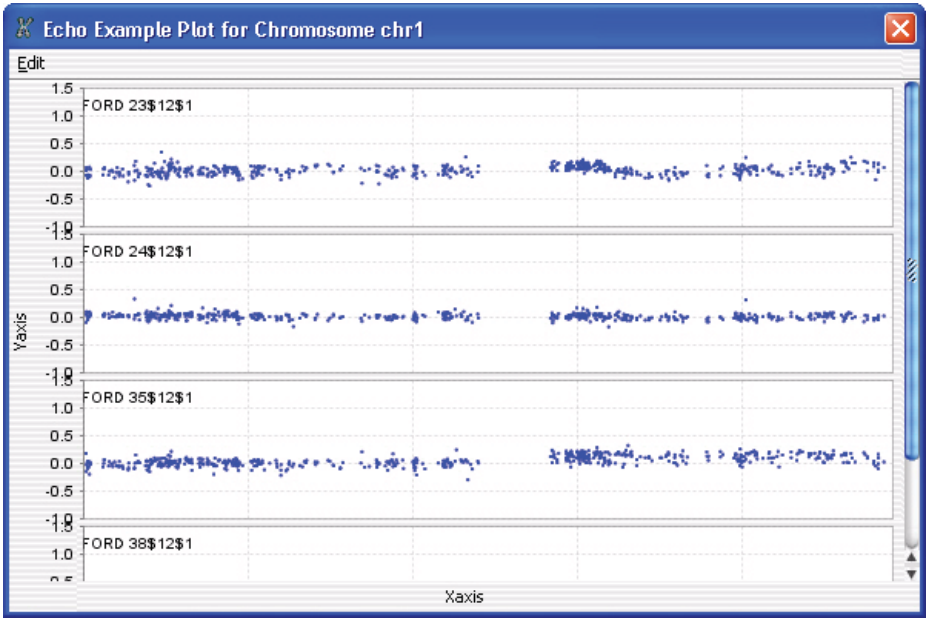


Figure 105 Echo Example Plot

Purpose: The Echo Example Plot is the output of the Echo Example plug-in. It displays the log ratio data for the selected chromosome in the active experiment. Data from all of the selected arrays in the experiment appear as a series of stacked plots, one for each array.

To open: To run the Echo Example plug-in, select the desired experiment, select the desired chromosome in Genome view, then click **Tool > Plugin > Echo Example.**

Edit Opens a menu with a Copy plot(s) to clipboard command. This command copies all of the plots to the clipboard as an image. You can then paste the image into a document in another program.

Plots Each plot displays the log ratio data for the selected chromosome from an individual array in the experiment.

You can right-click anywhere within each plot to display these options:

Option	Description
Properties	Opens the Chart Properties dialog box, where you can change the appearance of many aspects of the plot. See "Chart Properties" on page 255.
Copy	Copies the chart to the Clipboard, where you can paste it into a word processing or other program.
Save as	Opens a Save dialog box, where you can select a location for a saved *.png image file of the plots.
Print	Opens the Windows Page Setup dialog box, where you can set page size, margins, and orientation, and select a printer. From there, you click **OK** to open the Windows Print dialog box, where you can set print options and print the plot.
Zoom In	Opens another menu that lets you zoom in the plot. You can zoom in several ways: • **Both Axes** – Zooms in the Domain (X) axis for all stacked plots, and zooms in the range Range (Y) axis for the specific plot in which you clicked. • **Domain Axis** – Zooms in the Domain (X) axis for all stacked plots. • **Range Axis** – Zooms in the Range (Y) axis for the specific plot in which you clicked. You can also drag across an area of one of the plots to select an area to zoom in on.

Option	Description
Zoom Out	Opens another menu that lets you zoom out the plot. You can zoom out several ways: • **Both Axes** – Zooms out the Domain (X) axis for all stacked plots, and zooms out the range Range (Y) axis for the specific plot in which you clicked. • **Domain Axis** – Zooms out the Domain (X) axis for all stacked plots. • **Range Axis** – Zooms out the Range (Y) axis for the specific plot in which you clicked.
Auto Range	Opens another menu that lets you zoom the plot to encompass the full range of the data. You can zoom in several ways: • **Both Axes** – Appropriately zooms both axes of the specific plot to show the full set of data. • **Domain Axis** – Zooms in the Domain (X) axis for all stacked plots so this axis encompasses the full range of X values of the data. • **Range Axis** – Zooms in the Range (Y) axis for the specific plot in which you clicked, to encompass the full range of Y values of the data.

Edit Array Color

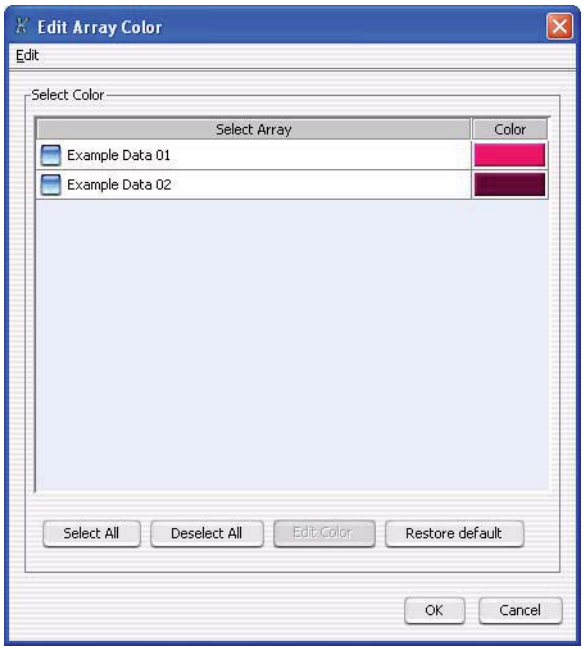


Figure 106 Edit Array Color dialog box

Purpose: To change the display color of the arrays in an experiment

To open: Right-click an experiment name, then click **Edit Array Color.**

Edit Opens a menu with a Copy command. If you click **Copy,** the program copies the list of arrays and their assigned colors to the Clipboard. You can then paste the list into a document in another program such as Word or PowerPoint.

Select Array Mark the check box for the array(s) whose color you want to change.

Color Column Next to the name of an array, click the color swatch in this column to change the display color for the specific array. The Select Color dialog box appears, where you can select the desired color. See "Select Color" on page 394.

Select All Marks all of the check boxes.

Deselect All Clears all of the check boxes.

Edit Color Click to change the color for the selected array(s).

Restore default Click to restore the original color(s) to the selected array(s).

Edit Array Order

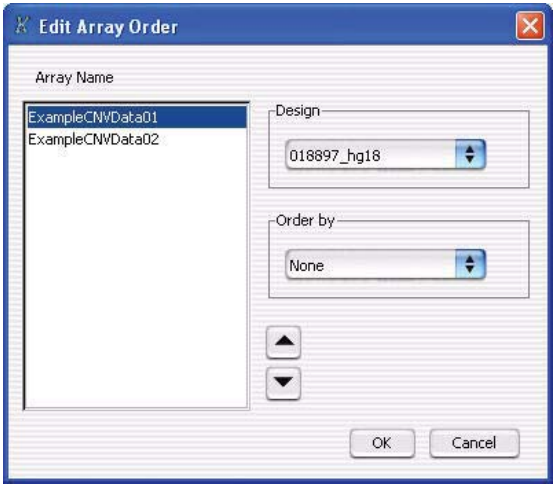


Figure 107 Edit Array Order dialog box

Purpose: This dialog box lets you change the order of arrays in an experiment, which defines the order in which the program displays arrays and array data in the Experiment pane of the Navigator and in Tab View. If you select the **Stacked** rendering style for scatter plots, the array order also determines the order in which the scatter plots for the arrays appear.

To open: Right-click the name of an experiment, then click **Edit Array Order** in the shortcut menu.

Array Name The arrays in the selected design, listed in their current order.

Design Select the name of a design. In Array Name, the program displays the arrays associated with the selected design.

Order by Select an attribute to use as a basis for ordering the list. For example, if you select Barcode, the program reodered that Array Name list based on Barcode.

Moves a selected array up in the Array Name list. To select an array in this list, click the name of the array.

Moves a selected array down in the Array Name list. To select an array in this list, click the name of the array.

OK Applies the new array order.

Cancel Closes the dialog box without making any changes to the array order.

Enrichment Analysis Setup

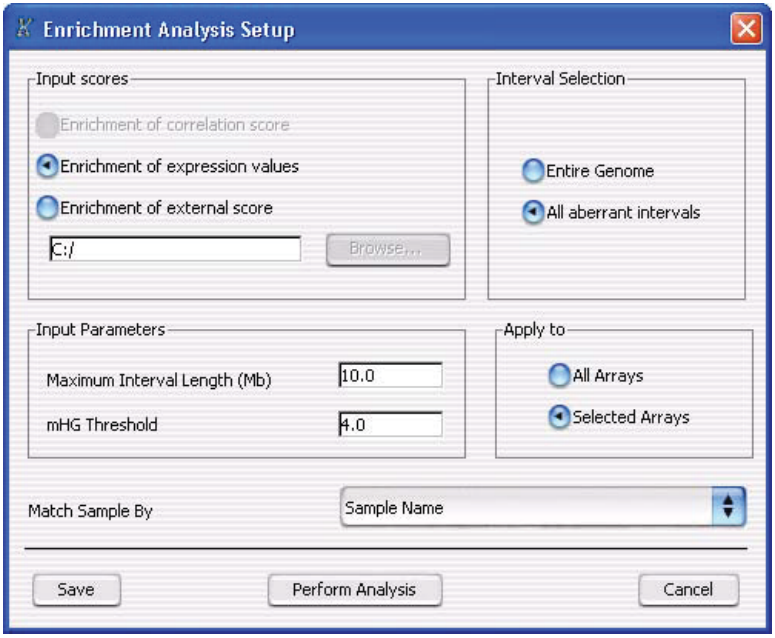


Figure 108 Enrichment Analysis Setup dialog box

Purpose: Enrichment analysis computes an enrichment score for each genomic interval. The intervals can represent the entire genome, or only detected aberrant regions. An enrichment score is a measure of the increase in the number of probes with high values of a selected type, as compared to what can be expected at random, for a given interval. The program can calculate an enrichment score for any sample for which you have both CGH and expression data. For details on this kind of analysis, see "Algorithms for Joint CGH-Expression Analysis" on page 526.

To open: The Enrichment Analysis Setup dialog box opens when you click **Discovery > Joint-Analysis > Enrichment Analysis > Setup.** It also opens when you click **Enrichment Analysis** in the Correlation Results dialog box (see "Correlation Results" on page 272).

Input scores The type of enrichment analysis, based on the source of the values to be analyzed. Select one of these options:

- **Enrichment of correlation score** – This type of analysis considers the correlation scores from a joint CGH-expression correlation analysis. This option is only available after you perform a correlation analysis.
- **Enrichment of expression values** – This type of analysis considers expression values in existing expression data.
- **Enrichment of external score** – This type of analysis considers values in a file you import. The values can be any quantitative information, listed by interval.

  The imported file must be a tab-delimited *.txt file with columns titled Chr, Start, Stop, and Score–<array name>. Chr is the number (or name) of the chromosome, and Start and Stop are the locations of the first and last base pairs of the interval, respectively. Score–<array name> is a decimal value.

  The first line of the file must be column headers, and rest of the file is data lines. For example, the first few lines of a file to analyze an array called BT4 could look like this:

```
Chr         Start       Stop        Score-BT4
1           779448      808099      -0.6
1           1001925     127708      -0.3
```

Interval Selection Select one of the options. You can analyze the entire genome, or only aberrant intervals.

Input Parameters Edit these parameters as desired:

- **Maximum Interval Length (Mb)** – The largest interval considered by the enrichment scoring algorithm. The default value is 10.0 Mb. The program ignores the interval in the enrichment analysis if it exceeds this size. If you set this parameter to 0, the program uses a default of 1000 Mb as the maximum interval length.
- **mHG Threshold** – The lowest enrichment score for an enrichment event to be considered significant. Increase this number to make the detection process more stringent. The default threshold score is 4.0.

Apply to Select one of these options:

- **All Arrays** – Analyzes all arrays in the active experiment.
- **Selected Arrays** – Analyzes only the selected arrays in the active experiment.

Match Sample By Select an array attribute. The analysis matches CGH arrays with expression arrays based on the values defined for this attribute.

Save Saves the analysis parameters internally. Later, when you run an enrichment analysis, the program uses the parameters you set. To run an enrichment analysis, click **Discovery > Joint-Analysis > Enrichment Analysis > Perform.**

Perform Analysis Opens the Match Sample dialog box, where you can view the paired CGH-expression arrays, and start the enrichment analysis.

Cancel Closes the dialog box without changing any parameters.

Experiment Properties

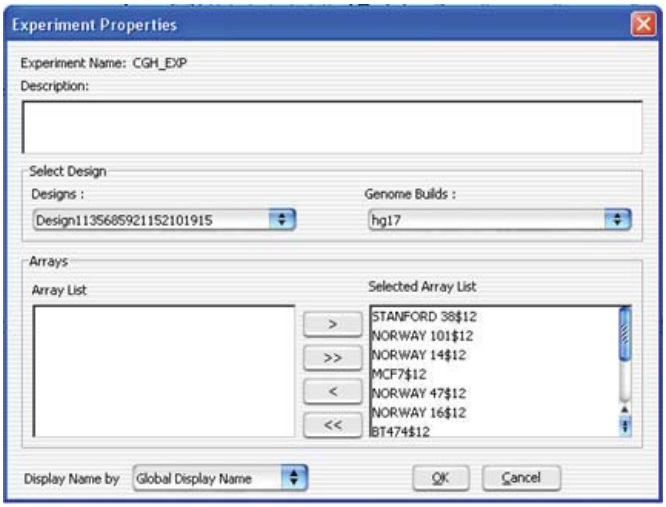


Figure 109 Experiment Properties dialog box

Purpose: Lets you select array designs and data to link to an experiment. See "To add arrays to an experiment" on page 62.

To open: In the Create Experiment dialog box, click **Properties**, or in the Experiment pane of the Navigator, right-click the name of an experiment, then click **Show Properties.**

Experiment Name (Read-only) The name of the selected experiment.

Description Description that was typed when the experiment was created.

Select Design

Designs Displays all of the designs available in the program. Select the design associated with arrays to add to the experiment.

Genome Builds Displays the genome build(s) associated with the design. Select the desired genome build to display the arrays associated with a single genome build.

Arrays

Array List Displays the arrays in the selected design that are available for this experiment.

- To select an array to move to the Selected Array List, click its name.
- To select additional arrays, hold down the **ctrl** key and click their names.
- To select a contiguous block of arrays, click the name of the first array, then hold down the **shift** key and click the name of the last one.

Selected Array List Displays the arrays that you have selected for this experiment.

> Moves the selected arrays in Array List to the Selected Arrays List. You can move arrays from as many designs as you like, as long as they are all associated with the same genome build.

>> Moves all of the arrays in Array List to the Selected Arrays List.

< Removes an array from the Selected Array List. To select an array for removal, click its name. If desired, you can re-add an array.

<< Clears the Selected Array List.

Display Name by Click to select how the array names will be displayed in the experiment. The Global Display name is the name assigned in Sample Manager for the array. See the *Sample Manager User Guide* for more information.

OK Adds selected arrays in the Selected Array List to the experiment, and closes the dialog box.

Cancel Closes the dialog box without adding any arrays to the experiment.

Export

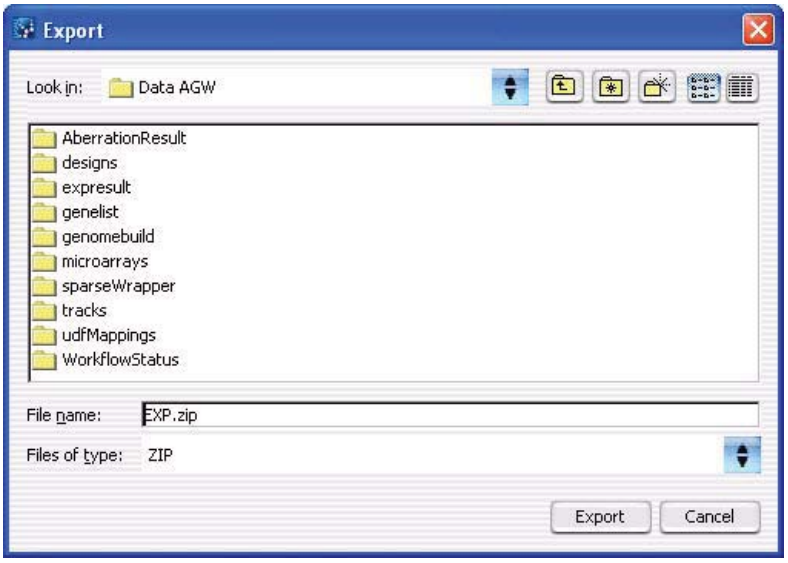


Figure 110 Export dialog box – Several types of file exports use this dialog box. This specific example exports selected experiment(s) as a ZIP format file.

Purpose: Lets you select a location for an exported file.

To open: This dialog box appears after you select specific experiment(s), track(s), or filter(s) to export. See "To export experiments" on page 83, "To export tracks" on page 84, or "To export filters" on page 85.

Look in Displays the folder or other location whose contents appear in the main pane of the dialog box. To select another folder or other location, click .

Moves to the next higher folder level.

Moves to the Desktop.

Creates a new folder in the selected location in *Look in*.

Displays the names only of folders, files, and other locations in the main pane of the dialog box.

Displays both the names and details of folders, files, and other locations in the main pane of the dialog box.

Main pane Displays the folders, files, and other locations in the selected location in *Look in*. The program limits listed files to the type selected in *Files of type*. To select a file, click its name. To open a folder or other location, double-click its name.

File name Displays the name of the file to which the exported content will be saved. To change the name, you can either select a file in the main pane of the dialog box, or type a new name.

Files of type Limits the files listed in the main pane to those of the appropriate type for your specific kind of export. To show all files, click  , then select **All Files.**

Export Saves the selected content to the location specified in the dialog box.

Cancel Cancels your selections and closes the dialog box.

Export Array Attributes

Purpose: This dialog box lets you select arrays whose attributes you want to export. It contains two tabs: an Array tab where you select the arrays, and an Attribute tab where you select the attributes of the selected arrays to export. See "To export array attributes" on page 82.

To open: In the Home command ribbon, click **Export > Array Attributes**, or in the Experiment pane of the Navigator, right-click the name of an experiment, then click **Export Attributes.**

Array tab

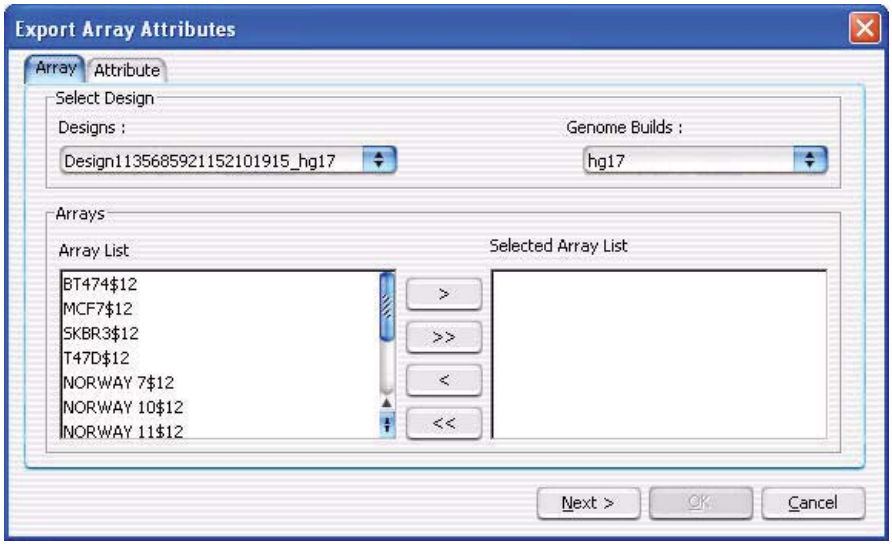


Figure 111 Export Array Attributes dialog box – Array tab

Select Design

Designs Displays all of the designs available in the program. Select the design associated with arrays whose attributes you want to export.

Genome Builds Displays the genome build(s) associated with the design. Select the desired genome build to display the arrays.

Arrays

Array List Displays the arrays in the selected design.

- To select an array to move to the Selected Array List, click its name.
- To select additional arrays, hold down the **ctrl** key and click their names.
- To select a contiguous block of arrays, click the name of the first array, then hold down the **shift** key and click the name of the last one.

Selected Array List Displays the arrays that you have selected for this experiment.

> Moves the selected arrays in Array List to the Selected Arrays List. You can move arrays from as many designs as you like, as long as they are all associated with the same genome build.

>> Moves all of the arrays in Array List to the Selected Arrays List.

< Removes an array from the Selected Array List. To select an array for removal, click its name. If desired, you can re-add an array.

<< Clears the Selected Array List.

Next Moves to Attribute tab for attribute selection. You must select one or more arrays before you move to the Attribute tab.

Cancel Closes the dialog box without selecting any arrays for export.

Attribute tab

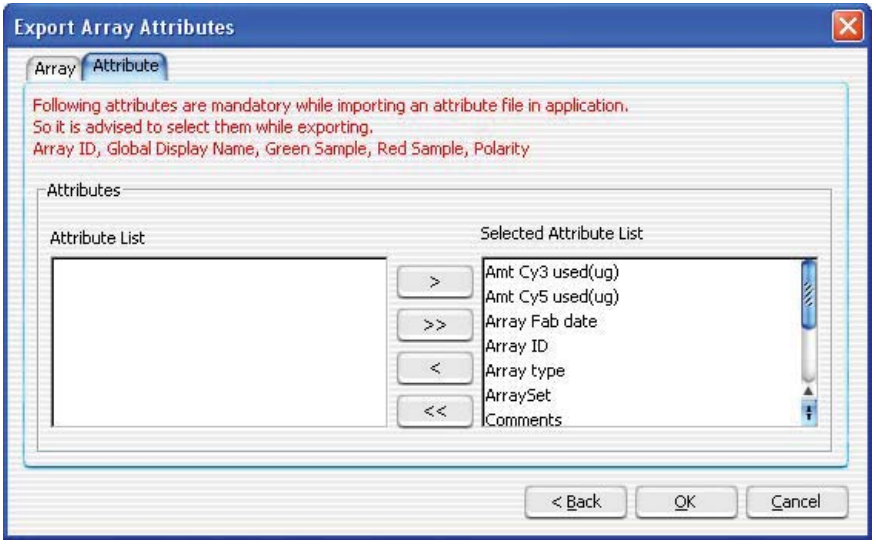


Figure 112 Export Array Attributes dialog box – Attribute tab

Attributes

Selected Attribute List For the selected arrays, displays the attributes to be exported if you click **OK**.

- To select an attribute to move to the Attribute List, click its name.
- To select additional attributes, hold down the **ctrl** key and click their names.
- To select a contiguous block of attributes, click the name of the first attribute, then hold down the **shift** key and click the name of the last one.

NOTE You must select the following mandatory attributes, or else the attribute file cannot be imported at a later time: Array ID, Global Display Name, Green Sample, Red Sample, Polarity.

Attribute List Displays the attributes that will not be exported for the selected arrays.

< Removes an attribute from the Selected Attribute List. To select an attribute for removal, click its name. If desired, you can re-add an attribute.

<< Clears the Selected Attribute List.

> Moves the selected attributes in the Attribute List to the Selected Attribute List.

>> Moves all of the attributes in the Attribute List to the Selected Attribute List.

Back Moves back to the Array tab for array selection or removal.

OK Opens the Export dialog box, where you select or type a name for the exported attribute file. See "Export" on page 315.

Cancel Closes the dialog box without exporting the attributes in the Selected Attribute List.

Export Experiments

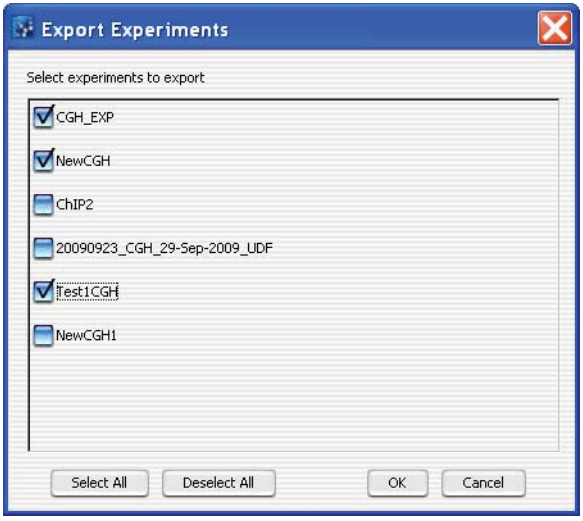


Figure 113 Export Experiments dialog box

Purpose: Lets you select experiments for export. The program exports all array designs and data associated with the experiments as a single ZIP file. This file does not include any parameter settings, array selections, or results. See "To export experiments" on page 83.

To open: In the Home tab, click **Export > Experiments.**

•

Select experiments to export Displays all experiments available for export. Mark each experiment you want to export.

Select All Selects all experiments for export.

Deselect All Clears all check boxes under Select experiments to export.

OK Opens an Export dialog box. See "Export" on page 315.

Cancel Cancels the export and closes the dialog box.

Export Filters

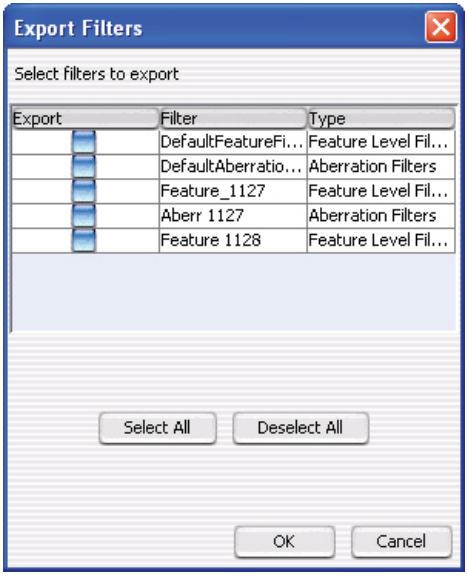


Figure 114 Export Filters dialog box

Purpose: Lets you select feature-level, array-level, design, metric set, and/or aberration filters, to export as a single *.xml file. See "To export filters" on page 85.

To open: In the Home tab, click **Export > Filters.**

Select filters to export Displays all of the filters available in the program. The table has these columns:

- **Export** – Mark the check box next to each filter to export.
- **Filter** – The name of each filter.
- **Type** – The type of content to which the program applies each filter.

For more information on filters, see "Changing Preprocessing Options" on page 126.

Select All Selects all available filters for export.

Deselect All Clears all of the check boxes under Select filters to export.

OK Opens the Export dialog box, where you can select a location for the exported *.xml file of filters. See "Export" on page 315.

Cancel Cancels the export and closes the dialog box.

Export Tracks

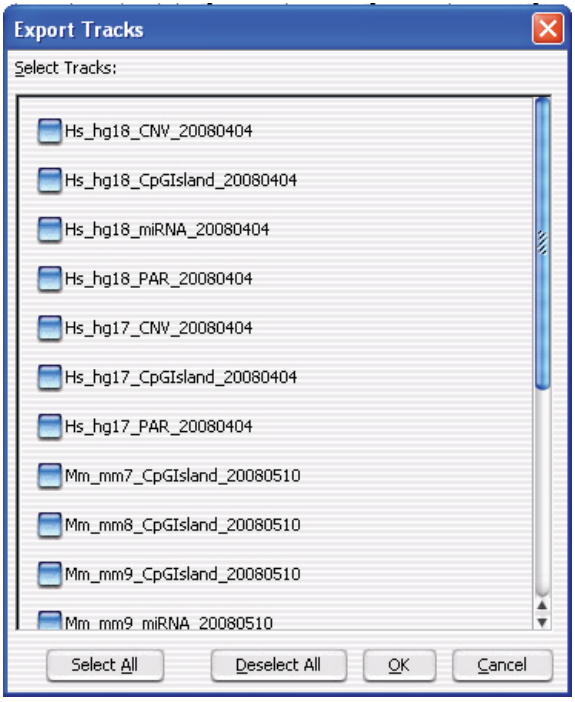


Figure 115 Export Tracks dialog box

Purpose: Lets you select tracks to export as a single BED format file. See "To export tracks" on page 84.

To open: In the Home tab, click **Export > Tracks.**

Select Tracks Displays all of the tracks available in the program. Mark the check box next to each track to export.

For more information about tracks, see "Create Track" on page 284.

Select All Selects all available tracks for export.

Deselect All Clears all of the check boxes under Select Tracks.

OK Opens the Export dialog box, where you can select a location for the exported BED format file. See "Export" on page 315.

Cancel Cancels the export and closes the dialog box.

Feature Level Filters

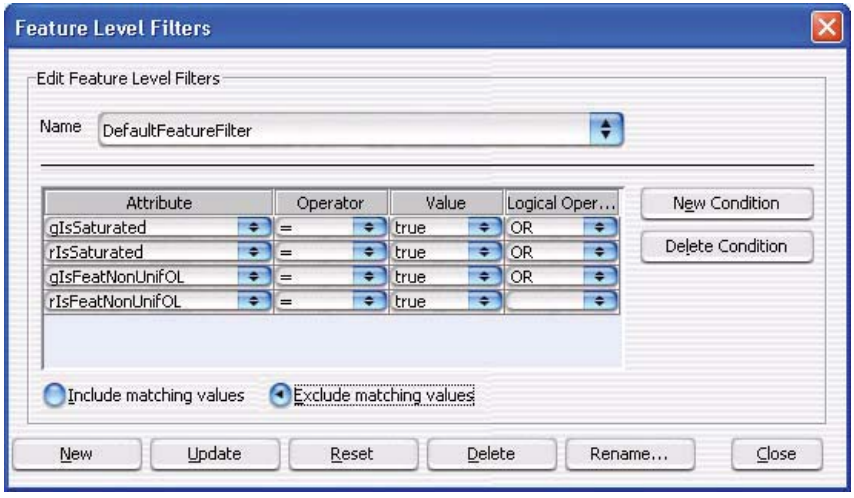


Figure 116 Feature Level Filters dialog box

Purpose: This dialog box allows you to create, change, or delete feature filters. Feature filters include or exclude data from certain microarray features based on specific conditions. See "To create or modify a feature filter" on page 127.

To open: Click **Preprocessing > Feature Filters > Edit Filter.**

Name Select the name of the feature filter you want to edit. To create a new filter, and add its name to the list, click **New.**

Filter conditions For each condition (row), select options from the lists. Specifically in **Value,** select an option from the list, if available. Otherwise, type a value, then press **Enter.** To add another row to the table, click **New Condition.**

Each condition has these elements:

- **Attribute** – The feature attribute evaluated by the filter.
- **Operator** – How the filter uses the entry in Value to evaluate arrays. For example, the **>=** operator configures the filter to include or exclude features where the selected attribute is greater than or equal to the entry in Value.
- **Value** – The value the filter uses as the basis to evaluate a feature. For example, if you select the **gIsSaturated** attribute, the **=** operator, and a value of **true,** the filter includes or excludes a feature if its gIsSaturated attribute is **true.**

Logical Operator – (Available only if you configure more than one condition.) The relationship between the condition and the next one in the list. For example, if you select **AND** in **Logical Operator** for the first condition, the filter includes or excludes an array if it passes both the first condition and the next condition.

New Condition Adds a new, blank condition (row) to the table.

Delete Condition Removes a condition from the list. To remove a specific condition, click anywhere within the condition, then click **Delete Condition.**

Include/Exclude matching values Select one of these options:

- **Include matching values** – If a feature passes the filter conditions, the program *includes* it in the analysis.
- **Exclude matching values** – If a feature passes the filter conditions, the program *excludes* it from the analysis.

New Opens an Input dialog box, where you can type a name for the new filter. To accept the name, click **OK.** The program creates the filter, and adds the new name to the Name list.

Update Saves any changes you make to the filter conditions without closing the dialog box.

Reset Restores the values of the filter conditions to what they were before you made any changes to them.

Delete Opens a Confirm dialog box that asks you if you want to delete the selected filter. To delete the filter, click **Yes.**

Rename Opens an Input dialog box where you can type a new name for the filter. To accept the name, click **OK.**

Close Closes the dialog box. If you created or changed a filter, but did not update it, a Confirm dialog box opens. Click **Yes** to accept the changes, **No** to reject the changes, or **Cancel** to return to the dialog box.

Find in column

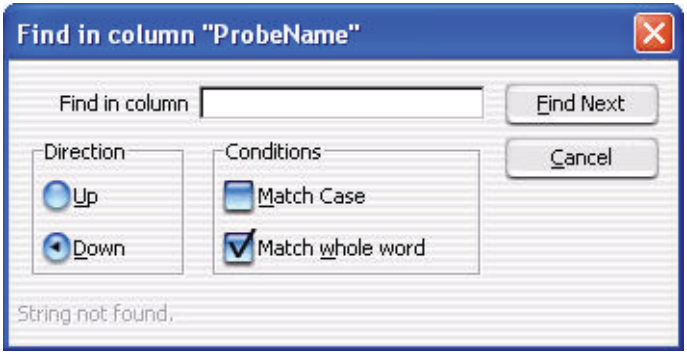


Figure 117 Find in column dialog box

Purpose: This dialog box allows you to set search parameters for a specific column entry for the selected chromosome. Based on these parameters, the program can highlight the row of the first entry that matches. The cursor then moves to the location defined in the row.

To open: Right-click any entry in a tab in Tab View other than a column heading, then click **Find in column** in the shortcut menu.

Find in column Type all or part of the entry you want to find.

Direction Select a search direction:

- **Up** – Sets the search to scan the column you clicked in an upward direction from the highlighted row.

- **Down** – Sets the search to scan the column you clicked in an downward direction from the highlighted row.

Conditions Mark any of these search options:

- **Match Case** – Mark this option to take case into account. For example, if you mark Match Case, and you type `aa351` in Find in column, the search finds the next entry in the column that contains **aa351.** It does *not* find entries that contain **AA351** or **Aa351.**
- **Match whole word** – Mark this option to only find entries in which the complete entry matches what you type in Find in column. For example, if you type `AA351` in Find in column, and mark **Match whole word,** the program finds the next **AA351** entry. It does not find entries such as **AA3512** or **AA351992.**

Find Next Finds the next matching entry in the selected column, and moves the cursor to the location defined in the row that contains the entry. The search is performed only for the chromosome selected in the Genome View.

Cancel Closes the dialog box.

Font Selection

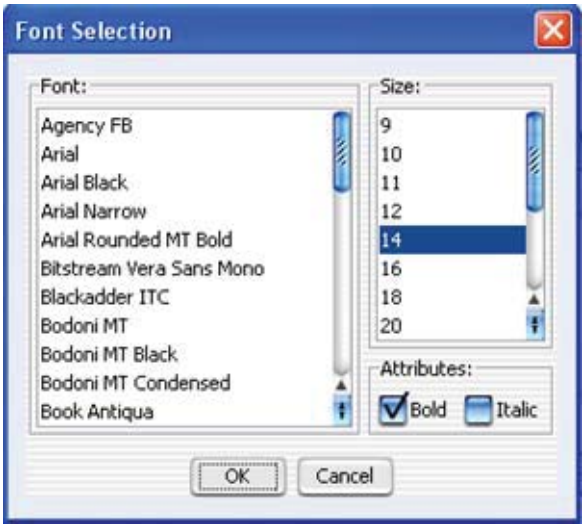


Figure 118 Font Selection dialog box

Purpose: This dialog box lets you select a font for a chart.

To open: In the Chart Properties dialog box, click either the **Title** or **Plot** tab and then next to Font, click **Select**.

Font Select the font to use for the chart.

Size Select the size to display the font.

Attributes Mark the check box next to the attribute for the font. You can select Bold or Italic, or both.

OK Accepts the selections and close the dialog box.

Cancel Closes the dialog box without making font changes.

Gene List

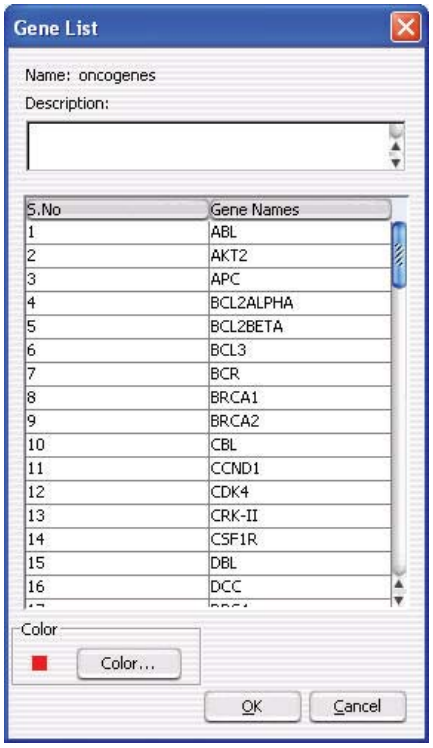


Figure 119 GeneList dialog box

Purpose: Lets you view the names of the genes in a specific gene list and to change the display color of the gene list. See "To display the genes in a gene list" on page 75.

To open: In the My Entity List pane of the Navigator, right-click the name of a gene list, then click **View in Table.**

Name (Read-only) The name of the gene list.

Description (Optional) Brief descriptive comments about the gene list, such as how it was created or the nature of the genes in the list. You can edit the description.

S. No The sort order number. This is the index number of each gene within the gene list.

Gene Names The names of the genes in the gene list.

Color Opens the Choose Gene List Color dialog box, where you can change the display color for the gene list. See "Chromosome Selection" on page 260.

OK Saves the gene list with any new description or display color, and closes the dialog box.

Cancel Closes the dialog box without making any changes to the gene list.

Generating Genotype Reference File

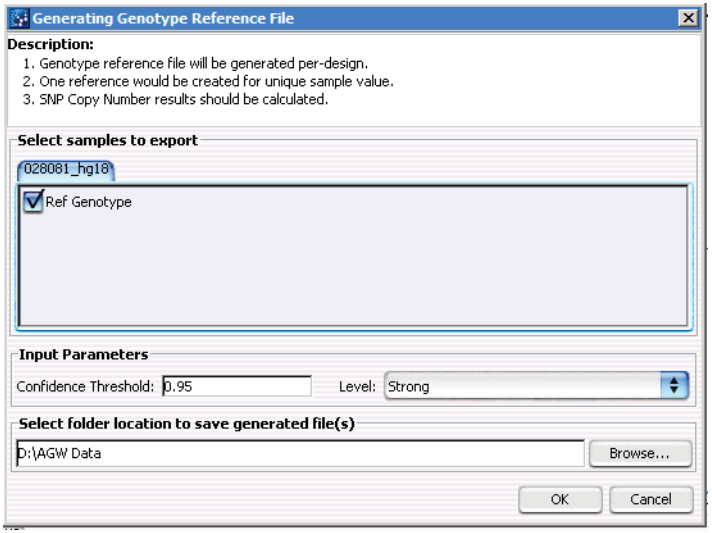


Figure 120 Generating Genotype Reference File

Purpose: Creates a genotype reference file using selected samples and parameters. You can import the new genotype reference file later to the Agilent Genomic Workbench database, and use it to analyze CGH+SNP data.

To open: In the Discovery tab, click **Generate Genotype Reference**.

Select samples to export

If the experiment has more than one reference set in the Red Sample (Green Sample for dye-flipped), mark the check box next to the sample(s) you want to genotype. Only the arrays containing the selected sample(s) are used.

Input parameters

Select the confidence parameters that must be met to include the genotype information.

Confidence threshold.

Type the confidence threshold to apply. A higher value makes the selection more stringent. For the default threshold of 95%, the following confidence assignments are made:

Confident: The software reports an AsCN within 0.05 of an integer. It is considered to have called the AsCN as the nearest integer, with 95% confidence.

Tentative: The software reports an AsCN which is *not* within 0.05 of an integer. It is considered to have called the AsCN as the nearest integer, but with low confidence.

No call: The software makes no call for the SNP in that sample.

Confidence level

Select a level of confidence for the selection.

- **Strong** – The SNP is called as Confident in at least one sample, and it is never called as Tentative. Also, all samples in which it is called report the same (integral) AsCN.
- **Weak** – All samples in which the SNP is called report the same (integral) AsCN. Some samples are called as Tentative.
- **Majority** – The SNP is called as Confident in at least one sample, but it is sometimes called with a different (integral) AsCN. All such calls are Tentative.
- **Contradictory** – The SNP is called with different (integral) AsCNs in different samples, but none of the calls are Confident.

Genotype Reference Details

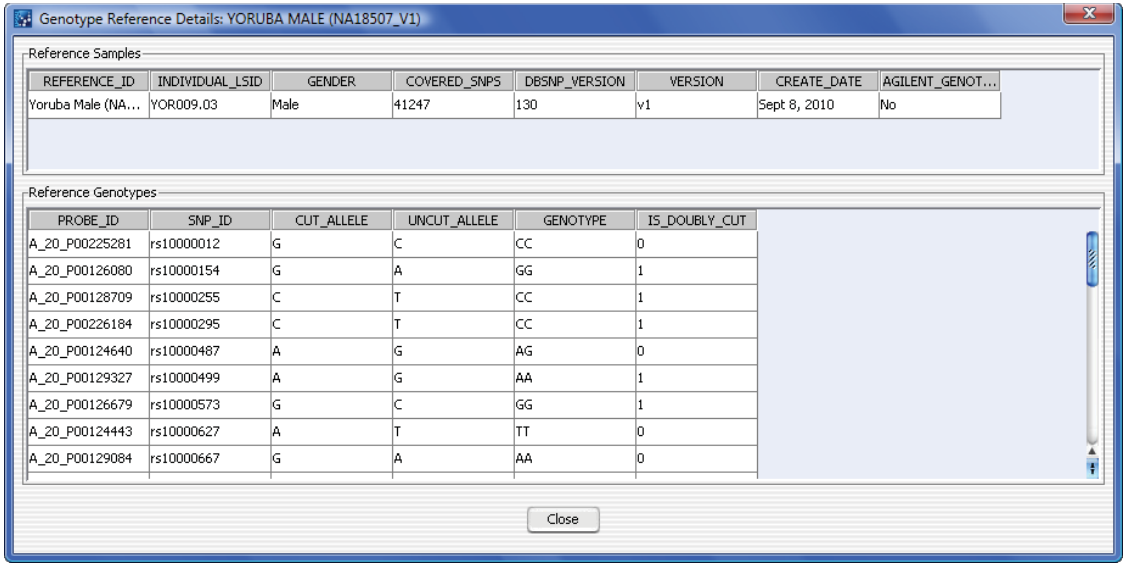


Figure 121 Genotype Reference Details dialog box

Purpose: Shows the details of the genotype reference selected in the Genotype pane of the Navigator.

To open: In the Genotypes pane of the Navigator, right-click on a genotype reference, and select **Show Properties**.

Genotype Reference Importer

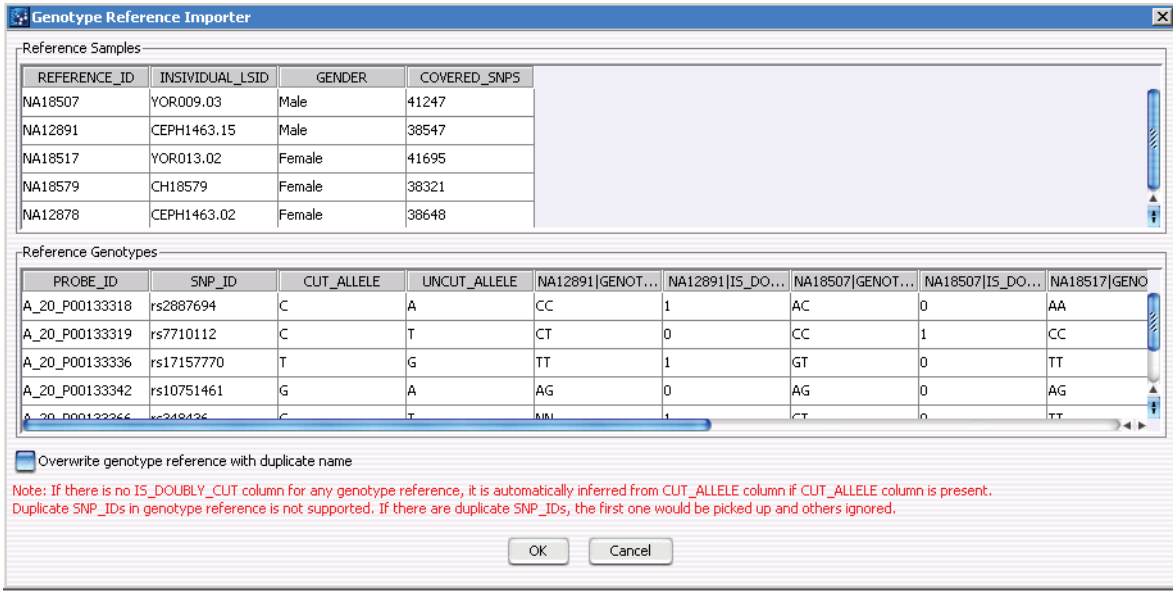


Figure 122 Genotype Reference Importer dialog box

Purpose: Displays the contents of a genotype reference file you want to import, and lets you choose to overwrite existing genotype references in the database when you import the file.

To open: From the Home tab, click **Import > Genotype References**. Or, in the Home tab, right-click the Genotypes folder, and select **Import Genotype Reference.** In the Import Genotype Reference Files dialog box, browse to a location and select the genotype reference file you want to import, then click **OK**.

You can import a genotype reference file you generated and/or exported from the program, or a tab-delimited text file in the required format. See "Generating Genotype Reference File" on page 329 and "Required format for genotype reference files" on page 333.

Reference Samples Displays a table of the samples in the file, including number of SNP probes covered by the sample.

Reference Genotypes Displays a table of the genotypes in the file. Duplicate SNP_IDs are not allowed. If there are duplicate SNP_IDs in the file, only the first SNP_ID is imported.

NOTE If the CUT_ALLELE column is present for a genotype reference, and there is no IS_DOUBLY_CUT column, the IS_DOUBLY_CUT column will be automatically inferred from the CUT_ALLELE column.

Required format for genotype reference files

A genotype reference file must be a tab delimited text file. It has the following format and requirements.

Header section – Contains description and information about the references.

- **Reference_id** – An identifier for the genotype reference
- **Individual_LSID** – ID for the individual
- **Gender** of the reference
- **Covered_SNPs** – how many SNPs in reference file
- **dbSNP_version** – the SNP version

Reference genotype table – each row corresponds to a Probe, with the following column headings

- **Probe_Id**
- **Species**
- **Snp_id** – The ID of SNP associated with probe. There may be more than one probe associated with a SNP. Duplicate SNP IDs are not allowed. If there are duplicates, only the first one is used.
- **Cut Allele**
- **Uncut Allele**
- **Reference ID|genotype** – genotype for the SNP in that reference.
- **Reference ID|is_doubly_cut** – a flag denoting whether the SNP is doubly cut in the reference (1 doubly cut, 0 not doubly cut)

#reference samples

reference_id	individual_LSID	Gender	covered_SNPs	dbSNP_version
NA18507	YOR009.03	Male	41247	130
NA12891	CEPH1463.15	Male	38547	130
NA18517	YOR013.02	Female	41695	130
NA18579	CH18579	Female	38321	130
NA12878	CEPH1463.02	Female	38648	130

#

#reference genotypes

probe_id	species	snp_id	cut_allele	uncut_allele	NA12891|genotype	NA12891|is_doubly_cut	NA18507|genotype	NA18507|is_doubly_cut
A_20_P00122122	Hs	rs2405741	G	C	CC	0	CC	0
A_20_P00224028	Hs	rs2405741	G	C	CC	0	CC	0
A_20_P00122159	Hs	rs1492169	G	A	AA	0	AG	0

Figure 123 Example of part of a genotype reference file (including five references)

Go To Gene/Genomic Location

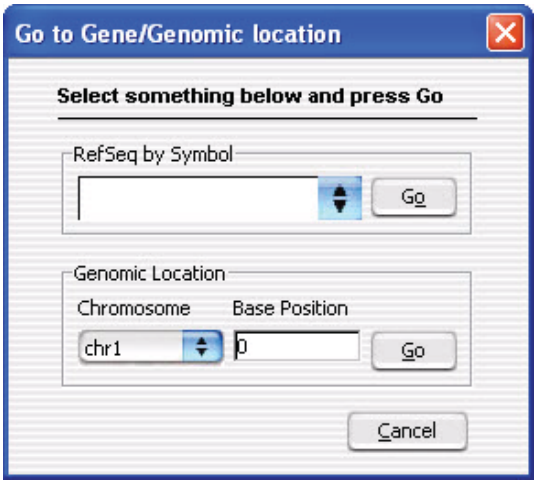


Figure 124 Go To Gene/Genomic location dialog box

Purpose: To find a specific gene location in Gene View by either selecting the RefSeq by Symbol or by selecting the Genomic Location.

To open: Click **Home > Go to Gene/Genomic location**.

RefSeq by Symbol Select the Reference Sequence accession symbol from NCBI, and click **Go.**

Genomic Location

- **Chromosome** – The chromosome number.
- **Base Position** – The position on the chromosome.

Click **Go** after selecting the chromosome number and the position of the gene on the chromosome.

Cancel Closes the dialog box.

Graphical Aberration Summary (report)

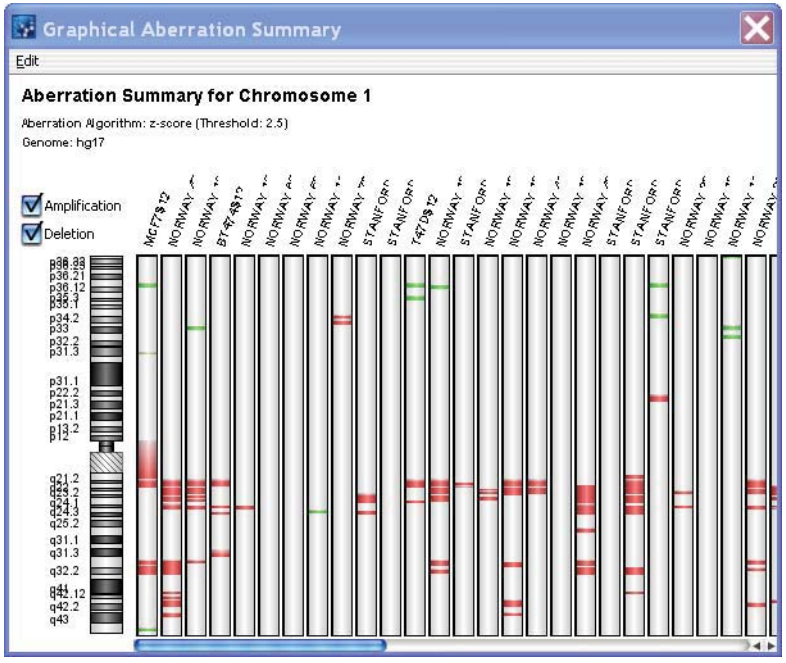


Figure 125 Graphical Aberration Summary report

Purpose: To display a graphical representation of aberrations from selected arrays in the current experiment.

To open: From the Reports tab, under Aberration, click **Graphical**.

Amplification Mark to display amplification aberrations. These regions appear in the plots in red. Clear to remove display of amplifications.

Deletion Mark to display deletion aberrations. These regions appear in the plots in green. Clear to remove display of deletions.

Edit Opens a menu with a Copy summary to clipboard option. If you select this option, the program copies the main pane of the window to the Clipboard as an image. You can then paste the image into a document in another program.

Graphical Aberration Summary (for common aberration analysis)

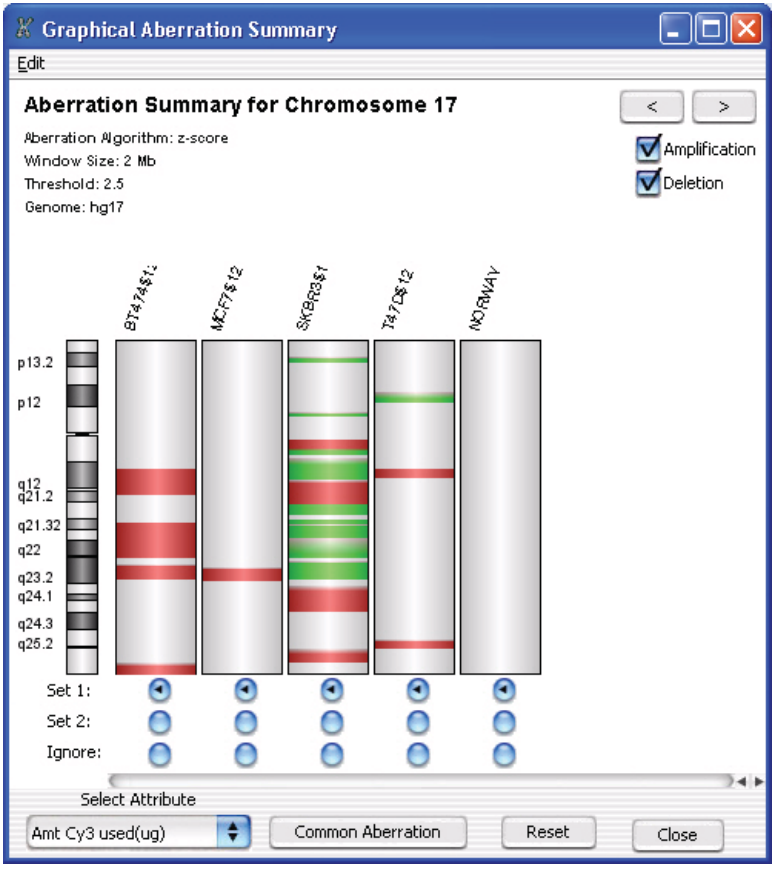


Figure 126 Graphical Aberration Summary (Used to set up a graphical common aberration analysis)

Purpose: This Graphical Aberration Summary allows you to display detected aberrations for the individual arrays in the analysis, and to group them together for common aberration analysis.

To open: Click **Calculate Aberration** in the Graphical Common Aberration Setup dialog box.

Aberration Summary Heat map-style plots for each array appear next to an ideogram of the current chromosome. Within the plots, significant amplifications appear in red, and significant deletions appear in green.

Edit Opens a menu with a Copy summary to clipboard option. If you select this option, the program copies the main pane of the window to the clipboard as an image. You can then paste the image into a document in another program.

< Displays plots for the previous chromosome.

> Displays plots for the next chromosome.

Amplification Mark this option to display chromosomal regions with significant amplifications for all selected arrays. These regions appear in the plots in red.

Deletion Mark this option to display chromosomal regions with significant deletions for all selected arrays. These regions appear in the plots in green.

Set 1/Set 2/ Ignore Lets you select up to two specific subsets of arrays for analysis. The program analyzes each subset separately. Select **Set 1** for the arrays to include in the first set, and **Set 2** for the arrays to include in the second set. To exclude an array from the analysis, select **Ignore.**

Select Attribute Lets you select an array attribute. The value of this attribute for a given array appears as a tool tip when you place the pointer over the corresponding graphical cylinder.

Common Aberration Calculates the common aberrations for each specified grouping of arrays, and displays the results in the Graphical Common Aberration Summary window. See "Graphical Common Aberration Summary" on page 341.

Reset Restores all of the grouping option buttons to **Set 1.**

Close Closes the dialog box without calculating common aberrations.

Graphical Common Aberration Setup

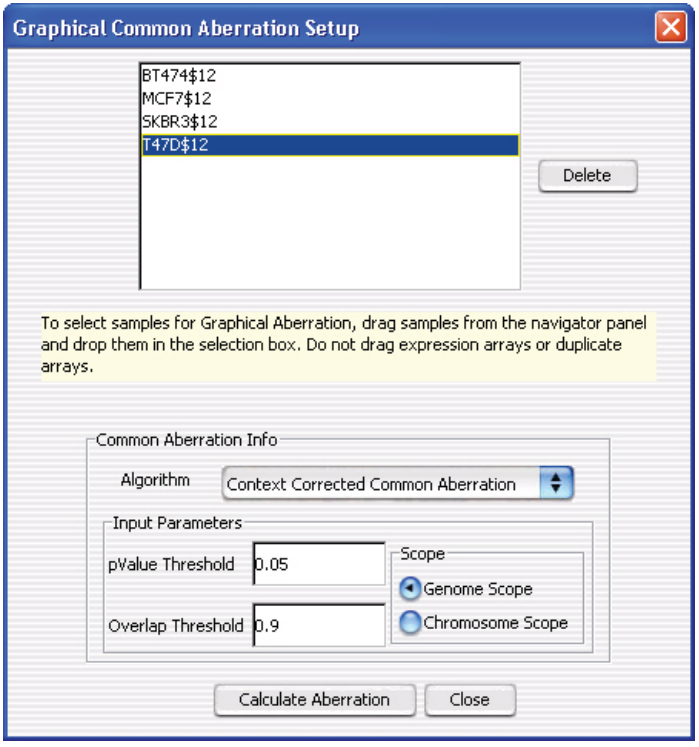


Figure 127 Graphical Common Aberration Setup dialog box

Purpose: This allows you to select arrays and perform a common aberration analysis independent of the selected experiment, and create graphical results in a new window. Common aberration analysis detects common aberrant regions across multiple samples. For instructions on how to do a common aberration analysis, see "To display common aberrations" on page 152.

To open: In the Discovery tab, under Common Aberration, click **Graphical.**

Top pane The initially empty pane in the dialog box stores the names of the arrays to analyze. To add data to the analysis, drag arrays from either the **Data** or **Experiment** panes of the Navigator to this pane of the dialog box.

Delete To delete arrays from the analysis, click the name of an array in the top pane of the dialog box, then click **Delete.**

Algorithm The algorithm used to detect common aberrations among multiple samples. For a detailed description of the available algorithms, see "Common Aberration Analysis" on page 508. Select one of these options:

- **Context Corrected Common Aberration** – This algorithm detects common aberration in a subset of samples to maximize the score. Also, it takes into account the probability of finding a common aberrations by considering all of the other aberrations present in the genome.
- **T-Test Common Aberration** – For more information, go to http://www.springerlink.com/content/j627358450r77722/

Input Parameters Type values for these parameters for multi-sample experiments:

- **pValue Threshold** – The *p*-value is used as a significance threshold for the T-Test.
- **Overlap Threshold** – When this is set at 0.9 (the default value), it means for intervals that are 90% common, the program will report only the one with the highest score. This reduces the size of the list of common aberrant intervals, and reports the most significant ones.

Scope (Available if you select the Context Corrected Common Aberration algorithm.) Select one of these options:

- **Genome Scope** – Displays results for the entire genome.
- **Chromosome Scope** – Displays results on a chromosome-by-chromosome basis.

The scope of the T-Test Common Aberration algorithm is always the entire genome.

Calculate Aberration Analyzes each of the arrays for aberrations, then displays them in a Graphical Aberration Summary window. "Graphical Aberration Summary (for common aberration analysis)" on page 337. You then use this window to configure and start the common aberration analysis.

Close Closes the dialog box without generating common aberrations.

Graphical Common Aberration Summary

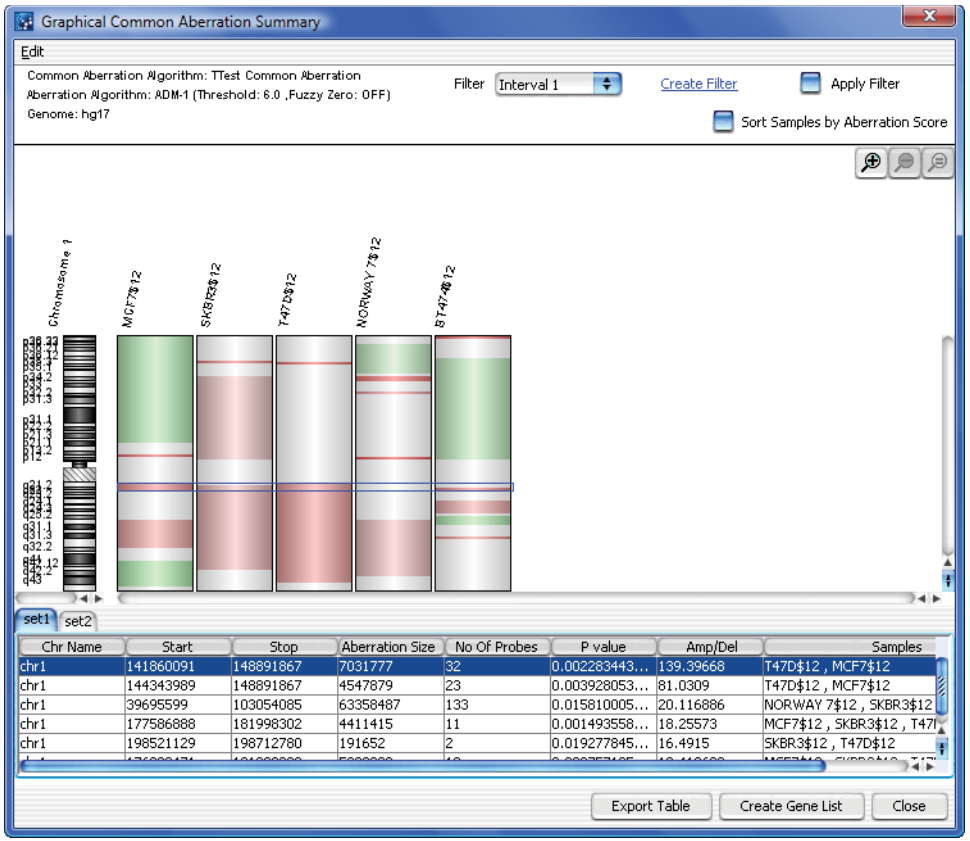


Figure 128 Graphical Common Aberration Summary

Purpose: To display the Graphical Common Aberration Summary.

To open: The Graphical Common Aberration Summary appears when you click **Common Aberration** in the special Graphical Aberration Summary dialog box that is part of the graphical common aberration analysis process. For instructions on how to perform this kind of analysis, see "To display common aberrations" on page 152.

Main Pane

The main pane of the summary shows heat-map style plots of the aberrant regions of each individual array next to an ideogram of the applicable chromosome. Detected amplifications appear in red, and detected deletions appear in green.

A *common aberration,* as detected by the analysis, appears as a blue box around the given region across all of the plots. To display each of these common regions, and to move among the chromosomes, click a row of the results table that appears at the bottom of the window.

To zoom the view, click any of these zoom buttons, located at the top right of the main pane:

 – Zooms in the view.

 – Zooms out the view.

 – Resets the zoom to its initial setting (zoomed all the way out to see the entire chromosome).

Results table(s)

At the bottom of the window, the detected common aberrations appear in a table. If you configured two sets of arrays for the analysis, results for each set appear in separate tabs. Click any row of the table to move to the appropriate chromosome, and center the graphical plots in the main pane on the particular common aberration.

These columns appear in the table(s):

Column	Description
Chr Name	The name of the chromosome on which the common aberration is found.
Start	The location of the first base pair in the common aberrant region.
Stop	The location of the last base pair in the common aberrant region.
Aberration size	The number of base pairs in the aberrant region.
No Of Probes	The number of probes upon which the common aberration call is based.
P value	The likelihood that the region is a common aberrant region. (Lower p-values indicate a greater likelihood.)

Column	Description
Amp/Del	The magnitude of the detected common aberration. Amplifications are positive numbers. Deletions are negative numbers.
Samples	The arrays on which an aberration was found in this region.

Other commands The window also offers these additional commands:

- **Edit** – Opens a menu with a Copy summary to clipboard command. If you click this command, the program copies the entire summary to the clipboard as an image. You can then paste the image into a document in another program.
- **Filter** – Displays the interval filters available for common aberration results. Interval filters include or exclude certain aberrant intervals based on specific conditions. Select an interval filter. To apply the filter, mark **Apply Filter.** To create a new interval filter, click **Create Filter** (see below).
- **Create Filter** – Click this link to open the Interval Filter dialog box, where you can create, change, or delete interval filters. See "Interval Filter" on page 361.
- **Apply Filter** – Mark this option to apply the selected interval filter to the results.
- **Sort Samples by Aberration Score** – (Available if you used the Context Corrected Common Aberration algorithm to detect common aberrant regions.) Sorts the plots in the main pane in ascending order based on aberration score.
- **Export Table** – Opens an Open dialog box, where you can type a name and select a location for a *.xls file that contains the list of common aberration analysis settings, and the complete table of detected common aberrations.
- **Create Gene List** – Opens the Create Gene List dialog box, where you can create a gene list that contains the genes in the common aberrant region selected in the results table. The program pre-sets the chromosome start and stop positions in this dialog box. Type a name for the gene list, then click **OK.** See "Create Gene List" on page 279.
- **Close** – Closes the Graphical Common Aberration Summary.

Graphical Differential Aberration Summary



Figure 129 Graphical Differential Aberration Summary dialog box

Purpose: This shows the detected differentially aberrant regions both graphically and as a table. See "To set up a differential aberration analysis" on page 158.

To open: The Graphical Differential Aberration Summary dialog box opens when you click **Run** in the Graphical Aberration Setup dialog box that is part of the differential aberration analysis process.

Analysis Summary The top of the dialog box shows information about the original aberration analysis, including the aberration detection algorithm that was used, and other associated analysis settings.

Line Plot(s) The line plots in the dialog box represent the *p*-values for gains and losses in both comparison sets as a function of genomic location. When you select a row in the table of differentially aberrant regions at the bottom of the dialog box, the line plots for the corresponding chromosome appear. Also, the blue line in the line plots reflects the genomic location of the selected row.

In addition, when you right-click anywhere within the line plots, a shortcut menu opens with these options:

- **Properties** – Opens the Chart Properties dialog box, where you can customize the plots, and add and format a title and a legend. See "Chart Properties" on page 255.
- **Save as** – Opens a Save dialog box, where you can type a name and select a location for an exported *.png image file of the plots.
- **Print** – Opens a Page Setup dialog box, which allows you to set page and printer options for the printing of the plots. After you click **OK** in this dialog box, the Print dialog box appears, where you can set additional options, and send the plots to your printer.
- **Zoom In** – Opens a menu that allows you to zoom in the scale of all of the line plots. These options are available:
  - **Both Axes** – Zooms in both axes of all plots
  - **Domain Axis** – Zooms in only the chromosomal location axis
  - **Range Axis** – Zooms in only the *p*-value axis
- **Zoom Out** – Opens a menu that allows you to zoom out the scale of all of the line plots. These options are available:
  - **Both Axes** – Zooms out both axes of all plots
  - **Domain Axis** – Zooms out only the chromosomal location axis
  - **Range Axis** – Zooms out only the *p*-value axis
- **Auto Range** – Adjusts the scale of both axes for optimal viewing of all of the data associated with a given chromosome.

Cylinder plots

The main pane of the summary shows heat-map style plots of the aberrant regions of each individual array next to an ideogram of the applicable chromosome. Detected amplifications appear in red, and detected deletions appear in green.

This pane has several additional features:

- A *differential aberration*, as detected by the analysis, appears as a blue box around the given region across all of the plots. To display each of these regions, and to move among the chromosomes, click a row of the results table that appears at the bottom of the dialog box.
- To identify the array that a particular plot represents, place the pointer over the plot. A ToolTip shows the name of the array.
- To zoom the view, click any of these zoom buttons, located at the top right of the pane:
  -  – Zooms in the view.
  -  – Zooms out the view.
  -  – Resets the zoom to its initial setting (zoomed all the way out to see the entire chromosome).

Table of regions

At the bottom of the window, the detected differentially aberrant regions appear in a table. Click any row of the table to shift the plots to the appropriate chromosome. In the *p*-value line plots, a blue line shows the specific chromosomal location of the selected region. In the aberration plots, a blue box identifies the differentially aberrant regions.

These columns appear in the table:

Column	Description
Chr Name	The name of the chromosome on which the differential aberration is found.
Start	The location of the first base pair in the differentially aberrant region.
Stop	The location of the last base pair in the differentially aberrant region.
Size	The number of base pairs in the differentially aberrant region.
No of Probes	The number of probes upon which the differential aberration call is based.
Num Gains in Set 1	Number of aberration gains for the differentially aberrant region in first group.

Column	Description
Num Losses in Set 1	Number of aberration losses for the differentially aberrant region in the first group.
-log10(PVal) Gain in Set 1	The likelihood that the region represents a differential gain for set 1. (Lower *p*-values indicate a greater likelihood.)
-log10(PVal) Loss in Set 1	The likelihood that the region represents a differential loss for set 1. (Lower *p*-values indicate a greater likelihood.)
Num Gains in Set 2	Number of aberration gains for the differentially aberrant region in second group.
Num Losses in Set 2	Number of aberration losses for the differentially aberrant region in the second group.
-log10(PVal) Gain in Set 2	The likelihood that the region represents a differential gain for set 2. (Lower *p*-values indicate a greater likelihood.)
-log10(PVal) Loss in Set 2	The likelihood that the region represents a differential loss for set 2. (Lower *p*-values indicate a greater likelihood.)

Other commands

The dialog box also offers these additional commands:

- **Edit** – Opens a menu with a Copy summary to clipboard command. This command copies the entire summary to the clipboard as an image. You can then paste the image into a document in another program.
- **Filter** – Displays the interval filters available for differential aberration results. Interval filters include or exclude certain intervals based on specific conditions. Select an interval filter. To apply the filter, mark **Apply Filter.** To create a new interval filter, click **Create Filter** (see below).
- **Create Filter** – Click this link to open the Interval Filter dialog box, where you can create, change, or delete interval filters. See "Interval Filter" on page 361.
- **Apply Filter** – Mark this option to apply the selected interval filter to the results.
- **Export Table** – Opens an Open dialog box, where you can type a name and select a location for a *.xls file that contains the list of aberration analysis settings, and the complete table of detected differentially aberrant regions.

- **Create Gene List** – Opens the Create Gene List dialog box, where you can create a gene list that contains the genes in the differentially aberrant region selected in the results table. The program sets the chromosome start and stop positions in this dialog box. Type a name for the gene list, then click **OK.** See "Create Gene List (from common or differential aberration results)" on page 282.
- **Close** – Closes the dialog box.

Graphical Penetrance Summary

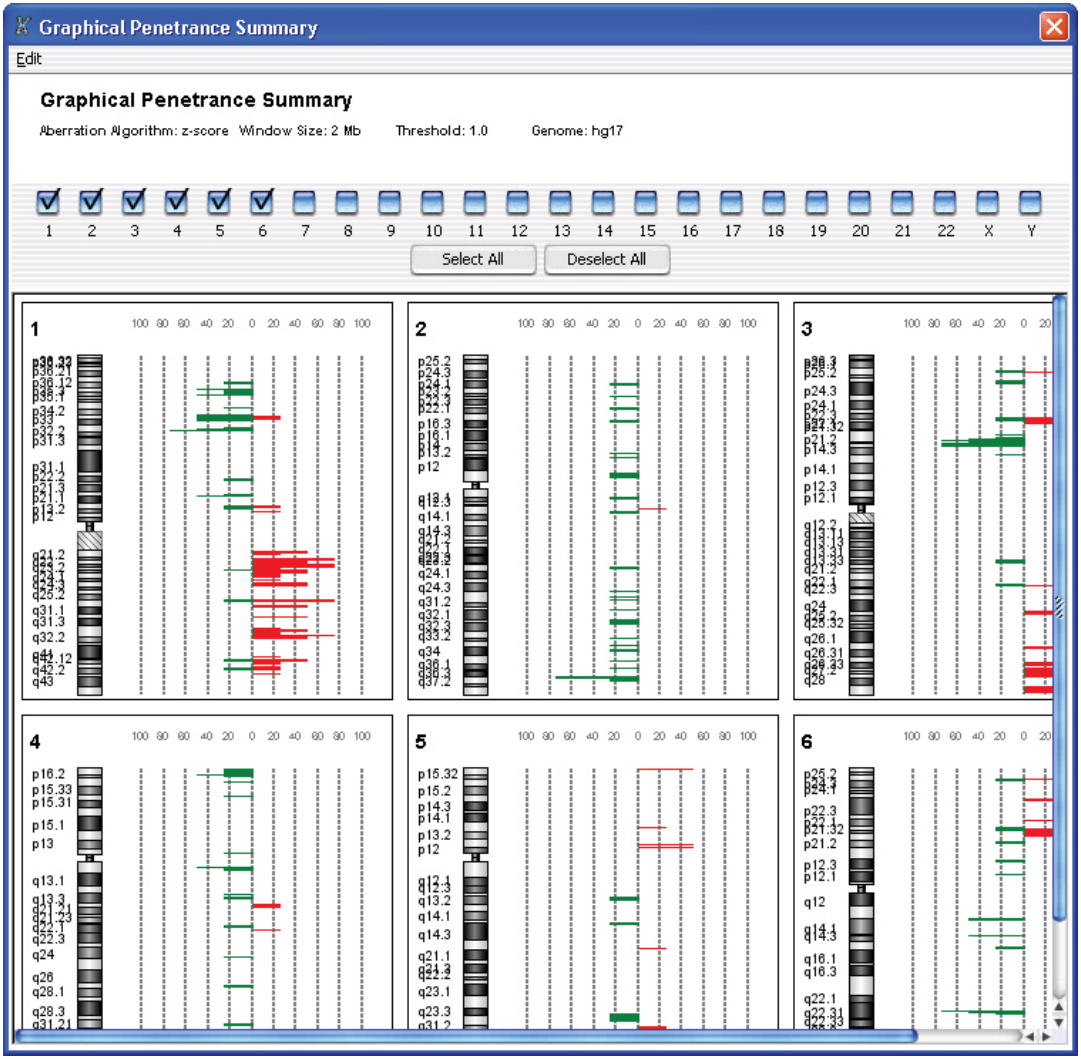


Figure 130 Graphical Penetrance Summary

Purpose: This window allows you to display the results of probe penetrance analysis for each chromosome in the genome.

To open: In the Discovery tab, under Graphical Penetrance, click **Probe.**

Probe penetrance analysis produces penetrance scores for each probe position across multiple arrays. A probe penetrance score reflects the percentage of the selected arrays in the experiment that show a significant amplification or deletion at the position. For a more detailed description of penetrance analysis, see "Penetrance" on page 506.

Plots Penetrance plots for selected chromosomes appear in the main pane. For each probe position on a given chromosome, the percentage of arrays that show a significant deletion appear in green. The percentage of arrays that show a significant amplification appear in red. Amplifications and deletions reflect the results of the experimental analysis as configured in the main window of the program.

When you double-click within the plot area of any penetrance plot or within any chromosome, the cursor in Genome, Chromosome, Gene, and Tab Views in the main window moves to the new location.

Select Chromosomes The check boxes at the top of the window correspond to the chromosomes of the applicable genome. Mark as many as you like. The program displays a probe penetrance plot for each selected chromosome.

Select All Produces probe penetrance plots for all chromosomes in the genome.

Deselect All Clears all of the check boxes, and removes all of the probe penetrance plots from the window.

Graphical Interval Penetrance Summary

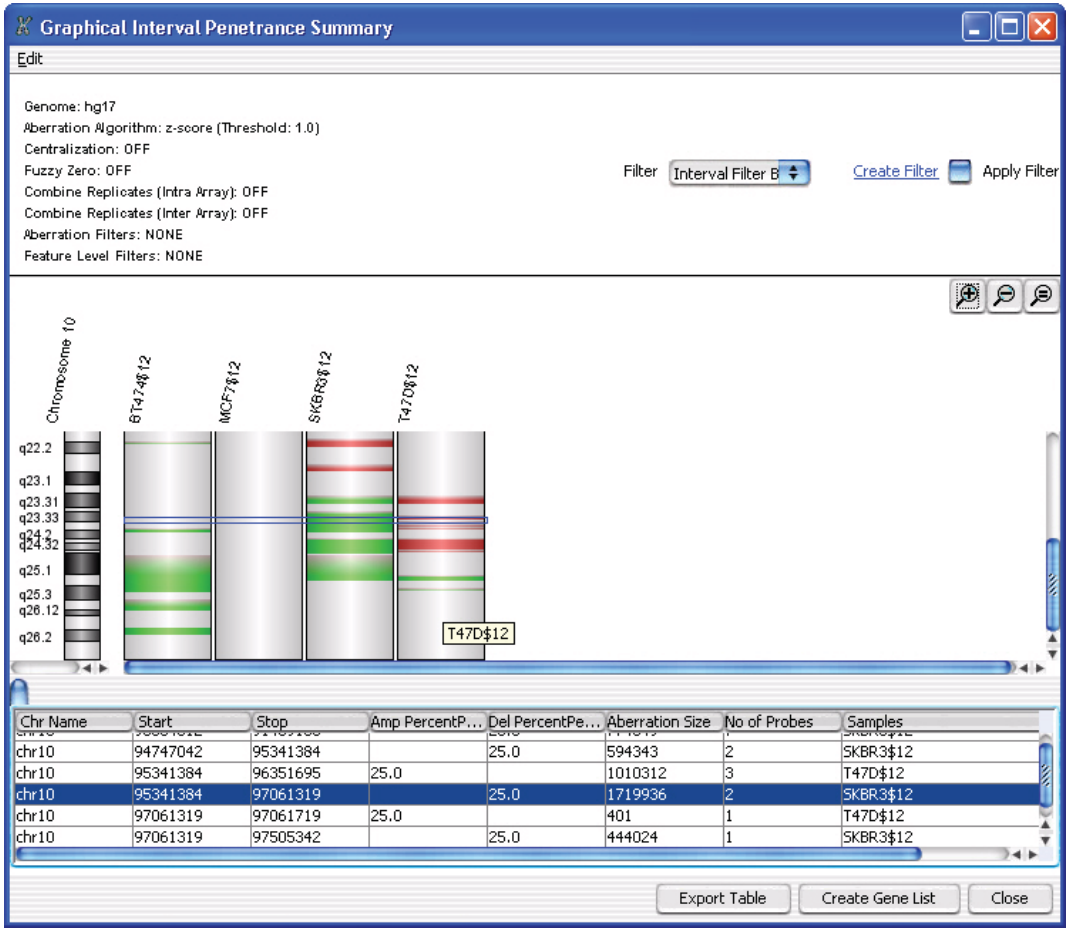


Figure 131 Graphical Interval Penetrance Summary

Purpose: This window allows you to display and save interval penetrance results for the selected arrays in the current experiment. Interval penetrance analysis produces scores for detected aberrant intervals across multiple samples. These scores reflect the percentage of samples that show each region to be aberrant. For a more detailed discussion of penetrance analysis, see "Penetrance" on page 506.

To open: In the Discovery tab, under Penetrance, click **Interval.**

Main Pane

The main pane of the summary shows heat-map style plots of the aberrant regions of each individual array next to an ideogram of the applicable chromosome. Detected amplifications appear in red, and detected deletions appear in green.

Each interval scored by the analysis appears as a blue box around the given region across all of the plots. To display each of these regions, and to move among the chromosomes, click a row of the results table that appears at the bottom of the window.

To zoom the view, click any of these zoom buttons, located at the top right of the main pane:

 – Zooms in the view.

 – Zooms out the view.

 – Resets the zoom to its initial setting (zoomed all the way out to see the entire chromosome).

Results table

At the bottom of the window, the scored aberrant intervals appear in a table. Click any row of the table to move to the appropriate chromosome, and center the graphical plots in the main pane on the particular aberrant interval.

These columns appear in the table:

Column	Description
Chr Name	The name of the chromosome on which the aberrant interval is found.
Start	The location of the first base pair in the aberrant interval.
Stop	The location of the last base pair in the aberrant interval.
Amp Percent Penetrance	The percentage of arrays that show a significant amplification for the interval.
Del Percent Penetrance	The percentage of arrays that show a significant deletion for the interval.
Aberration Size	The number of base pairs in the aberrant interval.
No of Probes	The number of probes used to make the aberrant interval call.
Samples	The arrays on which an aberration was found in this interval.

Other commands The window also offers these additional commands:

- **Edit** – Opens a menu with a Copy summary to clipboard command. If you click this command, the program copies the entire summary to the clipboard as an image. You can then paste the image into a document in another program.
- **Filter** – Displays the interval filters available for the interval penetrance results. Interval filters include or exclude certain aberrant intervals based on specific conditions. Select an interval filter. To apply the selected filter, mark **Apply Filter.** To create a new interval filter, click **Create Filter** (see below).
- **Create Filter** – Click this link to open the Interval Filter dialog box, where you can create, change, or delete interval filters. See "Interval Filter" on page 361.
- **Apply Filter** – Mark this option to apply the selected interval filter to the results.
- **Export Table** – Opens an Open dialog box, where you can type a name and select a location for the exported table. The program saves the results table that appears at the bottom of the window as a *.xls file that you can view in Microsoft Excel.
- **Create Gene List** – Opens the Create Gene List dialog box, where you can create a gene list that contains the genes in the aberrant interval selected in the results table. The program pre-sets the chromosome start and stop positions in this dialog box. Type a name for the gene list, then click **OK.** See "Create Experiment" on page 278.
- **Close** – Closes the Graphical Interval Penetrance Summary.

Import

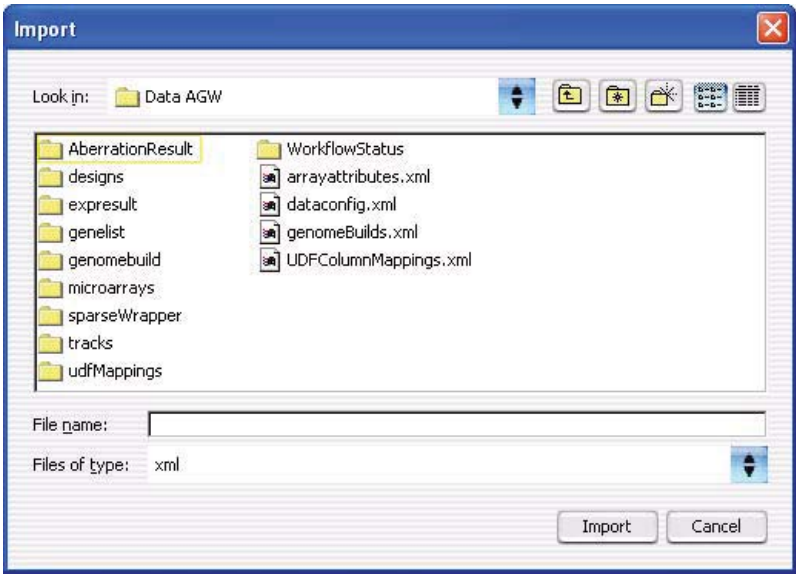


Figure 132 Import dialog box

Purpose: Lets you select files for import into Agilent Genomic Workbench. The title of this dialog box changes depending on the type of file to import.

To open: In the Home tab, click **Import,** then select any kind of import except Genome Build or Track. The type of file to be imported appears in the title of the dialog box. To import a gene list, right-click the **Gene List** folder in the My Entities List pane of the Navigator, then click **Import Gene List.**

Use the standard Windows Explorer commands in the dialog box to select a file for import.

For some imports, you can select multiple files. Click the name of the first file, then hold down the **ctrl** key and click the names of additional files. To select a contiguous block of files, click the name of the first file in the block, then hold down the **shift** key and click the name of the last one.

File name Displays the name of a file you select for import.

Files of type The program limits the list of files to the specific types expected for the import. To display all files, click ,then select **All Files.**

File type	Extension
FE array file	*.txt
Axon array file	*.gpr
UDF file	*.txt
Design file (GEML)	*.xml
Axon design file	*.gal
Array attributes	*.txt
Experiments	*.zip
Filters	*.xml
Gene list	*.txt

Import Imports the file into the program. In some cases, the name of this button is *Open*, rather than *Import*. Also, when you click **Import,** in many cases one or a series of additional dialog box(es) allows you to further define the content for import. See the instructions for each specific type of import in Chapter 2.

Cancel Cancels the import and closes the dialog box.

Import (filters)

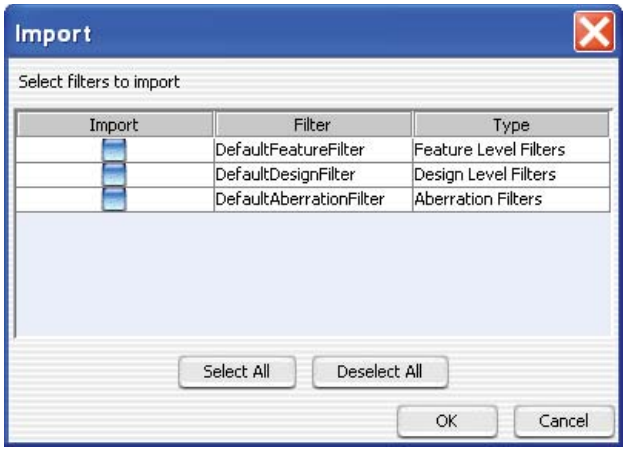


Figure 133 Import (for filters) dialog box

Purpose: Lets you select the specific filters within a .zip exported filter file to import into the program. See "To import filters" on page 58.

To open: In the Home tab, click **Import > Filters.** In the dialog box that appears, select the desired ZIP exported filter file, then click **Import.**

Select experiments to import

These columns appear:

- **Import** – Mark the check box for the experiment(s) to import.
- **Filter** – The names of the filters available for import in the .zip filter file.
- **Type** – The type of filter

Select All Selects all of the filters in the .zip file for import.

Deselect All Clears all of the check boxes under Import.

OK Imports the selected filters into the program. If the name of a filter matches one that is already available in the program, the Confirm overwrite dialog box appears, where you can select the filters that you want to overwrite. See "Confirm Overwrite" on page 268.

Cancel Cancels the import and closes the dialog box.

Import (experiments)

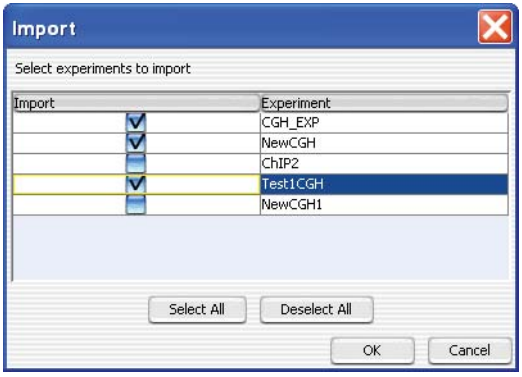


Figure 134 Import dialog box (for experiments)

Purpose: Lets you select the specific experiments within a ZIP format experiment file to load into the program. See "To import an experiment file" on page 57.

To open: In the Home tab, click **Import > Experiments.** In the dialog box that appears, select the desired ZIP format experiment file, then click **Import.**

Select experiments to import These columns appear:

- **Import** – Mark the check box next to the experiment(s) to import.
- **Experiment** – The names of the experiments available for import in the ZIP format experiment file.

Select All Selects all of the experiments in the ZIP file for import.

Deselect All Clears all of the check boxes under Import.

OK Imports the selected experiments into the program. If the name of an imported array design or data file matches one that is already available in the program, the Confirm overwrite dialog box appears, where you can select the data and/or design files to overwrite. See "Confirm Overwrite" on page 268.

Cancel Cancels the import and closes the dialog box.

Import GEML design files

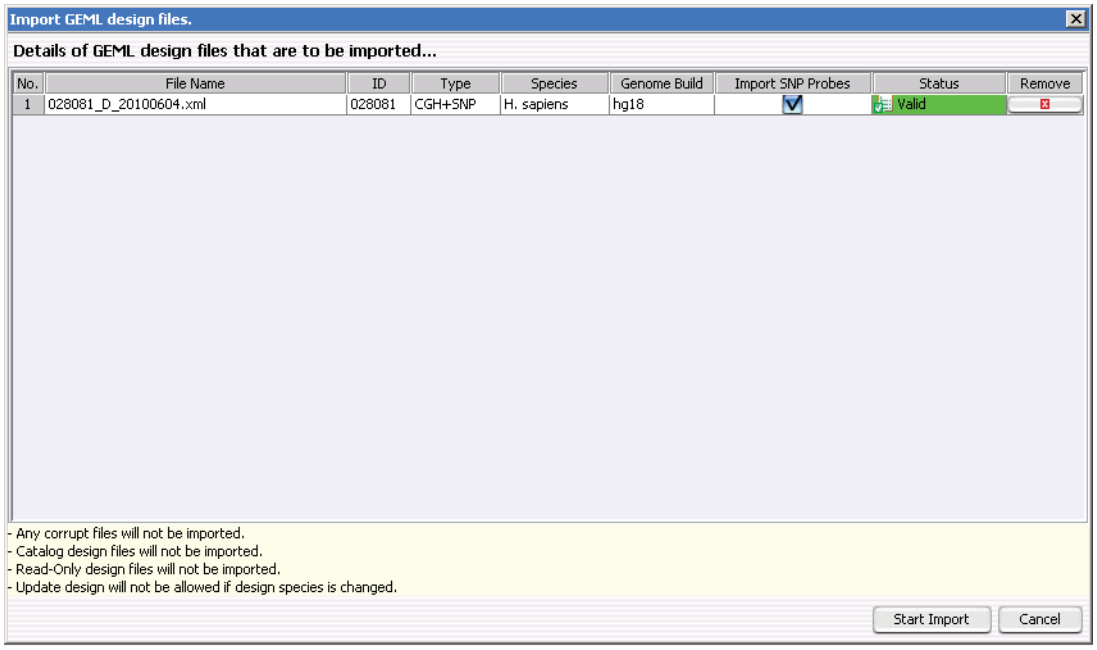


Figure 135 Import GEML design files dialog box

Purpose: To display information in the design file and to remove any files that you don't want to import.

To open: In the Home tab, click **Import > Design Files > GEML File.** Select the desired *.xml design files, then click **Open.**

File Name The name(s) of the design file(s) to be imported.

ID The Agilent ID number for the design file.

Type The design type, which can be CGH, ChIP, or gene expression.

Species The species for the genome build. This appears automatically when the Genome Build is selected.

Genome Build The genome build for the design. If the genome build is not read automatically, a "?" appears. Click **Genome Build** and select the correct value from the list.

Import SNP Probes Available for CGH+SNP designs. Mark the box to include SNP probes with the design data.

Status

- **Not Set** – Appears if Genome Build and Species information is not shown.
- **Not Allowed** – Appears if a Genome Build is selected that does not match the design.
- **Overwrite** – Appears when the design file has been updated and will overwrite any existing one of the same name.
- **Valid** – Appears when the file is new.
- **Corrupt** – Appears when the file is corrupt.

Remove Click  to remove a specific design file from the list.

Start Import Starts the import of the design files in the list.

Cancel Cancels the transfer and closes the dialog box.

Import Genome Build

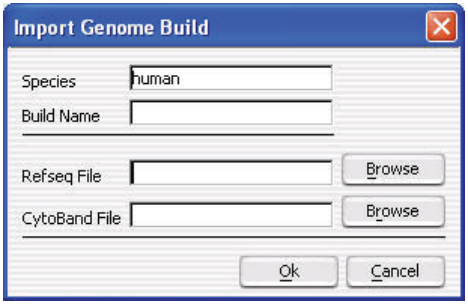


Figure 136 Import Genome Build dialog box

Purpose: To import a new set of genome build files into Agilent Genomic Workbench. See "To import a genome build" on page 50.

To open: In the Home tab, click **Import > Genome Build.**

Species The genome's species of origin.

Build Name The name of the build to be imported.

Refseq File The location of the RefSeq database file. This file contains chromosomal locations of genes. To select a Refseq file, click **Browse.**

CytoBand File The location of the applicable cytoband file. This file contains graphical cytoband information for Gene View and Chromosome View. To select a cytoband file, click **Browse.**

OK Imports the genome build and closes the dialog box.

Cancel Cancels the import and closes the dialog box.

CAUTION Import only Agilent-provided genome build files.

Import Track

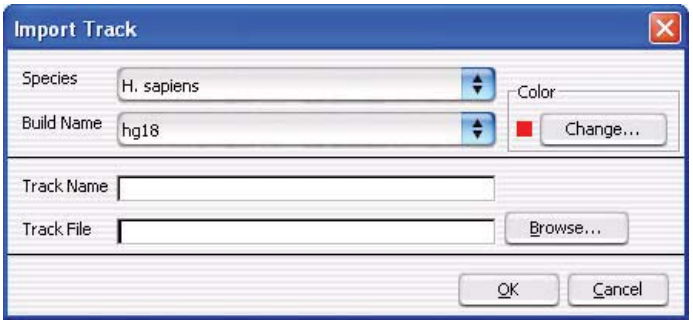


Figure 137 Import Track dialog box

Purpose: Lets you import a BED format track file. See "To import tracks" on page 55. Track information can appear in Gene view. See "To show tracks in Gene View" on page 111.

To open: In the Home tab, click **Import > Track.**

Species Select the species to which the track relates.

Build Name This list contains the available genome builds for the selected species. Select the desired genome build.

Color Shows the assigned display color for the track. To change this color, click **Change.** For details, see "To change the appearance of tracks" on page 112 for instructions for changing the color of gene lists or tracks.

Track Name Type a name to identify the imported track.

Track File Type the location of the BED track file to import, or click **Browse** to select a file.

Browse Opens an Open dialog box, where you can select the BED track file to import.

OK Imports the track into the program.

Cancel Cancels the import and closes the dialog box.

Interval Filter

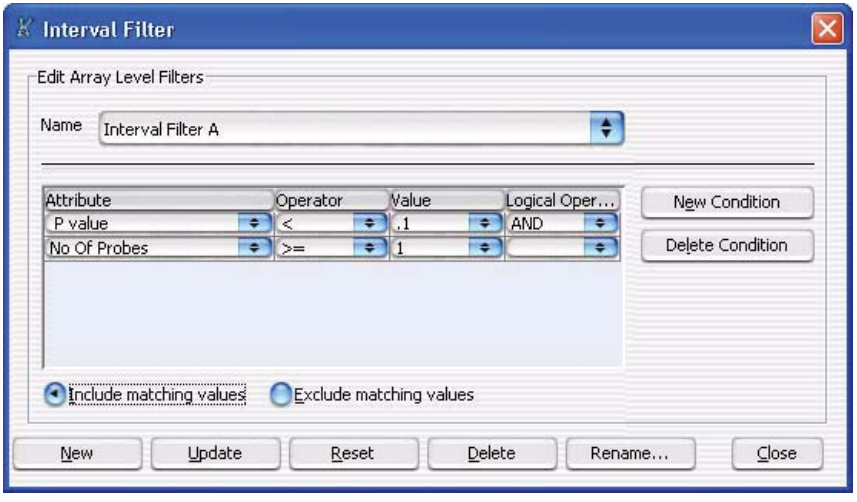


Figure 138 Interval Filter dialog box

Purpose: You use the Interval Filter dialog box to create, change, or delete interval filters. An interval filter includes or excludes certain detected intervals from the results, based on specific conditions.

To open: The Interval Filter dialog box opens when you click **Create Filter** in any of these dialog boxes:

- Graphical Common Aberration Summary – See "Graphical Common Aberration Summary" on page 341.
- Graphical Differential Aberration Summary – See "Graphical Differential Aberration Summary" on page 344.
- Graphical Interval Penetrance Summary – See "Graphical Interval Penetrance Summary" on page 351.

Name Select the name of the interval filter you that want to edit. To create a new filter, and add its name to the list, click **New.** You can apply interval filters to both common aberration results and interval penetrance results. However, because the available attributes for interval filters in these two situations are not the same, the filters you create in one context are not available in the other.

Filter conditions For each condition (row), select options from the lists. Specifically in **Value,** select an option from the list, if available, or type a value, then press **Enter.** To add another row to the table, click **New Condition.**

Each condition has these elements:

- **Attribute** – The attribute of the intervals evaluated by the filter. The attributes available in this list vary with the specific context of the filter.
- **Operator** – How the filter uses the entry in Value to evaluate common aberrant intervals. For example, the `>=` operator configures the filter to include or exclude features where the selected attribute is greater than or equal to the entry in Value.
- **Value** – The value the filter uses as the basis to evaluate a feature. For example, if you select the **Aberration Size** attribute, the `>` operator, and a value of **1000,** the filter includes or excludes an interval if it is greater than 1000 bp in size.

- **Logical Operator** – (Available only if you configure more than one condition) The relationship between the condition and the next one in the list. For example, if you select **AND** in **Logical Operator** for the first condition, the filter includes or excludes an interval if it passes both the first condition and the next condition.

Include/Exclude matching values Select one of these options:

- **Include matching values** – If an interval passes the filter conditions, the program *includes* it in the final result.
- **Exclude matching values** – If an interval passes the filter conditions, the program *excludes* it from the final result.

New Condition Adds a new, blank condition (row) to the table.

Delete Condition Removes a condition from the list. To remove a specific condition, click anywhere within the condition, then click **Delete Condition.**

New Opens an Input dialog box, where you can type a name for a new filter. To accept the name, click **OK.** The program creates a new filter, and adds its name to the Name list.

Update Saves any changes you make to the filter conditions.

Reset Restores the values of the filter conditions to what they were before you made any changes to them.

Delete Opens a Confirm dialog box that asks you if you want to delete the selected filter. To delete the filter, click **Yes.**

Rename Opens an Input dialog box where you can type a new name for the filter. To accept the name, click **OK.**

Close Closes the dialog box. If you created or changed a filter, but did not update it, a Confirm dialog box opens. Click **Yes** to accept the changes, **No** to reject the changes, or **Cancel** to return to the dialog box.

Manually Reassign Peaks

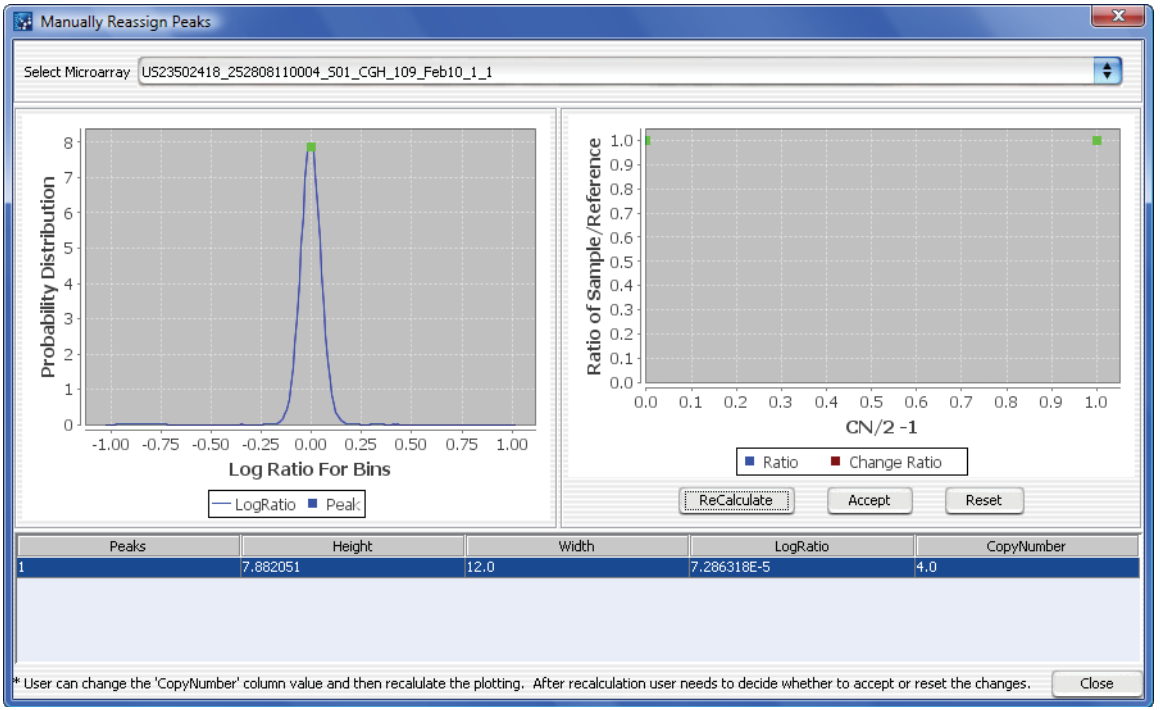


Figure 139 Manually Reassign Peaks dialog box

Purpose: To view and change copy number for peaks.

To open: In the Analysis tab, under SNP Analysis, mark the **Manually Assign Peaks** check box.

Select Microarray Click the arrow to select the microarray to display.

Recalculate After you change the copy number, click to recalculate the data and redraw the graphs.

Accept Click to accept recalculated data.

Reset Resets the data and graphs to the original values.

Close Closes the dialog box.

Matched Sample

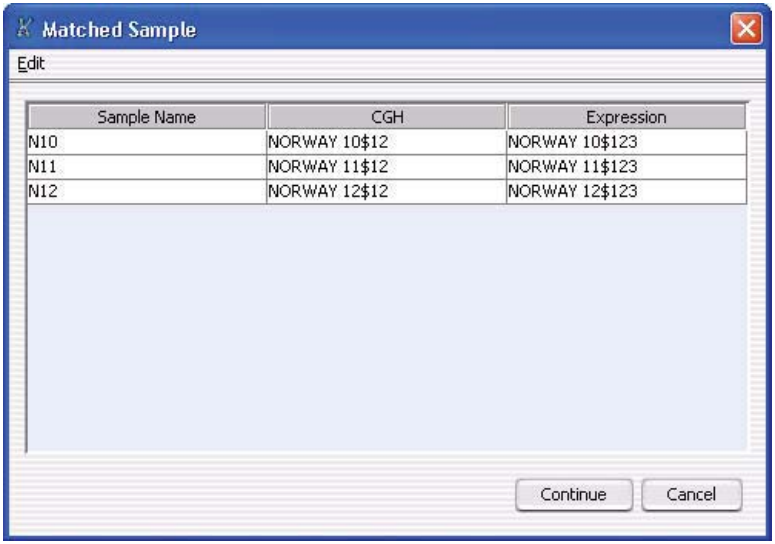


Figure 140 Matched Sample dialog box

Purpose: The Matched Sample dialog box displays the paired CGH-Expression arrays that the program will use in the correlation analysis.

To open: The matched sample dialog box opens when you click **Discovery > Joint-Analysis > Correlation Analysis > Perform,** or when you click **Perform Analysis** in the Correlation Analysis Setup dialog box (see "Correlation Analysis Setup" on page 270).

Edit Opens a menu with a Save Table Data option. If you select this option, a Save dialog box opens, where you can select a location for the saved file. The program saves the table data in a *.xls format file that you can open with Microsoft Excel.

Array table (Read-only) A table of the CGH-Expression array pairs that the program will use for the correlation analysis. The table has these columns:

- First column – When you define the parameters of the correlation analysis, you select an array attribute to use as the basis to match CGH and expression arrays. The values for this attribute for the arrays in the analysis appear in this column. By default, the program uses the Green Sample attribute to pair matching arrays.
- CGH – The CGH samples used in the correlation analysis.
- Expression – The expression samples used in the correlation analysis.

Microarray Properties

Purpose: Displays the properties associated with an array. You can also edit the values of specific attributes. To add attributes to the list, see the *Sample Manager User Guide*.

To open: For any array in the Data folder or Experiments folder, right-click the array name, then click **Show Properties.** For non-Agilent arrays, only the Attribute tab appears.

Attribute tab

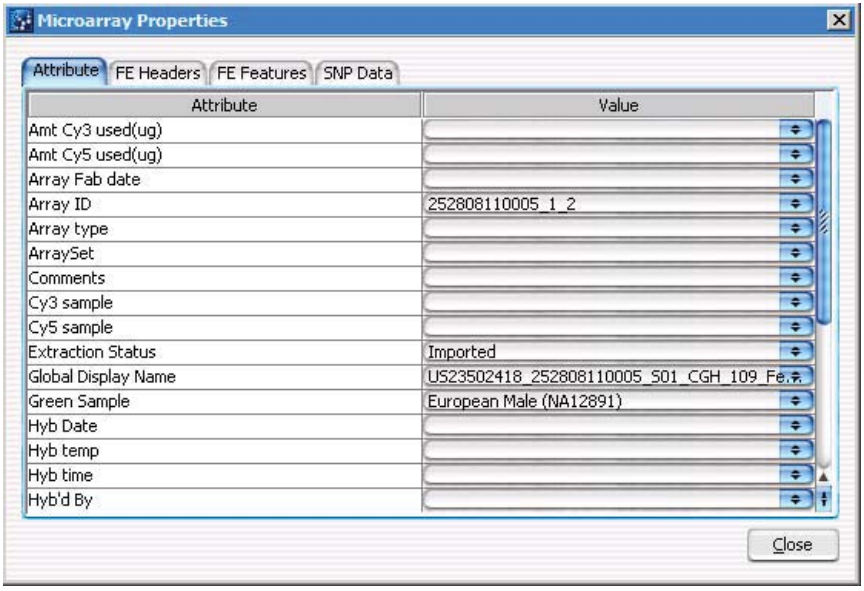


Figure 141 Microarray Properties dialog box with list of Attributes and their values

- **Attribute** – Displays the attributes in the array by name.
- **Value** – Indicates the values, if any, for each array. To edit the value of an attribute, select a new value for it under Value. Alternatively, click , then type or edit the value.

Close Closes the dialog box.

FE Headers tab

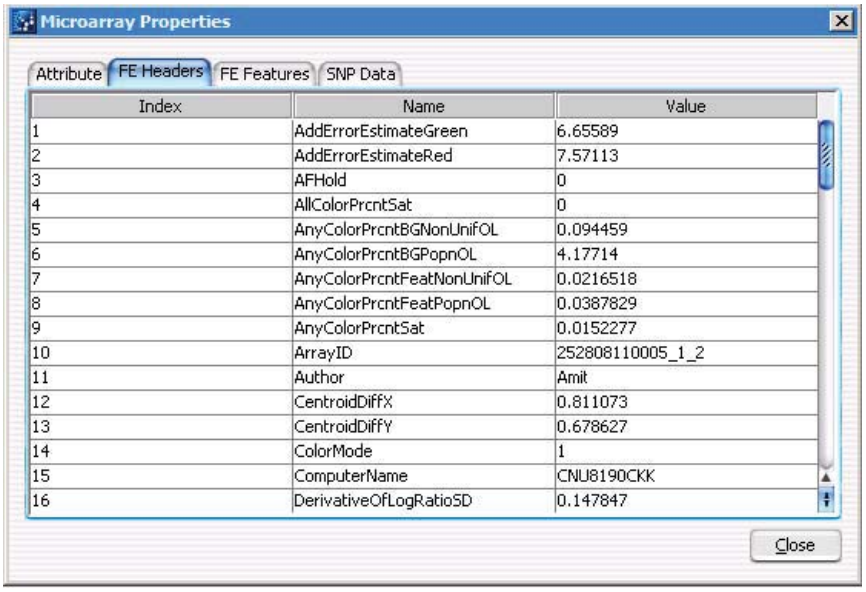


Figure 142 Microarray Properties dialog box with list of FE Headers their values

Index Displays a sequential index to help identify FE properties.

Name Displays feature parameters, statistics, and constants for the whole array.

Value Displays the value for each parameter, statistic, and constant.

Close Closes the dialog box.

FE Features tab

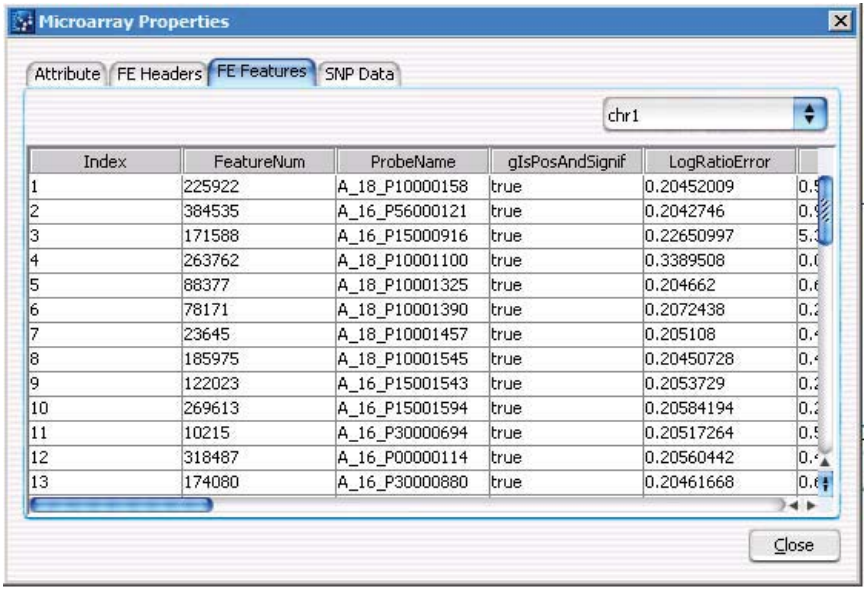


Figure 143 Microarray Properties dialog box with list of FE Features and associated data

Selection List Select the chromosome whose feature information you want to display.

List Box Displays FE features and the associated data.

SNP Data tab

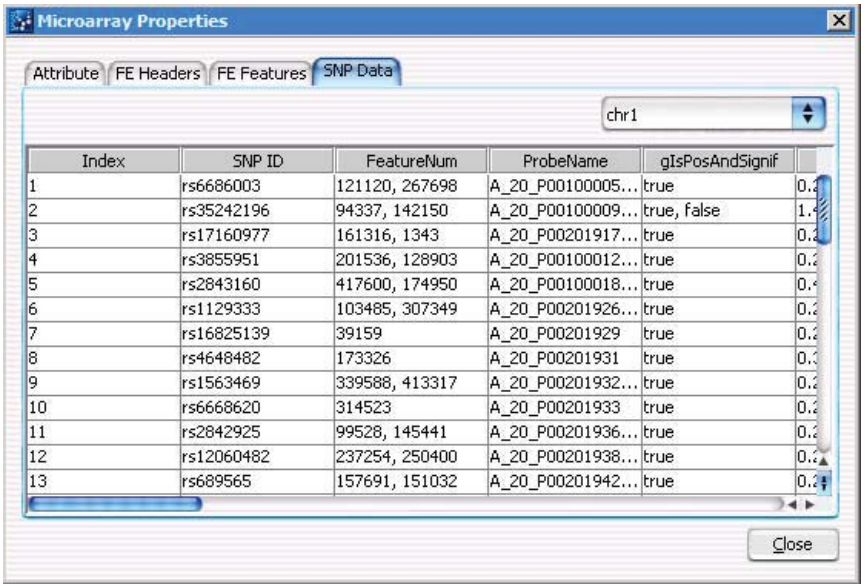


Figure 144 Microarray Properties dialog box with list of SNP data and values

Selection List Select the chromosome whose SNP information you want to display.

List Box Displays SNP data.

MovAvg Example Parameters

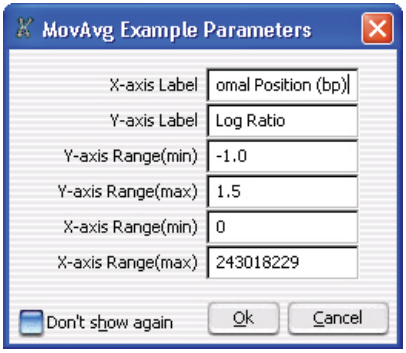


Figure 145 MovAvg Example Parameters dialog box

Purpose: It allows you to set display parameters for the MovAvg Example plug-in. The plug-in computes a 10-point moving average of the data from the selected chromosome in the selected arrays of the active experiment. It displays stacked plots of moving averages for the all arrays, one plot per array.

To open: Click **Tool > Plugin Settings > MovAvg Example.** This dialog box also opens when you click **Tool > Plugin > MovAvg Example,** if **Don't show again** is cleared.

Parameters Set any of these parameters:

Parameter	Description
X-axis Label	The text that appears beneath the X-axis of the plot as a label.
Y-axis Label	The text that appears beside the Y-axis of the plot as a label.
Y-axis Range(min)	The minimum value on the Y-axis.
Y-axis Range(max)	The maximum value on the Y-axis.
X-axis Range(min)	The minimum value on the X-axis.
X-axis Range(max)	The maximum value on the X-axis.

Don't show again Mark this option to prevent the appearance of this dialog box in the future when you click Tool > Plugin > MovAvg Example. To restore the dialog box so it appears again, click **Tool > Plugin Settings > MovAvg Example,** then clear **Don't show again.**

OK Click to accept the parameters and prepare the plot. You can further refine the appearance of the plot once the plug-in generates it.

Cancel Discards any changes you made, and closes the dialog box.

How to modify the plugin

The plug-in program itself (**MovAvg Example.pl**, located in the Plugins folder of the Agilent Genomic Workbench installation folder on your computer) is a short Perl program. It is a good example of how computed columns are processed. You must have Perl installed on your computer to use this plug-in.

The plot is very simple, but the simple plug-in architecture allows you to write your own computational methods to analyze data from selected arrays in the CGH module.

Within the code of the plug-in, you can add text strings to column headers to control format.

To create a line graph instead of a scatter plot, you append `-plotline` to a column header.

To prevent the plug-in from including a specific column, you append `-noplot` to the column heading. Note that the plug-in removes this extra text from the header before it displays it on the plot. The extra text does not appear in figures, and is only used to control the format of the plot.

MovAvg Example.pl shows how column-naming can be used. As you read the first line (which contains the header text), you can add text to the existing headers or add text to the headers for your generated columns, as well, to give you a small amount of formatting control.

MovAvg Example Plot

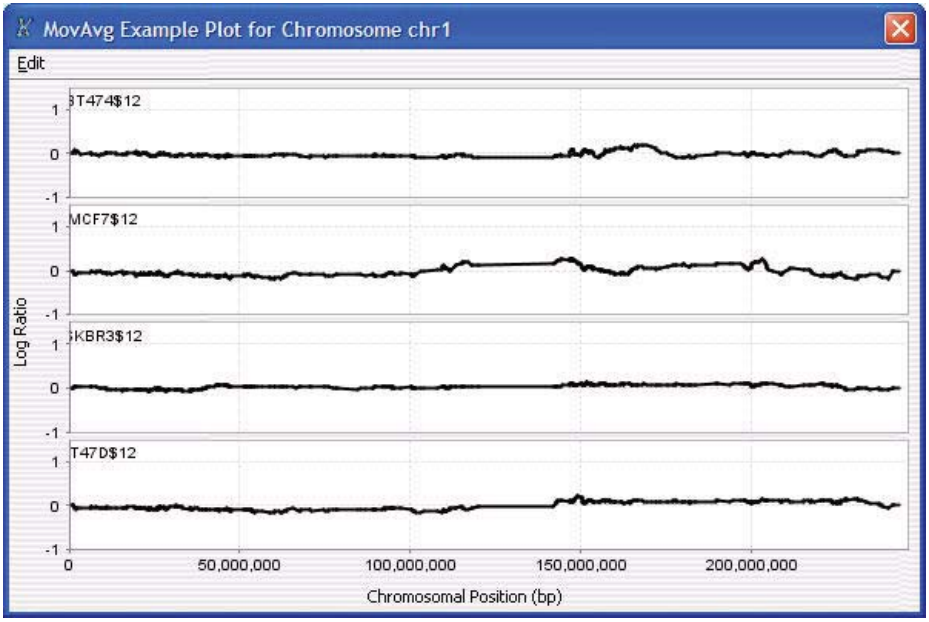


Figure 146 MovAvg Example Plot

Purpose: This window displays the output of the MovAvg Example plug-in. The plug-in computes a 10-point moving average of the data from the selected chromosome in the selected arrays of the active experiment.

To open: Click **OK** in the MovAvg Example Parameters dialog box. See “” on page 370.

Plot(s) The main plot area shows moving average line plots for the selected chromosome. A separate plot appears for each array.

When you right-click anywhere within the plot area, a shortcut menu opens with these options:

Option	Description
Properties	Opens the Chart Properties dialog box, where you can change the appearance of many aspects of the plot. See "Chart Properties" on page 255.
Save as	Opens a Save dialog box, where you can select a location for the *.png image file of the plots.
Print	Opens a Page Setup dialog box, where you can set page size, margins, and orientation, and select a printer. From there, you click **OK** to open the Print dialog box, where you can set print options and print the plot.
Zoom In	Opens another menu that allows you to zoom in the plot. You can zoom in several ways: • **Both Axes** – Zooms in the Domain (X) axis for all stacked plots, and zooms in the range Range (Y) axis for the specific plot in which you clicked. • **Domain Axis** – Zooms in the Domain (X) axis for all stacked plots. • **Range Axis** – Zooms in the Range (Y) axis for the specific plot in which you clicked. You can also drag across an area of one of the plots to select an area to zoom in on.
Zoom Out	Opens another menu that allows you to zoom out the plot. You can zoom out several ways: • **Both Axes** – Zooms out the Domain (X) axis for all stacked plots, and zooms out the range Range (Y) axis for the specific plot in which you clicked. • **Domain Axis** – Zooms out the Domain (X) axis for all stacked plots. • **Range Axis** – Zooms out the Range (Y) axis for the specific plot in which you clicked.
Auto Range	Opens another menu that allows you to zoom the plot to encompass the full range of the data. You can zoom in several ways: • **Both Axes** – Appropriately zooms both axes of the specific plot to show the full set of data. • **Domain Axis** – Zooms in the Domain (X) axis for all stacked plots so this axis encompasses the full range of X values of the data. • **Range Axis** – Zooms in the Range (Y) axis for the specific plot in which you clicked to encompass the full range of Y values of the data.

Edit Opens a menu with a Copy Plot(s) to Clipboard option. If you select this option, the program copies the plots to the clipboard as an image. You can then paste the image into a document in another program.

Properties (for common aberration result)

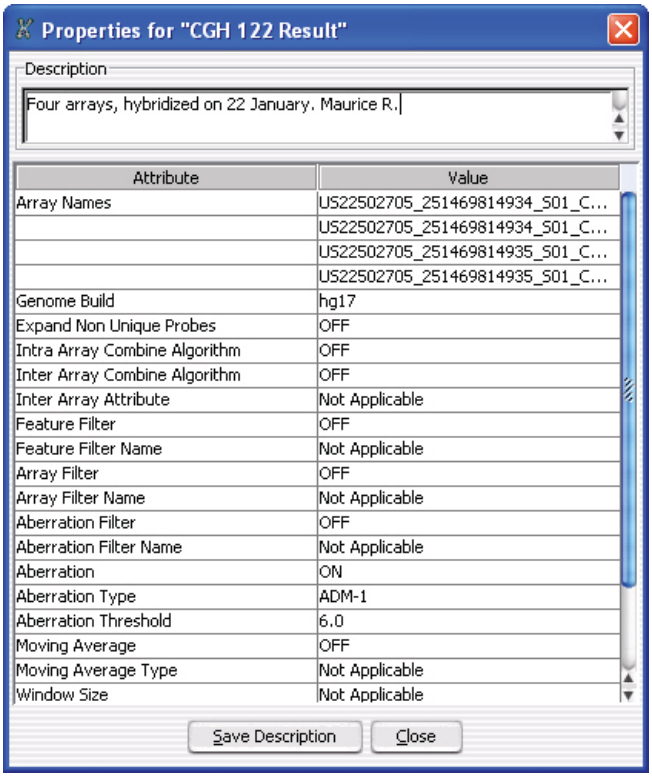


Figure 147 Properties dialog box for common aberration results

Purpose: This dialog box allows you to display the attributes of the common aberration results. You view the actual detected common aberrations in Genome, Chromosome, and Gene views.

To open: Right-click the name of the desired common aberration result in the Experiment pane of the Navigator, then click **View Details** in the shortcut menu.

Description Descriptive text that you can save with the properties.

Attribute table (Read-only) Displays the attributes of the common aberration results, such as the common aberration analysis settings, the names of the arrays included in the analysis, and the general analysis settings.

Save Description (Available if you make changes to the description) Saves the description. The new description now appears whenever the dialog box opens.

Close Closes the dialog box.

Properties (of an experiment result)

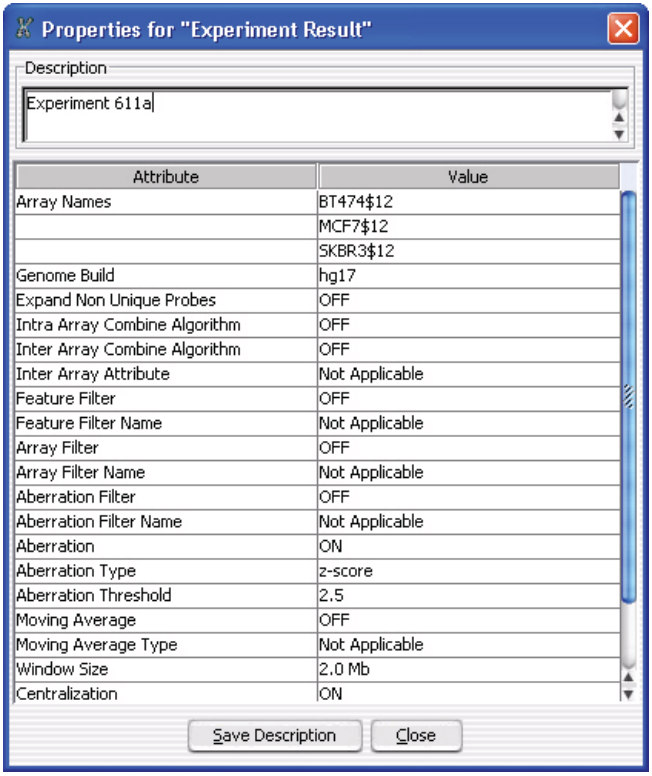


Figure 148 Properties dialog box (for experiment results)

Purpose: This dialog box displays the attributes of a saved experiment result.

To open: Right-click the name of the result in the **Results** folder of an experiment.

Description Shows the description associated with the result. You can type a new description, or change an existing one. Click **Save Description** to save the description with the experiment result.

Table Shows the attributes of the experiment result and their corresponding values.

Save Description (Available after you add or change a description) Saves the current description with the experiment result.

Close Closes the dialog box.

QC Metrics Table

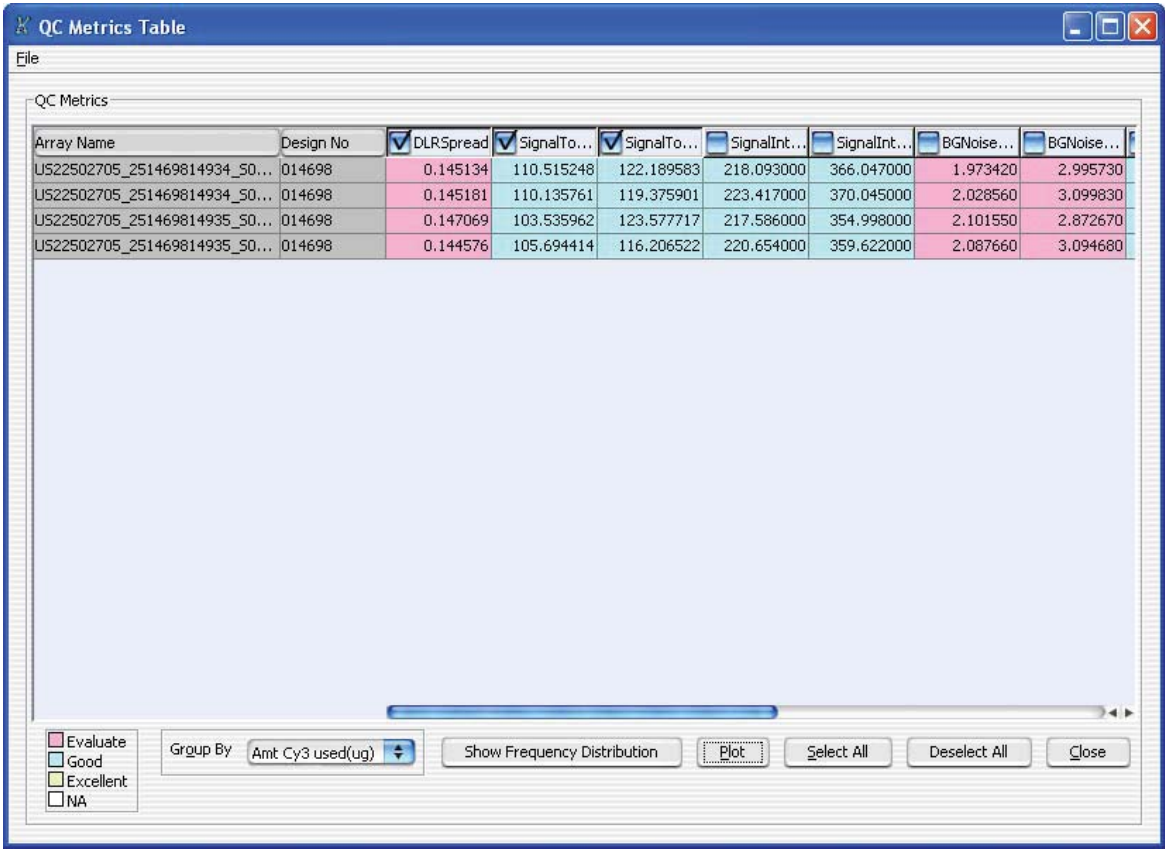


Figure 149 QC Metrics Table

Purpose: The QC Metrics Table shows the available metrics for one or more arrays. With this table and the available plots, you can evaluate the quality of your microarray results, and assign a manual QC status to each microarray. Some metrics come from the Agilent FE program, while others are calculated by the CGH module itself. These metrics are available only for Agilent microarrays. See "To display QC metrics of arrays and set array QC status" on page 134.

To open: Click **Preprocessing > QC Metric**, or right-click an experiment or an array and click **QC Metrics**.

File Lets you save the QC Metrics Table as a Microsoft Excel (*.xls) format file. When you click **File,** a menu opens with an Export command. This command opens a Save dialog box, where you can select a location and type a name for the exported file.

Table The values of the QC metrics for arrays appear under QC Metrics, one array per row. The table has many columns:

- **Microarray Name** – Displays the names of microarrays. Because you can open the QC Metrics Table in several ways, the list can contain an individually-selected microarray, or those associated with an experiment or with a design.
- **Design No.** – Identifies the Agilent designID for each microarray.

Metrics – The program evaluates each metric, and assigns it a rating of Excellent (yellow), Good (turquoise), Evaluate (pink), or NA (white). The name of each metric appears as a column heading. Mark the check box next to the name of the metric to include it in the available plots. Drag the column heading of a metric horizontally to change its position in the table.

- **ManualQCFlag** – Lets you set the QCMetricStatus attribute of the array. Status can be Pass, Fail, Marginal, or NA. Later, you can filter arrays based on this attribute.

Group By The program displays the arrays in the table, and in the QC Metrics Graph, grouped by the array attribute you select here.

Show Frequency Distribution Opens the Frequency Distribution dialog box. This dialog box contains line plots of the distribution of each selected metric over the all of the arrays in the QC Metrics Table. See "QC Metrics – Frequency Distribution" on page 381.

Plot Opens the QC Metrics Graph dialog box. This dialog box contains plots of each selected metric for each array. See "QC Metrics Graph" on page 382.

Select All Marks the check boxes of all metrics.

Deselect All Clears the check boxes of all metrics.

Close Closes the QC Metrics Table.

QC Metrics Thresholds

The following table lists the QC Metrics calculated by the CGH program.

Metric Name	Lower Threshold	Lower Threshold Warning	Upper Threshold Warning	Upper Threshold
DLRSpread			0.2	0.3
SignalToNoiseRed	30	100		
SignalToNoiseGreen	30	100		
SignalIntensityGreen	50	150		
SignalIntensityRed	50	150		
BGNoiseGreen			5	15
BGNoiseRed			5	15
ReproducibilityGreen	0		0.05	0.2
ReproducibilityRed	0		0.05	0.2

Figure 150 QC metrics threshold settings for CGH

QC Metrics – Frequency Distribution

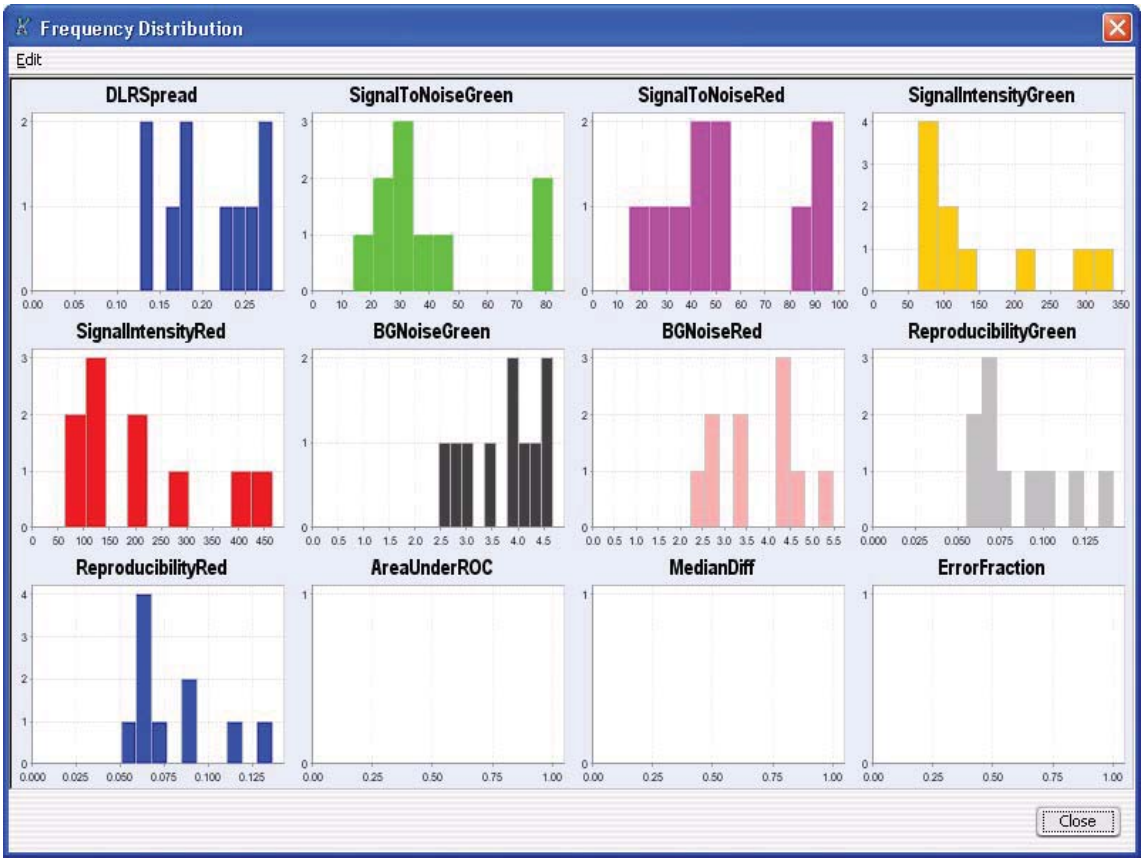


Figure 151 QC Metrics Frequency Distribution Plot

Purpose: The plots in this dialog box represent the selected columns in the QC Metrics Table. Each plot shows the number of arrays within each value range for a metric. See "QC Metrics Table" **on page 378**.

To open: Click **Show Frequency Distribution** in the QC Metrics Table.

Edit Opens a menu with a Copy command that copies the plots in the dialog box to the Clipboard. You can then paste the image into a document.

Close Closes the dialog box.

QC Metrics Graph

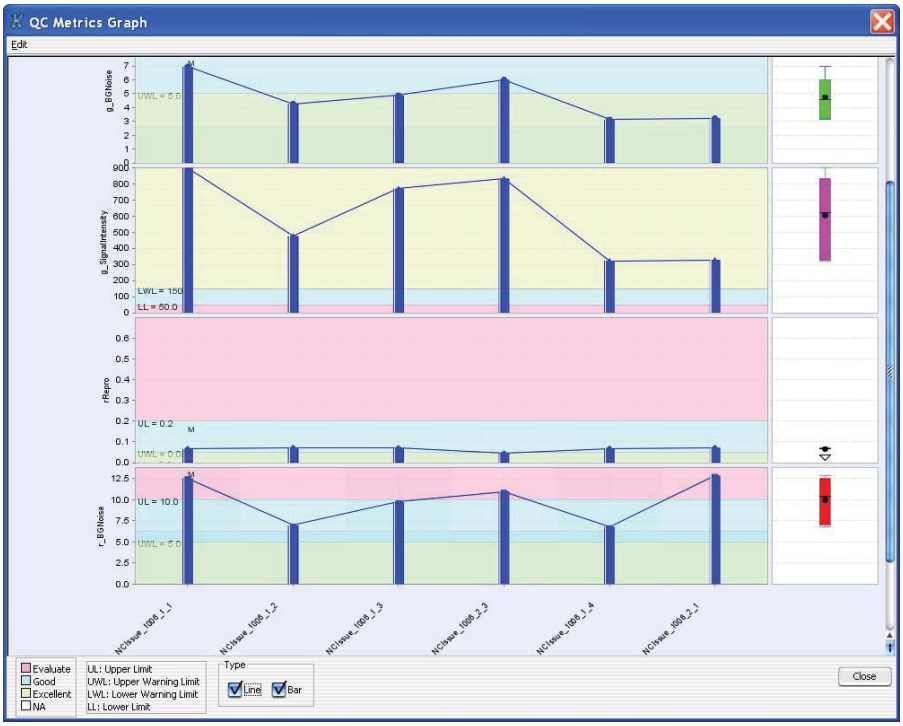


Figure 152 QC Metrics Graph

Purpose: To display plots of selected metric(s) for all arrays.

To open: Click **Plot** in the QC Metrics Table. See "QC Metrics Table" on page 378.

Main Plots The plots in this dialog box represent the selected columns in the QC Metrics Table. Each plot shows the value of a given metric for all arrays.

These plots have several additional features:

- The background colors in each plot correspond to quality guidelines developed by Agilent, based on normal ranges observed for analyses of well-established cell lines using standard Agilent protocols. See the descriptions of each of these metrics in "QC Metrics Table" on page 378.
- A "Box & Whisker" plot appears to the right of the main plot for each metric. See "'Box & Whisker' Plot," below.
- The program can plot the data as a line graph, a bar graph, or both. See "Line" and "Bar," below.
- If you set your view preferences to show ToolTips (see "View Preferences" on page 422), a tool tip appears when you place the pointer over any bar. The ToolTip shows the value of each bar and the name of the corresponding array. See Figure 153.
- You can right-click any bar to open a shortcut menu for the corresponding array. The options in the shortcut menu let you set the QCMetricStatus attribute for the array. See Figure 153.
- The QCMetricStatus attribute for each array appears over all of the corresponding bars of the main plot. The four possible values for QCMetricStatus are: (P)ass, (F)ail, (M)arginal, and (N)A.

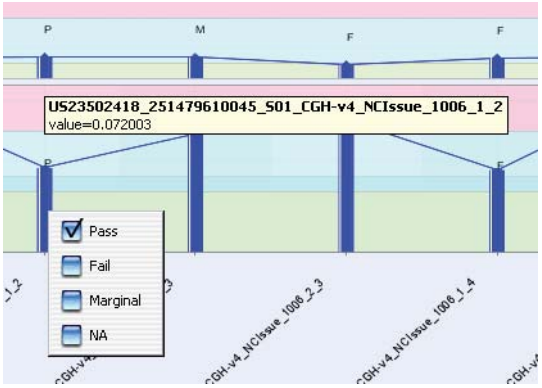


Figure 153 Portion of the QC Metrics Graph, showing a ToolTip (values in white box), and a shortcut menu. You use the shortcut menu to set the QCMetricStatus attribute for the array. Right-click any bar to open the shortcut menu for the corresponding array.

"Box & Whisker" Plots

A small plot appears to the right of each of the main plots. It represents the overall distribution of values for the metric. Two examples appear in Figure 154. The symbols have the following meanings:

- The lower and upper edges of the box represent the 25th and 75th percentiles, respectively.
- The black horizontal line in the box is the median.
- The black dot is the mean.
- The "whiskers" represent the range of values that are not outliers. An outlier is a point that is out of the 25th to 75th percentile range by more than 150%.
- Open circles represent outliers, and an open triangle represents outliers that plot beyond the available space on the graph.

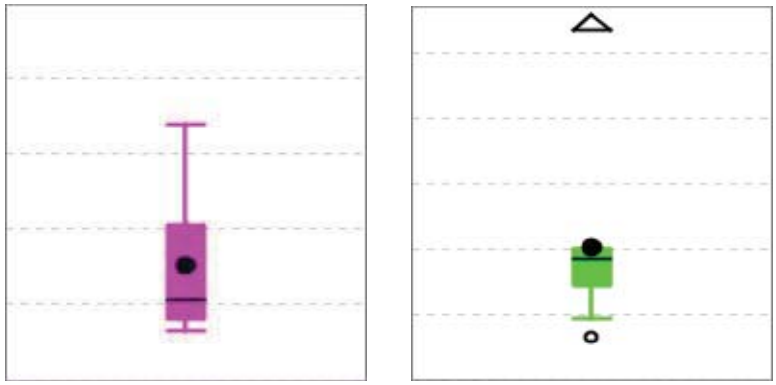


Figure 154 Two "Box & Whisker" plots

Line Mark this option to display each metric as a line graph.

Bar Mark this option to display each metric as a bar graph.

Edit Opens a menu with a Copy command. This command copies the plots in the dialog box to the Clipboard as an image. You can then paste the image into a document in another program.

Close Closes the dialog box.

Report Access Details

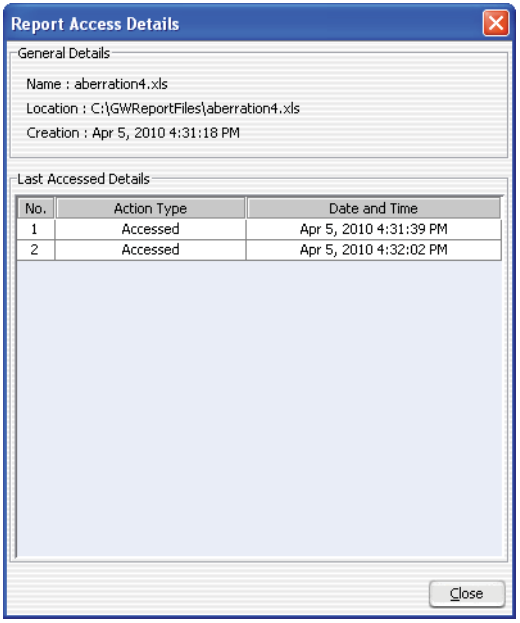


Figure 155 Report Access Details dialog box

Purpose: Displays details about access to reports.

To open: From the Report Management dialog box, click **More Details**. See "Report Management" on page 386.

General Details Displays name, location, and creation date and time for the selected report.

Last Accessed Details Displays a list of the action type and the date and time of the action.

Report Management

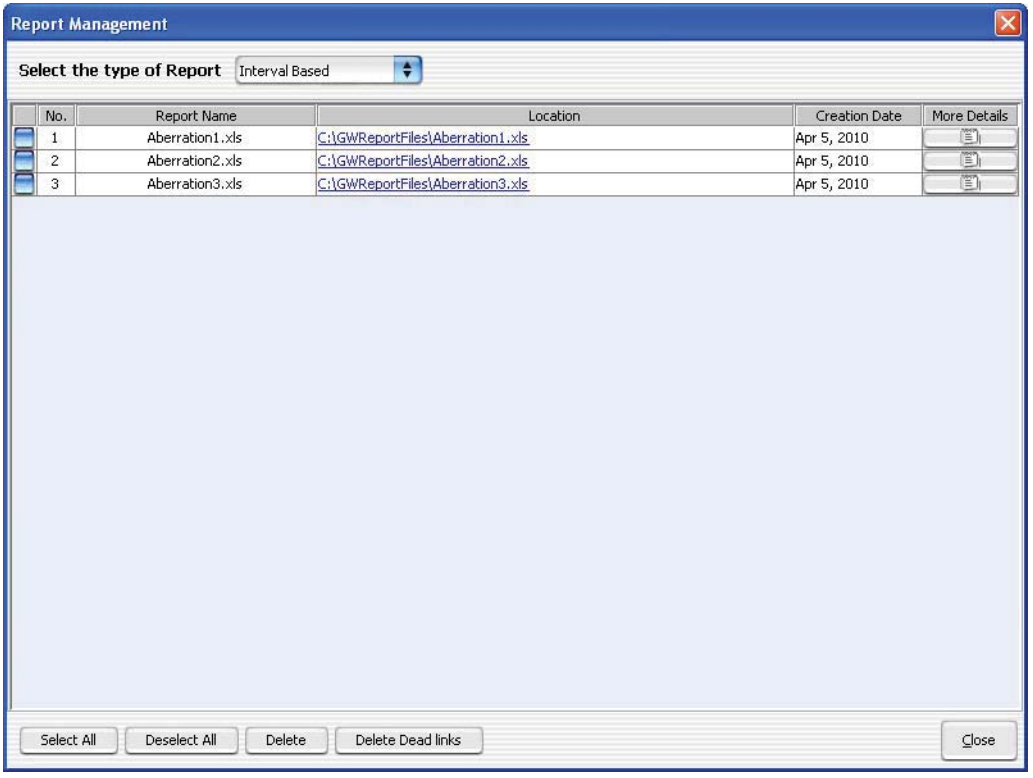


Figure 156 Sample Management dialog box

Purpose: This dialog box lets you display a list of reports created in Agilent Genomic Workbench, and manage their storage.

To open: From the Reports tab, click **Report Manager**.

Select the type of Report Click the arrow and select the type of report you want to show in the list.

No. Displays an index number for the report.

Report Name Displays the file name for the report.

Location Gives the path name for the location of the report. Click the link to open the report.

Creation Date Shows the date the report was created.

More Details Click to open the Report Access Details dialog box, which shows access information for the report. See "Report Access Details" on page 385.

Select All Click to select all of the reports in the list.

Deselect All Click to deselect all of the reports in the list.

Delete Click to delete all of the selected reports.

Delete Dead links Click to delete only the reports where the displayed link is not functional.

Run Cytogenetic Report

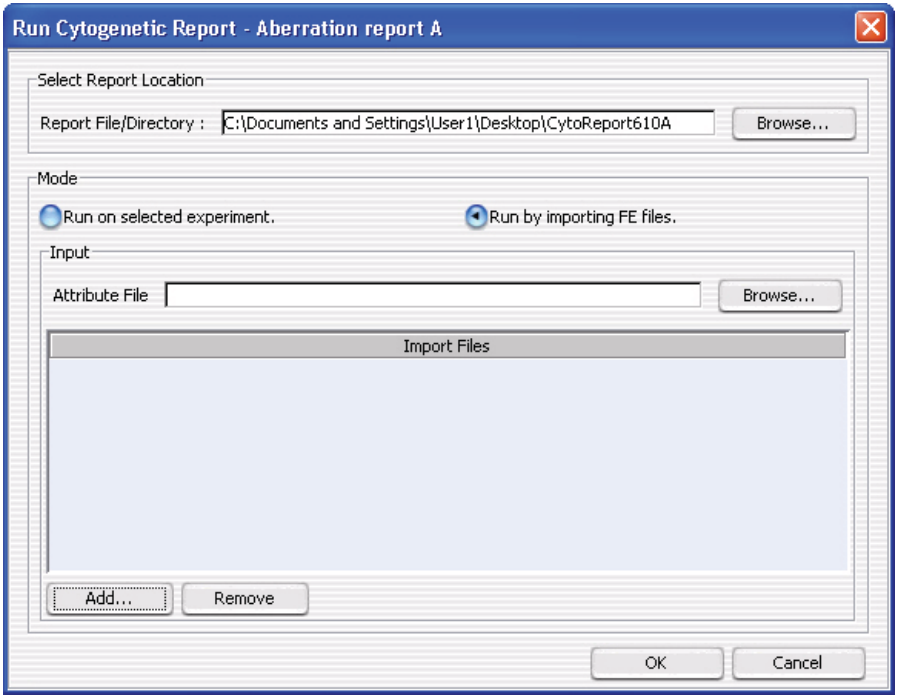


Figure 157 Run Cytogenetic Report dialog box

Purpose: The Run Cytogenetic Report dialog box allows you to select a location for the report and, optionally, select data and array attribute files as input for the report. Cytogenetic reports summarize analysis settings and detected aberrations by array.

To open: Click **Reports > Cyto Reports,** then select the name of the desired report template. By default, no templates appear. To create a template, see "To create a new Cyto Report template" on page 169.

Select Report Location Defines the location for the cyto report file(s). Type or edit the location in Report File/Directory. Alternatively, click **Browse** to open the Select report folder dialog box. Select a location for the file(s), then click **Open.**

Mode Defines the source of array data for the report. Select one of these options:

- **Run on selected experiment** – The program uses the selected arrays in the active experiment as input for the report.
- **Run by importing FE files** – The program imports Agilent FE data files that you select as input for the report. If you select this option, the commands under Input become available

Input Use these settings to select Agilent FE files to import as input for the report:

- **Attribute File** – (Optional) Shows the location of the array attributes file to be imported with the data files. To select an array attributes file for import, click **Browse.** In the dialog box that appears, select the desired file, then click **Open.**

  An array attributes file contains specific attributes of your imported input arrays. See the *Sample Manager User Guide* for details.

NOTE For CGH+SNP arrays, the attributes file must contain the genotype reference in the Green Sample field (or Red Sample field for dye-flipped arrays.)

- **Import Files** – (Read-only) Displays the Agilent FE files selected for import.
- **Add** – Opens the Select files to import dialog box. To select a data file for import, use the Windows navigator commands in the dialog box to locate the desired file, then click its name. To select multiple files, hold down the **ctrl** key and click their names. When you are finished selecting files, click **Open.**
- **Remove** – Removes a selected file from Import Files. To remove a file, click its name under **Import Files,** or hold down the **ctrl** key and click multiple files, then click **Remove.**

OK Runs the cytogenetic report. The program generates a separate PDF file for each array.

Cancel Closes the dialog box without generating a cytogenetic report.

Sample Attributes

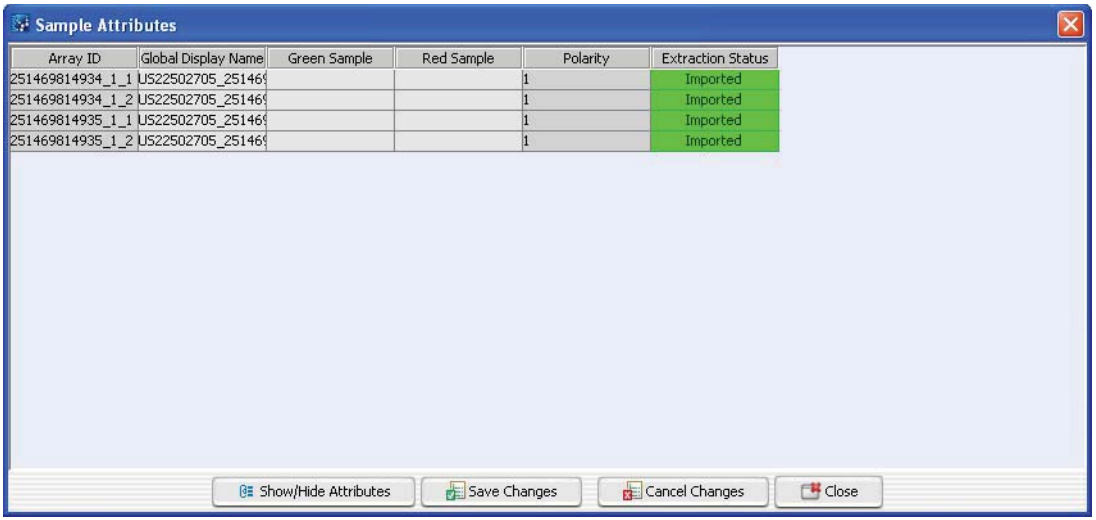


Figure 158 Sample Attributes dialog box

Purpose: To show, hide, or edit array attributes in the experiment.

To open: In the Experiment Pane of the Navigator, right-click the experiment, and click **Sample Attributes**.

This dialog box lets you type or change the existing values for the attributes listed for the arrays in the experiment. You can also show or hide attribute columns. The columns that appear initially are the default columns (Array ID, Global Display Name, Green Sample, Red Sample, Polarity and Extraction Status), plus any that you selected to show. See the *Sample Manager User Guide* for more information on array attributes.

NOTE Changes to array attributes you make in this table are applied globally to the arrays and are changed in the Sample Manager Sample Utility Tab.

Scroll to Column

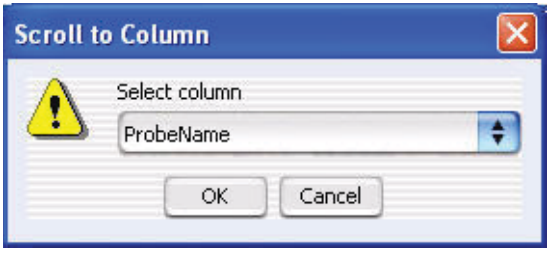


Figure 159 Scroll to Column dialog box

Purpose: Lets you select a "Scroll to" column. The program then scrolls the tab so that you can see the selected column.

To open: Right-click a column heading in Tab View, then click **Scroll To Column** in the shortcut menu.

Select column Displays the columns available in the selected tab. Select the one you want to display.

OK Scrolls the current tab so that you can see the selected column.

Cancel Closes the dialog box.

Search Probes in eArray

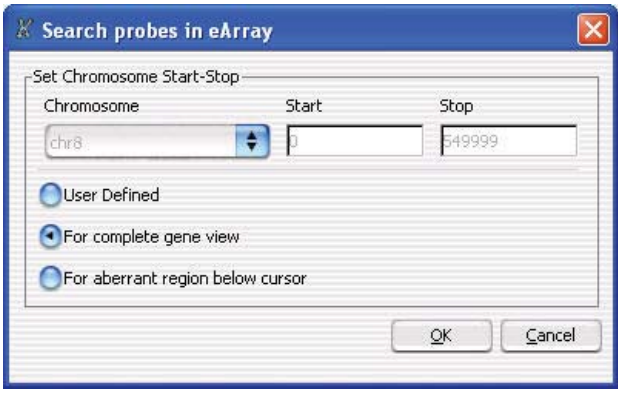


Figure 160 Search probes in eArray dialog box

Purpose: To select the probes you want to update in eArray

To open: Right-click Gene View, and click **Search probes in eArray**.

Select a chromosome and a region in Chromosome View for selecting the probes related to the genes in this region.

User Defined Select to choose the region from which the probes to be searched in eArray will be selected. The chromosome selection list and the Start and Stop positions on the Y axis are activated when this option is selected.

For complete gene view All the probes related to the genes in Gene View will be searched.

For aberrant region below cursor Selects those probes for the genes that appear just below where the cursor sits in Gene View.

Chromosome If you select User Defined, you can select a different chromosome than had been selected before opening this dialog box.

Start/Stop If you select User Defined, you can type in Start and Stop positions for defining the region contained the genes to be in the list.

Select Chromosome Intervals

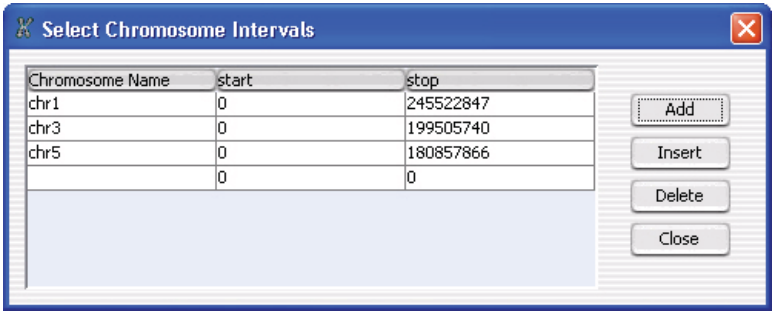


Figure 161 Select Chromosome Intervals dialog box

Purpose: The Select Chromosome Intervals dialog box allows you to define genomic intervals for a correlation analysis.

To open: The Select Chromosome Intervals dialog box opens when you select **Genomic Location** and click ... in the Correlation Analysis Setup dialog box (see "Correlation Analysis Setup" on page 270).

List of intervals Lets you view and edit the chromosomal intervals defined for the correlation analysis. By default, no intervals are defined. To add an interval, click **Add.** The list has these columns:

- **Chromosome Name** – The name of the chromosome defined for the interval. To define the name, click the desired cell under **Chromosome Name,** then select the chromosome from the list.
- **start** – The first base pair in the interval. Double click the number, edit it, then press **Enter**.
- **stop** – The last base pair in the interval. Double click the number, edit it, then press **Enter**. If you don't select a stop location, the program sets this value to the last base pair on the chromosome.

Add Adds an empty row to the list of chromosomal intervals.

Insert Select any row; click **Insert** to add a duplicate row above the selected one.

Delete Select any row, then click **Delete** to remove the row.

Close Closes the dialog box and accepts the list of chromosomal intervals.

Select Color

Purpose: To select a color. Three tabs are available for selecting colors:

- Swatches tab - select colors based on samples (swatches)
- HSB tab - select colors based on an HSB schema (Hue, Saturation, and Brightness)
- RGB tab - select colors based on an RGB schema (Red-Green-Blue)

To open: This dialog box opens when a function allows you to change a color. For example, right-click an array in an experiment, click **Edit Array Color** and click the **Swatches, HSB,** or **RGB** tab.

Swatches tab

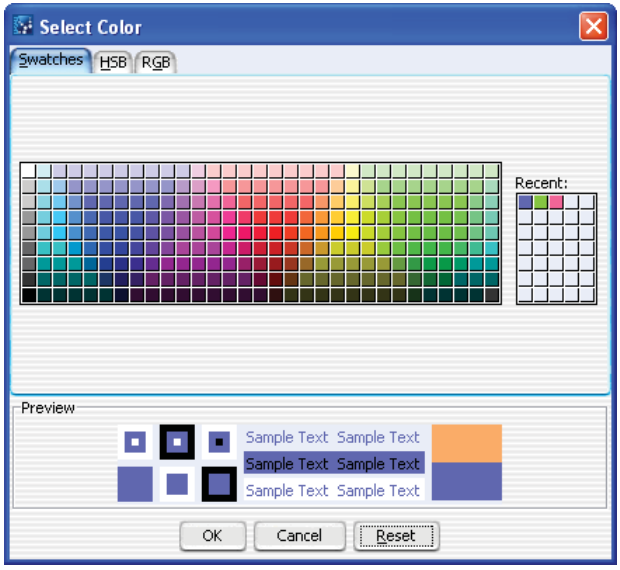


Figure 162 Select Color - Swatches tab

This tab is used to select a color based on color samples (swatches).

Preview The Preview area shows how the selected color appears. When you change the color, the original color appears at the top of the color box on the right.

Recent: Choose a recent color selection.

OK Click to select the color and close the dialog box.

Cancel Click to close the dialog box without changing the color.

Reset Click to change swatches, HSB, and RGB colors back to the default colors.

HSB tab

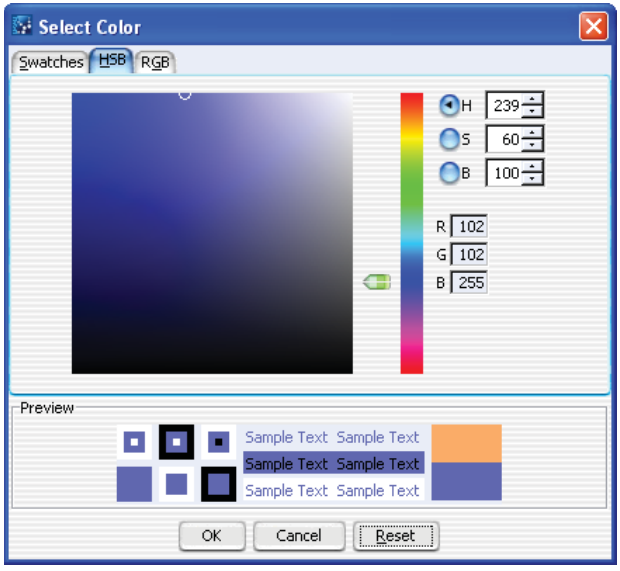


Figure 163 Select Color - HSB tab

In this tab, you can select a color based on an HSB schema (Hue, Saturation, and Brightness).

Hue Click the **H** button, and move the slider up and down, or go up and down the list of numbers, to select the hue or color of the array.

Saturation Click the **S** button, and move the slider up and down, or go up and down the list of numbers, to select the saturation level for the color.

Brightness Click the **B** button and move the slider up and down, or go up and down the list of numbers, to select the brightness level for the color.

RGB Numbers Reflect the amount of red, green and blue in the resulting color.

Preview The Preview area shows how the selected color appears. When you change the color, the original color appears at the top of the color box on the right.

OK Click to select the color and close the dialog box.

Cancel Click to close the dialog box without changing the color.

Reset Click to change the swatches, HSB, and RGB colors back to default values.

RGB tab

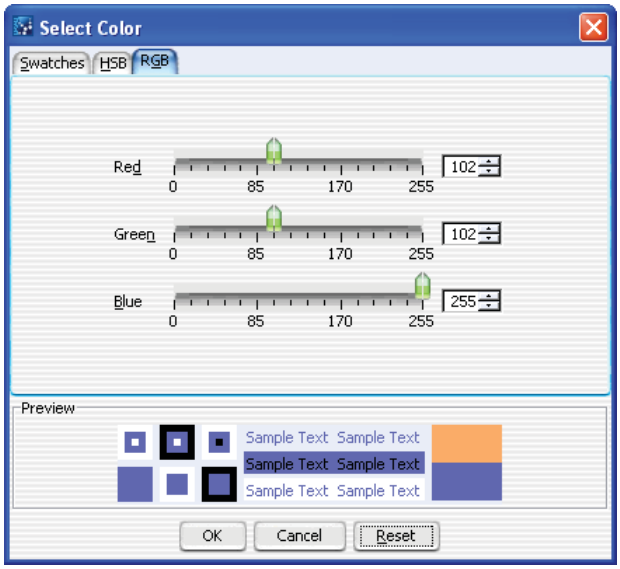


Figure 164 Select Color - RGB tab

This tab is used to select a color based on an RGB schema.

Red Move the slider to change the amount of red in the color. Or, click the up or down arrow to select a number.

Green Move the slider to change the amount of green in the color. Or, click the up or down arrow to select a number.

Blue Move the slider to change the amount of blue in the color. Or, click the up or down arrow to select a number.

Preview The Preview area shows how the selected color appears. When you change the color, the original color appears at the top of the color box on the right.

OK Click to select the color and close the dialog box.

Cancel Click to close the dialog box without changing the color.

Reset Click to return the swatches, HSB, and RGB colors back to default values.

Select data type for experiments

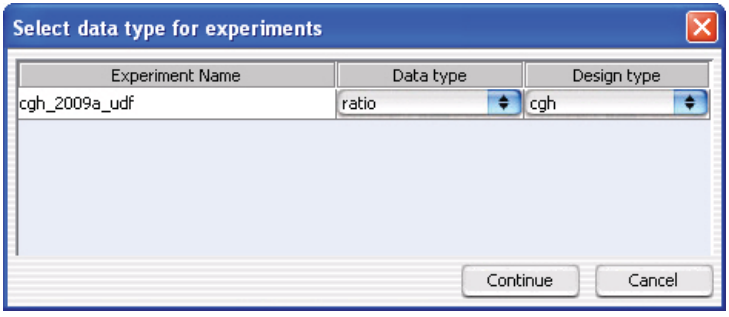


Figure 165 Select data type for experiments dialog box

Purpose: Lets you select the mathematical form of the data in an imported UDF file, and its associated application type. See "To import a UDF file" on page 52.

To open: In the Home tab, click **Import > Array Files > UDF File.** In the dialog box that appears, select the desired UDF file, then click **Open.**

Experiment Name By default, the experiment name is the name of the imported UDF file. To change the name, double-click it, then edit it as desired.

Data type Select the mathematical form of the array data in the UDF file. The options are:

- **ratio**

- **$\log_2$ ratio**
- **$\log_{10}$ ratio**
- **ln ratio** (base e)

Design type Select the application type (CGH or CH3, for example) associated with the array data in the UDF file.

Continue Accepts your selections, and goes to the next step in the UDF import process.

Cancel Cancels the UDF import.

Set genome build and species for Axon design files

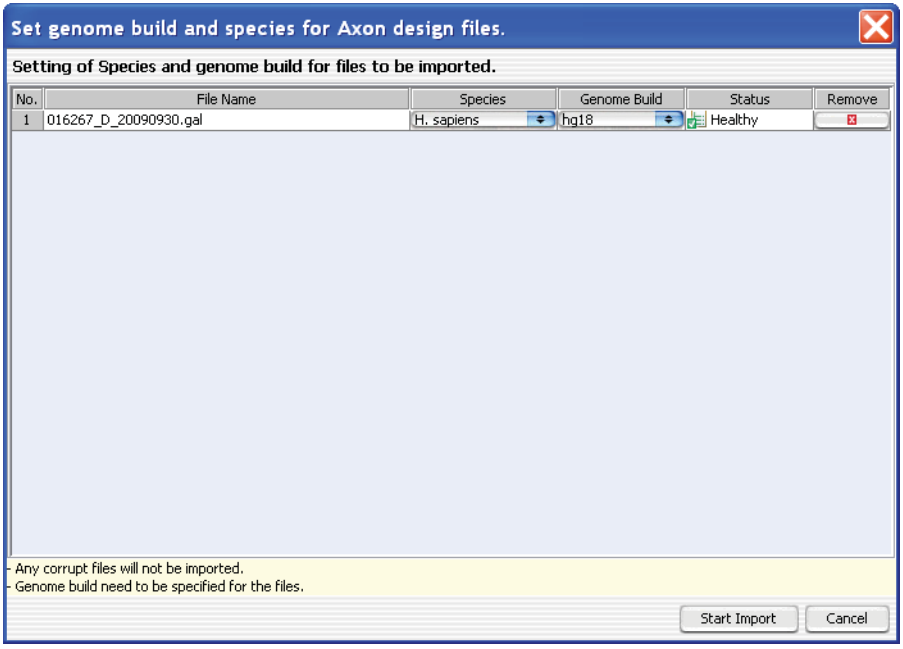


Figure 166 Set genome build and species for Axon design files dialog box

Purpose: Lets you set the species and genome builds associated with imported Axon design file(s), and to remove specific designs files from the import, if necessary. See "To import Axon design files" on page 49.

To open: In the Home tab, click **Import > Design Files > Axon File.** In the dialog box that appears, select at least one Axon design file, then click **Import.**

No. An index number within the dialog box for each Axon file.

File Name The names of each Axon design file selected for import.

Species The species associated with each design file. If a species is incorrect, select the correct one from the appropriate list.

Genome Build The genome build associated with each of the design files. If a genome build is incorrect, select the correct one from the appropriate list.

Status The status of the file is one of the following:

- **Valid** – The file is a new file that can be imported.
- **Healthy** – The file passes validation and can be imported.
- **Not Set** – Appears if Genome Build and Species information is not shown.
- **Not Allowed** – Appears if a Genome Build is selected that does not match the design.
- **Overwrite** – The file is a valid design file, but when you import it, it will replace an existing design that has the same name.
- **Corrupt** – The file failed validation. When you start the import process, the program ignores the file.

Remove Click  to remove a specific design file from the list. This can be useful if you select a design file in error, or if you do not want to overwrite an existing one.

Start Import Imports the file(s) and closes the dialog box.

Cancel Cancels the import and closes the dialog box.

Setting Clustering Method and Parameters

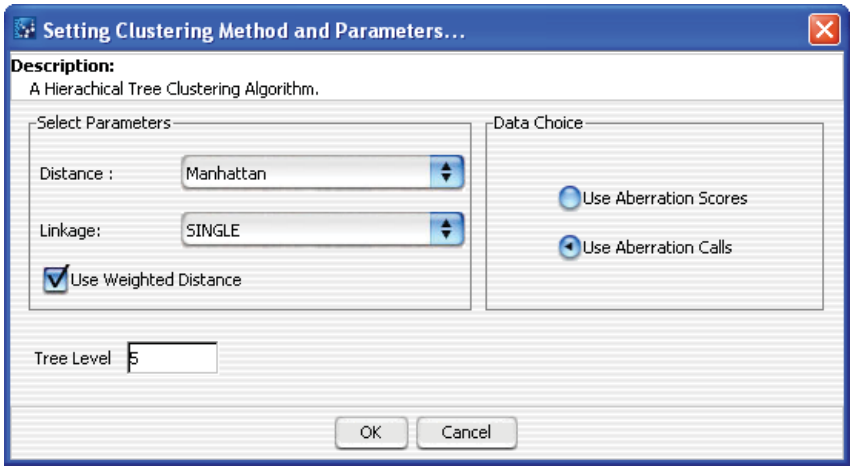


Figure 167 Setting Clustering Method and Parameters dialog box

Purpose: This allows you to set up a clustering analysis of the aberrations that were detected in the selected arrays of the active experiment. See "To set up a clustering analysis" **on page 161**. Cluster analysis is used to identify new classifications, or groupings, of samples based entirely on the aberration profile information.

To open: this dialog box, click **Discovery > Clustering.**

Select Parameters Select the desired distance metric, as described in the table below. By default, the program uses Manhattan distance. Also by default, the program uses weighted distances. To use unweighted distances, clear **Use Weighted Distance.**

Distance Metric	Description
Manhattan	Manhattan distance is also called City Block distance, or Minkowski's *L1* metric. It is the distance in multidimensional space between two points as taken by following each dimensional axis.
Euclidean	Euclidean distance is also called Minkowski's *L2* metric. It is the distance in multidimensional space between two points connected by a straight line, and is therefore shorter than the Manhattan distance for all cases where the number of dimensions is greater than 1 (that is, a straight line).
Correlation	Distance is defined using Pearson correlation, that measures the angle of the vectors from two arrays around the mean of the aberration vectors. Using the Pearson correlation often results in more negative correlations than using the cosine correlation.
Distance Similarity	Normalized Euclidean distance.
COSINE Similarity	The cosine of the angle between two vectors *A* and *B* measures how related the two samples are across all intervals containing aberrations.

Linkage	Description
SINGLE	The distance between two clusters is calculated as the smallest distance between pairs of members from the two clusters. Single linkage maximizes the connections of a cluster, but in doing so may tend to find spread-out chain-like clusters that contain dissimilar elements.
AVERAGE	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.
COMPLETE	The distance between two clusters is calculated as the largest distance (dissimilarity) between any two members in the two clusters. Complete linkage maximizes the maximum within-cluster distance, and so tends to find discrete, compact clusters. However, it may overemphasize small differences between members.

Data Choice The source of the data for the clustering analysis. Select one of these options:

- **Use Aberration Scores** – Uses the actual score as determined by one of the aberration detection algorithms as the basis to measure distance or similarity in cluster analysis.
- **Use Aberration Calls** – Uses a simplified numerical representation of whether an aberration was present (1 for amplification, –1 for deletion, 0 for a region of no change) as the basis to measure distance or similarity in cluster analysis.

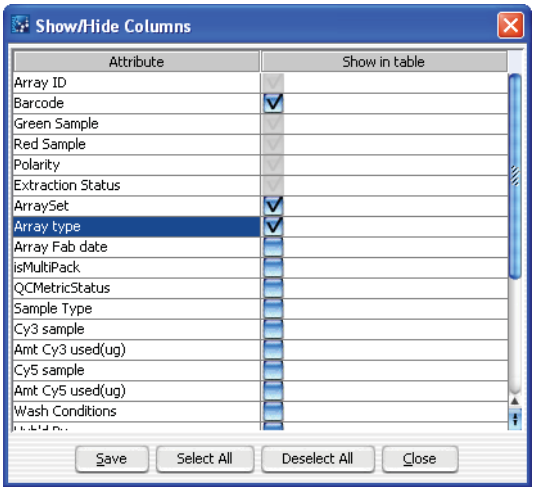
Tree Level This parameter decides how many levels of the tree will be rendered. Agilent recommends that you leave the setting at the default of 5. The program determines the number of clusters by cutting the tree at a particular level.

OK Begins the clustering analysis on the elected arrays in the active experiment.

Cancel Closes the dialog box without starting a clustering analysis.

NOTE Before you start a clustering analysis, you must first set up an experiment with the desired arrays and apply an aberration detection algorithm.

Show/Hide Columns



Show/Hide Columns dialog box

Purpose: Used to select the attributes to be displayed in the Sample Attributes dialog box and the Sample Utility tab. The Sample Utility tab is available when you go to Sample Manager. See the *Sample Manager User Guide* for information about Sample Manager.

To open: This dialog box appears when you click **Show/Hide Attributes** at the bottom of the Sample Attributes dialog box.

All available attributes are shown in the Attributes column. Attributes with a check-mark next to them will be displayed in the Sample Attributes and Sample Utilities tab. To select an attribute for display, mark the **Show in Table** box next to it. To hide an attribute, clear the **Show in Table** box again.

Save Saves the current list of selected attributes and updates the Sample Utility table, based on the selections.

Select All Selects all the attributes in the list.

Deselect All Clears all check marks from attributes in the list.

Close Closes the dialog box. If changes have been made, the program asks if you want to save your changes before closing.

Show Intensity Bar Charts

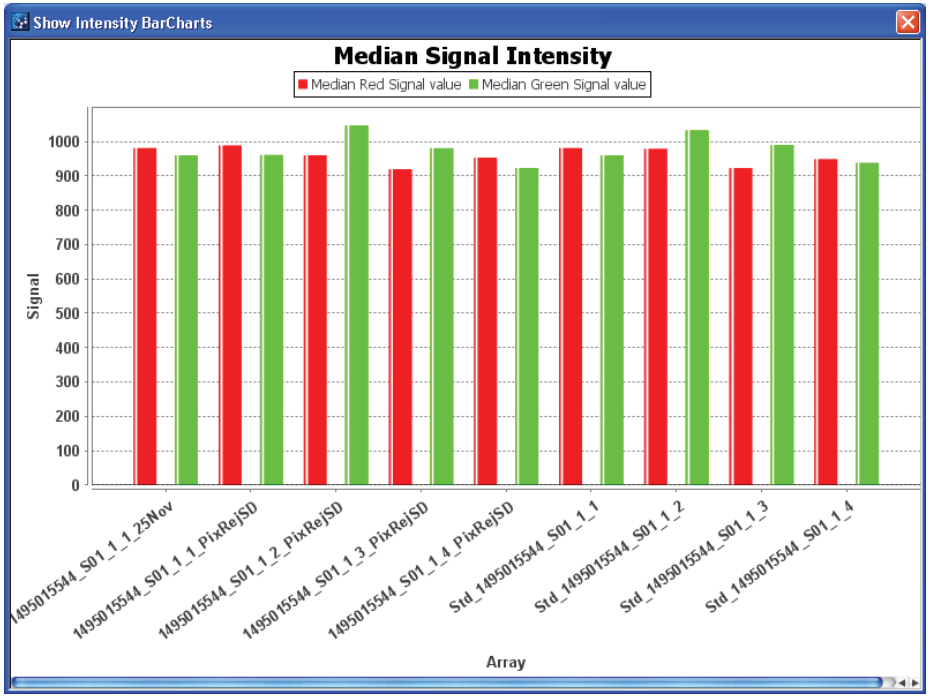


Figure 168 Show Intensity Bar Charts display

Purpose: This window displays a signal intensity bar chart.

To open: This display appears when you create a signal intensity bar chart from Gene View. See "To create a Signal Intensity Bar Chart" on page 104.

Right-click the chart and select from the following options.

Copy Click to copy the bar chart to the Clipboard. You can then paste the bar chart image into another application that supports the Clipboard.

Save as Click to open the **Save as** dialog box that lets you save the bar chart in a file on your hard disk.

SNP CN QC Metrics Table

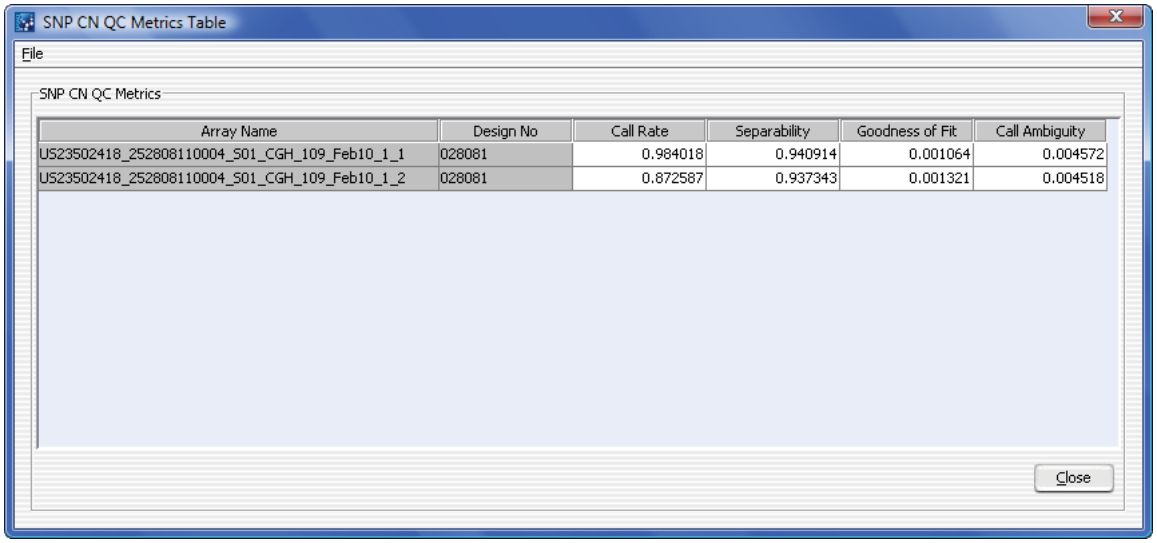


Figure 169 SNP CN QC Metrics Table

Purpose: To display calculated metrics for selected arrays in a CGH+SNP experiment.

To open: In the Navigator, right-click on a selected CGH+SNP experiment, and then click **SNP QC Metrics**.

File Opens the Export command, where you can choose to export the displayed metrics table to a file on your hard disk.

Array Name (Read only) The name of the microarray.

Design No (Read only) The microarray design number.

Call Rate Call rate is the number of SNPs that receive a genotype call at a 95% threshold divided by the total number of SNPs for which there is signal in the reference. For high-quality DNA samples hybridized against a genotyped and qualified reference, this value should exceed 90%.

Separability The separation between the CN peaks. It is defined as the distance between the average log ratio of SNPs with CN 1 and 2, which is a measure of the log ratio compression. For high-quality DNA samples this value should exceed 0.85. Lower values can indicate that the hybridization or wash stringency was too low, or the hybridization time was too short.

Goodness of Fit The error in the Gaussian fit for the peaks. It is obtained from the difference between the observed log ratio distributions and the modeled Gaussian distributions. It is a measure of the quality of the curve fitting model. For high-quality DNA samples this value should not exceed 2%.

Call Ambiguity This measures the overlap between the CN 1 and 2 Gaussian peaks. If a probe falls into this region, the CN call for that probe will not be made with high confidence. In other words, it is a measure of the degree of ambiguity of copy number calls. For high-quality DNA samples this value should not exceed 2%.

Close Closes the dialog box.

SNP Genotype Report Setup

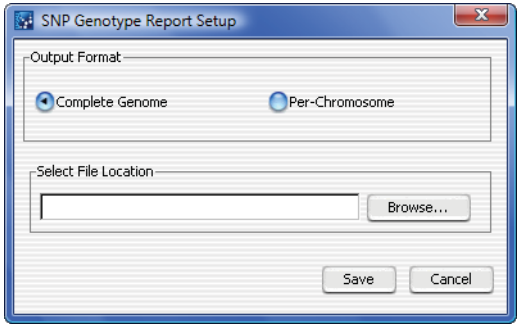


Figure 170 SNP Genotype Report Setup dialog box

Purpose: Used to select how to report SNP genotype text results, and to select a location for the report.

To open: In the Reports tab, under SNP Genotype, click **Text**.

Output Format

Complete Genome Select to create a single report that contains SNP genotype information for the entire genome.

Per-Chromosome Select to create a separate SNP Genotype text report for each chromosome.

Select File Location

Shows the location to which the program saves the report file(s). To select a location for the report(s), click **Browse.** In the Open dialog box that appears, select a location and type a name for the report, then click **Open.**

Save Saves the selected report in the indicated location.

Close Closes the dialog box without generating a report.

Stroke Selection

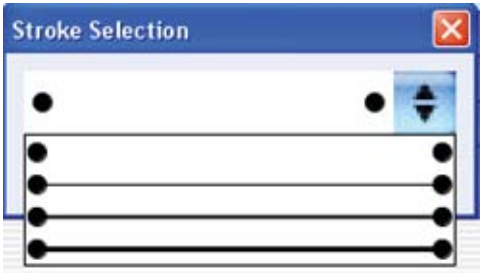


Figure 171 Stroke Selection dialog box

Purpose: This dialog box is used to select the line width for outline in a chart.

To open: In the Plot tab of the Chart Properties dialog box, select the **Appearance** tab and then next to Outline stroke, click **Select**.

Click the arrow and select the desired line width, then click **OK** to close the dialog box.

Text Penetrance Summary Setup

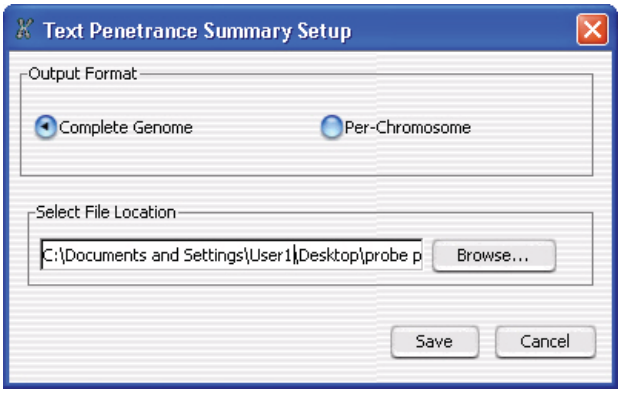


Figure 172 Text Penetrance Summary Setup dialog box

Purpose: This dialog box allows you to configure the Text Penetrance Summary Report, and select a location for it. This report displays each probe associated with a significant aberration, and gives the percentage of selected arrays that show a significant deletion or amplification in the region associated with each probe. The workflow creates one or more *.xls format files that you can work with in Microsoft Excel.

To open: In the Reports tab, under Penetrance**,** click **Probe.**

Output Format Select one of these options:

Option	Description
Complete Genome	Creates a single report file for the entire genome.
Per-Chromosome	Creates a separate report file for each chromosome in the genome.

Select File Location Shows the location to which the program saves the report file(s). To select a location for the report(s), click **Browse.** In the Open dialog box that appears, select a location and type a name for the report, then click **Open.**

Save Creates the Text Penetrance Summary report, and saves it in the specified location.

Cancel Closes the dialog box without creating a Text Penetrance Summary Report.

Track

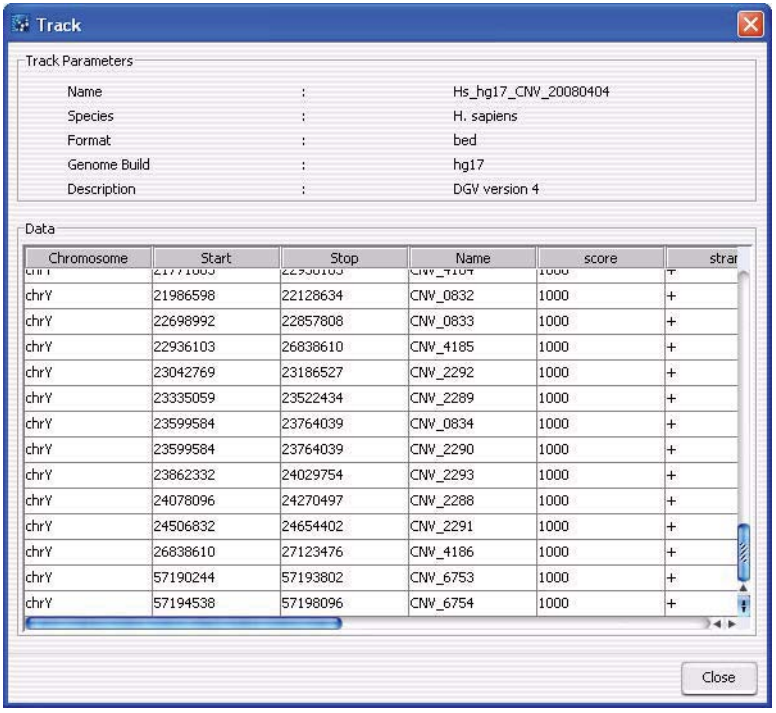


Figure 173 Track details

Purpose: This dialog box allows you to display the parameters and data of the track.

To open: Click the **Details** link next to the desired track in the **Tracks** tab of the User Preferences dialog box. See "Tracks tab" on page 415.

Track Parameters These parameters appear:

Parameter	Description
Name	The name of the track.
Species	The species to which the track applies.
Format	The format of the track data. Agilent Genomic Workbench supports the BED format.
Genome Build	The specific genome build of the species to which the track applies.
Description	Descriptive text saved with the track.

Data Tracks must contain entries for at least these four columns in the table:

Column	Description
Chromosome	The name of the chromosome.
Start	The first base pair of the particular feature in the chromosome.
Stop	The last base pair of the particular feature in the chromosome.
Name	The name of the feature. This name appears next to the defined region for the feature.

The other columns are additional BED track file columns that can appear for some tracks. Agilent Genomic Workbench does not render these.

Close Closes the Track dialog box.

UDF Import Summary

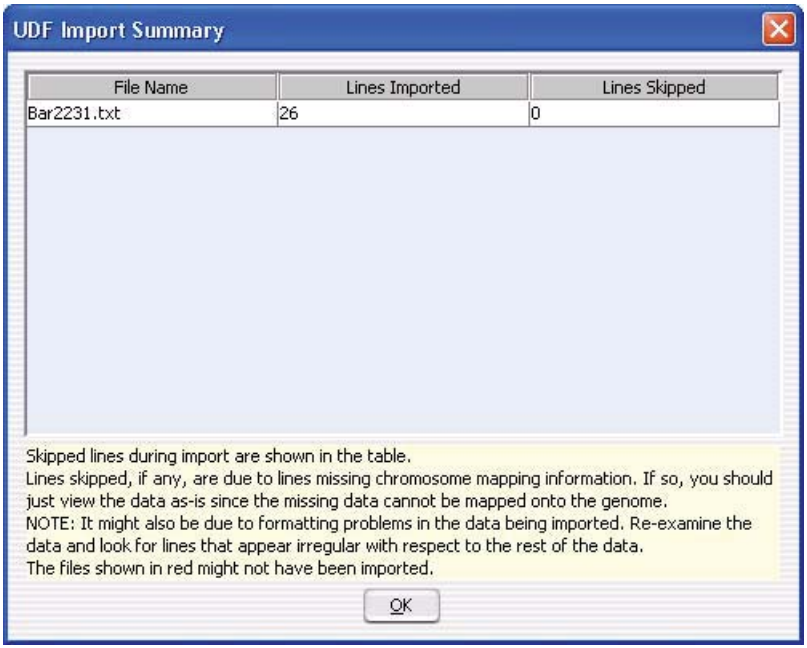


Figure 174 UDF Import Summary dialog box

Purpose: Reports how many lines of data were successfully imported from a UDF file, and how many lines were skipped. Skipped lines can be caused by missing chromosome mapping information, or improper formatting of the UDF file.

To open: Import a UDF file (see "To import a UDF file" on page 52). This dialog box appears after you map the columns of the UDF file.

Table Displays the file name of the imported UDF file, the number of lines that were successfully imported, and the number of lines, if any, that were skipped during import. If many lines were skipped, re-examine the data for improper formatting or missing chromosome mapping information.

OK Closes the dialog box.

Universal Data Importer - Map Column Headers

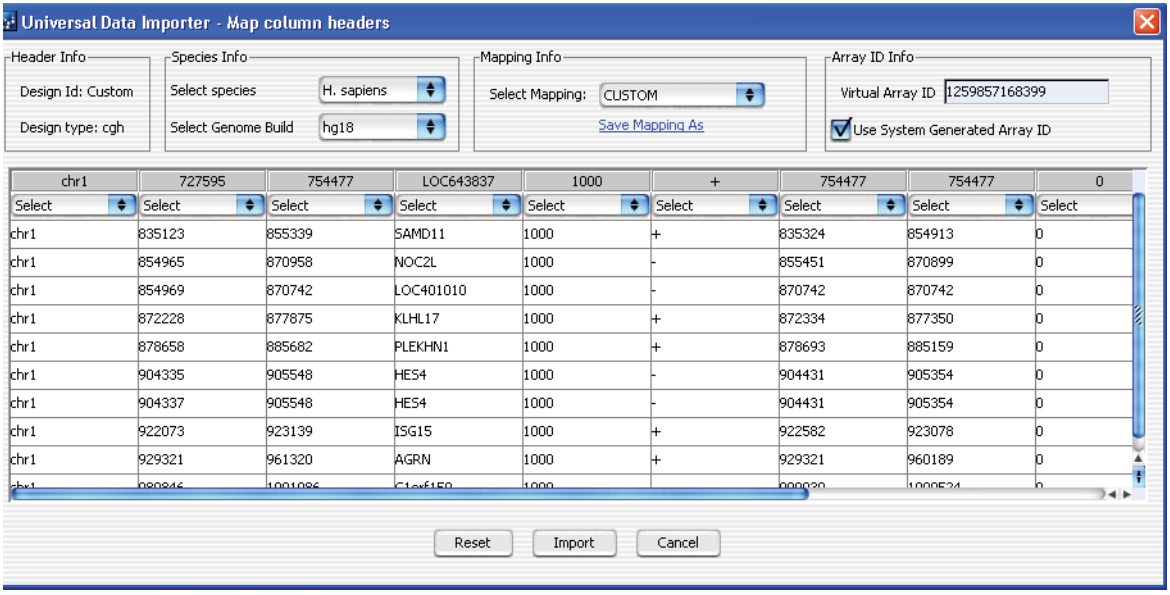


Figure 175 Universal Data Importer - Map Column Headers dialog box

Purpose: Lets you set up a universal data file (UDF) for import. You select several properties associated with the UDF, and identify the contents of each column of data in the file. You can also save column mappings for reuse.

To open: As you go through the UDF import process (see "To import a UDF file" on page 52), in the Select data type for experiments dialog box, click **Continue.** See "Select data type for experiments" on page 397.

Species Info

Select Species Select the species associated with the array data in the UDF.

Select Genome Build Sets the species-specific build to use.

Mapping Info

Select Mapping Applies a previously saved column map to the current UDF. A column map identifies the contents of each column of data. To create a new column map for the current UDF, select **CUSTOM.**

Save Mapping As Saves the column map under a new name. Opens an Input dialog box, where you can type a name for the new map.

Array ID Info

Virtual Array ID A number that uniquely identifies the data in the UDF. Typically, an Agilent microarray slide has a physical barcode that Agilent Genomic Workbench uses to generate an Array ID. The Array ID is used to track the data from the slide as it goes through the steps of an analysis workflow. A "virtual" Array ID is, by default, a system-generated ID that serves the same purpose for data from UDFs. You can also create your own virtual Array ID.

Use System Generated Array ID By default, the virtual barcode assigned to the array data in a UDF is a number that is created by the program. To create your own barcode, clear **Use System Generated Array ID,** then type a new number in **Virtual Array ID.**

Table

This table allows you to identify the contents of the columns of data in the UDF. The first row of the table shows the column heading information from the UDF. The second row contains labels that you apply to each column, and the rest of the table displays lines of data from the UDF. If the UDF contains data from Agilent CGH arrays, the column headings exactly match the labels in the lists.

In the list below each column heading, select the applicable label. You must use each of the labels exactly once, except LogRatio, which you can use more than once. These labels are available:

Label	Description
ProbeName	The column contains names of probes.
ChrName	The column contains names of chromosomes.

Label	Description
Start	The column contains the first chromosomal location to which each probe is designed.
Stop	The column contains the last chromosomal location to which each probe is designed.
Description	The column contains text annotation related to the probe.
LogRatio	The column contains array data values that correspond to each probe. You can use this label more than once.

NOTE If you select a saved column mapping, then change or reset the column labels in the table, the program changes or resets the saved column map as well.

Reset Clears all the column labels in the second row of the table. If you have selected a saved column mapping, this command also clears the labels in the saved map.

Import Imports the UDF file with the specified parameters, and opens the UDF Import Summary dialog box (see "UDF Import Summary" on page 411).

Cancel Cancels the import and closes the dialog box.

User Preferences

Purpose: This dialog box is used to set up preferences for display of tracks, data storage locations, and licenses.

To open: From the Home tab, click **User Preferences**. Or, right-click in the Gene View, Chromosome View, or Genome View, and click **User Preferences**.

Tracks tab

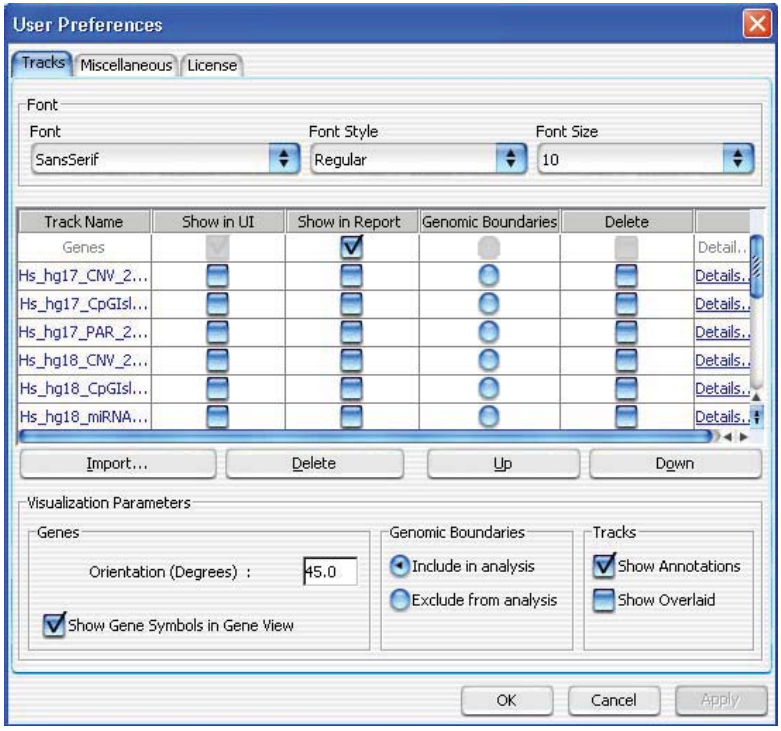


Figure 176 User Preferences dialog box - Tracks tab

Purpose: To import and set up the appearance of tracks next to the Gene View. Tracks are additional graphic displays of genomic information loaded from an external file. They align with genomic coordinates in Gene View.

To open: In the User Preferences dialog box, click the **Tracks** tab.

Font Options

Select the font type, style and size for the gene annotations that appear in the selected tracks.

Tracks List

Track Name Name of the track already loaded or imported

Show in UI Mark the check box to display the track next to Gene View.

Show in Report Mark the check box to display the track information in all the reports.

Genomic Boundaries Click to use the track to define only the regions that aberration detection algorithms will run. You can choose to do this for only one track.

Delete Mark the check box to delete the track from the list. Then, click **Delete** to delete the track from the list.

Details Click to display all the chromosome locations defined in the track.

Import Click to import new tracks.

Delete Click to delete the tracks selected in the Delete column.

Up Click to move a track up the list.

Down Click to move a track down the list.

Visualization Parameters

Genes These options affect the appearance of the Track and Gene View.

- Orientation – Type a number to set the angle at which the Gene Symbols appear in Gene View and the Track Annotations appear in the tracks.
- Show Gene Symbols – Mark to show gene symbols in Gene View, and clear the check box to hide them.

Genomic Boundaries These options let you include or exclude the Genomic Boundaries from the analysis.

Tracks These options affect the appearance of the Track Views.

- Show Annotations – Mark to show the names of the gene regions for the tracks, and clear to hide them.
- Show Overlaid – Mark to overlay all the tracks that appear next to Gene View, and clear the check box to display the information in separate tracks.

Miscellaneous tab

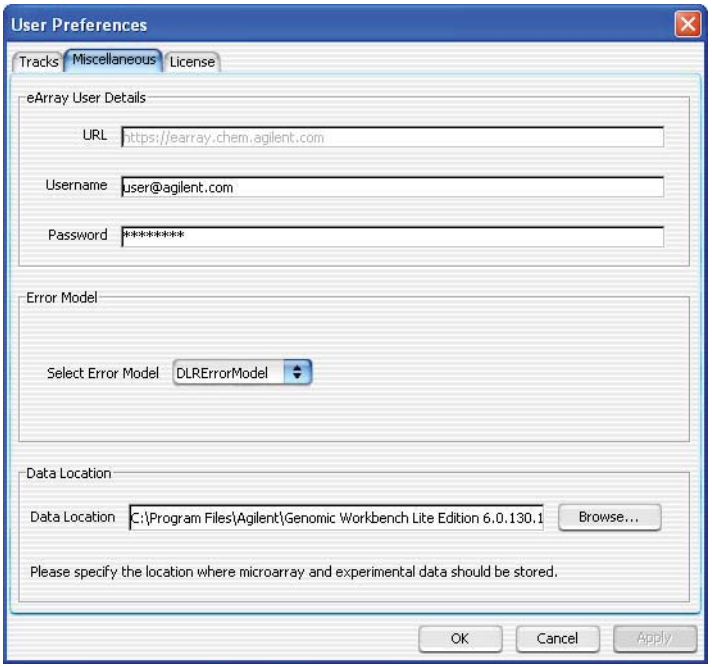


Figure 177 User Preferences dialog box – Miscellaneous tab

Purpose: For data/content set-up, this dialog box allows you to set up eArray access and to change the location for data.

To open: In the User Preferences dialog box, click the **Miscellaneous** tab.

eArray User Details Sets login details for the Agilent eArray Web site.

- **URL** – At present, https://earray.chem.agilent.com
- **Username** – The name registered on the eArray site.
- **Password** – The password registered on the eArray site.

Error Model The DLRErrorModel (Derivative Log Ratio) is the only selection. This measures noise in the data for CGH analyses.

Data Location The folder where the program stores array data and design files. To select a location, click **Browse.**

Apply Applies any changes to the preferences.

OK Accepts any changes and closes the dialog box.

Cancel Cancels all changes and closes the dialog box.

License tab

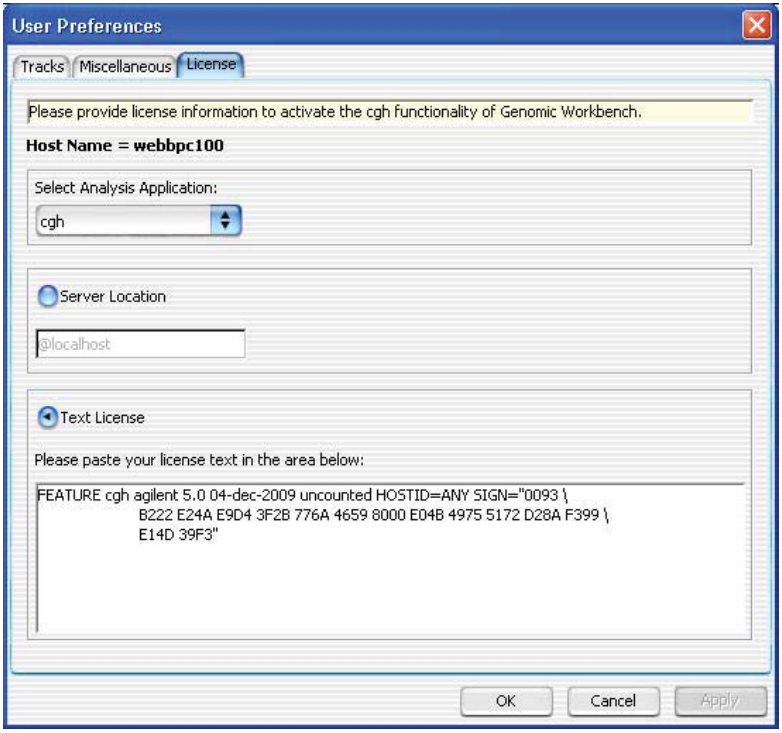


Figure 178 User Preferences dialog box – License tab

Purpose: The License tab allows you to display and update your CGH module license. This license enables the CGH application, and allows you to use it to analyze array data.

To open: In the User Preferences dialog box, click the **License** tab.

Host Name	Displays the host computer name automatically.
Select Analysis Application	Select the Agilent Genomic Workbench DNA Analytics module for which you have a license.
Server Location	Select this option if you have a concurrent user license, or want to use a server location for your user license(s). To edit this name, select **Server Location,** then type or copy the name of the computer used as your license server, or the fully qualified path name for the folder where your license text files are located. If you select this option, the Text License option is unavailable.
Text License	Select this option if you have a DNA Analytics module license (CGH, ChIP, CH3). To change the license, delete the old license text, and paste the new license text in the box.
OK	Accepts any changes you have made, and closes the dialog box.
Cancel	Closes the dialog box without changing any license information.
Apply	Accepts any changes you have made, but does not close the dialog box.

View coordinates in UCSC browser

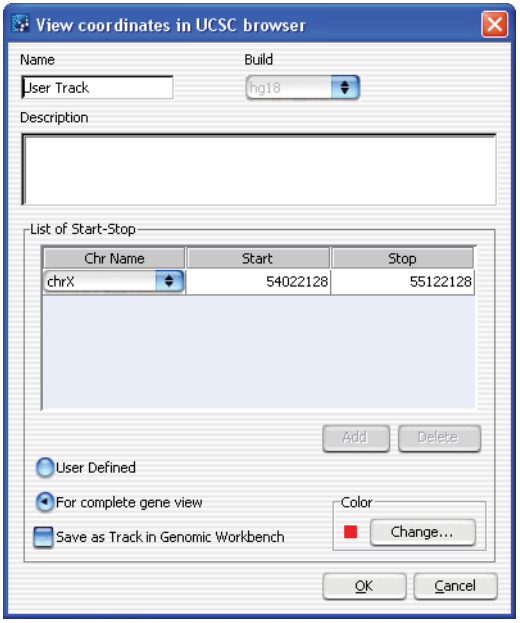


Figure 179 View coordinates in UCSC browser

Purpose: Defines a track to upload to the UCSC Web site so that you can see the information in the UCSC Genome Browser.

To open: Right-click in the Gene View, and select **Show in UCSC**.

Name Type a name for the track. This name identifies the track when it appears in lists and displays.

Build (Available if you select **User Defined** in **Set Chromosome Start-Stop.**) Select the genome build with which to associate the track.

Description Type descriptive text to attach to the track for reference.

Set Chromosome Start-Stop This parameter defines the region of the chromosome for which the track will be defined. Select one of these options:

- **User Defined** – Lets you define an arbitrary region of any chromosome. If you select this option, select the desired chromosome in **Chromosome,** then type the beginning **(Start)** and end **(Stop)** locations of the desired interval.
- **For complete gene view** – The chromosomal region that appears in Gene View.

Save as Track in Genomic Workbench Mark the check box to save this track in the Tracks folder in the My Entity List pane of the Navigator.

Change Click to open the Choose Track Color dialog box to select the color to use for display of the track in the Tracks folder. See "Select Color" on page 394.

OK Creates the track and opens the UCSC Web site, where you can display the track and associated information. For information on using the UCSC Web site, see the help and information provided there.

Cancel Closes the dialog box without creating a track.

View Preferences

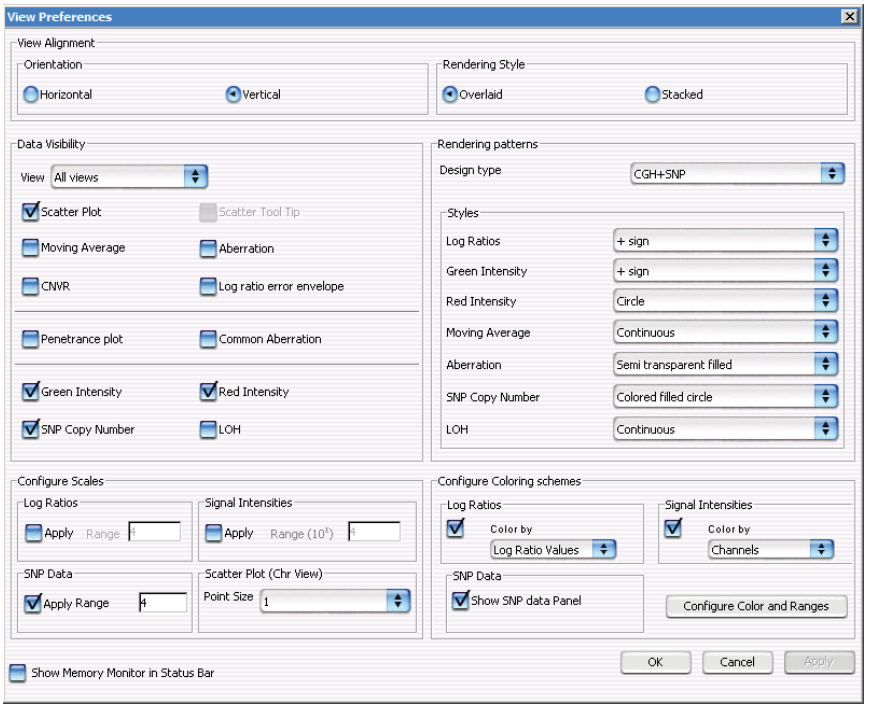


Figure 180 View Preferences dialog box for CGH

Purpose: This dialog box allows you to configure how data and results appear in Genome, Chromosome, and Gene views.

To open: In the View tab, click **View Preferences.** Or, right-click in any of the views and select **View Preferences**.

View Alignment Selects the orientation and rendering style (described below).

Option	Description
Orientation	
Horizontal	Stacks Genome, Chromosome, and Gene views horizontally in the main program window. Genomic locations appear across the bottom of each view.
Vertical	Displays Genome, Chromosome, and Gene views from left to right as side-by-side panes in the main program window.
Rendering Style	
Overlaid	In Chromosome View and in Gene View, displays data and results as a single, combined pane for all arrays. (Default)
Stacked	In Chromosome View and in Gene View, displays a separate pane for each array.

Data Visibility For each view, or all views, selects the kind(s) of data and results to display.

In **View,** select the view you want to configure. To set availability of display items for all views, select **All views.** Some display items are only available for certain views and modules. When you select a display item, it enables the item for display; for some items, you must take additional steps to display them. For example, you may need to configure a specific algorithm in the toolbar.

Mark any of the following options, as available:

Option	Description/Comments
Scatter Plot	The plot(s) of individual log ratio, intensity, or probe score data points.
Scatter Tool Tip	The ToolTips that appear when you place the pointer over specific data points on the scatter plot(s) in Gene View. The tool tip shows the array of origin and the numerical log ratio value for the data point.
Moving Average	The result of the Moving Average algorithm. See "To set up a moving average calculation to smooth the data" on page 136.
Aberration	The result of the selected aberration detection algorithm. See "To apply the aberration algorithm" on page 137.

Option	Description/Comments
CNVR	Detected copy number variant regions. See "To visualize a CNVR (copy number variant region)" on page 156.
Log ratio error envelope	The log ratio error envelope is a visual representation of the log ratio error calculated by Feature Extraction.
Penetrance plot	The probe penetrance plot for the active experiment. If you select this option, all other display items are unavailable. In addition, because the probe penetrance plot takes into account all arrays, this option overrides the stacked rendering style.
Common Aberration	The results of a common aberration analysis. To display this, you must first perform a common aberration analysis. See "To display common aberrations" on page 152. In addition, if you select this option, all other display items are unavailable.
Green Intensity	Mark the check box to display green raw signal intensity.
Red Intensity	Mark the check box to display red raw signal intensity.

Rendering Patterns

These options control the specific appearance of data and results in Genome, Chromosome, and Gene views. You configure these options separately for each type of application design.

- **Design Type** – Select the application design type for which you want to define rendering patterns.
- **Styles** – Select the display style for each of these elements:

Option	Description/Comments
Log Ratios	Select the symbol used for log ratio data points in the scatter plots in Chromosome and Gene views.
Green Intensity	Select the symbol to use for display of the green raw signal intensity.
Red Intensity	Select the symbol to use for display of the red raw signal intensity.
Moving Average	Select the line style for the moving average display. Lines appear in the display color defined for each array. See "To set up a moving average calculation to smooth the data" on page 136. • **Continuous** – A solid line. • **Dashed** – A dashed line. • **Dotted** – A dotted line. • **Do not show area** – No line.

Option	Description/Comments
Aberration	Select the rendering style for detected aberrations. • **Semi transparent filled** – Solid, colored regions (in the display colors defined for each array, if applicable). • **Hatched** – Cross-hatched colored lines (in the display colors defined for each array, if applicable). • **Do not show area** – Aberrations do not appear.
SNP Copy Number	Select the symbol to use for showing SNP Copy Number.
LOH	The only selection for showing regions of LOH is "continuous".

Scatter Plot (Chr View) Point Size

Select a point size to use for display of scatter plot data points in the Chromosome View.

NOTE

Rendering scatter plots for more than 10 high density arrays in the Chromosome View may take significant time. Selecting filled circles as the rendering style for CGH scatter plots can also decrease performance. For faster performance, change the rendering style for CGH data from the filled circle to the plus (+) or cross hair sign.

Configure Scales

For Log Ratios or Signal Intensities plots, mark **Apply** to enable the custom scale. In Range, type the value to use as the range for the scatter plot.

Configure Coloring schemes

Use these options to change the display of the scatter plot in the Gene View. These options are the same as those displayed in the Scatter Plot box in the Gene View.

To do this	Follow these steps
Show or hide the log ratio values in the Log Ratios plot	• To show the data points - Mark the **Log Ratios** check box and select **Log Ratio Values** from the list. • To hide all data points - Clear the **Log Ratios** check box.
Show or hide LogRatios color-coded by Probe Score Values in the Log Ratios plot	• To show the data points - Mark the **Log Ratios** check box and select **Probe Score Values** from the list. • To hide the data points - Clear the **Log Ratios** check box.

To do this	Follow these steps
Show or hide Intensity values in the Signal Intensities plot	• To show the data points - Mark the **Signal Intensities** check box and select **Intensity Values** from the list. • To hide all data points - Clear the **Signal Intensities** check box.
Show or hide Signal Intensities color-coded by Channels in the Signal Intensities plot	• To show the data points - Mark the **Signal Intensities** check box and select **Channels** from the list. • To hide the data points- Clear the **Signal Intensities** check box.
Show or hide Signal Intensities color-coded by Probe Score values in the Signal Intensities plot	• To show the data points - Mark the **Signal Intensities** check box and select **Probe Score Values** from the list. • To hide the data points- Clear the **Signal Intensities** check box.
Show or hide SNP data panel	• To show the SNP data panel - Mark the **Show SNP Data Panel** check box. • To hide the SNP data panel - Clear the **Show SNP Data Panel** check box.
Change the ranges and colors for scatter plot and signal intensities panels	• Click **Configure Color and Ranges** to enter ranges and change colors. See "Configure Coloring Ranges and Shades" on page 264 for more information.

Show Memory Monitor in Status Bar Displays a memory usage monitor in the eighth cell of the status bar. For information about the Status Bar, see "Status Bar" on page 238.

OK Applies the changes you made to all preferences and closes the dialog box.

Cancel Closes the dialog box without applying changes.

Apply Applies changes without closing the dialog box.

Reports

You can produce various standardized reports using the CGH module of Agilent Genomic Workbench that help you examine and analyze results. You can also create customized Cytogenetic Reports, where you select exactly what you want to include in the report. This section gives examples of the different standardized reports you can create.

Text Aberration Summary Reports

Each report shown below also includes header information that contains the parameters used for the data preprocessing and analysis. The following reports were generated using the same microarray and analysis parameters, with common intervals or regions highlighted for comparison, if applicable.

NOTE The *Stop* position listed in the report is the start of the last probe in that interval.

Probe Based

AberrationNo	CytoBand	ChrName	ProbeName	Start	Stop	Description	Genes		Logratio	Amplification	Deletion
1.188	p34.1	chr1	A_16_P30147946	45753224	45753275	Homo sapiens peroxi	PRDX1		1.1566998	1.0346224	
1.189	p33	chr1	A_16_P30155942	47678098	47678147	Homo sapiens forkhe	FOXD2		-1.7148536		-1.71485
1.19	p33	chr1	A_16_P30158058	51198005	51198049	Homo sapiens Fas (TN	FAF1		-1.890943		-1.89094
1.191	p32.3	chr1	A_16_P00064465	51583516	51583560	Unknown			-1.2975296		-1.29753
1.192	p32.3	chr1	A_16_P00068800	54861780	54861824	Homo sapiens acyl-C	ACOT11 ,FAM151A		1.1017393	1.1017393	
1.193	p32.3	chr1	A_16_P15126656	55469748	55469807	Unknown			-0.6178095		-0.67304
1.194	p32.3	chr1	A_16_P15126656	55469748	55469807	Unknown			-0.7367969		-0.67304
1.195	p32.3	chr1	A_16_P15126656	55469748	55469807	Unknown			-0.6244842		-0.67304
1.196	p32.3	chr1	A_16_P15126656	55469748	55469807	Unknown			-0.7957025		-0.67304
1.197	p32.3	chr1	A_16_P15126656	55469748	55469807	Unknown			-0.5904218		-0.67304
1.198	**p32.2**	**chr1**	**A_16_P00071447**	**56817487**	**56817534**	**Homo sapiens phospl**	**PPAP2B**		**-1.91642**		**-1.91642**
1.199	p31.1	chr1	A_16_P15174860	75698956	75699015	Homo sapiens solute	SLC44A5		-1.050432		-1.05043
1.2	p31.1	chr1	A_16_P00101977	78284071	78284130	Homo sapiens cDNA,	FLJ97370.		-3.43538		-3.43538
1.201	p31.1	chr1	A_16_P00102728	78879706	78879765	Homo sapiens interfe	IFI44L		1.202869	1.202869	
1.202	p31.1	chr1	A_16_P30203405	82194201	82194260	Homo sapiens latrophi	LPHN2		-1.1051049		-1.1051

Figure 181 CGH Aberration Report - Probe Based

Interval Based

AberrationNo	Chr	Cytoband	Start	Stop	#Probes	Amplification	Deletion	pval	Gene Names
36	chr1	p34.1	45045016	45045016	1	0	-2.29112	6.11E-44	TCTEX1D4
37	chr1	p34.1	45752746	45753224	2	1.034622	0	4.55E-19	PRDX1
38	chr1	p33	47678098	47678098	1	0	-1.71485	1.73E-25	FOXD2
39	chr1	p33	51198005	51198005	1	0	-1.89094	1.43E-30	FAF1
40	chr1	p32.3	51583516	51583516	1	0	-1.29753	2.19E-15	
41	chr1	p32.3	54861780	54861780	1	1.101739	0	1.38E-11	ACOT11, FAM151A
42	chr1	p32.3	55469748	55469748	5	0	-0.67304	4.58E-20	
43	**chr1**	**p32.2**	**56817487**	**56817487**	**1**	**0**	**-1.91642**	**2.40E-31**	**PPAP2B**
44	chr1	p31.1	75698956	75698956	1	0	-1.05043	1.09E-10	SLC44A5
45	chr1	p31.1	78284071	78284071	1	0	-3.43538	3.89E-96	
46	chr1	p31.1	78879706	78879706	1	1.202869	0	1.79E-13	IFI44L
47	chr1	p31.1	82194201	82194201	1	0	-1.10511	1.20E-11	LPHN2
48	chr1	p31.1	83664313	83712963	3	0.575142	0	8.77E-10	
49	chr1	p22.3	85235056	85235056	1	0	-0.99296	9.88E-10	MCOLN2

Figure 182 CGH Aberration Text Report - Interval Based with interval highlighted

Probe and Interval Based

AberrationNo	CytoBand	ChrName	ProbeName	Start	Stop	Description	Genes		Logratio	Amplification	Deletion
1.192	p32.3	chr1	A_16_P00068800	54861780	54861824	Homo sapiens a	ACOT11 ,	FAM151A	1.101739	1.1017393	
1.193	p32.3	chr1	A_16_P15126656	55469748	55469807	Unknown			-0.61781		-0.67304
1.194	p32.3	chr1	A_16_P15126656	55469748	55469807	Unknown			-0.7368		-0.67304
1.195	p32.3	chr1	A_16_P15126656	55469748	55469807	Unknown			-0.62448		-0.67304
1.196	p32.3	chr1	A_16_P15126656	55469748	55469807	Unknown			-0.7957		-0.67304
1.197	p32.3	chr1	A_16_P15126656	55469748	55469807	Unknown			-0.59042		-0.67304
1.198	**p32.2**	**chr1**	**A_16_P00071447**	**56817487**	**56817534**	**Homo sapiens p**	**PPAP2B**		**-1.91642**		**-1.91642**
1.199	p31.1	chr1	A_16_P15174860	75698956	75699015	Homo sapiens s	SLC44A5		-1.05043		-1.05043
1.2	p31.1	chr1	A_16_P00101977	78284071	78284130	Homo sapiens c	DNA, FLJ97370.		-3.43538		-3.43538
1.201	p31.1	chr1	A_16_P00102728	78879706	78879765	Homo sapiens i	IFI44L		1.202869	1.202869	
1.202	p31.1	chr1	A_16_P30203405	82194201	82194260	Homo sapiens l	LPHN2		-1.1051		-1.1051
1.203	p31.1	chr1	A_16_P15194313	83664313	83664372	Unknown			0.664638	0.5751422	
1.204	p31.1	chr1	A_16_P00109785	83695596	83695655	Unknown			0.403801	0.5751422	
1.205	p31.1	chr1	A_18_P10204499	83712963	83713022	Unknown			0.656988	0.5751422	
1.206	p22.3	chr1	A_16_P30206844	85235056	85235100	Homo sapiens r	MCOLN2		-0.99296		-0.99296
1.207	p22.3	chr1	A_16_P35209254	86173546	86173605	Homo sapiens c	COL24A1		1.004279	1.004279	

Figure 183 CGH Aberration Text Report - Probe and Interval Based with region highlighted

SNP Reports

SNP Genotype Report

Index	ArrayName	ProbeID	SNP ID	Chr	SNP Position	Genotype	p-Val	Log Ratio
4	US23502418_25	A_20_P00100012, A_20_P00201	rs3855951	chr1	1794161	TC	0.001024	-0.82333
5	US23502418_25	A_20_P00100018, A_20_P00201	rs2843160	chr1	2298941	TT	0.001024	-4.57129
6	US23502418_25	A_20_P00201926, A_20_P00100	rs1129333	chr1	2325536	GG	0.001024	0.027611
7	US23502418_25	A_20_P00201929	rs16825139	chr1	2416458	NN	0.001024	-0.21117
8	US23502418_25	A_20_P00201931	rs4648482	chr1	2739780	TTT	0.001024	3.259384
9	US23502418_25	A_20_P00201932, A_20_P00100	rs1563469	chr1	2776007	GA	0.001024	-0.72524
10	US23502418_25	A_20_P00201933	rs6668620	chr1	2784397	NN	0.001024	-0.67233
11	US23502418_25	A_20_P00201936, A_20_P00100	rs2842925	chr1	2876218	NN	0.001024	0.107138
12	US23502418_25	A_20_P00201938, A_20_P00100	rs12060482	chr1	2960792	NN	0.001024	0.092238
13	US23502418_25	A_20_P00201942, A_20_P00100	rs689565	chr1	3153814	GGG	0.001024	0.799873
14	US23502418_25	A_20_P00100039	rs13374875	chr1	3190196	NN	0.001024	0.276925
15	US23502418_25	A_20_P00201947, A_20_P00100	rs12409315	chr1	3219119	NN	0.001024	0.17788
16	US23502418_25	A_20_P00201948	rs7516150	chr1	3243749	CCC	0.001024	2.675796
17	US23502418_25	A_20_P00201949	rs882430	chr1	3248975	NN	0.001024	1.891996
18	US23502418_25	A_20_P00100046, A_20_P00201	rs12085231	chr1	3287500	CA	0.001024	-0.96495
19	US23502418_25	A_20_P00100047	rs2483225	chr1	3314896	NN	0.001024	0.134196
20	US23502418_25	A_20_P00201954, A_20_P00100	rs6670123	chr1	3316656	CCC	0.001024	2.4159

Figure 184 SNP Genotype Report

SNP Aberration & LOH Report

This report is similar to the Text Aberration Probe Based Summary Report, except that cytobands are not shown, and the Amplifications and Deletions column are combined into one column “Type” that indicates AMP or DEL. Log ratio values are listed in the AveCGHLR column for both aberration types.

Index	ArrayName	Class	Chr	Start	Stop	Type	#Probes	p-Val	AvgCGHLR	Gene Names
37	US23502418	CGH	1	45752746	45753224	AMP	2	4.55E-19	1.034622	PRDX1
38	US23502418	CGH	1	47678098	47678098	DEL	1	1.73E-25	-1.714854	FOXD2
39	US23502418	CGH	1	51198005	51198005	DEL	1	1.43E-30	-1.890943	FAF1
40	US23502418	CGH	1	51583516	51583516	DEL	1	2.19E-15	-1.29753	
41	US23502418	CGH	1	54861780	54861780	AMP	1	1.38E-11	1.101739	ACOT11, FAM151A
42	US23502418	CGH	1	55469748	55469748	DEL	5	4.58E-20	-0.673043	
43	**US23502418**	**CGH**	**1**	**56817487**	**56817487**	**DEL**	**1**	**2.40E-31**	**-1.91642**	**PPAP2B**
44	US23502418	CGH	1	75698956	75698956	DEL	1	1.09E-10	-1.050432	SLC44A5
45	US23502418	SNP	1	75999074	79472657	LOH	95	8.351616		ACADM, RABGGTB, SNORD45C,
46	US23502418	CGH	1	78284071	78284071	DEL	1	3.89E-96	-3.43538	
47	US23502418	CGH	1	78879706	78879706	AMP	1	1.79E-13	1.202869	IFI44L
48	US23502418	CGH	1	82194201	82194201	DEL	1	1.20E-11	-1.105105	LPHN2
49	US23502418	CGH	1	83664313	83712963	AMP	3	8.77E-10	0.575142	

Figure 185 SNP Aberration & LOH Report with interval highlighted

Configuration Settings

This section details the configuration settings for Agilent Genomic Workbench. You do not need to change any settings or make modifications to configuration files, except in cases where noted in the Guide.

CGH Properties File

The main configuration file for Agilent Genomic Workbench is the **CGHAnalytics.properties** file located within the installation folder in the folder named config. On Windows machines, the default installation path is:

C:\Program Files\Agilent\Agilent Genomic Workbench Lite Edition <VERSION>\config

The CGH properties file contains function settings. These function settings list the function name followed by the settings assigned to that function. For example:

```
CGHDataDirectory=data/
```

The file defines the following function settings, in this order:

Setting	Description
CGHDataDirectory	The folder into which all CGH data is stored. The folder must end in '/'. The default location is data/.
Version	Version number of the software. Used to set the name of the user settings file.
cacheSize	The maximum number of chromosomal data to cache. Default=500.
INITIAL_CAPACITY_OF_COLUMNS_IN_FE	The number of columns supported for import of Feature Extraction (FE) files. Default=175.
INITIAL_CAPACITY_OF_ROWS_IN_FE	The number of rows supported for import of FE files. Default=45000.
COLUMN_CAPACITY_INCREMENT	The increment for column capacity. Default=25.

Setting	Description
TEXT_SUMMARY_HEADERS	The column headers for text summaries. Default=ChrName, ProbeName, Start, Stop, Description
bundleBaseName	The message bundle. The bundle's directory hierarchy must be defined in the classpath. Default=com.agilent.cgh.ui.CGHMessageBundle
language	Internationalization property. Sets the languaged used in the application. Default=en.
country	Internationalization property. Sets the country where the application is used. Default=US.
ErrorCodesNotToSuppress	List of error codes. Listed errors will display error message, unless modified by exclude behavior, shown below.
exclude	Chooses whether or not to exclude errors listed in this file by the following conditions: • If exclude=true and code is listed, msg suppressed • If exclude=false and code is listed, msg is displayed • If exclude=true and code is not listed, msg is displayed • if exclude=false and code is not listed, msg is suppressed
FrequencyDistributionBin	The default number of bins for calculating frequency distributions. Default=10.
eArrayHost	The URL of Agilent eArray service. Default=https://earray.chem.agilent.com.
FilesInDataFolder	A list of the file types to be stored in the data folder defined above. Includes tracks, designs, gene lists, etc.
maxIterations	The maximum number of iterations used in heuristics such as centralization. Default=5.
MIN_MEAN_DIFF	The minimum difference in the mean of nested intervals. Applies to the step gram (ADM-1 and ADM-2) algorithms. Default=0.0.

Setting	Description
topFrac	The fraction of top input scores used in calculations. Default=0.1.
CommonAberrationAlgo	The list of available algorithms to use to identify common aberrations, such as the t-test and the Context Corrected Common Aberration ('CoCoA'). Default=TTestCommonAberrationAnalyzer,CoCoA.
CNVRAlgo	The list of available algorithms for CNV identification. In Agilent Genomic Workbench, there is one CNVR algorithm. Default=CNVRAnalyzer.
EnrichmentThreshold	The threshold value for identifying significant interval scores. Used in the Z-score algorithm. Default=4.
LICENSE_STRING	FlexLM license constants.
SaveResults	Indicates whether Agilent Genomic Workbench will prompt the user to save results when changing the application mode or exiting the application. Default=true.
hg, rn, mm, etc.	The species mapping for the "search probes in eArray" function. If there is no species available for some genome build, then by default *H. sapiens* will be used, that is, hg=H. sapiens.
helpFileName	The name of the User Guide PDF file.
QuickStartFileName	The name of the Product Overview PDF file.
TutorialFileName	The name of the Agilent Genomic Workbench Tutorial PDF file.
defaultEncoding	The default file encoding. Default=UTF8.
LOGGER_CONFIG_FILE_FOR_BACKGROUND_PROCESS	A separate log file for background processes. Default=config/processlog4j.lcf
PROCESS_FOLDER	The folder where the FE file importer function writes files. Default=process.
FE_CHUNK_SIZE	The buffer size allocated to the FE file importer function. Default=50000000.

Setting	Description
cacheSizeForSparseMap	The heapsize in sparse manager. Default=4.
memorySizeForSparseMap	The memory size allocated to sparse manager (the number of SparseWrapper objects to be serialized). Default=10.
CentralisationBinSize	The bin size used to speed up the centralization algorithm. Default=10.
scatterplotOnGenomeViewFlag	Sets whether the scatterplot is displayed by default in the genome view. Default=false.
ENRICHMENT_MAX_LENGTH_CEIL	Allows the user to control the maximum length of the interval for enrichment. Default=10.
filterPartial	Specifies whether full or partial filtering is applied in the generation of cytogenetic reports. Default=true.
imageHeight	Specifies the default image height in the report. Default=300.
imageWidth	Specifies the default image width in the report. Default=270.
showHeader	Specifies whether the header is displayed in the report. Default=false.
headerFontSize	Sets the font size of the report header, if the header is displayed. Default=12.
showBaseLine	Specifies whether the baseline is visible. Default=false.
showAberrationInArrayColor	Specifies whether aberrations are color-coded to coincide with the color assigned to the array. Default=false.
showAberrationType	Specifies which aberration type is displayed by the following codes: • -1: no aberration displayed • 0: aberration • 1: graphical aberration • 2: penetrance • 3: common aberration Default=0.

Setting	Description
showAberrationbar	Specifies whether the sidebar line that shows aberration length is displayed. Default=false.
KEEP_XML_REPORT	Specifies whether to keep the XML report. Default=false.
workflowStatusDir	Sets the folder to hold the status of the current workflow. Default=temp.
OnlyInternalIntervals	Specifies whether to create a track from intervals (aberration / CNVR intervals) that are within the user-defined start and stop locations. Default=false.
SHOW_IMAGE_IN_PROPERTY_DIALOG	Specifies whether to display the image in the property dialog. Default=false.
MAX_MULTIPLE_WORKFLOWS_32_BIT	Sets the maximum number of workflows that can run concurrently on 32-bit machines. Default=1.
MAX_MULTIPLE_WORKFLOWS_64_BIT	Sets the maximum number of workflows that can run concurrently on 64-bit machines. Default=3.
NON_UNIQUE_PREFIX_STRING	Nonunique probe settings. The prefix value to identify nonunique probes should not contain any special characters. Default=X.
NON_UNIQUE_APPLY_SETTINGS	Nonunique probe settings. Specifies whether the nonunique settings are applied. Default=true.
NON_UNIQUE_SET_PREFIX	Nonunique probe settings. Specifies whether the prefix should be set for nonunique probes. Default=true.
NON_UNIQUE_SET_COUNTER	Nonunique probe settings. Specifies whether the counter should be set for nonunique probes. Default=true.
NON_UNIQUE_STRIP_OFF	Nonunique probe settings. Default=false.
PROBE_NAME_COLOR	The color settings in RGB for the probe name. Default=126,178,244.
NON_UNIQUE_COLOR_TABLE_BACKGROUND	Nonunique probe settings. Sets whether the color table background is used. Default=true.

Setting	Description
FEATURE_NO_COLOR	The color for feature numbers in the table when intra-array normalization is applied. Default=CYAN.
FEATURE_NO_COLOR_TABLE_BACKGROUND	Sets whether the color table background for feature numbers is used. Default=true.
BLANK_CELL_COLOR	Sets the color for filtered-out or blank log ratio cells in Tab View. Default=GRAY.
AMPLIFIED_CELL_COLOR	Sets the color (in RGB) for amplified log ratio cells in Tab View. Default=255,100,128.
DELETED_CELL_COLOR	Sets the color (in RGB) for deleted log ratio cells in Tab View. Default=103,252,125.
USE_NEW_TABLE_CACHE	Sets whether to use control coding for amplified or deleted probes in Tab View. Default=true.
COLOR_AMP_DEL_PROBES	Sets whether to use color for amplified or deleted probes in Tab View. Default=true.
AMP_DEL_COLOR_TABLE_BACKGROUND	Specifies whether the color table background is applied. Default=true.
INTERVAL_MARGIN	The maximum margin (in bp) to be added to the ends of the interval. Default=200.
MAX_NO_OF_LINES_IN_REPORT	Sets the maximum number of rows in a report. The default setting is used to create reports that are compatible with earlier versions of Microsoft Excel, which can load 64k rows. Default=64000.
PROCRESS_CONFIG_FEIMPORT	Process configurations are different on different operating systems. Windows default=config/config_FEimport.properties.
PROCRESS_CONFIG_WORKFLOW	Process configurations are different on different operating systems. Windows default=config/config_workflow.properties.

6 Statistical Algorithms

Overview of CGH Algorithms 439
- Preprocessing algorithms 440
- Aberration detection algorithms 441
- Visualization algorithms 445
- Algorithms for multi-array analysis 445
- SNP analysis algorithms 447
- Definitions Used in the Statistical Algorithms Sections 450

Preprocessing Algorithms 452
- Centralization Algorithm 452
- GC Correction Algorithm 455
- Fuzzy Zero 458
- How the data centering algorithms can affect aberration calls 461
- Error Model and Combining Replicates 462

Aberration Detection Algorithms for CGH Analysis 465
- Z-Scoring for Aberrant Regions 466
- ADM-1 468
- ADM-2 471
- HMM 472
- CBS 483
- Noise Estimation – the Derivative Log Ratio Spread 491
- Interpretation and Visualization of the Detection Algorithms 493

Visualization Algorithms 502
- Triangular Smoothing 502

Algorithms for Multi-array Analysis 506
- Penetrance 506
- Common Aberration Analysis 508
- Differential Aberration Analysis 512
- Cluster Analysis 515

Copy Number Variation Region 518
Algorithms for Joint CGH-Expression Analysis 526
Enrichment of extreme expression values in a single sample 527
Regional correlation scores with multiple-matched CGH and Expression samples 528
Enrichment of CGH/Expression correlation 530
Enrichment of external gene annotation 531
Algorithms for CGH+SNP Analysis 533
ASCN (SNP CN) – Allele-specific copy number detection algorithm 534
Assignment of SNP genotypes 541
LOH (Loss or lack of heterozygosity) algorithm 544
Appendix 555
Enrichment Analysis - the Hypergeometric distribution 555
Cluster Analysis - Distance-based methods 557
Cluster Analysis - Similarity-based methods 559
Cluster Analysis - Linkage methods 561
Segmentation- Discrete Haar Wavelet Transform and Gaussian mixtures for HMM 564
References 566

This chapter describes the algorithms in the CGH application of Agilent Genomic Workbench. The program uses many different algorithms to perform the statistical analyses that are needed for comparative genomic hybridization (CGH), copy number variation (CNV), and single nucleotide polymorphism (SNP) studies. Additional algorithms let you find common aberrant regions, and also correlate CGH and expression data.

For an overview of the available algorithms, see "Overview of CGH Algorithms" on page 439. More detailed information on each algorithm follows later in the chapter. Algorithms are organized by general purpose. An appendix provides necessary background information for some of the more complex algorithms.

Overview of CGH Algorithms

Figure 186 summarizes the algorithms that are available in the CGH application of Agilent Genomic Workbench, and their relationship to each other.

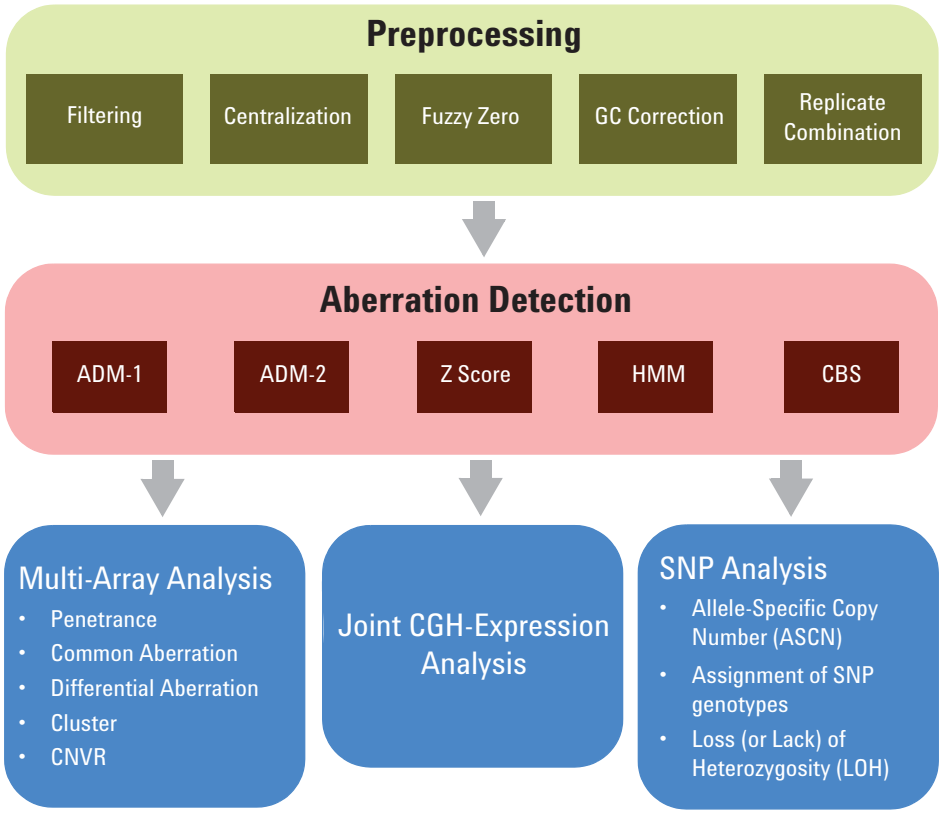


Figure 186 Overview of CGH Application algorithms. The relationship among the algorithms is shown. Certain preprocessing algorithms are options for only some of the types of downstream analyses.

The algorithms in the CGH application fall into several general categories. The list below describes these categories, and tells you where you can go in this section for brief descriptions of the algorithm(s) in each category. More detailed descriptions of the algorithms appear later in the chapter.

- **Preprocessing algorithms** prepare array data for downstream analyses. They include data correction and data centering algorithms, as well as algorithms that combine replicates and establish error models. For a brief description of these algorithms, see "Preprocessing Algorithms" on page 452.
- **Aberration detection algorithms** define the boundaries and magnitudes of regions of DNA loss or gain. For a brief description of these algorithms, see "Aberration detection algorithms" on page 441.
- **Visualization algorithms** smooth data so that you can view it with less complexity. For a brief description of these algorithms, see "Visualization algorithms" on page 445.
- **Algorithms for multi-array analysis** combine and evaluate aberrations across multiple samples. For a brief description of these algorithms, see "Algorithms for multi-array analysis" on page 445.
- **Algorithms for joint CGH-expression analysis** correlate the results from CGH and gene expression arrays. For a brief description of these algorithms, see "Algorithms for joint CGH-expression analysis" on page 446.
- **Algorithms for SNP analysis** analyze data from Agilent CGH+SNP microarrays, which combine CGH and SNP probes on the same array. For a brief description of these algorithms, see "SNP analysis algorithms" on page 447.

Preprocessing algorithms

Centralization

Many statistical algorithms for aberration detection assume that log ratio values are centered around zero if no aberration occurs. This is a reasonable assumption if there is no difference between the reference and sample channels. However, for samples with a high aberration percentage, this assumption can lead to erroneous results as the measured center of the data can deviate from a log ratio value of zero. To re-center the data, the centralization algorithm finds a constant value to add to or subtract

from all log ratio measurements. This ensures that the zero-point reflects the most-common-ploidy state. See "Centralization Algorithm" on page 452.

GC Correction

A frequent observation in aCGH profiles is a "wavy" technical artifact that correlates with the local GC content of genomic regions. The exact cause of this artifact has not been ascertained, but its presence can interfere with detection algorithms and lead to inaccurate aberration, SNP, and LOH calls. The GC correction algorithm corrects the log ratio value for each CGH and SNP probe on the array for GC-correlated "wavy" artifacts. The algorithm bases its corrections on values in the array design file that give the GC content of genomic segments that are centered on the locations of CGH probes. See "GC Correction Algorithm" on page 455.

Fuzzy Zero

ADM-1 and ADM-2 scores can identify extended aberrant segments with low absolute mean ratios, especially in data sets with a varying baseline. Often such aberrations represent noise, and are detected because of a high number of probes in the region. If long, low aberrations are detected in an analysis, you can apply the fuzzy zero algorithm to correct for the reliance on segment probe number. See "Fuzzy Zero" on page 458.

Replicate Combination

To combine replicate probes within an array (intra-array replicates) or among multiple arrays (inter-array replicates), the program calculates a weighted average of the log ratio values for the replicated probes. The weight given to each log ratio value is proportional to quality, which in turn is based on quality measures from the Agilent Feature Extraction (FE) program. Agilent Genomic Workbench uses an error model that is based on the log ratio error values in the FE files. See "Error Model and Combining Replicates" on page 462.

Aberration detection algorithms

The aberration detection algorithms in the CGH application form the basis of CGH and CNVR analysis. These algorithms detect DNA gain and loss events, show them in the Genomic Viewer, and make them available in reports. You can select from among many different aberration detection algorithms in Agilent Genomic Workbench.

You can refine the analysis of CGH and CNV data with several options that are available in the program. Some functions estimate the global or local noise in the data. Others centralize or filter the data before the

aberration regions are identified. Additional options let you filter out aberrations and do further analysis after aberrations are detected. The algorithms can be limited to known genomic regions and the output from the aberration detection algorithms can be correlated with gene expression data.

A threshold can be applied to each algorithm (except CBS and HMM). The optimal threshold depends upon several factors. These factors can include the DNA sample type, the DNA sample quality, or other measures, and can be explored interactively in the program.

NOTE A recommended threshold starting point for both the ADM-1 and ADM-2 algorithms is a value of 6.

The aberration detection algorithms cover a range of computational complexity and are best suited to different stages of analysis. For example, the Z-score algorithm is a straightforward test that can quickly identify regions with aberrant probes. The ADM-1 and ADM-2 algorithms are more sophisticated ways to identify aberrant regions, and the CBS algorithm can be used to perform segmentation of the whole genome.

Further, the output of the aberration detection algorithms can form the basis for further exploration. For example, the aberration calling algorithms ADM-1 and ADM-2 supply the input for the CNVR discovery and multi-array analysis algorithms.

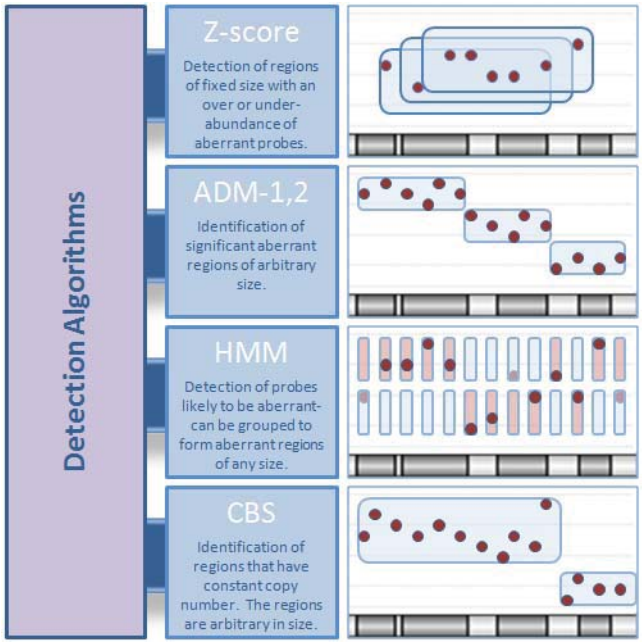


Figure 187 Overview of the aberration detection algorithms

Z-score The Z-score algorithm is a quick method of detecting aberrant regions. It finds genomic intervals with an over- or under-abundance of probes with log ratios that deviate significantly from baseline. To score intervals, the Z-score algorithm uses a sliding window of a fixed size that you set. Results from the Z-score identify regions of enriched probe log ratio change, and can suggest aberrant intervals. See "Z-Scoring for Aberrant Regions" on page 466. Because the Z-score algorithm only considers regions of fixed size in which there is enrichment of probe over- or under-abundance, it is not the preferred method for aberration identification. The ADM-1 or ADM-2 algorithms, described below, are recommended for such identification.

ADM-1 The Aberration Detection Method 1 (ADM-1) algorithm identifies all aberrant intervals in a given sample with consistently high or low log ratios based on a statistical score. The ADM algorithms automatically determine the optimal size of a statistically significant aberration. See "ADM-1" on page 468.

ADM-2 The Aberration Detection Method 2 (ADM-2) algorithm generates a similar statistical score to that produced by ADM-1 analysis, but ADM-2 incorporates quality information about each log ratio measurement. Use of the probe log ratio error in addition to the log ratio values makes ADM-2 more robust than ADM-1 when the data has noisy probes and you want to identify small aberrant regions. See "ADM-2" on page 471.

HMM The Hidden Markov Model (HMM) algorithm identifies all aberrant intervals in a given sample based upon the individual likelihood of such signals in a genomic context. A Hidden Markov Model partitions a large number of observations into a smaller number of (hidden) states. In the context of aCGH data, the observations are the log ratios of the probes and their positions ordered along a chromosome, and the hidden states are the aberration states of the data (*i.e.* amplifications, deletions, and no change). After it assigns states to the probes, the algorithm groups probes that have the same states to form intervals on the chromosome. In this way, the HMM algorithm segments the chromosome into genomic regions with different aberration states. The HMM algorithm differs from other detection algorithms in that it identifies local probabilities in the data. See "HMM" on page 472.

CBS Circular Binary Segmentation (CBS) is a method that identifies all genomic change points where the mean log ratio score changes between intervals. CBS therefore provides a reliable check on all intervals that should contain constant copy numbers. Since the output of the CBS algorithm is comprehensive and unranked, other methods, especially ADM-1 and ADM-2, are better suited for aberration classification. In addition, the CBS algorithm is processor intensive and can take a long time to run. See "CBS" on page 483.

Derivative Log Ratio Spread (*dLRsd*) To make aberration calls, a detection algorithm needs a measure of probe-to-probe noise. A measure of the minimum log ratio difference is needed to make reliable amplification or deletion calls. The *dLRsd* algorithm is a robust method that estimates noise from the sample array alone. It calculates the spread of the log ratio differences between consecutive probes along all chromosomes. See "Noise Estimation – the Derivative Log Ratio Spread" on page 491.

Visualization algorithms

To help you visualize large datasets, Agilent Genomic Workbench lets you use a sliding window across the genome to smooth the data points for detailed examination of probe log ratios. These functions are available independent of the main detection algorithms. A change in window size in the visualization algorithms does not affect aberration calls.

Moving Average The moving average algorithm sets a fixed window size around every point of interest. The average of that point and neighboring points within the window boundaries are reported as the value for that point. The moving average algorithm is available in two modes: Linear and Triangular Smoothing.

Triangular Smoothing The triangular smoothing algorithm is a shaped smoothing algorithm based on a moving average. As the moving average centers on a point of interest, a maximum weight is applied and that weight falls off with increasing distance. The point of interest is then adjusted as the weighted mean of itself and neighboring points. See "Visualization Algorithms" on page 502.

Algorithms for multi-array analysis

The CGH Application has algorithms that can combine the results of the aberration algorithms in biologically and statistically meaningful ways. These algorithms identify and summarize commonly aberrant regions within multiple samples, which can supply a robust report of aberration coverage. The Copy Number Variant Region (CNVR) Report in the CGH Application uses the CNVR Algorithm to combine aberrant regions that overlap from multiple samples into putative CNVRs.

Penetrance The probe-based and interval-based penetrance algorithms give a count score for probes and regions, respectively, which show aberration across multiple samples. The count is the percentage of samples which show an aberration in the same direction (over- or under-abundance). The penetrance algorithms use as input the aberrant regions found using any of the detection algorithms. See "Penetrance" on page 506.

Common/ Differential Aberration The algorithms to find genomic regions sharing common aberrations or having significant difference in aberrations across multiple samples use as input the list of per-sample aberrant regions as determined by either the ADM-1 or ADM-2 algorithm. The aberrant regions considered in both the

common aberration algorithm and the differential aberration algorithm are therefore subject to the user-defined ADM threshold. For the common aberration analysis, the probability that any set of regions in different samples are in common is then calculated. Conversely, the probability that a given genomic interval has far more (or less) aberrations between two groups of samples is calculated in the differential aberration analysis. See "Common Aberration Analysis" on page 508 and "Differential Aberration Analysis" on page 512 for more information.

Cluster Analysis The clustering algorithm is a data mining method for partitioning multiple samples into groups based upon their aberration profiles. Arrays are grouped into clusters such that the arrays in one cluster have aberrations that are more similar to each other than to arrays belonging to a different cluster. Such groupings can aid in discerning significant experimental parameters that may account for variations between the aberration profiles in the two groups. The clustering algorithm uses the list of aberrant regions across samples as determined by either any of the aberration detection methods such as the ADM-1 and ADM-2 algorithms. Because of this, the aberrant regions considered in the clustering algorithm are subject to the user-defined ADM threshold. The clustering algorithm uses a hierarchical method and has options for different measures of cluster similarity and linkage. See "Common Aberration Analysis" on page 508.

CNVR The copy number variation region (CNVR) algorithm uses output from an aberration detection algorithm for multiple samples. It generates a list of potential intervals that cover the same or similar genomic regions, and joins the intervals together into a putative CNV region. See "Copy Number Variation Region" on page 518.

Algorithms for joint CGH-expression analysis

Agilent Genomic Workbench provides tools for joint analysis between aberrant region detection and output from expression microarray experiments. These analysis tools allow for correlation between CGH and expression arrays and enrichment analysis based upon expression levels from genes found in an aberrant or CNV region. See "Algorithms for Joint CGH-Expression Analysis" on page 526.

Enrichment of extreme expression values in a single sample

CGH-Expression Analysis can be computed for any sample in which both CGH and expression are measured. Genomic Workbench computes an enrichment of the number of probes (genes) with high log ratios in a given region. The enrichment is a measure of the increase in the number of probes (genes) with a high log ratio as compared to what may be expected at random. The output is a list genomic intervals for each sample with a statistical score showing the enrichment of over- and under-expressed probes in that interval. This analysis can be performed either on all aberrant intervals or on the whole genome. See "Enrichment of extreme expression values in a single sample" on page 527

Regional correlation scores for multiple matched CGH and EXP samples

These scores are computed for all probes for which expression was measured and there is a well-defined known genomic location. For this analysis the designs for the CGH and the expression arrays need not be the same, but the sample set must contain at least six pairs of matching CGH and EXP arrays. See "Regional correlation scores with multiple-matched CGH and Expression samples" on page 528.

Enrichment of CGH/Expression correlation

This can be computed for a set of samples for which both CGH and expression were measured. The output consists of a list of genomic intervals and a statistical score of the enrichment of CGH/expression correlation in each interval. See "Enrichment of CGH/Expression correlation" on page 530.

Enrichment of external gene annotation

You can compute this annotation for any external quantitative information about genes. Samples do not have to be matched. The output consists of a list of genomic intervals and, for each, a statistical score of the enrichment of the external quantitative annotation in that interval. See "Enrichment of external gene annotation" on page 531.

SNP analysis algorithms

Agilent Genomic Workbench supports CGH+SNP arrays, which combine CGH probes and SNP probes on the same array. The inclusion of SNP probes allows the determination of allele-specific copy number for specific SNP sites (ASCN), the assignment of genotypes for specific SNP sites, and the detection of regions of loss or lack of heterozygosity (LOH).

The Agilent CGH+SNP platform uses restriction enzyme cleavage to differentiate between alleles at a given SNP site. At the SNP sites that can be detected by this method, one allele is cut by the enzymes, while the

other is not. The method also uses a known genotyped reference. Since the reference genotype is known, the raw $\log_2$ ratios are "reference adjusted" to values that reflect a hypothetical ASCN of 2 for the reference at all SNP sites.

Allele-specific Copy Number (ASCN)

For each SNP site targeted on the array, the ASCN algorithm calculates an expectation value for the copy number of the uncut SNP allele. In Agilent Genomic Workbench, this algorithm is known as the **SNP CN** (SNP Copy Number) algorithm.

A normal, diploid region of the genome typically contains a distribution of SNP genotypes with SNP copy numbers of 0, 1 and 2 uncut alleles. Three distinct ASCN states will appear in the SNP CN panel. However, aberrations can affect this distribution:

- In a diploid genomic region that comprises a copy-neutral LOH or UPD aberration, the SNP probes only report alleles that are homozygously cut and uncut (0 and 2 uncut copies). Only two states appear.
- For a region of the genome that is affected by a hemizygous deletion, two states also appear. These states represent SNPs that have only one copy of an allele, either cut or uncut (0 or 1 copy of the uncut allele).
- The amplification of a region or an entire chromosome adds one or more states. In this case, the SNP probes report four or more states that correspond to 0, 1, 2, 3 (or more) copies of the uncut allele.

Table 7 summarizes the relationship of genotype to SNP status.

Table 7 Relationship of genotype to SNP status (number of uncut alleles)

Genomic Status	Genotype	Number of uncut alleles
Normal diploid genome	AA, AB, BB	0, 1, 2
Diploid Genome with copy-neutral LOH or UPD	AA, BB	0, 2
Hemizygous deletion	A, B	0, 1
Amplification (*e.g.* trisomy)	AAA, AAB, ABB, BBB	0, 1, 2, 3

To calculate ASCN expectation values, the algorithm first adjusts the measured log ratios to account for the (known) ASCN of the reference sample. It finds the peaks in the probability density distribution of log

ratio values, and assigns the most likely copy number to each peak. It then fits a separate Gaussian distribution to each peak. This creates a model that gives the likelihood that a given log ratio value corresponds to each possible copy number. See Figure 217 on page 537. The expectation value that is reported for the ASCN of each SNP is the most likely copy number, given the observed log ratios. See "ASCN (SNP CN) – Allele-specific copy number detection algorithm" on page 534.

Assignment of SNP genotypes

The program reports the genotypes of the targeted SNP sites. To do this, it uses SNP allele information from the known genotyped reference, and expectation values for CGH copy number and the allele-specific SNP copy number (ASCN). See "Assignment of SNP genotypes" on page 541.

Loss or Lack of Heterozygosity (LOH)

The LOH algorithm identifies genomic regions that report a statistically significant scarcity of heterozygous SNP calls. Identification of such regions can inform constitutional cytogenetic studies. The algorithm reports LOH for the regions that are also detected as deletions in standard CGH analysis. However, it can also detect LOH in amplified and copy-neutral regions.

The algorithm uses total and allele-specific copy numbers to label each SNP site as homozygous or heterozygous. It then uses a binomial probability distribution to report regions that contain an unusually high fraction of homozygous SNPs. See "LOH (Loss or lack of heterozygosity) algorithm" on page 544.

Definitions Used in the Statistical Algorithms Sections

The following abbreviations and mathematical symbols are utilized in this chapter, listed in order of appearance:

Abbreviation or symbol	Definition
L	The log ratio of signal obtained from comparing each channel for each probe. The log ratio is obtained from the Feature Extraction file by converting the ratio to log base 2.
LE	The error associated with the log ratio value of a feature. The Log Ratio error (LE) is obtained from Feature Extraction files after converting it to log base 2.
Z	The *Z*-normalized value of the log ratio.
w	The moving average window. This is a user-defined window width. The average (or weighted average) is calculated from every probe within that region. The sliding window is then advanced across the genome and a new average is calculated for the new range covered by that window.
h	The height of a measured aberrant region. *h* is the unsigned magnitude, or the absolute value of the average log ratio for the aberrant region.
I	A genomic interval used for calculation of the magnitude of a aberrant region.
S(I)	The aberration score for interval *I*.
t	A user defined threshold for *S(I)*. Intervals with scores greater than *t* are marked significant and retained as aberrant regions.
q	The weight of a probe in further calculations, equal to the inverse of *LE* squared.
G	An entire genome, chromosome, or defined genomic boundary.
e	The probe calculated error, defined as either *LE* or *dLRsd*, whichever is greater.
$x_i \in \{s_1 s_2 s_3 ..\}$	Hidden Markov Model states by probe genomic position *i*.

Abbreviation or symbol	Definition
$p(s_{i+1} \mid s_i)$	Hidden Markov Model state transition probabilities from position *i* to position *i+1*. The transition probabilities form a matrix representing how the states change as the position is advanced along the chromosome.
$O_i \in \{O_1 O_2 O_3 ..\}$	Hidden Markov Model observations by probe genomic position *i*.
$p(O_i \mid s_i)$	Hidden Markov emission probabilities. The likelihood of a probe log ratio value given a particular state.
$\mid\mid A \mid\mid$	For a vector *A*, this is the magnitude of the vector.
$A.B$	For two vectors *A* and *B*, this is the product of the vectors: $A \cdot B = \sum_{i=1}^{n} A_i B_i$

Preprocessing Algorithms

Centralization Algorithm

Given a data vector for a single sample or entire genome, this algorithm attempts to find the best way to center the data by adding or subtracting the same constant to or from all log ratio measurements. Doing so will make the most-common-ploidy the new zero-point.

Purpose Many algorithmic approaches to aberration detection on aCGH data assume that the data points are distributed around some zero value if no aberration occurs. Typically, aCGH data fluorescence ratios for each array are normalized by setting the average log fluorescence ratio for all array elements to zero. This may lead to erroneous aberration calls for highly-aberrant genomes such as those found in tumor samples.

Use The centralization algorithm is a normalization algorithm which is used as a parameter for detecting aberrant regions or regions of constant copy number using ADM-1, ADM-2, or CBS.

Algorithm Define a score S for a possible centralization value c where $S(c)$ equals the number of probes that are not included in all aberrations as called by aberration finding routines applied to the original log-ratios, shifted by c.

Try to find the value of c that minimizes score $S(c)$. That is, a value that minimizes the number of probes that are called aberrant. The search for the optimal value of c can be time consuming because you must run ADM-1 on each possible value.

In order to speed up the computation, without affecting the performance, contiguous probes are binned across the genome. In the user interface, you can choose a bin size for this algorithm (the default bin size is 10). In the default case each ten contiguous probes are averaged to reduce the number of probes used in the centralization procedure.

In the following example for the given array, the log ratio values are plotted in Figure 188. The plot of score $S(c)$ for different values of c is generated, and the plot is shown in Figure 189. The centralization algorithm defines the new zero where the center of the highest peak lies in Figure 189.

Interpretation The centralization algorithm affects the output from the aberration algorithms. See "ADM-1" on page 468.

Visualization See "ADM-1" on page 468 for more information about displaying the ADM algorithms.

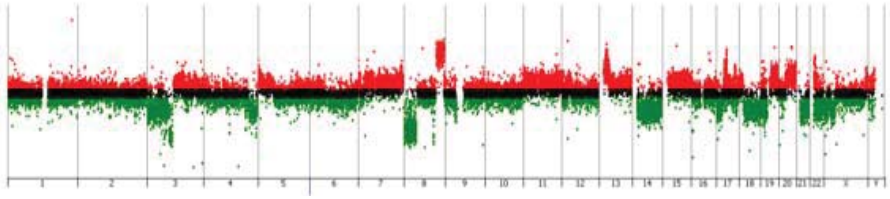


Figure 188 Log ratio values of an HT29 cell line in Genome view

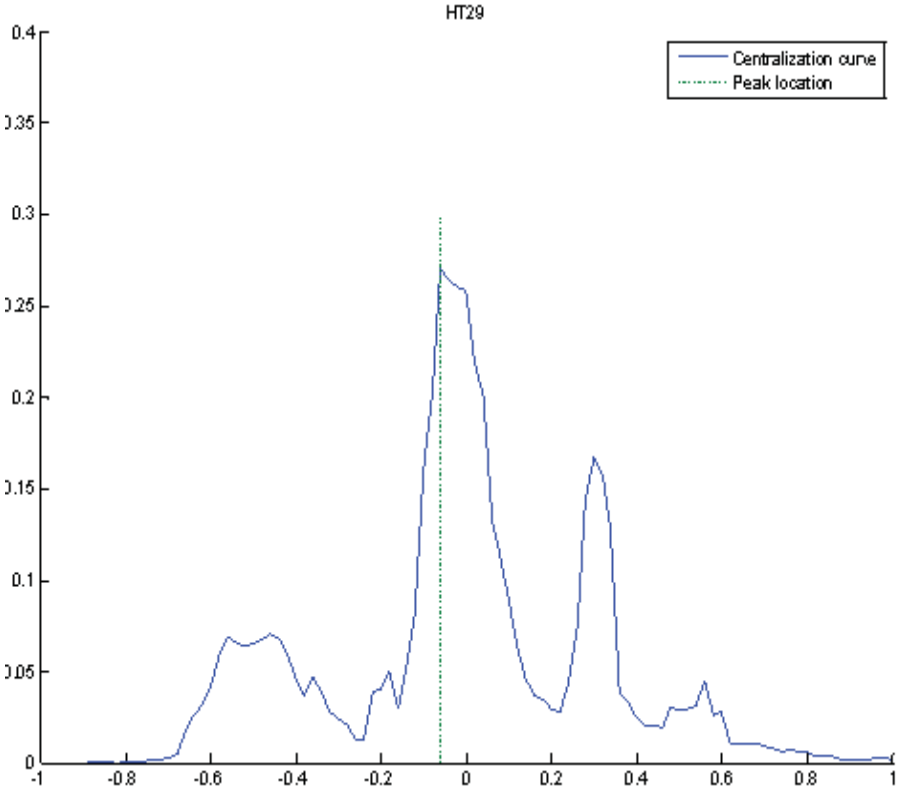


Figure 189 The plot of the score S(c) and the location of the peak for this centralization curve. For this example the log ratios will be shifted after centralization by 0.06, the absolute value of the x-coordinate of the peak.

GC Correction Algorithm

The GC Correction algorithm corrects aCGH log ratio data for the presence of “wavy” artifacts. This specific type of technical artifact correlates with the local GC content of genomic regions. The exact cause of these artifacts has not been ascertained, but their presence can interfere with detection algorithms and lead to inaccurate aberration, copy number, and LOH calls.[1]

The algorithm only corrects for GC-correlated “wavy” artifacts. The effects of the algorithm are often negligible, but it can help considerably in many cases when the baseline shows these artifacts.

Agilent recommends that you apply GC Correction when you perform CGH+SNP analysis. For the algorithm to work properly, the array design file must contain GC content values for the genomic regions that flank each CGH probe. The design files for Agilent Catalog arrays, for example, currently contain GC content values for window sizes of 2 kb, 10 kb, and 40 kb surrounding each CGH probe. In Agilent Genomic Workbench, a window size of 2 kb is selected by default. If GC content data for the selected window size is not present in the design file, the program prompts you to select another value. If no GC content data is present in the design file, the algorithm cannot run.

Algorithm

The GC correction algorithm first removes significant aberrations from the log ratio data for the CGH probes on the array. The residual log ratio values are then correlated with local GC content, and the result is used to correct the log ratio values for all CGH and SNP probes on the array.

Steps

Step 1
Find correlation with GC content

In this step, the algorithm first calls aberrations in the log ratio data from the CGH probes on the array. It removes the aberrations from the data, which leaves a set of residual log ratio data.

It then assumes a linear relationship between the residual log ratio data and local GC content. The array design file supplies values for local GC content. GC content values are typically available for several window sizes around each CGH probe, and you select the desired window size when you set up GC correction. Although the algorithm is not particularly sensitive to window size, Agilent recommends that you select the smallest window size of 2 kb.

The algorithm computes a robust linear regression of log ratios on GC content. See Figure 190.

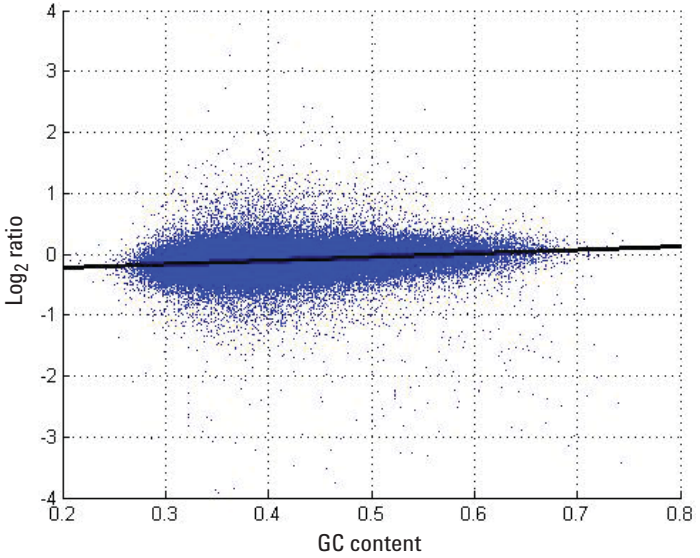


Figure 190 Linear relationship between local GC content and log ratio signal intensity from residual artifacts. This relation supplies correction factors for all probes on the array. For each probe, the GC content value reflects a region ±1 kb around the location of the probe.

Step 2 Subtract correction factors

From the log ratio value of each CGH and SNP probe on the array, the algorithm subtracts the linear trend determined by the regression. If GC Correction is selected, the corrected log ratio values are used in all downstream analysis and visualization routines for the array.

Visualization

The CGH module of Agilent Genomic Workbench lets you view a plot of the probability density for the probes on an array both before and after GC Correction is applied. See "To apply GC Correction" on page 129. Figure 191 shows an example of this output. When you apply GC Correction, check the final distributions to assure that the correction is reasonable.

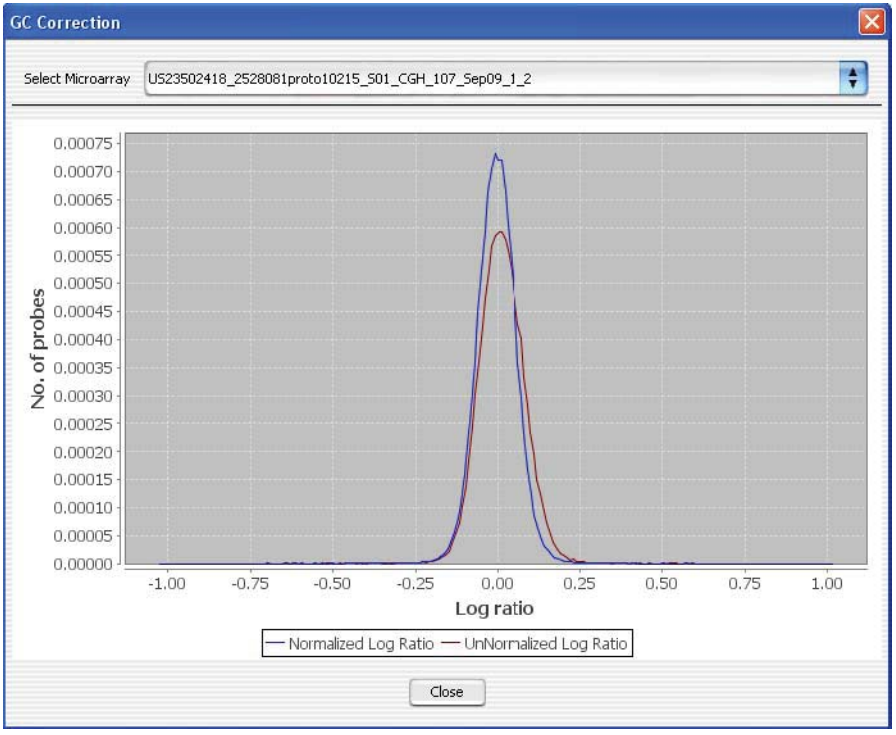


Figure 191 Distribution of $\log_2$ ratio values for an array before (red) and after (blue) the application of GC Correction. To view these distributions, in the Preprocessing tab, under Normalization, click **Plot Distribution.**

Fuzzy Zero

The Fuzzy Zero algorithm is an improved error model that explicitly includes the effects of long-range log ratio correlations. In this model, we assume that there are two independent sources of noise contributing to the total log ratio variation: the local uncorrelated probe-to-probe noise:

σ^1_{Local}

and a global noise describing baseline variation:

σ_{Global}

Purpose The ADM-1 and ADM-2 algorithms estimate the standard deviation of the mean log ratio of an interval using a statistical error model that treats probe to probe errors as independent. In many samples, the assumption that the log ratio errors of successive probes are independent is not in fact valid. The errors of the probes are often correlated over wide genomic intervals, and the ADM algorithms therefore underestimate the error for long intervals. Long aberrations with low average log ratios are thus often incorrectly deemed significant.

Use Fuzzy zero correction applies a "Global error model" to all aberrant intervals identified in ADM-1or ADM-2 analysis. The global error model uses a more realistic error model to avoid erroneous aberration calls when the errors are correlated.

Algorithm For the global error model, we assume that there are two independent sources of noise contributing to the total noise of the intervals. A local probe-to-probe noise, σ^1_{Local}, which is not correlated among different probes along the interval as described above, and a global noise, σ_{Global}, which is correlated among probes in an interval. The global noise component, σ_{Global}, is calculated as the variation of the average log ratios in large genomic intervals. As local probe-to-probe noise, σ^1_{Local}, is not correlated between different probes, when k probes are averaged, we assume that the local noise is reduced by a factor of $1/\sqrt{k}$. Thus,

$$\sigma^k_{Local} = \sigma^1_{Local} / \sqrt{k} \quad \textbf{(1)}$$

The score of interval I under the global error model, $S_g(I)$, is

$$S_g(I) = \frac{h}{\sqrt{(\sigma^k_{Local})^2 + \sigma^2_{Global}}} \quad (2)$$

here h is the average log ratio of all probes in the interval I. If the ADM-2 algorithm is used, h is the quality weighted average log ratio of all probes in the interval I.

Using α to denote $\left(\frac{\sigma_{Global}}{\sigma^1_{Local}}\right)^2$ we derive

$$\sigma^k = \sigma^1_{Local}\sqrt{\frac{1}{k} + \alpha} \quad (3)$$

and

$$S_g(I) = \frac{h}{\sigma^1_{Local}\sqrt{\frac{1}{k} + \alpha}} \quad (4)$$

Fitting the model parameters

For a given log ratio vector v of length N (for a particular sample), σ^1_{Local} equals the Derivative Log Ratio Spread (*dLRsd*):

$\sigma^1_{Local} = dLRsd(v_0)$, where v_0 is the vector of the individual probe log ratios L.

See "Noise Estimation – the Derivative Log Ratio Spread" on page 491 for a description of *dLRsd*.

Then α is estimated using the following iterative procedure:

1 Start with an initial estimate of $\alpha_0 = 0.01$

Start with an initial value of the vector v as v_0.

2 At each iteration i, new values of α and v are calculated as α_i and v_i:

a Find all aberrant intervals I in v_i with the score $S_g(I)$ (**4**) above the user defined threshold, T (the default recommended starting threshold is 6).

Note that the score $S_g(I)$ (**4**) depends on the current value of α_i. This set of aberrant intervals is considered as the signal component of the data.

b Compute the residual vector v^r.

To compute v^r, we subtract from v_i the heights of each aberrant interval I. Namely, we subtract from each probe in I the height h of the aberration containing the probe. The resulting vector v^r represents the current estimate of the noise in the data.

c Estimate the combined noise σ_i^k from the residual vector v^r.

To estimate the combined noise σ_i^k, we bin consecutive probes into bins of size $k = \sqrt{N}$. Then we derive a binned vector u_k, where each element of u_k is the average log ratio of all probes in one bin. We estimate σ_i^k by computing $dLRsd(u_k)$. To make this estimation more robust we repeat the binning using 10 different starting positions of the first bin. The final estimation of σ_i^k is the median of these 10 different estimations.

d Compute the new α_{i+1} based on the current estimate of σ_i^k.

e Set $i+1 = v^r$.

3 Continue the iterations until the process converges, *i.e.* $|\alpha_i - \alpha_{i-1}| < 0.001$, or 10 iterations were made.

How the data centering algorithms can affect aberration calls

The data centering algorithms are used in the aberration detection algorithms. Output from application of the detection algorithms is visible in Genome, Chromosome, and Gene views. Figure 192 and Figure 193 show how the application of the data centering algorithms can affect aberration detection.

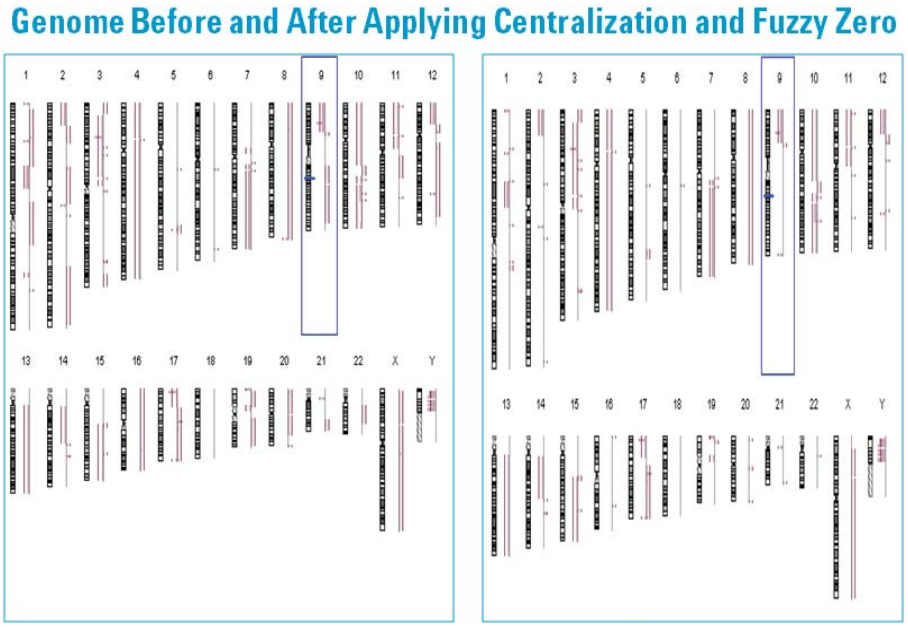


Figure 192 Aberration results before (left panel) and after (right panel) the application of data centering algorithms. Before application of these algorithms, aberrant regions (dark red lines) are present in regions without a clear over- or under-abundance of log ratios (light red blocks). In the right panel, most spurious aberrant regions have been removed, while the robust aberrant regions remain. A detailed view of chromosome 9 appears in Figure 193.

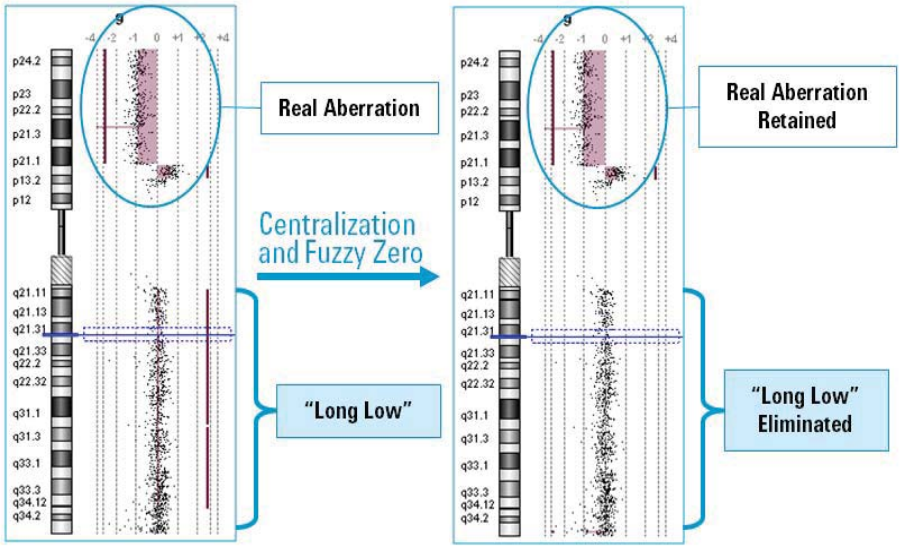


Figure 193 Application of the data centering (centralization and fuzzy zero) algorithms, illustrated. Re-centering the data allows identification of robust aberrant regions while removing spurious aberrations that are reported when the log ratios of individual probes are not completely independent.

Error Model and Combining Replicates

Error modelling is a way to separate events measured from biological signals such as protein-DNA binding from signals measured from systematic variations in the technology. The parameters for error calculation use probe log ratio and quality measures from Agilent Feature Extraction (FE).

Error Model

Purpose The purpose of the error model is to identify which probes are most reliable in the calculation of aberrant regions. The confidence of such events is calculated based on the log ratio error of each probe. For more information regarding the FE log ratio error calculation, see the *Agilent Feature Extraction User Guide*.

Use The error model works by selecting robust probe signals for inclusion in the detection algorithms and is therefore on by default. The error model is especially important whenever there are replicate probes within an array or between samples.

Algorithm To combine log ratios of replicated features, a weight is applied first. The weight is proportional to quality, and quality is defined as the inverse of square of log ratio error.

To combine replicated probes in an array (intra array) or within replicated arrays (inter array), combine the log ratio and the log ratio error as follows:

1 Define a weight q_i for *each* probe to be $q_i = 1/\ e_i^2$ That is, the noisier a given probe is, the smaller is its weight. The error, e_i, is defined as the maximum between the log ratio error, LE_i, of that probe and the spread of derivative of log ratio, *dLRsd*, for that array, i.e. $e_i = max(LE_i, dLRsd)$. The *dLRsd* is described under "Noise Estimation – the Derivative Log Ratio Spread" on page 491.

2 Define the quality-weighted average log ratio for replicated probes as:

$$L_{ave} = \frac{\sum_{i \in I} q_i L_i}{\sum q_i} \quad (5)$$

3 Estimate the Log Ratio Error of the above mean:

$$LE_{ave} = \frac{1}{\sqrt{\sum_{i \in I} \frac{1}{e_i^2}}} = \sqrt{\frac{1}{\sum q_i}} \quad (6)$$

Combining Dye Swap

Algorithm When combining dye-swapped arrays, Genomic Workbench separately combines any replicates using the original dye channels and the dye-swapped chips according to the above equations. This yields both the average weighted sample and the combined error measurement for each polarity.

The error model for the dye-swapped arrays is the sum of the Log Ratio Error of the two polarities in quadrature:

$$\sigma_{combined}(dyeSwap) = sqrt((\sigma^2_{polarity1} + \sigma^2_{polarity2})/\ 2) \quad \textbf{(7)}$$

NOTE

The current release of Agilent Genomic Workbench lets you combine multiple arrays into one larger, "virtual" array. However, the program does not support combined virtual arrays (nested combinations of arrays, or arrays that were combined and then combined again).

For example, suppose you use a particular array attribute to combine several arrays into a virtual array, and you then use another array attribute to combine arrays again. The program decouples the first virtual array, and uses the second attribute chosen to recombine the arrays.

Aberration Detection Algorithms for CGH Analysis

This section describes in detail the purpose, comparative use, and step-by-step methods used by the aberration detection algorithms for CGH analysis in Agilent Genomic Workbench. These algorithms interpret log ratios and probe quality measures from the microarray to identify contiguous genomic regions that correspond to chromosomal aberrations or copy number variations. At the end of this section is a guide to the interpretation and visualization of the detection algorithm output. Table 8 lists the topics available in this section.

Table 8 Detection algorithms topics

Subject	See these topics
Z-score algorithm	"Z-Scoring for Aberrant Regions" on page 466
Aberration Detection Method (ADM) algorithms	"ADM-1" on page 468 "ADM-2" on page 471
Hidden Markov Model (HMM) algorithm	"HMM" on page 472
Circular Binary Segmentation (CBS) algorithm	"CBS" on page 483
Noise estimation	"Noise Estimation – the Derivative Log Ratio Spread" on page 491
Interpretation and Visualization	"Interpretation and Visualization of the Detection Algorithms" on page 493 "How the data centering algorithms can affect aberration calls" on page 461

NOTE For descriptions of the algorithms for the analysis of CGH+SNP arrays, see "Algorithms for CGH+SNP Analysis" on page 533.

Z-Scoring for Aberrant Regions

This method identifies all aberrant regions in a given sample using statistical analysis based on hypergeometric Z-scores.

Purpose The Z-score algorithm is a quick method of detecting aberrant regions. It finds intervals using a sliding window of fixed size, specified by the user. Enrichment in the number of probes with high log ratios for any given window size yields an aberrant region.

Use The exploratory phase of an analysis in the CGH Application may include visual inspection of regions with an over abundance of probes with log ratios which pass a scoring threshold. This abundance is compared to the number of probes expected to pass the same threshold at random, yielding a Z-score for each window.

Algorithm The scoring method has essentially two steps. In the first step it identifies the total number of probes with log ratios significantly different from zero in the sample array. These probes are referred to as outlier probes.

NOTE Ideally, the probe statistics would be computed for samples that contain no genetic anomalies, so that μ and σ represent the distribution of a normal sample.

In the second step, the method determines if the actual sample of interest has a significantly higher proportion of aberrant probes in any given genomic region than the proportion of total aberrant probes in the sample used in the first step. If it identifies any such region, then this region gets a higher score and it is called as an aberrant region. The two steps are explained in more detail in the following sections:

Step 1: Calibration

For each probe, the log ratio is *Z*-normalized by computing the usual formula:

$$Z(L) = \frac{L - \mu}{\sigma} \quad \textbf{(8)}$$

where L is the signal log ratio, μ is the mean and σ is the noise level of the population of such log ratios. Chromosomes X and Y are not included in the calculation of μ and σ since gender differences between arrays can offset the statistics.

Each Z-value can be classified as significantly above or below the mean by using a Z cutoff, Z_C. This cutoff can be supplied as a user-specified value. In essence, you are simply stating that you consider log ratios greater than Z_C to be outliers from the normal population of log ratios.

NOTE

Z_C is not a cutoff used to *filter* data. It is a cutoff for *classifying* data as being significantly above or below the mean. To avoid reinforcing the idea that this value filters Z-scores, the CGH Application User Interface (UI) refers to Z_C as threshold.

As part of the computation, the number of entries in each of three classes are counted using the threshold setting Z_C from the UI:

- R = number above positive cutoff (*i.e.* the number of outliers above Z_C in a normal array after proper normalization)
- R' = number below negative cutoff (*i.e.* the number of outliers below $-Z_C$ in a normal array after proper normalization)
- N = the total number of measurements

These Z values and counts are pre-computed and reserved for calculations in **step 2**. The values computed in **step 1** would only need to be recomputed if a different Z_C were desired. Even so, μ and σ can still be reused without computation.

If you want to determine the pre-computed statistics more accurately, you can select the specific arrays to be used in this step.

Step 2: Computation

To compute a moving average, log ratios are averaged over a small subset of points in the genome. This moving average window, w, may be simply a number of adjacent measurements or it may be over a positional window (such as every megabase). The objective is to analyze the over- or under-abundance of log ratios within the window that deviate significantly from the mean from step 1. For this smaller subset, the same three counts as in step 1 are computed using exactly the same cutoff values, but in this case, only for the points within the averaging window, w:

- r = the number above the positive cutoff (Z_C) in w
- r' = the number below the negative cutoff ($-Z_C$) in w
- n = the total number of measurements in w

Now, compute an exact Z-score that measures the significance of this over abundance or under abundance in w of significant positive deviations as:

$$Z(w) = \frac{\left(r - n\frac{R}{N}\right)}{\sqrt{n\left(\frac{R}{N}\right)\left(1 - \frac{R}{N}\right)\left(1 - \frac{n-1}{N-1}\right)}} \quad \textbf{(9)}$$

You can compute the same formula for r' to obtain a score for negative deviations. This score can be plotted in a manner analogous to a moving average. This would identify statistically significant groups of probes that appear to deviate from the typical distribution of values for the given microarrays. In this way, it provides some predictive power to call amplification or deletion events in CGH studies.

ADM-1

Aberration Detection Method 1 (ADM-1 or "adam-one") is a detection algorithm that identifies all aberrant intervals in a given sample with consistently high or low log ratios based on the statistical score.

Purpose The ADM-1 algorithm searches for intervals in which the statistical score exceeds a user specified threshold. The statistical score is calculated based on the average log ratios of the probes and the number of probes in the interval. ADM-1 reports contiguous genomic regions of arbitrary size as aberrant regions.

Use In contrast to the *Z*-score algorithm, the ADM algorithms do not rely upon a set window size, instead sampling adjacent probes to arrive at a robust estimation of the true range of each aberrant segment. The output differs from that of the CBS algorithm by reporting statistically significant aberrant regions, allowing rapid genomic assessment. The ADM-1 algorithm may be limited to a subset of genomic locations which have annotated features. This restriction, termed 'genomic boundaries', is available in the CGH Application UI.

Algorithm The ADM-1 statistical score is computed as the average normalized log ratios of all probes in the genomic interval multiplied by the square root of the number of these probes. It represents the deviation of the average of the normalized log ratios from its expected value of zero.

The ADM-1 score is proportional to the height h (absolute average log ratio) of the genomic interval, and to the square root of the number of probes in the interval. Roughly, for an interval to have a high ADM-1 score, it should have high height or/and include large number of probes.

Before calling the ADM-1 routine, the log ratios are normalized in the following way:

Step 1: Normalization

In the normalization step, the expected average μ is subtracted from all log ratios L, and then these modified log ratios are divided by the estimated variance σ. This transforms the log ratio scores into a normal Z distribution with a mean of 0 under the null model assumption:

$$Z(L) = \frac{L - \mu}{\sigma} \quad \textbf{(10)}$$

where μ is the mean and σ is the noise level of the population of such log ratios.

Once the data are transformed the following score is assigned to each interval I:

$$S(I) = \frac{\sum_{i \in I} Z_i}{\sqrt{N_I}}, \quad \textbf{(11)}$$

where N_I is the number of probes in the interval, and $S(I)$ represents the number of standard deviations that the sum of values in I deviates from its expected value of 0 under the null model.

Step 2: Iteration

A call to the ADM-1 function starts a recursive process. The first step is to identify the interval I for which $S(I)$ is maximal and exceeds a predefined threshold parameter, t, specified in the user interface. Then the process is called on the interior of this interval, using the interval median as a mean for re-centering the values, as well as on the two intervals, one to the left and the other to the right flanking I, towards the two ends of the chromosome.

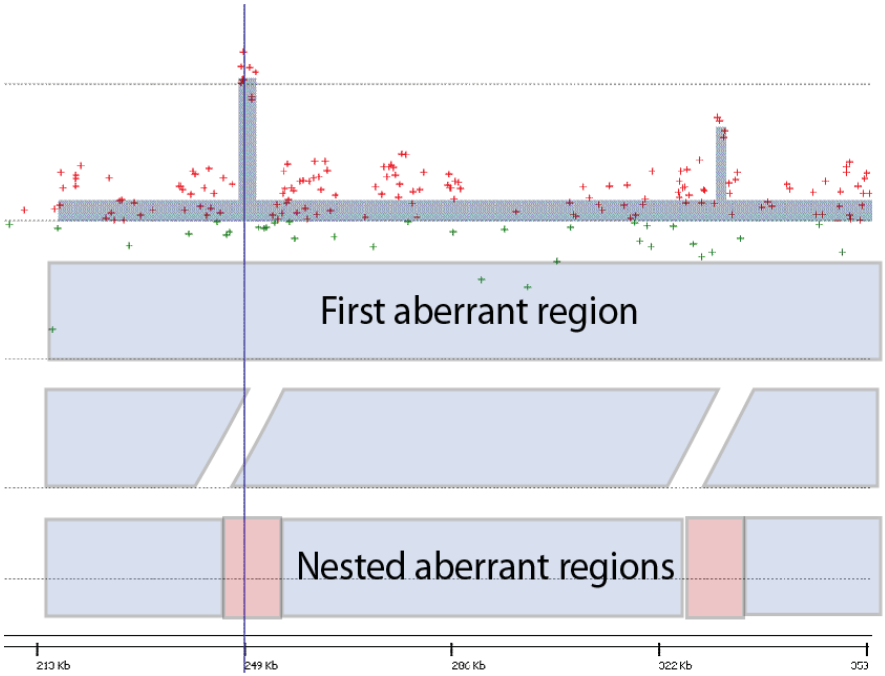


Figure 194 Aberration region identification by recursive interval scoring. For each genomic interval *I*, Genomic Workbench computes an interval score, *S(I)*.

The overall recursive structure of the algorithm is:

Given a data vector for a single sample, single chromosome, and a statistical threshold value:

1. Find the most significant interval *I* in the chromosome.
2. If *S(I)* >= *t*, mark *I* a significant interval.
3. Add *I* to the list of intervals.

Search recursively for more intervals (a) to the left of *I*, (b) to the right of *I*, and after normalizing, (c) inside *I*.

NOTE A text report can also be generated which reports the p-value corresponding to each interval. The p-value is calculated using the normal probability distribution function and the score of that interval.

ADM-2

The Aberration Detection Method 2 (ADM-2 or "adam-two") algorithm identifies all aberrant intervals in a given sample with consistently high or low log ratios based on a statistical score.

Purpose

The ADM-2 algorithm searches for intervals in which a statistical score based on the average quality weighted log ratio of the sample and reference channels exceeds a user specified threshold. ADM-2 reports contiguous genomic regions of arbitrary size as aberrant regions.

Use

In contrast to the *Z*-score algorithm, the ADM algorithms do not rely upon a set window size, instead sampling adjacent probes to arrive at a robust estimation of the true range of the aberrant segment. The output differs from that of the CBS algorithm by reporting statistically significant aberrant regions, allowing rapid genomic assessment. ADM-2 differs from ADM-1 by using probe quality information to weight the log ratios before calculating the score for the interval. The ADM algorithms may be restricted to a subset of genomic locations which have annotated features. This restriction, termed 'genomic boundaries', is available in the CGH Application UI.

Algorithm

ADM-2 uses the same iterative procedure as ADM-1 to find all genomic intervals with the score above a user specified threshold. In ADM-2, the score represents the deviation of the weighted average of the log ratios from its expected value of zero. This score is similar to the statistical score used in ADM-1 analysis, but ADM-2 incorporates quality information about each probe measurement.

The Quality-Weighted Interval Score algorithm (ADM-2) computes a set of aberrations for a given sample. The overall recursive structure of the algorithm is the same as it is in ADM-1.

Step 1: Log ratio error model

The only difference between ADM-1 and ADM-2 is in the definition of the score of the interval. ADM-1 considers only the log-ratios, while in ADM-2 you also consider the log-ratio error information, hence the name Quality-Weighted Interval Score.

The following describes the ADM-2 score:

1 Input is a vector of pairs $(L_1,LE_1),(L_2,LE_2),\ldots,(L_n,LE_n)$, where

- L_i is the log-ratio signal for the i-th probe
- LE_i is the log-ratio error for the i-th probe ordered

- If *dLRsd* is greater than LE_i, then LE_i is set to the *dLRsd* value.

Define $q_i = 1/(LE_i)^2$ **(12)**

Assume that under the null model, $L_i \sim N(0, 1/q_i)$ and the different L_i are independent of each other.

2 Consider the weighted sum, for an interval *I*:

$$\sum_{i \in I} q_i L_i \quad \textbf{(13)}$$

3 Compute the variance of *S(I)*:

$$var[S(I)] = var(\sum q_i L_i) = \sum q_i^2 var L_i) = \sum q_i \quad \textbf{(14)}$$

4 Compute the ADM-2 interval score:

$$S(I) = \frac{\sum q_i L_i}{\sqrt{\sum q_i}} \quad \textbf{(15)}$$

This score reflects the deviation of the weighted sum from its expected value (0) in units of standard deviation. If the quality weight of each probe is the same as the probe to probe noise of the array, then the score will be same as ADM-1.

NOTE A text report can also be generated which reports the p-value corresponding to each interval. The p-value is calculated using the normal probability distribution function and the score of that interval.

HMM

This method identifies all aberrant regions in a given sample using a rigorous probabilistic analysis based upon the Markov chain principle. Hidden Markov models are a stochastic process which measure the likelihood of observing a probe with aberrant signal in the presence of uncertainty in the signal measurement. The probes are sequentially binned along the chromosome according to whether they are likely to measure an amplification, a deletion, or a region in which no copy number change occurred. The result of an analysis with an HMM is therefore a series of

states where each state reflects how the individual probe was binned. Aberrant regions of any length are then detected by syntenic probes sharing a common aberration state. Genomic Workbench offers an unsupervised version of an HMM, which optimizes the stochastic parameters directly from the data using what is known as a Baum-Welch learner.

Mathematically, a Hidden Markov Model of the observed data is defined by four components:

1. The number of states (K): This is the number of distinct aberration states into which the observed data is to be partitioned. For example if K = 3 then this would correspond to amplification, deletion and "no change" aberration status. This is a parameter of the algorithm to be specified at the outset in the user interface.
2. The initial state distribution (π): The initial estimate of the distribution of the aberration states of the probes.
3. The transition matrix (A): each element a_{ij} of the transition matrix is the probability of changing the state i to state j as the genomic position is incremented along each chromosome. For example, if the number of states (K) is equal to 3 then this is a 3x3 matrix. However, for array CGH data, it depends not only on the log ratios but also how close or far apart the probes are (see following paragraph).
4. The distribution of probes with a given state (B). This is assumed to be Gaussian with unknown mean and variance. These parameters are then estimated from the data.

Given the observed log ratios and number of states, the aim is to estimate the optimal value of parameters $P = (A,B,\pi)$ given the data. In other words, the aim is to maximize the likelihood $L(P|logratios)$. Once the likelihood has been maximized and parameters estimated, the distributions are then used to assign to each probe the probability of being in a given state. The state with maximum probability is then assigned to the probe.

Because the entire process is Markov, each state determination is also dependent upon which state the model is in prior to encountering a probe signal. This allows states unlikely to be encountered in the data to be assigned less frequently, while common states are more likely to remain in the same state through a series of probe signals. Such a measure of conditional probability is called the transition probability between any two states in a HMM.

The hidden Markov model based algorithms therefore attempt to optimize the probabilities based on the data. This procedure is iterative and usually slow. However, if there exists a good estimation of the initial state distribution and the expected values of the parameters of Gaussian distribution(s), then a speed improvement is achieved and convergence of the data-based probabilities is faster. To achieve this, Agilent Genomic Workbench uses a two step approach. In the initial step, a fast estimation of aberrant intervals is done. This allows a good estimation of the initial probabilities and of the initial parameters for the distributions to be optimized. This is done using Haar transforms for segmentation. These estimates are then used in HMM optimization procedure in the second step.

Purpose The Hidden Markov Model aberration call method is a visualization and report analysis which predicts the most likely assessment of the state of a given probe given the log ratio values. The state is the aberration status and can be magnitude changes in amplification or deletion, or represent no change. A putative aberration region is a continuous number of probes which share a common state. Such aberrations can be of any size, and a probabilistic measure of confidence can be obtained for every state (and therefore probe) position.

Use The HMM method is used to identify genomic intervals that are most likely to have a given aberration state as determined from probabilities observed in the data. HMMs are useful when there is noise in the signal measurements and use a probabilistic framework which allows for further confidence ratio calculations. After calculating aberration states, visualization of groups of sequential probes sharing an aberration state can be used to identify aberrant genomic intervals. Confidence scores are available as a report.

Algorithm The HMM detection analysis first sets the number of possible states. This value is defined in the Genomic Workbench window. See "To apply the aberration algorithm" on page 137 for more information. The initial transition probabilities of the states, called the stationary distribution (π), is then estimated using the Haar Wavelet. The transition and emission probabilities are optimized using an expectation-maximization (E-M) algorithm termed the Baum-Welch learner. Finally, the most likely sequence of states is determined by a dynamic programming algorithm called the Viterbi algorithm. The report contains the most likely state assigned to each probe.

NOTE The number of states can be set to either 3, 5, or 7. For 3 states, the HMM measures the likelihood of observing a probe value in either an amplification, deletion, or no change region. When set to either 5 or 7 states, the HMM is able to distinguish between states which vary in magnitude in either the amplification or deletion ranges.

Step 1: Initial segmentation

In this step an initial segmentation is calculated for genomic regions. This identifies genomic intervals consisting of different aberration status. For each probe, there exists a log ratio and the raw measurements of red and green intensity ratios. The signal vector $y[n]$ consists of values of log ratios for n probes ordered along the chromosome (arm). Segmentation is a procedure for recovering a piece wise constant signal $x[n]$ from the signal $y[n]$ which may contain noise. Each piece then reflects the probes with a particular aberration status and $x[n]$ is called a segmentation of $y[n]$.

The segmentation procedure consists of the following steps:

1. Apply a undecimated discrete wavelet transform to the data using the Haar wavelet. The Haar wavelet is the usual method for recovering a piece wise constant signal[5].
2. The wavelet transform consists of a series of bands. Obtain the detail sub bands.
3. Find the local maxima of each sub band.
4. Threshold each sub band separately using a FDR threshold supplied as a parameter.
5. Gather the thresholded maxima from each sub band and Unify them to obtain a set of breakpoints.
6. Reconstruct the segmentation result from the subbands by using an inverse transform.
7. From the segmentation results keep only the intervals that are significant.
8. Calculate the probabilities and probability distribution parameters for different aberration states and pass them to the second step (HMM).

See "Segmentation- Discrete Haar Wavelet Transform and Gaussian mixtures for HMM" on page 564 for more information.

Step 2: The Hidden Markov Model

The Hidden Markov Model consists of the following steps:

1. Select the model structure and initialize the Hidden Markov Model.
2. Find the optimal model by considering all possible models through application of the Forward-Backward procedure and Baum-Welch Expectation Maximization.
3. Assign optimal states through application of the Viterbi algorithm.

HMM construction

HMMs in general rely on probabilities to describe how the model moves from one position to another. The model itself can be described as a probabilistic mapping between the observed data (*i.e.* the log ratios) and the hidden states (the aberration states).

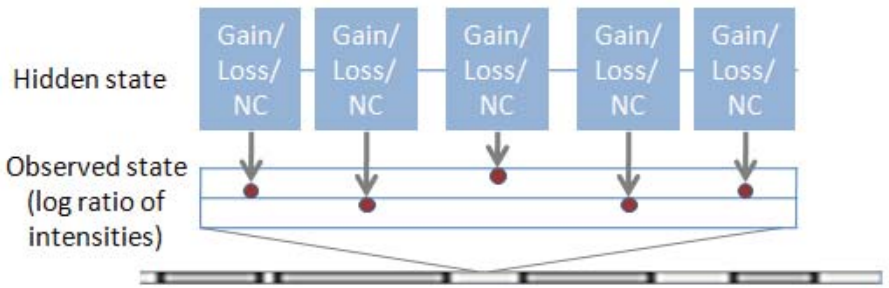


Figure 195 HMM overview for aCGH analysis. The aim of the HMM algorithm is to assess how likely it is to observe a log ratio of intensities from each of the possible hidden state assignments.

In CGH analysis, this translates into the question of whether or not a region is aberrant given the probe log ratio scores and is written:

$$p(H|O)$$

The probability of the hidden event (aberration status) is the probability of the data forming a model (*i.e.* aberrant or non aberrant). However, in aCGH analysis, it is the probability that a given log ratio value is part of a amplification or deletion region that is of interest. Assignment of the probability of amplification or deletion to given log ratios from a model is therefore necessary. It is difficult to calculate this probability based directly upon the data. However, it is relatively straightforward to calculate the probability of observing the data given each competing model H_i:

$$p(O|H_i)$$

i.e. the conditional probability asks which of a number of models for assigning probabilities is best compatible with the data. This model is then used to assign probabilities.

Since the conditional probabilities describe the intersection of two events occurring, the two conditional probabilities can be related through Bayes law:

$$p(H|O) = \frac{p(O|H)p(H)}{p(O)} \quad \textbf{(16)}$$

The Hidden Markov Model itself as described above is specified by four components:

1. Posterior probability $p(H|O)$: The probability of a hidden state or model occurring given observed data.
2. Likelihood probability $p(O|H)$: The probability of observed data given a hidden state or model.
3. Prior probability $p(H)$: The probability or prior "belief" that a hidden state or model is preferred over others.
4. Normalization $p(O)$: The summation of the probability of the data over all possible hidden states or models. Probabilities measure the chance of an event occurring out of all possible outcomes; the normalization measures the probability of the observed data given all competing models and is written:

$$\sum_{\forall H_i} p(O|H)p(H) \quad \textbf{(17)}$$

However, for array CGH data, the transition probabilities (*i.e.* whether the next probe belongs to same aberration state) depends not only on the log ratios but also on the distance between the probes. The transition matrix therefore has to incorporate this information.

For three states this becomes:

$$A_i = \begin{pmatrix} 1-(p_1+p_2)(1-e^{-f_i}) & p_1(1-e^{-f_i}) & p_2(1-e^{-f_i}) \\ p_3(1-e^{-f_i}) & 1-(p_3+p_4)(1-e^{-f_i}) & p_4(1-e^{-f_i}) \\ p_5(1-e^{-f_i}) & p_6(1-e^{-f_i}) & 1-(p_5+p_6)(1-e^{-f_i}) \end{pmatrix} \quad \textbf{(18)}$$

Where f is the distance between consecutive probes and the p_i are the parameters to be estimated. This can be rewritten as:

$$A_i = A_1 + e^{-f} A_2 \quad \textbf{(19)}$$

Where:

$$A_1 = \begin{pmatrix} 1-p_1-p_2 & p_1 & p_2 \\ p_3 & 1-p_3-p_4 & p_4 \\ p_5 & p_6 & 1-p_5-6 \end{pmatrix} \quad \textbf{(20)}$$

And:

$$A_2 = \begin{pmatrix} p_1+p_2 & -p_1 & -p_2 \\ -p_3 & p_3+p_4 & -p_4 \\ -p_5 & -p_6 & p_5+6 \end{pmatrix} \quad \textbf{(21)}$$

Select the optimal model

As described above, the problem is simpler if transformed into simply choosing one optimal model from the set of all models. The following section details this transformation.

The set of probe ratio signals is a series of observations from which the underlying state assignments must be inferred. The state assignments can transition from one state to any other state at each probe location. A measure of probability for any given path can be calculated simply by normalizing by the number of possible paths. For a series of n probes and m states, this is m^n possible paths through the state assignments. However, because of the Markov chain property of the HMM, a local calculation can be applied at each state position which reduces the computational complexity involved in enumerating all possible paths.

The Markov chain property allows local calculations at each state (probe) position to describe the entire likelihood of encountering the total series of states. A Markov chain is a stochastic process that traverses a series of states x_1 x_2 x_3 ... with the probability:

$$p(x_1x_2x_3..) = p(x_1)p(x_2|x_1)p(x_3|x_2)... \quad \textbf{(22)}$$

In other words, the probability of each state is only dependent on the previous state, beginning with one state. From this, the probability of a particular sequence of states S_1 S_2 S_3... can be calculated:

$$p(S_1S_2S_3..) = \prod_{i-1} p(x_{i+1}|x_i) = p(S_1)p(S_2|S_1)p(S_3|S_2)... \quad \textbf{(23)}$$

Because the observations and the states are related through the emission probabilities, the probability of a specific sequence of states S_i given a sequence of observations O_i can be calculated using conditional probabilities. Through Bayes' Law, calculation of the posterior probability of the states given the observations is calculated by normalizing by all possible states as follows:

$$p(S_1S_2..|O_1O_2..) = \frac{p(O_1O_2..|S_1S_2..)p(S_1S_2..)}{\sum_{\forall S_i} p(O_1O_2..|S_1S_2..)p(S_1S_2..)} \quad \textbf{(24)}$$

These calculations form the basis for the Forward algorithm, which consists of the following steps:

1. Start at observation i, $i=1$.
2. For each possible value of $S_i = \theta_1, \theta_2, ... \theta_N$:
   a. For each possible incoming path to state S_i from any state S_{i-1}:

      Sum the incoming paths as determined from the product of the prior, transition, and emission probabilities.
3. Advance to observation $i+1$ and repeat **Step 2**.
4. Sum the states S at $i=N$ to determine the total Forward probability, or the probability of the observations given all possible states.

This calculation is known as the Forward algorithm, which measures the posterior probabilities up to and including a state S_i. In order to assess the probability for each state, a mirror image of the Forward algorithm called the Backward algorithm is applied, which measures all state paths coming from a particular state S_i.

The Backward algorithm consists of the following steps:

1 Start at observation i, $i=N$.

2 For each possible value of $S_i=\theta_1, \theta_2, \ldots \theta_N$:

   a For each possible outgoing path to state S_i-1 from any state S_i:

   Sum the outgoing paths as determined from the product of the prior, transition, and emission probabilities.

3 Advance to observation $i-1$ and repeat **Step 2**.

4 Sum the product of the states S and the emission probabilities at $i=1$ to determine the total Backward probability, or the probability of the observations given all possible states.

The combination of the two algorithms yield the Forward-Backward algorithm, which measures the posterior probability of state S_i, or the number of paths which pass through state S_i normalized by all paths in the HMM.

The Forward-Backward algorithm consists of the following steps:

1 Start at observation i, $i=1$.

2 For each possible value of $S_i=\theta_1, \theta_2, \ldots \theta_N$:

   a Take the product of the Forward and Backward calculations for S_i.

   b Normalize by the total Forward probability $p(O)$ to get the posterior probability for each state S_i.

3 Advance to observation $i+1$ and repeat **Step 2**.

NOTE The total Forward and Backward probabilities are both equal to $p(O)$, the probability of the observations under all possible states.

To illustrate how Genomic Workbench used the Forward and Backward algorithms, the likelihood for Gain state G_i in the Forward algorithm (for a three state system consisting of either Gain G, No Change NC, or Loss L) in log space is:

$$p(G_i) = sum \begin{cases} p(G_{i-1})p(G_i|G_{i-1})p(N[\mu,\ \sigma]|G) \\ p(NC_{i-1})p(G_i|NC_{i-1})p(N[\mu,\ \sigma]|G) \\ p(L_{i-1})p(G_i|L_{i-1})p(N[\mu,\ \sigma]|G) \end{cases} \quad (25)$$

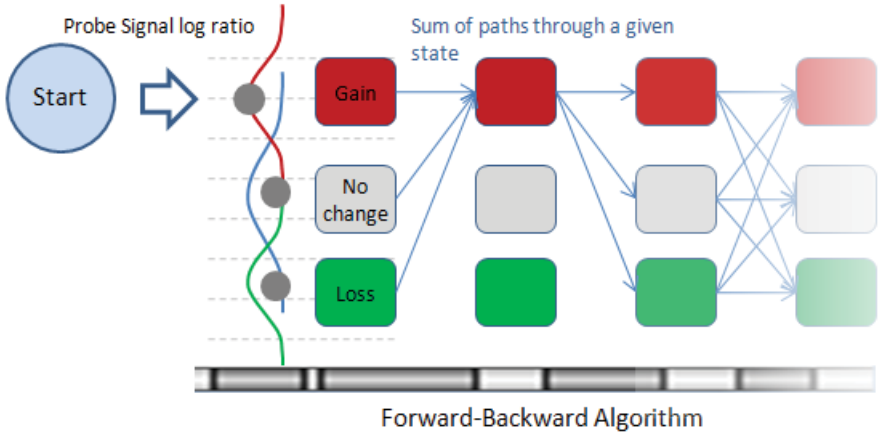


Figure 196 HMM Forward-Backward algorithm - all paths that include a Gain state are summed.

The posterior probabilities from the Forward-Backward algorithm are used in an iterative procedure termed Baum-Welch learning to optimize the HMM conditional probabilities to the data, ensuring maximal discrimination between state assignments. The algorithm consists of the following steps:

1. Calculate posterior probabilities $p(S_i|O)$ using the Forward-Backward algorithm.
2. Calculate emission probabilities as an expectation value based on the posterior probabilities. For the distribution of signal ratios, calculate the emission probability for each discrete state as the Gaussian score conditioned on the posterior probability for that state.
3. Calculate transition probabilities. For every possible state transition, the Forward algorithm to state S_i and the Backward algorithm coming from state S_i*+1* is multiplied by the estimated transition and emission probabilities. This product is normalized by the entire Forward algorithm probability.

4 Repeat steps 1-4 using the updated emission and transition probabilities until both probabilities converge to a stationary value.

State assignment using the Viterbi algorithm

Once the optimal HMM has been chosen, the most likely sequence of states for each probe is determined. These states are then assigned to probes and form basis of reports and visualization. The Viterbi algorithm is very similar to the Forward algorithm. However, for each state position S_i the product of the prior, transition, and emission probabilities from every possible state S_i-1 is replaced with a function that retains only the maximum path.

The likelihood for No Change state NC_i in the Viterbi algorithm (for a three state system consisting of either Gain G, No Change NC, or Loss L) in log space is:

$$p(NC_i) = max\begin{pmatrix} p(G_{i-1})p(NC_i|G_{i-1})p(N[\mu,\ \sigma]|NC) \\ p(NC_{i-1})p(NC_i|NC_{i-1})p(N[\mu,\ \sigma]|NC) \\ p(L_{i-1})p(NC_i|L_{i-1})p(N[\mu,\ \sigma]|NC) \end{pmatrix} \quad \textbf{(26)}$$

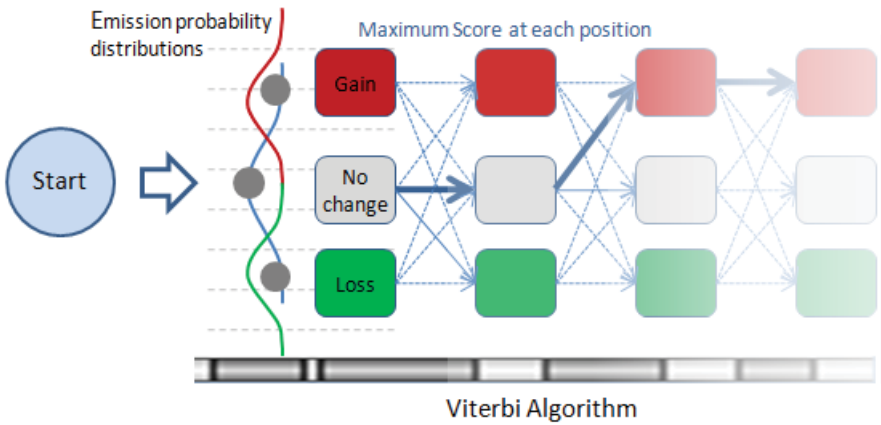


Figure 197 HMM Viterbi algorithm - all paths are compared to find the maximum from the product of the prior, transition, and emission probabilities.

The Viterbi algorithm consists of the following steps:

1 Start at observation i, $i=1$.

2 For each possible value of $S_i = \theta_1, \theta_2, \ldots \theta_n$:

   a For each possible incoming path to state S_i from any state S_{i-1}:

      Select the incoming path with the highest probability as determined from the prior, transition, and emission probabilities.

      Save the incoming path origin (S_{i-1}) to record the path.

3 Advance to observation $i+1$ and repeat **Step 2**.

4 Determine the state S at $i=N$ with the maximal probability.

5 Trace back the states $i-1$ and record each optimal state for observation i.

CBS

Agilent Genomic Workbench provides a new, faster version of the Circular Binary Segmentation[6] (CBS) algorithm developed by *Olshen et. al.*[7] for identification of genomic locations with significant adjacent mean copy number changes.

Purpose While similar in scope to the ADM algorithms, CBS differs from aberration calling algorithms in general by partitioning the set of probes on the array corresponding to a given chromosome into subsets that share the same copy number. This yields a robust map of chromosomal copy number change points and is useful for putative aberration characterization, copy number estimates, and downstream analysis.

Use In contrast to the Z-score algorithm, the CBS algorithm does not rely upon a set window size, instead sampling adjacent probes to arrive at a robust estimation of the true range of the aberrant segment. The output differs from that of the ADM algorithms by reporting regions with constant copy numbers. This approach allows for statistical analysis of the segments reported by the Z-score algorithm or the aberrant regions reported by the ADM algorithms.

Algorithm An effective approach to ensure that all change points are considered during segmentation is to recurse over any given segment length. Recursion within any given segment ensures that the relative locations of change points are preserved.

Step 1: Binary Segmentation

The approach of chromosomal segmentation for aCGH analysis is to recursively split a position-specific set of log ratio intensities into two candidate subsets and compute the difference of means between the subsets as shown in the first candidate split in Figure 198.

Let i be an index of probe location across a chromosome of length n such that $1 \leq i \leq n$ with any subset denoted by $\{s,t\}$ such that $1 \leq s \leq i \leq t \leq n$. Additionally, let the score X_i be the log ratio intensity of probe i. If we let the mean score for any subset be proportional to the copy number, any position i may denote a change point log ratio means, and therefore in copy number. Such an assignment will create two syntenic regions with a difference in mean scores, where the range of the first subset is $\{s..i\}$ and the second is $\{i+1..t\}$. The respective ranges become the new segment boundaries for a new nested candidate breakpoint. This definition is applied recursively in each subset until no further segment can be split. Figure 198 illustrates the application of binary segmentation of a given chromosome.

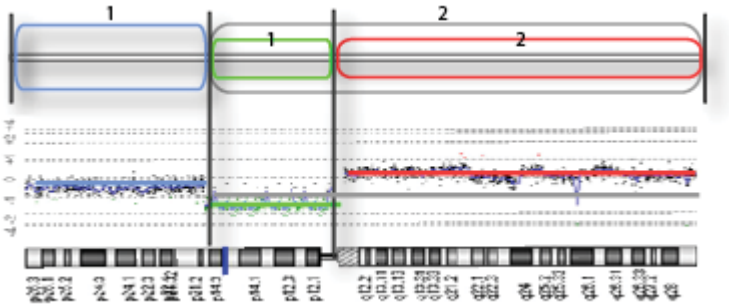


Figure 198 Binary Segmentation of a chromosome. The diagram shows a plot of log ratio values by chromosomal position. The top illustration shows successive iterations of binary segmentation. In the outermost rounded rectangles, segmentation of log ratio means (indicated by matching colored bars in the middle panel) leads to two daughter regions (split by the first vertical bar and indicated by outer labels *1* and *2*), each potentially having uniform copy number. In the second iteration (indicated by the inner colored rounded rectangles), the segment originally labelled as segment *2* is reassigned by a new change point into two new segments (*1,2*), each with potentially uniform copy numbers. This process is iteratively applied until no more change points are found.

Step 2: Circular Binary Segmentation

A potential complication of binary segmentation is that the algorithm as described may be insensitive to abrupt copy number changes where the range of such a change is minimized relative to the parent $\{s,t\}$ segment[6]. Under binary segmentation, such an event is modeled according to the relative means of candidate subsets of the partition, which, for very small variances, may not pass threshold. This problem is associated with the strategy of identifying one change point at each iteration, and the decreased resolution becomes limiting in CGH copy number change map analysis. A more robust model can be achieved by specifying the presence of two change points defined as the start and end of a region of differential mean value.

Circular Binary Segmentation extends the concepts of binary segmentation to allow both binary and ternary segmentation of a segment $\{s,t\}$. Let $\{i..j\}$ be a range of probe locations across a chromosome of length N such that $1 \leq s \leq i < j \leq t \leq N$. This definition allows the creation of a nested candidate subset $\{i..j\}$ with two flanking candidate regions. If the split is ternary in this way, the two end candidate subsets are joined together, treating the data as if it were a circle and allowing a mean value comparison between the $\{i..j\}$ subset and the joined $\{s..i,j..t\}$ subsets. Figure 199 illustrates the application of binary segmentation of a given chromosome.

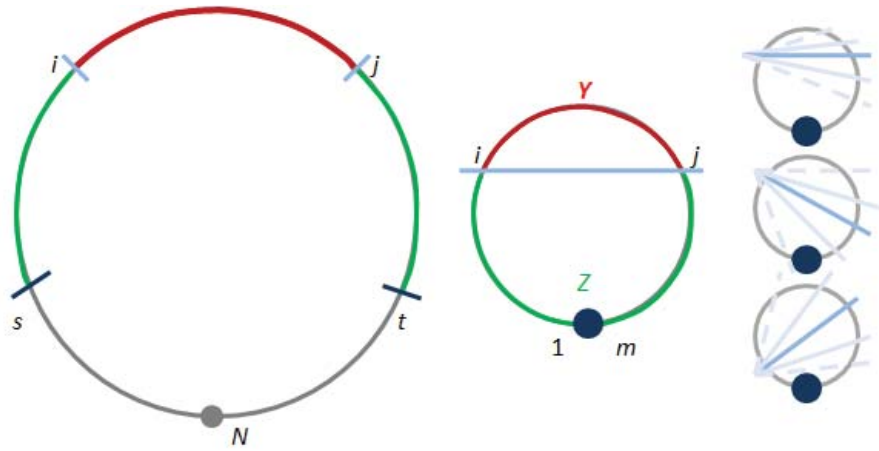


Figure 199 Circular Binary Segmentation of a chromosome. The diagram shows a circularized plot of log ratio scores by chromosomal position. The left illustration shows a first iteration of the CBS algorithm. The red arc denotes a candidate segment based upon log ratio mean values. Creation of this candidate segment leads to two candidate flanking arcs, colored green, with significant differences in mean from the red arc. The center figure shows how the end points of the green segment are joined to evaluate a combined mean log ratio score. Joining the green segments yields two arc segments *Y,Z* where *Y* includes points {*i...j*} and *Z* includes points {*1...i,j+1...m*}. This process is iteratively applied for all possible {*i...j*} to identify the points that yield the greatest discrimination between arc means.

CBS test statistic

CBS measures a difference of means between two adjacent candidate regions, one of which may be the result of a join of flanking regions resulting from a ternary split. If we assume the underlying model for copy number changes in partitioned subsets is based on the mean alone[8] (each region therefore having a common variance), a statistical test of the difference in means between syntenic regions formed by candidate partitioning is sufficient. The means of any two adjacent candidate regions suggested by the segmentation process can be computed from the partial sums of the member log ratio scores normalized by the length of the respective subsets (where that length may result from two joined regions flanking the putative change points). A t-statistic $T_{i,j}$ is calculated for a range given by $\{i,j\}$ for two candidate arcs Y and Z:

$$T_{ij} = \frac{\bar{Y}_{ij} - \bar{Z}_{ij}}{s_{ij}\{(j-i)^{-1} + (m-j+i)^{-1}\}^{\frac{1}{2}}} \quad (27)$$

The candidate t-statistic T is assigned to the $\{i,j\}$ segment such that

$$T = max_{1 \leq i \leq j \leq m} |T_{ij}| \quad (28)$$

The null hypothesis of no change between adjacent regions created by assignment T is tested under the assumption that all T_{ij} follow some unknown distribution that can be estimated from the data. If the p-value from the t-statistic exceeds a threshold α, T will split the parent region. Once a change point is detected in this manner, the procedure is repeated recursively in the newly created segment $\{i,j\}$ until no changes are detected in any of the segments.

Estimation of the log ratio score distribution

CBS generalizes the test statistic to non-parametric assumptions about the underlying distribution of the log ratio scores X^*_i by randomly permuting all X_i within $\{s,t\}$. If T^* is the maximum T^*_{ij} obtained from the shuffled log ratio scores, threshold parameters can be estimated without an assumption of normality. Generation of a robust p-value threshold can be achieved through a high number of permutations (B) but is computationally expensive. Because of this, Genomic Workbench uses a new hybrid strategy and permutation stopping rule for calculation of the p-values from the t-statistic.

The data points $X_1 \ldots X_m$ are partitioned into two groups by k such that $k<=m/2$. Depending on the assignments of $\{i,j\}$, the resulting arcs in the circle can have either less than or greater than k data points. If there are less than k data points, then the data are permuted to find the non-parametric distribution statistics. The calculation of the p-values are compared against a threshold α. Therefore, if the p-value exceeds B, the result is significant. Calculations may therefore be unnecessary beyond a given number of permutations if there is evidence in support of it. A new stopping rule[7] exists by sampling the number of times the permuted statistics exceeds the original statistic out of the running number of permutations. If this number is very rare, computation ends before the total $(1-\alpha)B$ usual calculations. If there are more than k data points, a parametric approximation is used which considerably speeds up the computation time for the algorithm.



Figure 200 Hybrid assessment of *p*-values from the *t*-statistic. For any *k*<=*m/2*, there exists two ways of assigning {*i...j*}. The left two illustrations show {*i...j*} such that *j-i*<=*k* or *j-i*>*m-k*. The right illustration shows {*i...j*} assigned such that *k+1*<=*j-1*<=*m-k*. Assignments of the first type result in permuting the data to find the distribution statistics. Assignments of {*i...j*} in the second category result in a parametric approximation of the distribution where both arcs have more than *k* observations.

Edge effect correction

CBS allows evaluation of candidate change points that are close to the parent segment boundaries. If the start change point for a given segment within $\{s,t\}$ given by the maximum T_{ij} is close to s or the stop change point is close to t, a binary split may be the true result in place of the suggested ternary split. Such an event may occur due to proximity of local noise to an established change point and will result in a false positive identification. To correct for this, all ternary splits are evaluated for viability in the context of overlapping segment data (allowing the start position i to be equal to s or the end position j to be equal to t). If any given i is not a viable change point by itself for the segment $\{s..j\}$, the change point is removed from further consideration. The same test is applied for the overlapping segment $\{i..t\}$.

Change point pruning

Local data trends may lead to false positive identification. Following segmentation, a list of C change points is evaluated. For each segment bounded by consecutive C values, a segment log ratio score average is calculated. The sum of squared deviations *SS(C)* around this average are used to compute the best set of change points *SS(1)..SS(C-1)*. These change points are used to compile the subset which minimizes a function c under a pre-defined constant γ:

$$c' = \min\{c:[SS(c)/SS(C)-1]<\gamma\} \quad \textbf{(29)}$$

where γ=0.05 or 0.10. The change points that give this minimum $SS(c')$ are retained.

Parameters used

The default parameters follow the R-language implementation of this algorithm[9] with the exception of nperm, which was changed to 100 permutations. The values are listed below with a brief summary of their function within CBS:

Parameter	Default	Used	Description
alpha	0.01	0.01	The significance level for the statistical test.
nperm	10000	100	The number of permutations to run.
p.method	"hybrid"	"hybrid"	The method used for computation of the *p*-value.
kmax	25	25	The maximum width of segments for permutation using the "hybrid" method.
nmin	200	200	The minimum length of data for which the approximation of maximum statistics is used under the hybrid method.
window.size	NULL	NULL	Size of window used to speed up computations when segment size is too large. Not applicable using the "hybrid" method.
overlap	0.25	0.25	Proportion of data that overlap for adjacent windows.

Parameter	Default	Used	Description
trim	0.025	0.025	Proportion of data to be trimmed for variance calculation for smoothing outliers and undoing splits based on the standard deviation.
undo.splits	"none"	"none"	A character string specifying how change-points are to be undone, if at all.
undo.prune	0.05	0.05	The proportional increase in sum of squares allowed when eliminating splits if **undo.splits** is applied with parameter "prune".
undo.SD	3	3	The number of standard deviations between means to keep a split if **undo.splits** is applied with parameter "sdundo".
verbose	1	1	The level of verbosity for monitoring the program's progress. The value 1 specifies that the program will print the current sample's progress.

Noise Estimation – the Derivative Log Ratio Spread

To make aberration calls, Agilent Genomic Workbench needs a measure of the log ratio noise for each array. A good estimate of noise is attained by calculating the robust standard deviation (spread) of the log ratio differences between consecutive probes (*dLRsd*) along all chromosomes, divided by $\sqrt{2}$ to compensate for noise averaging. The *dLRsd* is an estimate of noise that is robust even in the presence of highly aberrant samples.

Purpose

Even highly aneuploid samples have chromosomes with extensive stretches along which the genomic copy number is constant, or nearly so. In such constant-copy-number regions, the true log ratios are constant, although not necessarily zero. Estimations of log ratio error, and therefore the minimum log ratio difference required to make reliable amplification or deletion calls, is based on observations of the variation in such constant-copy-number regions. The *dLRsd* metric is an attempt to quantitate such "eyeball" estimates.

For normal samples, the *dLRsd* is the width of a self-self distribution, and should be below 0.2 log units. It will be somewhat greater for abnormal chromosomes because (a) the width of regions of constant copy number different from two will include both noise and the variable log ratio compression observed for many probes, (b) in deleted regions, the noise tends to be slightly higher, as expected from the error model.

To make *dLRsd* more robust and a true measure of noise, outliers are removed from the constant-copy-number regions. To remove outliers, IQR (Inter Quartile Range) statistics are used with appropriate scaling to calculate the spread of the distribution instead of calculating the standard deviation of the derivative of log ratio directly. See Figure 201.

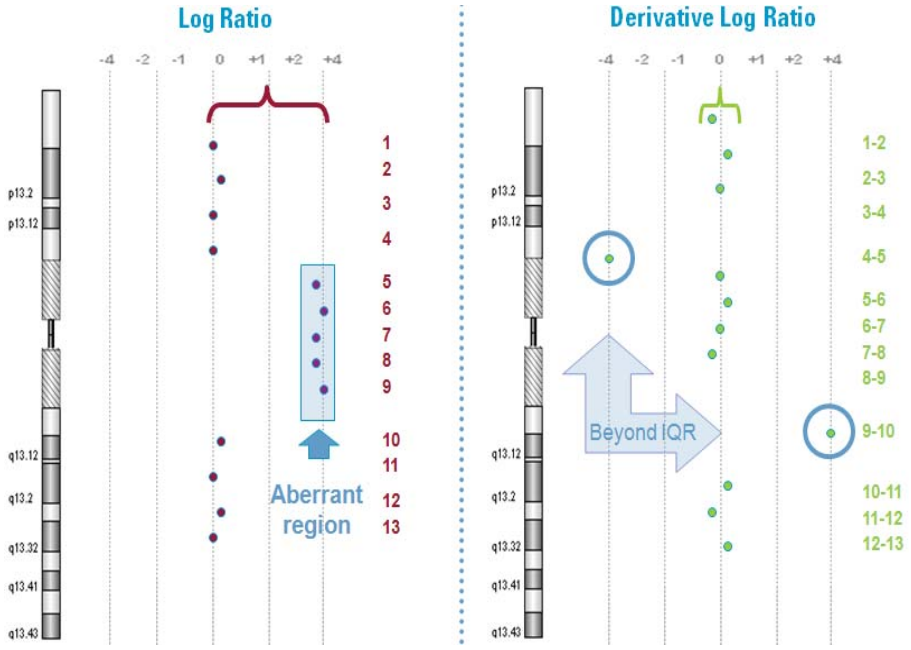


Figure 201 The difference between log ratios and derivative log ratios, illustrated. When using the log ratio spread as an estimate of array variance, aberrant regions can artificially inflate the noise of the array. The derivative log ratio calculates the difference between adjacent log ratios (the derivative log ratio, or DLR) as noise. Genomic Workbench uses the inter-quartile range (IQR) as 50% of the data around the mean of the DLR to estimate the variance.

Interpretation and Visualization of the Detection Algorithms

Output from application of the detection algorithms is visible in the genome, chromosome, and gene view panels. This section describes the interpretation of the visualization capabilities in Genomic Workbench.

Z-score algorithm:

Interpretation

For each moving window, the *Z*-score algorithm takes the points within that window as a sample and computes a hypergeometric *Z*-score that measures the significant number of outliers — points that lie above (positive values) or below (negative values) the threshold *Z*-score. Note that a point that is slightly beyond the threshold is counted the same as a point that is considerably beyond the threshold.

Z-scores are plotted to indicate statistically significant groups of probes that appear to deviate from the typical distribution of values for the given microarrays. They provide some predictive power to call amplification or deletion events in CGH studies. You should adjust the cut-off, *Z*, appropriately based on your visual analysis of amplified and deleted regions in the chromosomes.

Also note that the score that is plotted has nothing to do with the log ratio, and you should not expect the values to necessarily line up with the log ratios. It simply represents a statistical measure of aberration that can be used to track the distribution of outliers, which you can usually see by comparing the scatter plot to the *Z*-score plot.

One final interesting point is that if you set the threshold too low (i.e. cut-off is too small), most of the data points on an array are outliers (i.e. very high values of *R* and *R'*), and you will probably get a *Z*-score of zero. If on the other hand, you set the threshold too high, none of the points are outliers and again, the *Z*-score will be zero. Usually a threshold of 2-3 is the best setting. You can go slightly higher if you want to look for very deviant aberrations, but going too high will show no aberration.

Visualization

When the final *Z*-scores are computed, they can be plotted as a line graph similar to the moving average. To enhance the visibility of the plots and to distinguish them from the moving average, the graphs are filled from the origin. As a further refinement, the filling is alpha-blended for transparency. When plotting multiple microarrays simultaneously, this

minimizes obscuring of the data and allows you to detect overlaps. For two or three simultaneous plots, it is often possible to distinguish the various intersections based on the color blending.

The *Z*-score is also reduced by a factor of 10, thus allowing you to read the actual underlying value by interpolating the location on the graph scale (2, 4) and simply multiplying that value by 10. All *Z*-scores are positive, and those shown in the negative direction are actually positive *Z*-scores for decreased copy number.

It can still be difficult to read small segments of significant *Z*-scores, particularly in the overview. So as a further aid, *side-bars* are provided. Usually, these side-bars are not overlapped and provide a quick means for locating interesting anomalies in single microarrays. They also provide better separation when looking at multiple microarrays. Normally, the side-bars are stacked, but if there is insufficient room they may overlap. This allows you to see clear areas of interest. You can always manipulate the display to enlarge the available space in order to separate the side-bars as needed.

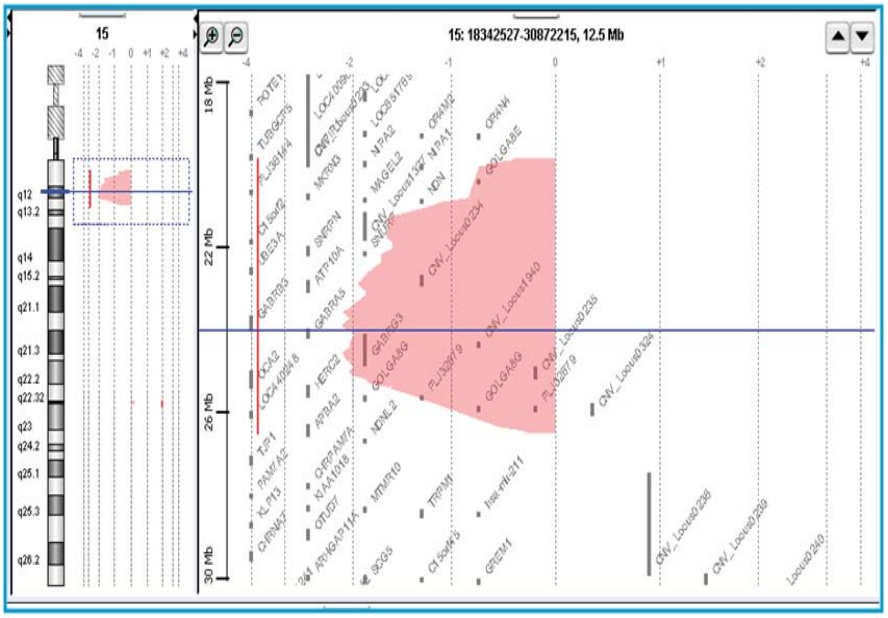


Figure 202 Regions with enriched over- or under-abundance probes as visualized with the Z-score algorithm output.

Interpretation

ADM algorithms:

The Aberration Detection Algorithms (ADM-1 and ADM-2) use a recursive process to identify contiguous stretches of aberrant probes. The recursion stops when no interval with *S(I)* exceeding the user-specified threshold *t* is found. All intervals found in this process are reported, and a plot is generated as output. The intervals are rendered as steps in the visualization panel. The height of each step is equal to the average log ratio of that interval. Steps are also extended on each side of an interval by a fixed distance (other than at the end of the chromosome and centromere). See Figure 203.

NOTE

The fixed interval is set to a default distance of halfway to a neighboring probe, or 200 base pairs (bp), whichever is less. The default value of 200 bp can be changed by accessing the settings in the CGHAnalytics.properties file. For more information, see "Configuration Settings" on page 431.

A threshold can be applied to the ADM algorithms. The optimal threshold depends upon several factors. These factors may be the DNA sample type, the DNA sample quality, or other measures, and can be explored interactively in Genomic Workbench. A recommended starting point for finding the optimum threshold is a value of 6.

Visualization

When the aberration regions are computed, they are plotted as a bar graph, colored by sample. To enhance the visibility of the plots and to distinguish them from the moving average, the graphs are filled from the origin. As a further refinement, the filling is alpha-blended for transparency. When plotting multiple microarrays simultaneously, this minimizes obscuring of the data and allows you to detect overlaps. For two or three simultaneous plots, it is often possible to distinguish the various regions based on the color blending.

It can still be difficult to read small segments of aberrant regions. By clicking on a chromosomal region of interest or by dragging a rectangular selection around a genomic region you can easily zoom in on that region.



Figure 203 An example of ADM output. All intervals are rendered as steps where the height of each step is equal to the average log ratio for that interval.

Interpretation **HMM algorithm:**

Genomic Workbench displays the Viterbi state path results from the HMM algorithm in chromosome and gene view. A report is generated which contains the posterior probabilities for each genomic probe location.

Genomic Workbench reports both the most likely sequence of states and the posterior probabilities for each probe. For the most likely series of states, for each state position S_i the product of the prior, transition, and emission probabilities from every possible state S_i- *1* is replaced with a function that retains only the maximum path.

The likelihood for No Change state NC_i in the Forward algorithm (for a three state system consisting of either Gain G, No Change NC, or Loss L) in log space is:

$$p(NC_i) = max\begin{cases} p(G_{i-1})p(NC_i|G_{i-1})p(N[\mu,\ \sigma]|NC) \\ p(NC_{i-1})p(NC_i|NC_{i-1})p(N[\mu,\ \sigma]|NC) \\ p(L_{i-1})p(NC_i|L_{i-1})p(N[\mu,\ \sigma]|NC) \end{cases} \quad \textbf{(30)}$$

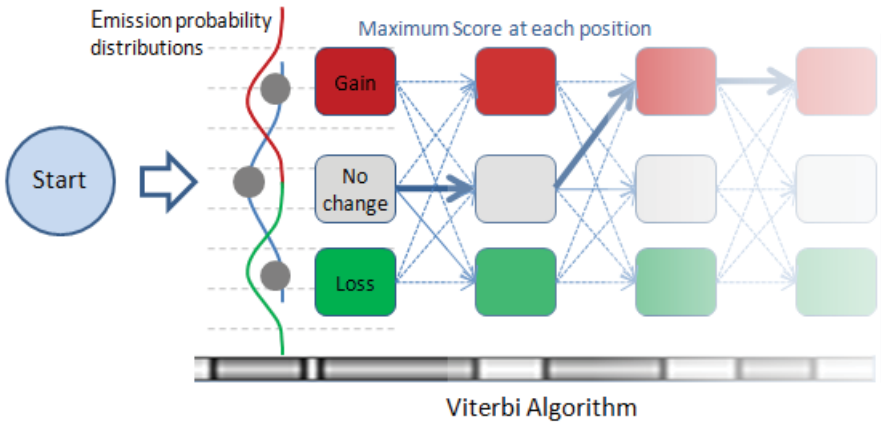


Figure 204 HMM Viterbi algorithm - all paths are compared to find the maximum from the product of the prior, transition, and emission probabilities.

To illustrate how Genomic Workbench used the Forward and Backward algorithms, the likelihood for Gain state G_i in the Forward algorithm (for a three state system consisting of either Gain G, No Change NC, or Loss L) in log space is:

$$p(G_i) = sum\begin{cases} p(G_{i-1})p(G_i|G_{i-1})p(N[\mu,\ \sigma]|G) \\ p(NC_{i-1})p(G_i|NC_{i-1})p(N[\mu,\ \sigma]|G) \\ p(L_{i-1})p(G_i|L_{i-1})p(N[\mu,\ \sigma]|G) \end{cases} \quad (31)$$

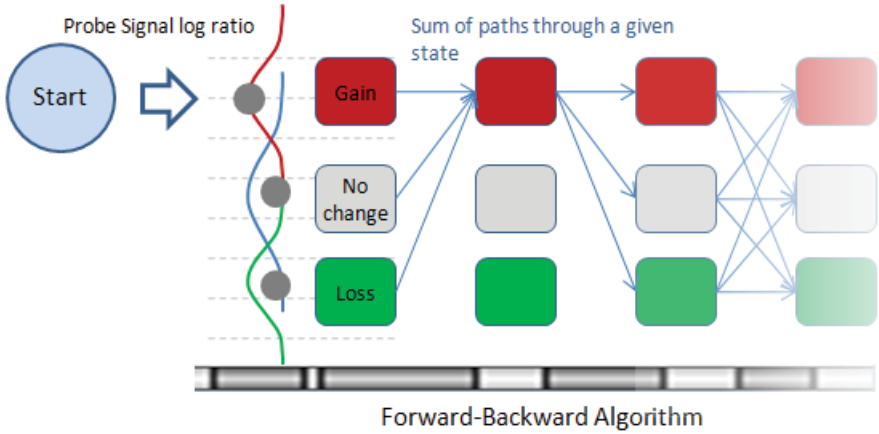


Figure 205 HMM Forward-Backward algorithm - all paths that include a Gain state are summed.

An output report is generated per chromosome. In addition to the most likely sequence of path states generated by the Viterbi algorithm, the report also contains the posterior probabilities generated from the Forward-Backward algorithm. The posterior probabilities from the Forward-Backward algorithm do not give the optimal sequence of states for the HMM given the data. Instead they give the discrimination of one state assignment over any other state assignment for every position in the sequence of states. They are therefore useful to assign confidence to the state assignment. In a two-state system consisting only of gains and losses, such a calculation of posterior probabilities using the Forward-Backward algorithm may look like:

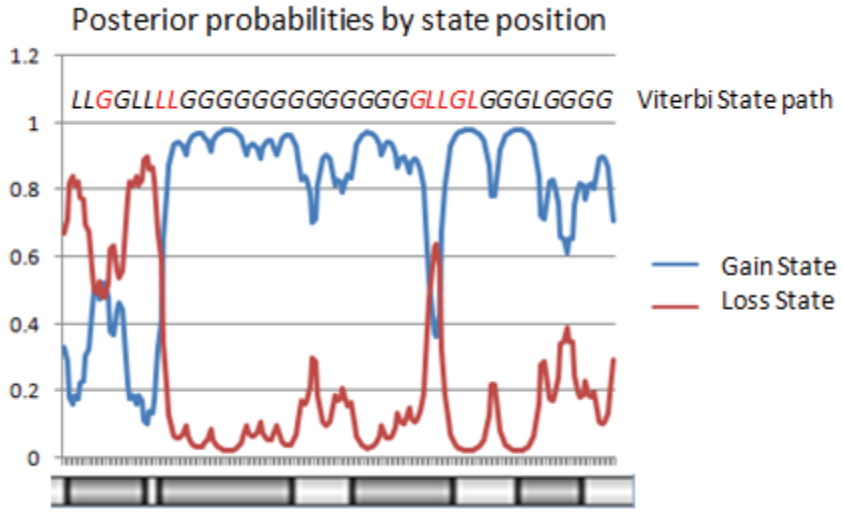


Figure 206 HMM Posterior probabilities plotted by chromosomal probe position. In this figure, optimum state paths from the Viterbi algorithm from a 2-state HMM are displayed above the graph. State assignments in black have a higher confidence than those in red based upon the relative posterior probabilities between the alternative states for each position.

Visualization

When the most likely state assignments are computed, they are plotted as a line graph, colored by sample. To enhance the visibility of the plots and to distinguish them from the moving average, the graphs are filled from the origin. As a further refinement, the filling is alpha-blended for transparency. When plotting multiple microarrays simultaneously, this minimizes obscuring of the data and allows you to detect overlaps. For two or three simultaneous plots, it is often possible to distinguish the various regions based on the color blending.

It can still be difficult to read small segments of aberrant regions. By clicking on a chromosomal region of interest or by dragging a rectangular selection around a genomic region you can easily zoom in on that region.

The number of possible levels in the line graph is set by the number of states for the HMM. Each level has a target value - therefore, the analysis reports state assignment for aberration status. While this may reflect the relative copy number of different regions, it does not determine the absolute copy number of the region. See Figure 203.

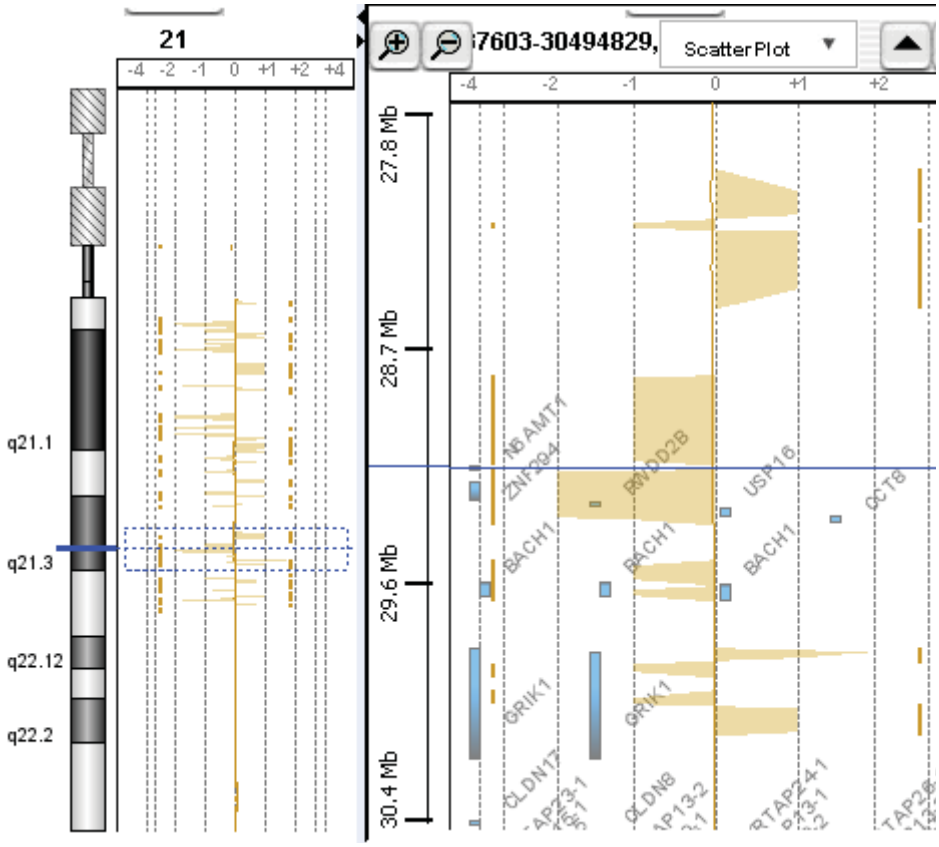


Figure 207 HMM output (5 states). The colored regions show state assignments that deviate from no change, detected through the Viterbi algorithm.

Interpretation

CBS algorithm:

The recursion stops when no additional change points can be found. All intervals with constant copy number bounded by these change points found in this process are reported, and a plot is generated as output. The intervals are rendered as bars in the visualization panel. Identification of these intervals can be used for further exploration of aberration regions or for copy change analysis.

Visualization

When the change points are identified using CBS, each region of nearly constant copy number is reported as a bar graph colored by sample and plotted by magnitude. Change points are then simply identified as breaks in the bar series for any given color. To enhance the visibility of the plots, the bars are alpha-blended for transparency. When plotting multiple microarrays simultaneously, this minimizes obscuring of the data and allows you to detect overlaps. For two or three simultaneous plots, it is often possible to distinguish the various regions based on the color blending.

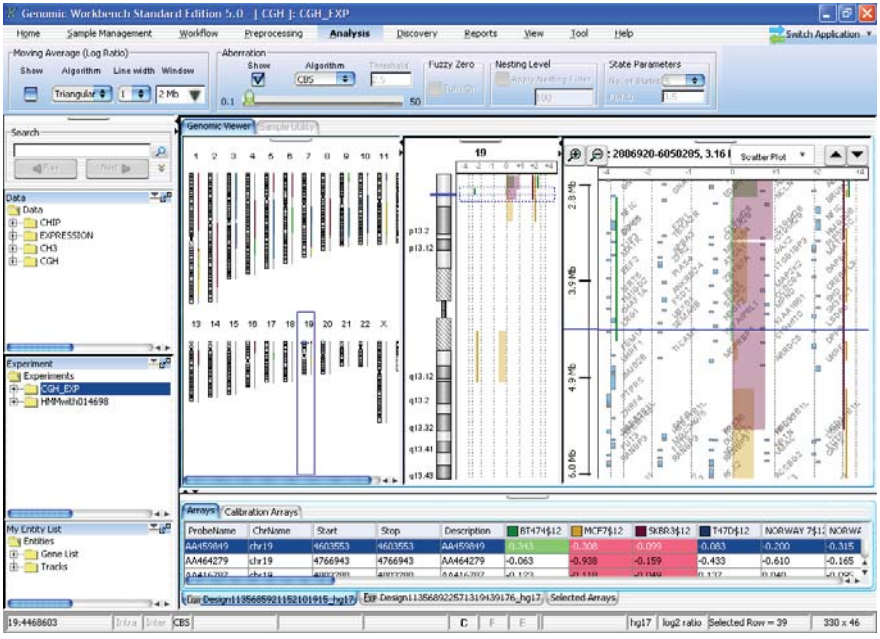


Figure 208 An example of CBS output. All intervals are rendered as bars indicating regions of constant copy number.

Visualization Algorithms

Before applying an aberration detection algorithm, it is useful to visualize the general trend of the data by using a moving average line plot. Genomic Workbench allows detailed inspection of probe-level log ratios through visualization. In order to reduce the complexity of the data from probe-to-probe noise, two smoothing algorithms can be applied to the data. One algorithm applies linear smoothing (the moving average algorithm, described earlier in the chapter), and the other applies a triangular smoothing function. These algorithms are generally applied independently of the aberration detection algorithms and use a sliding window of fixed size, set in the user interface (UI). Changing this window size does not affect aberration calls using the ADM, CBS, or HMM algorithms, but does change the window size used by the Z-score detection algorithm.

Triangular Smoothing

Agilent has introduced a triangular smoothing function that, like other shaped smoothing, has a peak (maximum weight) at the center and falls off to zero with increasing distance from the center. Triangular smoothing averages the value of a point of interest and a number of its neighboring points.

Purpose When visualizing or analyzing aCGH data, it is common to smooth the data using a moving average. However, the moving average approach is not the optimal means of reducing the noise associated with each independent point, since it can minimize localized copy number changes and obscure individual points. A good compromise is achieved between reducing the noise of the individual points, and still remain sensitive to true localized, or small-scale, variations in the data by using triangular smoothing.

Use The smoothing functions are used for visualization purposes only and do not affect the aberration algorithms or the aberration algorithm graphical output.

Algorithm The number of neighboring points to be used for smoothing depends on the type of moving average. If we have point input (Pt input) from UI, the number of points that are averaged is kept fixed, say at 3, 5, 7, 9, or 11 points, and each point is given equal weight. In other cases, a window of

constant width (specified from UI in Mb or Kb) moves across the data and centers on the point of interest. All points within its range are averaged to yield the moving average value for the point.

NOTE

In linear smoothing, the user defined point input (Pt input) uses the same number of probes as in triangular smoothing. However, an equal number of probes are taken in each direction (to the left and right of the center probe) to compute the moving average. This might not be ideal if the probes are placed at varying distances.

There are two potential problems in linear smoothing:

- First, using a fixed window width (by choosing a base pair from the UI) causes a variable number of probes to be averaged at each smoothed point. Therefore the degree of averaging varies from probe to probe, depending on how many probes happen to be in the fixed-width window. Since varying numbers of measurements contribute to each smoothed point, the degree of statistical noise reduction also varies for each point. This can complicate the error analysis.
- The second problem with linear smoothing arises on so-called "zoom-in" arrays, in which some genomic regions of interest are covered more densely than neighboring regions. For such arrays, the appropriate window size can vary greatly between the different genomic regions. Smoothing windows that are appropriate for sparsely tiled regions will obliterate all structure in densely tiled regions, whereas windows appropriate to densely tiled regions will perform practically no averaging at all in sparsely tiled regions.

These problems with linear smoothing can be avoided in triangular smoothing by using smoothing windows containing a fixed number of probes, regardless of the total range of sequences those probes span. This respects the fact that nearby probes are more relevant than distant probes to the average at any point.

The concept is illustrated in Figure 209. Fixed count smoothing includes the same number of points in each average, but weights probes far from the averaged point as much as points near to it. In the new triangular smoothing with Pt input, a symmetric window around the averaged point is enlarged until it contains the number of points chosen for the fixed size window. These points may be on one side of the averaged point or on

both, depending on the probe density around the averaged point. These points are then weighted appropriately for triangular smoothing function, depending on their distance from the averaged point.

When user inputs number of points from the user interface, it uses a variable window width, which is chosen to be the smallest window, symmetrical about the averaged point, which includes the specified number of points. Figure 209 illustrates the application of this method for Pt input.

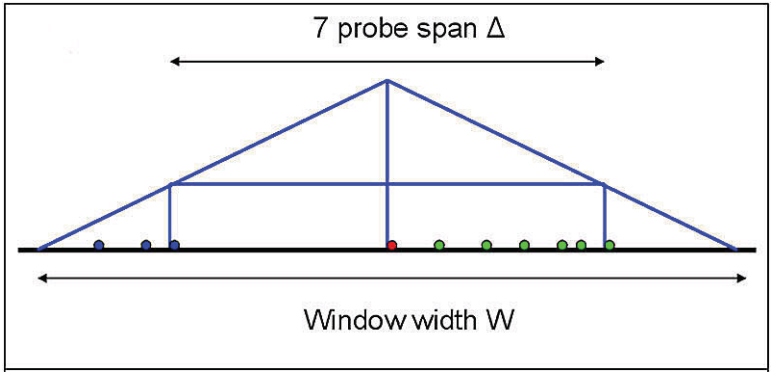


Figure 209 Triangular smoothing with Pt input.

Smoothing is applied to a region of varying probe density. The effective width of the smoothing window, W, depends on the length of the smallest symmetrical region (Δ) that includes the specified number of points. The weight given to each point is proportional to the height of the triangle at that point.

The weights assigned to the log ratio values of these probes are given by the equation:

$$w(x) = (W - |x|) / W^2 \quad \textbf{(32)}$$

where the effective window width, W, is determined by the span, Δ, of the symmetrical region spanning the specified number of points by the equation:

$$W = \Delta / (2 - \sqrt{2}) \quad \textbf{(33)}$$

Interpretation The smoothing algorithms affect the scatter plot in the chromosome and gene view.

Algorithms for Multi-array Analysis

Penetrance

Penetrance analysis is used to find the percentage of samples that share aberrations in a particular genomic region among multiple samples. Amplification and deletions are considered separately.

In probe based penetrance, for each aberrant probe, the percentage of samples that shows that amplification or deletion is calculated. The height of aberration is not considered for this calculation.

In interval based penetrance, for each interval, the percentage of samples that share that amplification or deletion is counted.

Purpose This algorithm is one of the algorithms that can identify aberrant regions that are common among multiple samples that use output from ADM-1, ADM-2, or CBS.

Use To generate a report showing the overlap of identified regions from ADM-1, ADM-2, or CBS across many samples. The algorithm differs from the common aberration analysis in that it does not recalculate the score of each candidate interval using the ADM algorithms, nor does it attempt to identify statistically significant common aberrations. Instead, it reports how many samples agree with a given aberrant probe or region prediction.

Interval Based Penetrance Algorithm Interval-based penetrance is the ability to calculate the percentage of samples that share regions of amplification or deletion. The analysis has the following steps:

1. Collect all start and end points from all the aberrant intervals in all the samples determined from any of the detection algorithms.
2. Sort by genomic position the resulting break points.
3. Construct a set of candidate genomic intervals using all break points listed in **Step 2**, such that there is no break point inside any candidate interval.
4. For each candidate interval A, count the number of agreements in amplifications and deletions separately, representing the penetrance of aberration A in sample set S.

5 Report candidate intervals with penetrance scores above a user-defined threshold (the default threshold is 6) and the names of the corresponding supporting samples.

NOTE For counting penetrance, amplification and deletion are considered separately. The aberration state is set by the aberration height of the deepest nested interval.

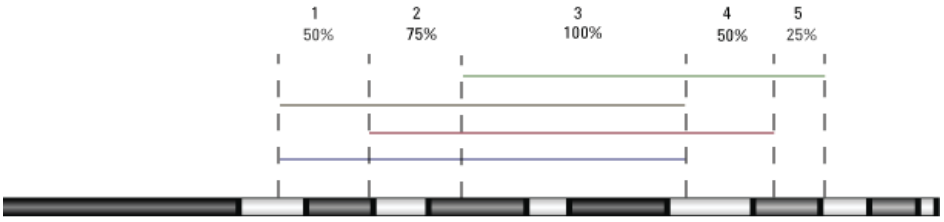


Figure 210 Scoring detected aberrant regions from different samples. Here each sample is color coded. Note that interval based penetrance does not take into account the height of the aberrant region. The vertical axis in the illustration serves only to differentiate the different samples, and not a difference in average log ratios.

Probe Based Penetrance Algorithm

Probe based penetrance calculates the percentage of samples that share a probe amplification or deletion. The analysis has the following steps:

1 Collect all aberrant probes from all samples.

2 For each aberrant probe, calculate the percentage of sample which also show that probe to be aberrant (in the same direction, amplification or deletion), without regard to the magnitude of the log ratio.

Common Aberration Analysis

Common aberration analysis is used with multiple samples to identify genomic intervals that have statistically significant common aberrations. The common aberration analysis can be used with ADM-1 and ADM-2 output, but not with output from the Z-score and CBS algorithms.

Purpose This algorithm is one of the reporting algorithms, which serve to combine and summarize the aberrant region output from the aberrant algorithms.

Use To generate a report showing the overlap of identified regions from the aberration algorithms across many samples.

Algorithm The report analysis has the following steps:

1. Apply one of aberration detection algorithms (ADM-1or ADM-2) to a set of samples, S_1, S_2, ... S_n, to identify a set of aberrant genomic intervals in each sample that have a score above a user-specified threshold.
2. Construct a set of candidate genomic intervals for common aberration analysis using all intervals identified in **Step 1**, and all possible intersections of those intervals.
3. For each candidate interval A, and each sample S, compute a statistical score representing the significance of aberration A in sample S.
4. Compute a combined common aberration score for each candidate interval A, using scores computed in **Step 3**, to test the hypothesis that A is a common aberration for samples S_1, S_2, ... S_n or a subset of these samples.
5. Report candidate intervals with common aberration scores above a user-defined threshold and the corresponding supporting samples. You can choose to report only the most significant common aberrant interval from each set of overlapping intervals to make the report more concise by using the filter on p-value column of the report. See "Properties (for common aberration result)" on page 375.

Steps 3 and 4 are implemented in different ways in the **t-test** and in the **context corrected** common aberration analyses (see below). The main difference between those two methods is that in the t-test analysis, the significance of each candidate interval, A, is computed independently of the aberration state of other parts of the genome. In contrast, the context-corrected statistic adjusts the significance of an interval A, at

sample *S*, according to the overall aberration state of *S*. Intuitively, the more aberrant *S* is, the less significant is any particular aberration in that sample.

T-test analysis

In **Step 3** of the common aberration detection framework, the significance of aberration *A* in sample *S*, *P(A,S)* is computed by applying interval scoring of the algorithm selected in Step 1.

In **Step 4**, the scores for all the samples for interval A are used to calculate significant commonly aberrant regions. Amplifications and deletions are not considered separately.

To test for commonly aberrant regions, the corresponding common aberration p-value, p is calculated as $Prob\{t>=T_n\}$, or based on the Student's t-distribution with n-1 degrees of freedom. The score of candidate interval *A* is the p-value from the t-test.

To compute the statistics T_n, let $\bar{\mu}$ and S^2 be the estimated mean and variance of per-sample statistical scores $\{P(A, S_1), P(A, S_2), ..., P(A, S_n)\}$:

$$\bar{\mu} = \frac{1}{n}\sum_{m=1}^{n} P(A, S_m) \quad \textbf{(34)}$$

$$S^2 = \frac{1}{n-1}\sum_{m=1}^{n} (P(A, S_m) - \bar{\mu})^2 \quad . \quad \textbf{(35)}$$

Then T_n is

$$T_n = \frac{\bar{\mu}}{S/\sqrt{k}} . \quad \textbf{(36)}$$

Context corrected analysis

In **Step 3** of the common aberration detection framework, the significance of aberration *A* in sample *S*, *P(A,S)* is computed using the chromosomal or whole genome context of aberrations distribution in sample *S*. Namely, for each candidate interval *A*, and each aberrant interval *I* (longer and higher than *A*) in sample *S*, compute the probability that *A* overlaps with *I*:

$$P(A \subseteq I) = \frac{|I| - |A| + \varepsilon}{|G| - |A| + \varepsilon}, \quad (37)$$

Where $|I|$ denotes the size of interval I in Mb or in number of probes, $|G|$ is the size of whole chromosome (chromosomal context) or the size if the whole genome (whole genome context), and ε is a small constant added to avoid dividing by zero, ε is proportional to $1/|G|$.

Context specific score of candidate interval A in sample S is computed as a probability that A is amplified in S, $P_{amp}(A, S)$ or as a probability that A is deleted in S, $P_{del}(A, S)$. To compute P_{amp}, consider the set $I = \{I_m\}$ of aberrant intervals in S with height (height is the ADM computed height of the interval) and length greater or equal than the height and length of the interval A, where height of A is the minimum height of all aberrant intervals in S that intersect A.

Then

$$P_{amp}(A, S) = \sum_{1} P(A \subseteq I_m) \quad (38)$$

$P_{del}(A, S)$ is defined is a similar way, using the set I of aberrant intervals in S with height less than or equal to the height of the interval A, and length greater than or equal to the length of the interval A. In this case, the height of A is the maximum height of all aberrant intervals in S that intersect A.

In Step 4 of the common aberration detection framework, you test for common amplification and common deletion. To test for common amplification, scores $P_{amp}(A, S_i)$ are sorted in increasing order. The probability that interval A is aberrant in all k samples, $S_{i1}, S_{i2}, \ldots, S_{ik}$, is estimated by applying Chernoff bound for each k between 1 and n. By Chernoff bound,

$$\text{Probability } (k \text{ or more aberrations}) \leq \frac{e^{(k-\bar{\mu})}}{(k/\bar{\mu})^k} = P_k, \quad (39)$$

where $\bar{\mu} = nP_{amp}(A, S_{ik})$

The score of candidate interval A is $max_k(-log10(p_k))$, and the corresponding set of supporting samples is the set $S_{i1}, S_{i2}, \ldots S_{ik}$ for k that yields this score.

To test for common deletion, apply this analysis to deletion scores $P_{del}(A, S_i)$.

Interpretation The common aberration analysis uses the output from the aberration algorithms. The report shows common aberrant regions using multiple samples as a measure of statistical significance.

Differential Aberration Analysis

Differential aberration is used with multiple samples which have been partitioned into two comparative groups. It is used to identify genomic intervals which have an overabundance of gains or losses in one group of samples compared to the other group in a given genomic region. This is in contrast to common aberration analysis (see "Common Aberration Analysis" on page 508) which identifies samples whose genomic regions have similar aberration profiles. This algorithm uses output from the aberration detection algorithms such as ADM-1 and ADM-2 and reports only statistically significant differences in aberration distributions between the two groups of samples.

Purpose The differential aberration method identifies genomic regions with statistically significant enrichment of gains or losses in one of the group of samples. The tool then allows one to select the regions of interest and save them for further investigation.

Use Differential aberration is used to identify genomic intervals that have statistically significant aberration differences across multiple samples. After calculating aberration profiles using one of the aberration detection methods, groups of interest are defined and the differential aberration is performed.

For samples which are naturally grouped into two groups by some characteristic then this tool can detect the aberrant frequencies in this partition and assign statistical significance to the differences. For example, assume you have two groups named *Group 1* and *Group 2*, both consisting of 100 samples. If Group 1 has 40 aberration amplifications (called gains) in a specific genomic region and *Group 2* has 5 gains in that same region, the statistical significance of observing the 40 gains in *Group 1* given observing only 5 gains in *Group 2* is calculated by using *p-values* from the Hypergeometric distribution. In this example, the *p-value* of observing 40 gains in *Group 1* is 1.3×10^{-9}, meaning that *Group 1* has a significant enrichment of gains compared to *Group 2*.

Algorithm The report analysis first finds all intervals defined by breakpoints for each sample in the study. Enrichment from the perspective of gains or losses from each of the two groups is calculated using the Hypergeometric distribution. It then selects significant results.

NOTE A minimum of three samples in each group is recommended for Differential Aberration analysis. The analysis method becomes statistically more powerful with increasing numbers of samples in each group.

Step 1: Enumerate aberrations by sample groups

The algorithm identifies genomic regions that share aberrations across sample members in each group.

1 Apply one of aberration detection algorithms (ADM-1or ADM-2) to a set of samples, S_1, S_2, ... S_n, to identify a set of aberrant genomic intervals in each sample that have a score above a user-specified threshold.

2 Construct a set of candidate genomic intervals for common aberration analysis using all intervals identified in **Step 1**.

3 Count the number of amplifications (gains) and deletions (losses) in each of the groups *1* and *2* for each candidate interval identified in **Step 2**. Total the number of gains and losses from the identical genomic intervals from the two groups.

NOTE The candidate intervals can be filtered based on the number of probes in the interval either before or after Differential Aberration analysis. Prior to Differential Aberration analysis an aberration filter can be applied before generating candidate genomic intervals. See "To create or modify a design filter" on page 127. Following the Differential Aberration analysis you can select the number of probes as an interval filter criterion. See "Interval Filter" on page 361

Step 2: Calculate probabilities using the Hypergeometric distribution

The enrichment of gains or losses for each group is calculated for each candidate interval. Enrichment of gains or losses is expressed as the area under the Hypergeometric tail (HGT). The Hypergeometric tail probabilities reflect the probability of observing y or more number of gains or losses in a group - *i.e.* it is a measure of enrichment of gains (losses) in a group. This is also referred to as a *p*-value, which is reported in Genomic Workbench as $-log_{10}(p)$; thus larger values are more significant. See "Enrichment Analysis - the Hypergeometric distribution" on page 555.

1 For each candidate interval compute the enrichment of gains in *Group 1*. This is the probability of observing a or more gains in *Group 1* where *Group 1* consists of A out of a total of N samples and there are n total aberrations. This cumulative probability is expressed as a

p-value calculated from the Hypergeometric Tail (*HGT*) in the following way:

$$'(X \geq a) = HGT(a, N, A, n) = \sum_{i=a}^{min(n, A)} \frac{\binom{n}{i}\binom{N-n}{A-i}}{\binom{N}{A}} \quad \textbf{(40)}$$

2 Compute the enrichment of losses in *Group 1*, and the enrichment of gains and losses in *Group 2* using the formula in **Step 1.**

As stated, the enrichment analysis is directional. This means each calculation considers only whether or not there is an overabundance of aberrations within an interval as compared to the other comparison group. Four total calculations must then be completed in order to assess whether there is enrichment of either gains or losses of *Group 1* and of *Group 2*. The following four p-values are therefore calculated for every interval:

- $p\text{-}value(G_1)$ = Enrichment of Gains in group *1* given the observed number of gains in group *1* and *2*.
- $p\text{-}value(G_2)$ = Enrichment of Gains in group *2* given the observed number of gains in group *1* and *2*.
- $p\text{-}value(L_1)$ = Enrichment of Losses in group *1* given the observed number of losses in group *1* and *2*.
- $p\text{-}value(L_2)$ = Enrichment of Losses in group *2* given the observed number of losses in group *1* and *2*.

Step 3: Select Significant Results

The four p-values for every candidate interval are available as results for further analysis and filtering. The filter applied is either the default filter described below, or a customized filter defined by the user. See "Interval Filter" on page 361.

The default filter selects intervals with a $-log_{10}(p\text{-value})$ greater than the selected threshold of $-log_{10}(p\text{-value})$. The default threshold is a p-value less than 0.001. Because Genomic Workbench reports the p-values in $-log_{10}$ values, the default threshold for enrichment is a value greater than *3*.

Cluster Analysis

It is often useful to find patterns from the data independently of existing experimental parameters. Such patterns may give insights into new classes, or aggregates of experimental parameters, which have an effect on the profile of observed aberrations. If the number of experimental parameters is high, or some phenotypic information is missing, a method of data mining called clustering allows the aberration profiles to be matched across samples. Clustering is therefore a class discovery method and is a powerful tool for discovering and assigning labels to samples. Cluster analysis is used with multiple samples following one of the aberration detection algorithms such as ADM-1 and ADM-2.

This algorithm creates a hierarchical tree which relates each sample by the distance and linkage of the aberration profile.

Purpose Cluster analysis is a report analysis which groups samples into clusters such that the arrays in one cluster have aberrations that are more similar to each other than to arrays belonging to a different cluster. Significant experimental parameters that affect the presence or absence of aberrations can be discovered with this tool.

Use Cluster analysis is used to identify new classifications, or groupings, of samples based entirely on the aberration profile information. The clustering technique used is a hierarchical method which results in a tree-like dendrogram relating every cluster, or group of samples to each other as well as the distances between each group of samples. By cutting the hierarchical tree at different levels, labels which describe attributes of the grouped samples can be discovered and saved for further data exploration.

Algorithm The analysis first finds all intervals without any internal breakpoints for each sample in the study. The data is then prepared for clustering by choosing either the aberration status (gain / loss / no change) or the aberration scores for use in calculating distances between samples. Lastly a hierarchical tree is created from distance and linkage measures.

NOTE A minimum of three arrays is required to effectively use the cluster analysis.

Step 1: Obtain aberrant intervals

The algorithm identifies genomic regions that share aberrations across sample members in each group.

1 Apply one of aberration detection algorithms (ADM-1, ADM-2, CBS or HMM) to a set of samples, S_1, S_2, ... S_n, to identify a set of aberrant genomic intervals in each sample that have a score above a user-specified threshold.

2 Construct a set of candidate genomic intervals for cluster analysis using all intervals identified in **Step 1**.

Step 2: Prepare data for clustering

The algorithm has a choice of either clustering the aberration status or the aberration scores. The series of aberrations present or absent in each genomic interval identified above is added to a list for each sample. This series of numbers is a vector which is then compared to other sample vectors to find similarity. This step therefore profiles the aberrations by one of the following methods:

- **Profile of aberration status:** Each interval is assigned a number based on the sign of the detected aberration. If the interval in a given sample contains amplification, it is assigned a value of +1, if it is a deletion it is assigned -1, and it is assigned zero if it is not called aberrant on that sample. Each sample is therefore a vector, characterized by a series of calls corresponding to the intervals identified in **Step 1** of the algorithm.
- Profile of aberration scores: Each interval identified in Step 1 is assigned a number based on the aberration score of the detected aberration. This score is assigned by the aberration calling algorithm (e.g. the ADM-2 score). Each sample is therefore a vector characterized by a series of scores corresponding to the intervals identified in Step 1 of the algorithm.

The algorithm also has a choice of weighting the interval call or score by the genomic size. If this option is selected, each interval is weighted by the percentage of the chromosome it occupies. This allows the algorithm to reduce the contribution of small intervals in calculating similarity. By default this option is selected.

Step 3: Obtain the distance matrix

Central to the clustering technique are the following:

- The measure of how close any two arrays or samples are to each other. This is determined by the *Distance* or *Similarity* metrics.
- The measure of how close any two clusters are to each other. This is determined by the *Linkage* metric.

Similarity and Distance measures

Similarity between samples is based on the correlation between the aberration profiles of the samples. Aberration profiles are vectors of values reflecting either the aberration status or the aberration score for each sample. Each correlation measure takes two aberration profiles and produces a number representing how similar or dissimilar the two samples are. Measures of similarity range from values approaching -1 (meaning that the samples are exactly opposite in their aberration profiles) to 1 (exactly the same aberration profile). For a measure of distance, the result varies from 0 (the same aberration profile) to infinity (different aberration profile).

If an aberration profile for any given genomic interval for either sample is missing, the measurements from that interval is not considered in the calculation of the similarity of the two arrays. Genomic Workbench currently offers these distance-based measures:

- Manhattan Distance
- Euclidean Distance
- Similarity Distance (Normalized Euclidean Distance)

It also offers these similarity-based measures:

- Cosine Similarity
- Correlation (Pearson)

See "Cluster Analysis - Distance-based methods" on page 557.

Linkage methods

When calculating distances between individual samples, the specified similarity metric is used directly. However, when the distance calculation involves a cluster, or group of samples, there is a need to specify how to extend the similarity metric to cover groups. In Genomic Workbench, the following linkage methods can be specified:

- Single linkage
- Average linkage
- Complete linkage

See "Cluster Analysis - Linkage methods" on page 561.

Step 4: Create the clustering tree

Hierarchical clustering in Genomic Workbench is an iterative partitioning method that produces tree-like structures relating the resultant clusters by distance. These resulting trees group samples together based on the similarity of their aberration profiles (either the vector of the aberration calls or the aberration scores) and allows for identification of groups

containing similar profiles. Hierarchical clustering does not produce a set of discrete clusters. Instead, a tree structure, in which individual samples are grouped together at the leaves, is assembled with larger clusters forming the tree root. Genomic Workbench uses an agglomerative method of hierarchical clustering. This is achieved using the following steps:

1 Assign all *N* samples to *N* discrete, unique clusters.

2 Calculate pair-wise distances (using the specified similarity metric) between all *N* samples.

3 Group the two samples with the most similar aberration profiles into one cluster, resulting in *N-1* clusters.

4 Recalculate the pair-wise distances (using the specified similarity metric). These distances may be between formed clusters, any remaining individual samples, or a combination of the two. If any clusters are used in calculating distances, use the specified linkage method.

5 Repeat **Step 3** and **Step 4** until there is one cluster consisting of *N* samples.

Copy Number Variation Region

CNVR analysis identifies genomic regions across multiple samples that have significant copy number variation. The CNVR algorithm uses output from an aberration detection algorithm for multiple samples, and generates a list of aberrant intervals that overlap. It joins the intervals together into a putative CNV region.

The algorithm differs from Common Aberration analysis and interval-based penetration analysis in that it computes combined neighboring overlap regions based on intervals discovered through aberration analysis.

Algorithm

This analysis joins together all overlapping aberrations into a larger region. The algorithm has these steps:

1 Collect all intervals discovered through aberration analysis.

2 For each interval *I*, find all intervals across all of the samples which overlap (even by one base pair) the genomic locations of *I*. Define that union of overlapping aberrant intervals as a CNVR. See Figure 211.

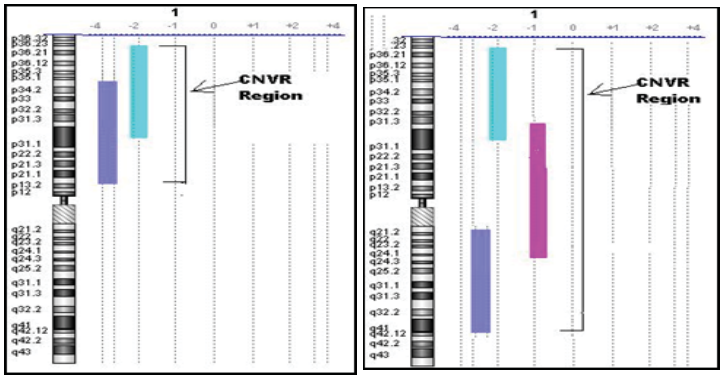


Figure 211 Construction of CNVR regions from aberration intervals.

Interpretation The CNVR analysis uses the output from the detection algorithms. CNVR regions are the union of any overlapping aberrant regions across multiple samples.

Interpretation and visualization of the algorithms for the comparison of commonly aberrant regions

Output from the comparison of common aberrant regions is generally visible in the algorithm-specific result windows. This section describes the interpretation of the visualization capabilities in Genomic Workbench for these analysis algorithms.

Differential Aberration analysis

Interpretation

For each candidate genomic interval, the differential aberration method identifies those regions with statistically significant enrichment of gains or losses in one of the group of samples. The tool then allows one to select the regions of interest and save them for further investigation.

For samples which are naturally grouped into two groups by some characteristic then this tool can detect the aberrant frequencies in this partition and assign statistical significance to the differences. For example, assume you have two groups named *A* and *B*, both consisting of 100 samples. If group *A* has 40 aberration amplifications (called gains) in a specific genomic region and group *B* has 5 gains in that same region, the statistical significance of observing the 40 gains in group *A* given observing only 5 gains in group *B* is calculated by using *p-values* from the Hypergeometric distribution. In this example, the *p-value* of observing 40 gains in group *A* is 1.3×10^{-9}, meaning that group *A* has a significant enrichment of gains compared to group *B*.

Visualization

The output of the differential aberration analysis is a graphical and tabular output which allows visual inspection of the aberration amplification (colored red) and deletion (colored green) $-log_{10}(p\text{-values})$. The topmost pane of the Graphical Differential Aberration Summary window is a line plot of chromosomal location by $-log_{10}(p\text{-value})$. The enrichment values comparing the first group to the two combined groups are on the left and the enrichment values from the second group is on the right. Below the line chart is a graphical representation of the chromosomes under consideration from the comparative groups. The top row of the graphical representation of chromosomes belongs to first group and the bottom row belongs to the second group. A scroll bar at the bottom allows aberration profiles of that chromosome from all samples to be displayed. The bottom-most pane of the Summary window contains the table lists the four comparative enrichment *p*-value scores of the candidate intervals chosen from one of the aberration detection

algorithms. In addition, the chromosome, the chromosomal position, and the interval size of the candidate intervals is displayed. Clicking on an interval in the interval table switches the graphical view to display the chromosome to which the selected interval belongs. On that chromosome, a blue line is superimposed to show the location (and relative size) of the selected interval. See Figure 212.

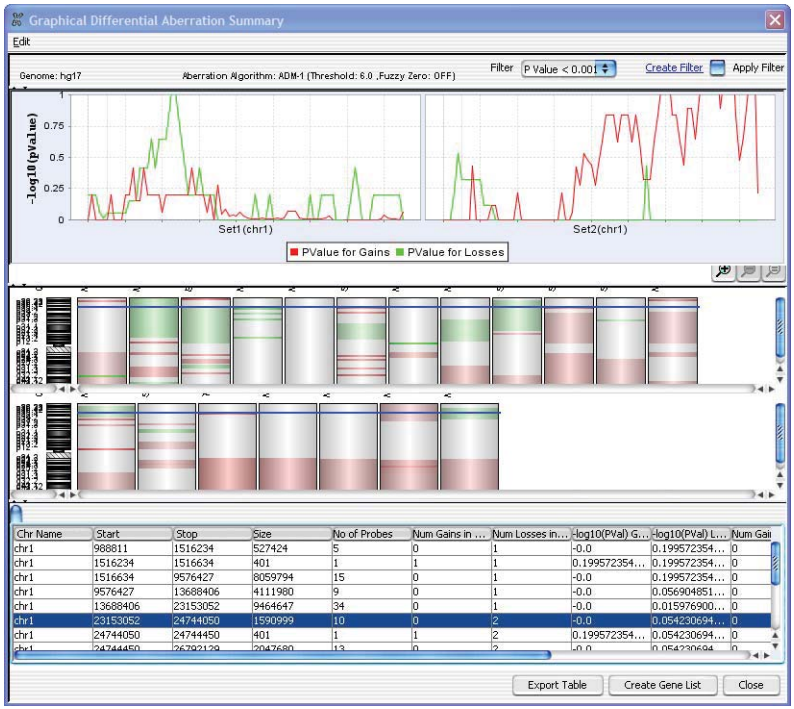


Figure 212 Per-chromosome results of applying the Differential Aberration algorithm.

Filters can be created and applied to broaden or narrow the range of differential intervals displayed based upon *p*-value or other thresholds. For example, if you are not interested in detecting short differentially aberrant intervals, you can apply a filter (like one shown below in Figure 213) to remove all intervals that contain less than a certain number of probes. In this example, a custom filter has been created by using the attribute drop-down menu to select “No of Probes” as a filter criteria. The greater-than inequality is used as an operator on the target value 10. Additional attributes, or filtering criteria, can be added as new

conditions, and the results can either display those results passing the filter, or those specifically not passing the filter. See Figure 213. See "Interval Filter" on page 361.

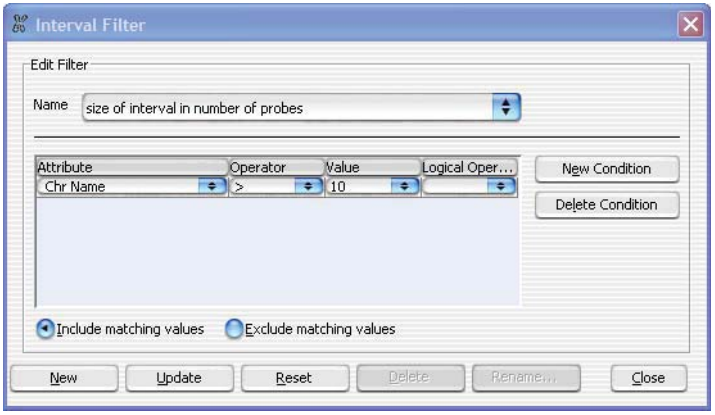


Figure 213 Creating a custom filter based on the number of probes in an interval for Differential Aberration analysis.

Applying either the default or a custom filter will change both the *p*-value plot for gains and losses and restrict the interval table to those intervals which pass the filter. The graphical representation of gains and losses in the chromosome diagrams will, however, remain unfiltered. This is useful for identifying regions that did not pass a filtering threshold. See Figure 214 for an example output of Differential Aberration analysis after applying the filter shown in Figure 213.



Figure 214 Per-chromosome results of applying the Differential Aberration algorithm with a custom filter to remove intervals containing less than 10 probes.

Cluster analysis

Interpretation

The results of hierarchical clustering on samples are displayed both as a tree and as a heatmap in the Clustering Results window. The Hierarchical tree can be cut at any desired level of clusters and a report generated for further exploration. For example, to partition the tree into two clusters, the tree is cut at level 2 (see Figure 215). The clustering results can be exported as a report. The export report tree is cut at the level indicated (if applicable) in the Clustering Results window. Cluster assignments are based on that cutoff level. Additionally, the entire tree and heatmap visualization can be copied to the clipboard and saved.

Visualization

Both a report and a Clustering Results window are available from application of Cluster analysis. The Clustering Results window consists of the following visualizations of the data:

- **Hierarchical clustering tree**

- **Aberration profile heatmap**
- **Aberration profile plot**

See Figure 215 for an example Cluster analysis output in the Clustering Results window.



Figure 215 Clustering Results window with tree cutting illustrated at level 2.

The left panel of the Clustering Results window shows the hierarchical tree obtained from clustering. Samples in which aberration profiles are more similar to each other will be shown closer together in the tree, and the branching structure shows where samples and groups of samples were joined during the clustering steps. By default, the “bottom” of the tree is shown on the right, where all samples exist as independent objects. Moving to the left shows the samples joined into larger branches until all the samples are joined. By selecting a node box within the tree diagram you can visualize all children nodes and their aberration profiles in the heatmap view.

The right panel of the Clustering Results window shows a false-color map of the aberration profiles (consisting of either the aberration status or aberration score values) of each node descending from the selected node in the tree view. Each sample is displayed along the y-axis and the chromosomal position of the aberrant intervals is displayed along the x-axis. The false-color map is colored by the aberration score where increasing luminosity in the red color spectrum range corresponds to increased values of amplifications, and increasing luminosity in the green color spectrum range corresponds to increased values of deletions.

The heatmap displays sample names as tooltips; clicking on a heatmap row corresponding to any sample displays the aberration profile for that sample in a line plot at the bottom of the Clustering Results window. Clicking again on the profile collapses the plot.

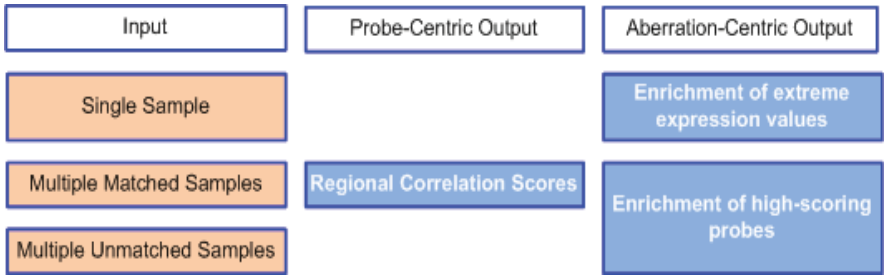
Algorithms for Joint CGH-Expression Analysis

Agilent Genomic Workbench provides tools for joint analysis between aberrant region detection and output from expression microarray experiments. These analysis tools allow for correlation between CGH and expression arrays and enrichment analysis based upon expression levels from genes found in an aberrant or CNV region.

It is often useful to compare DNA region variations with mRNA levels. This comparison yields information about the gene expression effects of the DNA variation and helps to identify aberrations and CNV regions which have a large impact on the function of the cell.

CGH arrays normally have a more uniform distribution of probes along the genome than expression arrays. This is due to the focus of each type of microarray. CGH arrays measure DNA aberrations, which may span both coding and noncoding genomic regions, while expression array probes fall mainly within known coding regions. As a result, joint analysis only takes into account those regions which have corresponding signal between the two array types. In this way, probes on each array don't have to match precisely; an aberrant region will enclose signal from probes on an expression array.

The following table shows the type of analysis this software provides for performing a joint analysis of CGH arrays and corresponding Gene Expression arrays.



Enrichment of extreme expression values in a single sample

CGH-Expression Analysis can be computed for any sample in which both CGH and expression are measured. Genomic Workbench computes an enrichment of the number of probes (genes) with high log ratios in a given region. The enrichment is a measure of the increase in the number of probes (genes) with a high log ratio as compared to what may be expected at random. The output is a list genomic intervals for each sample with a statistical score showing the enrichment of over- and under-expressed probes in that interval. This analysis can be performed either on all aberrant intervals or on the whole genome.

NOTE When probes have multiple representatives in the expression data, these multiples have to be unified into a single representative for each probe before performing an enrichment analysis. You must choose Intra Array replicate combine for those expression arrays.

Algorithm The following outlines the steps in this analysis.

Consider a single sample, S:

1. Rank all genes according to their expression values in S: g_1 (the highest level of expression in S), g_2, ... , g_n (the lowest level of expression in S).
2. Consider a genomic interval, I. This may be an aberrant region or the entire genome.
3. Compute the enrichment of over-expressed genes in I using the following process:
   - **a** Compute an occurrence vector v. This vector contains a place for every ranked probe g_1, ... , g_n within the genomic region I. For every place in the vector, if the log ratio value is over- or under-expressed, a value of 1 is generated. Otherwise a value of 0 is generated.
   - **b** Let B equal the number of 1s present in the occurrence vector.
   - **c** Compute the mHG score [1, 2] for the vector v. The mHG score represents the rank imbalance of genes from I in the ranked list g_1, ... , g_n:

$$(*) \quad mHG(I) = mHG(v) = min_{1 \leq i \leq n} HGT(b_i(v); n, B, i) \qquad \textbf{(41)}$$

where:

HGT is the probability of observing B or more 1s in the occurrence vector, given by the hypergeometric distribution:

$$HGT(b;n;B,\ i) = \sum_{k=b}^{min(i,\ B)} \frac{\binom{i}{k}\binom{n-i}{B-k}}{\binom{n}{B}} \quad \textbf{(42)}$$

r is a hard-coded boundary on the threshold optimization range, and:

$$b_i(v) = \sum_{k=1}^{i} v(k) \quad \textbf{(43)}$$

4 Using a symmetric procedure, rank the probes in reverse order. Repeating steps 2 and 3 from above yields intervals enriched with under-expressed probes.

5 Report intervals I with significant mHG(I), as determined by comparing to the user-defined threshold. Since the intervals are reported together with -log(mHG(I)), large numbers represent significant enrichment events.

Regional correlation scores with multiple-matched CGH and Expression samples

These scores are computed for all probes for which expression was measured and there is a well-defined known genomic location.

NOTE For this analysis the designs for the CGH and the expression arrays need not be the same, but the sample set must contain at least six pairs of matching CGH and EXP arrays.

It is assumed that a sample-by-sample aberration calling was run. The output will be transformed CGH data that, for each sample S and each genomic location x, there is an average value of log(R/G) signals in the interval that spans x, as determined in the aberration calling step. This

processed CGH data is maintained in a matrix, denoted by *M*. which is in fact, the basis for the aberration summary graphics. Rows of *M* represent genomic locations and columns represent samples.

Consider a probe *p*, and the vector of log ratios of *p* measured over the entire set of samples: *LE = LE(p)*.

Now, consider a row of *M* that corresponds to the genomic location *g* of probe *p*. This row forms a vector of DNA copy numbers at *g*, as denoted by *C = C(g)*.

The correlation between expression and DNA copy number at *g* is evaluated using two different scores:

Student-t-test approach:

1. Let *P* be the set of samples for which a region spanning g is called amplified.
2. Let *Q* be the set of samples for which a region spanning g is called deleted.
3. Let *U* denote the entire set of samples.
4. Calculate μ_P as the average expression value of g, for samples in *P*
5. Calculate μ_Q as the average expression value of g, for samples in *Q*
6. Calculate μ_{U-P} as the average expression value of g, for samples in *U-P* (the complement of *P*, or those samples in *U* but not *P*)
7. Calculate μ_{U-Q} as the average expression value, of g, for samples in *U-Q* (the complement of Q, or those samples in *U* but not *Q*)
8. Calculate the Student-t statistics and the corresponding p-values for the one-sided alternative hypotheses:
   a. The number of $\mu_P > \mu_{U-P}$ (where *g*'s expression levels in amplified samples are larger than those in non-amplified samples).
   b. The number of $\mu_Q > \mu_{U-Q}$ (where *g*'s expression levels in amplified samples are larger than those in non-amplified samples).

Pearson correlation approach:

The Pearson correlation of *C* and *E* is computed and a corresponding p-value is generated. This is a measure of the similarities between the cartesian coordinates of the two vectors.

Both scores are reported for every probe, *p*. Output formats include graphical output as well as a text report reflecting the bottom panel table.

Enrichment of CGH/Expression correlation

This can be computed for a set of samples for which both CGH and expression were measured. The output consists of a list of genomic intervals and a statistical score of the enrichment of CGH/expression correlation in each interval.

NOTE When genes have multiple representatives in the expression data these have to be unified into a single representative for each gene prior to performing the enrichment analysis.

Consider a set of samples for which regional correlation was computed as described above.

1. Rank all probes according to their correlation statistical score: g_1 (the most significant regional CGH/expression correlation), g_2, ... , g_n (the least significant regional CGH/expression correlation).
2. Consider a genomic interval, I. This may be an aberrant region or the entire genome.
3. Compute the enrichment of over-expressed probes in I using the following process:
   - **a** Compute an occurrence vector v. This vector contains a place for every ranked probe g_1, ... , g_n within the genomic region I. For every place in the vector, if the log ratio value is over- or under-expressed, a value of 1 is generated. Otherwise a value of 0 is generated.
   - **b** Let B equal the number of 1s present in the occurrence vector.
   - **c** Compute the mHG score [1, 2] for the vector v. The mHG score represents the rank imbalance of genes from I in the ranked list g_1, ... , g_n:

$$(*) \quad mHG(I) = mHG(v) = min_{1 \le i \le gn} HGT(b_i(v); n, B, i) \tag{44}$$

where:

HGT is the probability of observing B or more 1s in the occurrence vector, given by the hypergeometric distribution:

$$HGT(b; n; B, i) = \sum_{k=b}^{min(i, B)} \frac{\binom{i}{k}\binom{n-i}{B-k}}{\binom{n}{B}} \tag{45}$$

r is a hard-coded boundary on the threshold optimization range, and:

$$b_i(v) = \sum_{k=1}^{i} v(k) \quad \textbf{(46)}$$

4 Using a symmetric procedure, rank the probes in reverse order. Repeating steps 2 and 3 from above yields intervals enriched with under-expressed probes.

5 Report intervals I with significant mHG(I), as determined by comparing to the user-defined threshold. Since the intervals are reported together with -log(mHG(I)), large numbers represent significant enrichment events.

Enrichment of external gene annotation

You can compute this annotation for any external quantitative information about genes. Samples do not have to be matched. The output consists of a list of genomic intervals and, for each, a statistical score of the enrichment of the external quantitative annotation in that interval.

Example The external quantitative annotation can be differential expression scores, between tumor and normal samples, as measured in an independent study.

NOTE When genes have multiple representatives in the expression data these have to be unified into a single representative for each gene prior to performing the enrichment analysis.

1 Rank all genes according to their external quantitative annotation: g_1 (highest value, e.g. most over-expressed or most under-expressed in tumor), g_2, ... , g_n.

2 Consider a genomic interval, I. This may be an aberrant region or the entire genome.

3 Compute the enrichment of over-expressed probes in I using the following process:

   a Compute an occurrence vector v. This vector contains a place for every ranked probe g_1, ... , g_n within the genomic region I. For every

place in the vector, if the log ratio value is over- or under-expressed, a value of 1 is generated. Otherwise a value of 0 is generated.

b Let B equal the number of 1s present in the occurrence vector.

c Compute the mHG score [1, 2] for the vector v. The mHG score represents the rank imbalance of genes from I in the ranked list g_1, ... , g_n:

$$(*) \quad mHG(I) = mHG(v) = min_{1 \leq i \leq gn} HGT(b_i(v);n, B, i) \tag{47}$$

where:

HGT is the probability of observing B or more 1s in the occurrence vector, given by the hypergeometric distribution:

$$HGT(b;n;B, i) = \sum_{k=b}^{min(i, B)} \frac{\binom{i}{k}\binom{n-i}{B-k}}{\binom{n}{B}} \tag{48}$$

r is a hard-coded boundary on the threshold optimization range, and:

$$b_i(v) = \sum_{k=1}^{i} v(k) \tag{49}$$

4 Using a symmetric procedure, rank the probes in reverse order. Repeating steps 2 and 3 from above yields intervals enriched with under-expressed probes.

5 Report intervals I with significant mHG(I), as determined by comparing to the user-defined threshold. Since the intervals are reported together with -log(mHG(I)), large numbers represent significant enrichment events.

Algorithms for CGH+SNP Analysis

Agilent Genomic Workbench supports the analysis of Agilent CGH+SNP microarrays, which combine both CGH probes and SNP probes on the same array. Considered together, the data from these probes allow for the determination of SNP allele-specific copy number (ASCN), the assignment of SNP genotypes, and the identification of regions of loss or lack of heterozygosity (LOH). These results can be useful in constitutional cytogenetic studies.

ASCN (SNP CN) – Allele-specific copy number detection algorithm

The Agilent CGH+SNP platform distinguishes the two alleles of a SNP by whether or not the SNP site is cleaved by the AluI/RsaI restriction enzyme mixture that is used during the sample labeling process. One allele (the "cut" allele) is cleaved by the enzymes, binds poorly to the corresponding SNP probe(s), and produces very little signal. The other allele (the "uncut" allele) is not cleaved by the enzymes, binds strongly to the corresponding SNP probe(s), and produces a high signal level. See Figure 216.

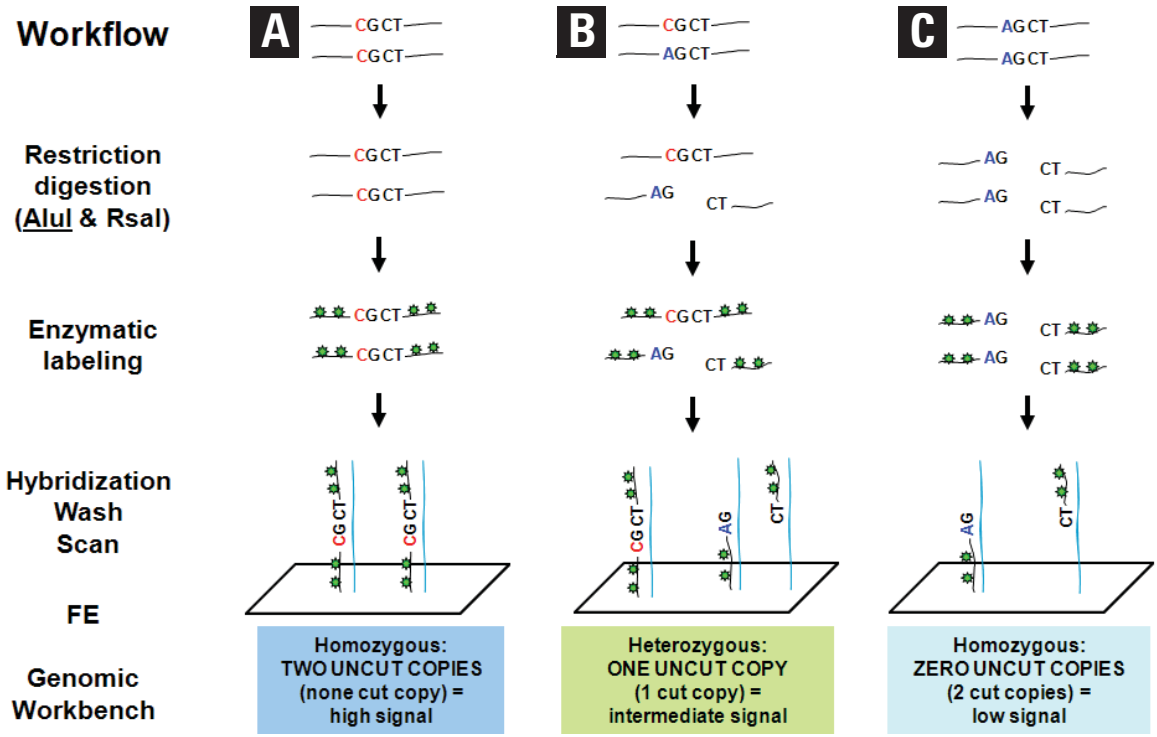


Figure 216 The Agilent CGH+SNP microarray workflow, with three possible cases for SNPs shown. In each of the three possibilities, homologous regions of two chromosomes are shown, with the bases of the SNPs in red and/or blue. **A** – Neither of the SNP sites is cut by AluI or RsaI, which lets the greatest amount of labeled material hybridize with the probes on the microarray. This yields the highest signal level. **B** – One of the SNP sites is cut by AluI or RsaI, which yields an intermediate (half) signal level. **C** – Both of the SNP sites are cut by AluI or RsaI, which yields the lowest signal level.

The ASCN detection algorithm finds the most likely copy number of the "uncut" SNP allele for each SNP that is interrogated on an Agilent CGH+SNP array. SNP calls are made from the $\log_2$ ratios of the sample signal versus the signal from a genotyped internal reference, which compensates for labeling and hybridization bias. Since the reference genotype is known, the raw $\log_2$ ratios are "reference adjusted" to values that reflect a hypothetical ASCN of 2 for the reference at all SNP sites. The details of these steps are described later in this section.

Allele-specific copy number values are used by the LOH detection algorithm. See "LOH (Loss or lack of heterozygosity) algorithm" on page 544. The output of the algorithm is also used to deduce the genotypes of SNP sites. See "Assignment of SNP genotypes" on page 541.

When you use the ASCN algorithm, you must also use Centralization, which is applied to the CGH probes on the array. See "Centralization Algorithm" on page 452.

Algorithm

The ASCN algorithm uses several main inputs:

- The $\log_2$ ratio values for all SNP probes on the array.
- CGH copy numbers, which are internally calculated expectation values for the copy numbers of the genomic regions including each CGH probe on the array.
- A file that contains information about the individual genotyped reference. For each SNP, this file includes the SNP ID, the SNP genotype at the site (*i.e.* the number of copies of each variant allele – see note below), and information about which SNP allele is cut by the AluI/RsaI enzyme mixture that is used in the genomic DNA labeling process. Agilent Genomic Workbench supports five standard HapMap references, as well as custom references. If you have a custom reference file, you can import it. You can also use Agilent Genomic Workbench to construct a custom reference file based on a prior experiment where the supported HapMap reference samples have been hybridized with the custom reference on the CGH+SNP microarray. See "To generate a genotype reference file" on page 164.

NOTE

Traditionally, the genotypes of SNPs are represented as two-character strings (*e.g.* "AA" or "AG."). However, when amplifications are identified, SNP genotypes will have additional characters correspondent to the extra alleles (e.g. "AAG" or "AGG".)

The algorithm constructs the distribution of $\log_2$ ratio values for all SNP probes on the array, and finds the peaks of the distribution. It fits a separate Gaussian distribution to each peak, and uses a Bayesian model to calculate an expectation value for the uncut SNP allele copy number at each SNP.

Steps

Step 1 Reference correct $\log_2$ ratios

Since the reference genotype is known, the raw $\log_2$ ratios are "reference adjusted" to values that reflect a hypothetical ASCN of 2 for the reference at all SNP sites. SNPs with zero copies of the uncut allele in the reference sample are ignored in this step, and the algorithm subtracts 1 from the $\log_2$ ratios of SNPs with one copy in the reference sample.

NOTE

The correction that is applied to the $\log_2$ ratios of SNPs with one copy in the reference sample is not precisely 1.0. It is modified by the measured log ratio compression, namely the Separability QC metric.

Step 2 Find peaks in distribution

In this step, the algorithm constructs the distribution of $\log_2$ ratio values for all SNP probes on the array, and identifies the peaks in the distribution. It then assigns an uncut allele copy number to each peak. Typically, the assigned copy numbers range from 0 to the measured total copy number in the general region that includes the SNP site.

Step 3 Fit Gaussians to peaks

In this step, the algorithm uses an iterative method to fit separate Gaussian distributions to each of the peaks in the overall distribution.

Typical results from the first three steps of the algorithm appear in Figure 217.

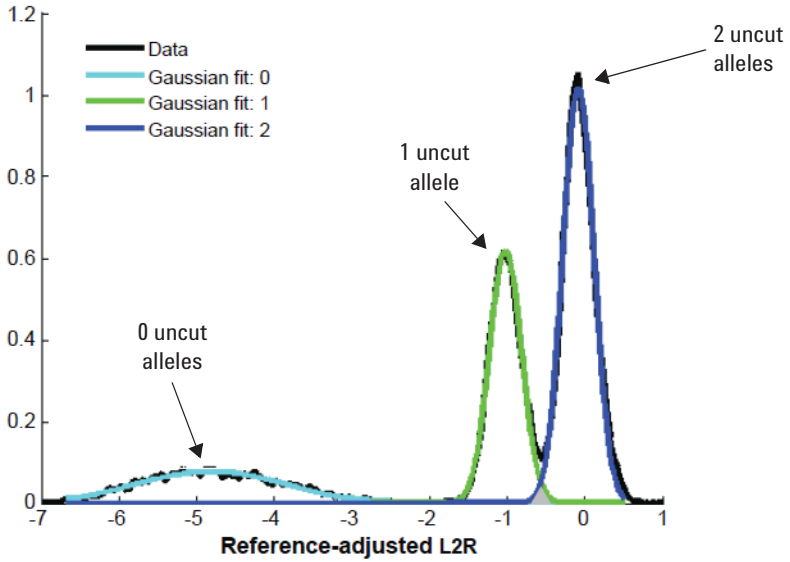


Figure 217 Example distribution of reference-adjusted $\log_2$ ratio values for SNP probes on an array (black). Gaussian distributions (blue, green, cyan) have been fit to each of the three peaks. Each peak, and thus each Gaussian distribution, has been assigned an uncut SNP allele copy number.

Step 4 Compute Bayesian likelihoods

In this step, the algorithm uses a Bayesian method to compute the likelihood that each SNP site has each possible number of uncut copies. The algorithm defines three terms for every possible combination of copy number (CN) and distinct $\log_2$ ratio value (LR) in the distribution:

- The prior probability of observing each $\log_2$ ratio value. This value, called P(LR), is obtained directly from the distribution of $\log_2$ ratio values.
- The probability that the specific $\log_2$ ratio value under consideration would be obtained given the specific copy number. This value, called P(LR|CN), is computed from the fitted Gaussian distribution that corresponds to each copy number.
- The prior probability that a given SNP has the copy number under consideration. This value, P(CN), is the area under the Gaussian curve that corresponds to each copy number.

Based on these terms, the algorithm uses Bayes's rule to calculate the likelihood that the uncut alleles would exist in each given copy number:

$$P(CN|LR) = \frac{P(CN) \cdot P(LR|CN)}{P(LR)} \quad \textbf{(50)}$$

Step 5 Adjust for outliers

In this step, the algorithm filters out outliers. Sometimes the measured log ratios of the probes violate the assumptions of the model. In particular, some probes very far from the fitted peaks have P(LR) very close to zero, which causes the model to produce a very high, spurious result (a division by zero in Equation 50). Also, since the distribution for CN=0 is much wider than the other distributions, probes reporting high $\log_2$ ratio values sometimes have a P(CN=0) that is higher than their P(CN=2). Both of these types of outliers are assigned an ASCN by linear interpolation, rather than by the Bayesian Equation 50.

Step 6 Rescue probes

Previously, in Step 1, the algorithm excluded data from probes targeting doubly-cut SNP sites of the known genotyped reference. In this step, the algorithm rescues the genotypes of some of the probes that were previously excluded.

SNPs can often be assigned an ASCN of zero based on the probe intensity alone, even when (as is the case when the ASCN of the reference sample is zero) the log ratio cannot be interpreted. Gaussian curves are fitted to the sample $\log_2$(signal) levels of SNP probes already assigned to ASCN=0 or ASCN<>0. Unassigned SNP probes which can be confidently assigned to the ASCN=0 distribution, based on their signal alone, are assigned a likelihood of having a copy number of zero. The confidence value of the probe is then reevaluated based on the likelihood that it represents a copy number of zero.

Step 7 Calculate ASCN expectation values

In the final step, the algorithm uses the likelihood values to calculate an expectation value, $\langle C_p \rangle$, for the uncut allele copy number for each SNP. SNPs whose maximum likelihood exceeds a user-specified threshold (0.95 recommended) are assigned a genotype based on the maximum likelihood ASCN. See "Assignment of SNP genotypes" on page 541. If no ASCN exceeds the likelihood threshold, the SNP genotype is not reported.

Visualization

When you apply the ASCN algorithm, ASCN assignments appear in the SNP CN panel in Gene View and in Chromosome View. The calculated expectation values for the ASCN of the SNPs are plotted against chromosomal location. A typical SNP array represents many different SNPs that cover all states of uncut SNP allele copy numbers. Because of this, separate bands of points are seen in the SNP CN panel for each possible uncut SNP allele copy number. See Figure 218.

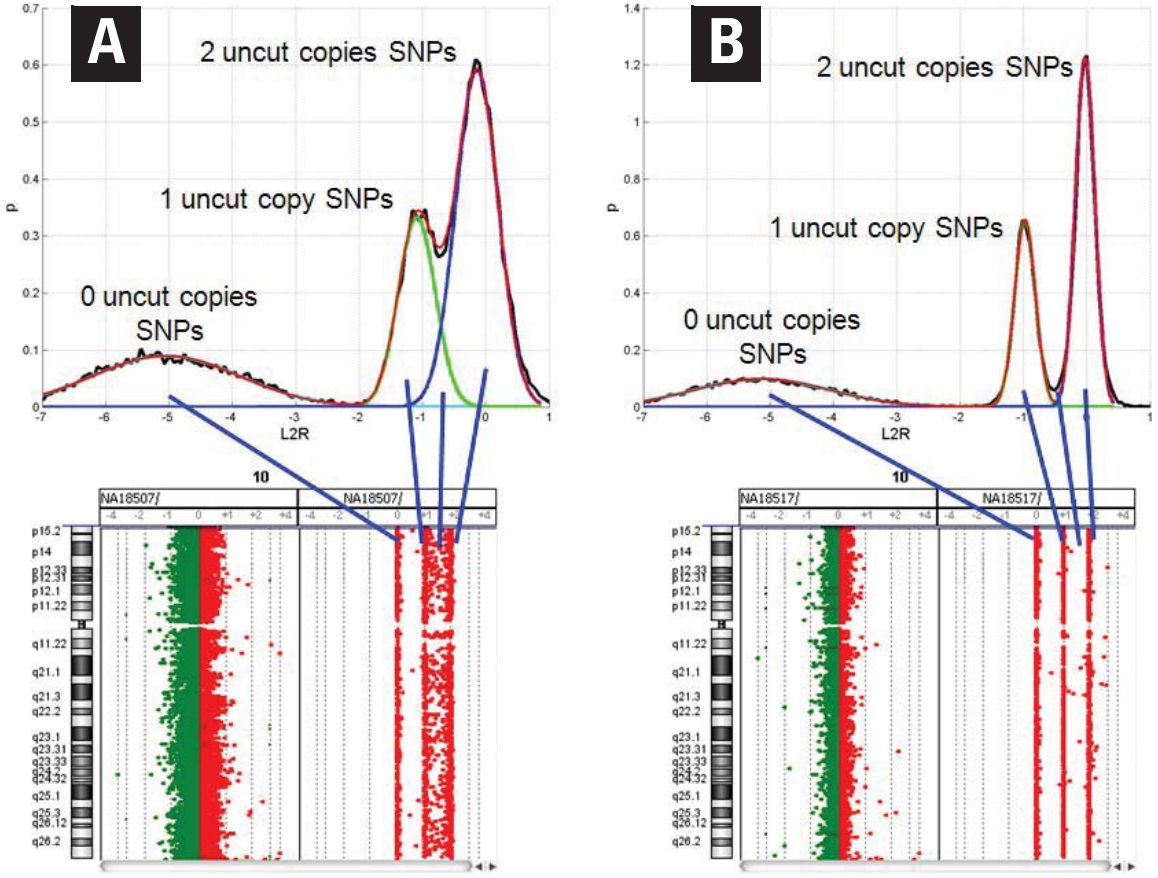


Figure 218 Two examples of SNP copy number calls. **A** – Upper plot: Copy number distribution plot for a sample with a poor derivative log ratio spread (DLRS = 0.33). Note how the peaks for copy numbers of 1 and 2 are not well-differentiated. Lower plot: Gene View. The right panel is the SNP CN panel. ASCN expectation values for probes are plotted against chromosomal location. Note how allele-specific copy number calls are not well defined, especially for copy numbers of 1 and 2.
B – Upper plot: Copy number distribution plot for a sample with a good derivative log ratio spread (DLRS = 0.15). Note the well-separated peaks for all copy numbers. Lower plot: Gene View. The right panel is the SNP CN panel. ASCN expectation values for probes are plotted against chromosomal location. Well-separated bands of points are seen for each uncut SNP allele copy number.

Assignment of SNP genotypes

The program reports the genotypes of the SNP sites targeted by SNP probes. To do this, it uses SNP allele information from the reference file, and expectation values for CGH copy number and ASCN. Examples of the logic that the program uses appear in "Example 1" on page 542 and "Example 2" on page 543.

To deduce the genotypes of SNP sites, the algorithm requires integer values for CGH copy numbers and ASCNs. To obtain these values, the algorithm rounds the measured expectation values for CGH copy number and ASCN (both of which can contain fractional values) to the nearest integer. Copy numbers are excluded from the analysis if the amount of rounding that is required is too great.

To assess the amount of rounding, and to decide whether or not to include a given copy number value in the SNP genotype analysis, the algorithm uses a threshold value that is entered by the user. It includes a given copy number value in the analysis if the following relation is true:

$$|\text{ASCN} - \text{rounded ASCN}| < (1 - \text{threshold value})$$

If the CGH copy number value and/or the ASCN value for a given region are not reported, the algorithm reports both alleles as "N" (not determined).

For genotype assignment and for visualization, you can set a threshold value for the confidence level for ASCN calls (0.95 recommended).

NOTE For the LOH algorithm, all SNP probes are considered without regard to the confidence threshold that you set. This threshold applies only to the visualization of copy number calls in the SNP CN panel, and to genotype assignment.

Example 1 In the example that appears in Figure 219, a given reference SNP site has a genotype of CT (a C base as one allele, and a T base as the other). Further, it is known that the C allele is the one that remains uncut by the AluI/RsaI enzyme mixture during sample preparation. If, for the sample, the CGH copy number for the region is 2, and the uncut allele copy number (ASCN) for the site is 2, the SNP genotype for the sample is CC.

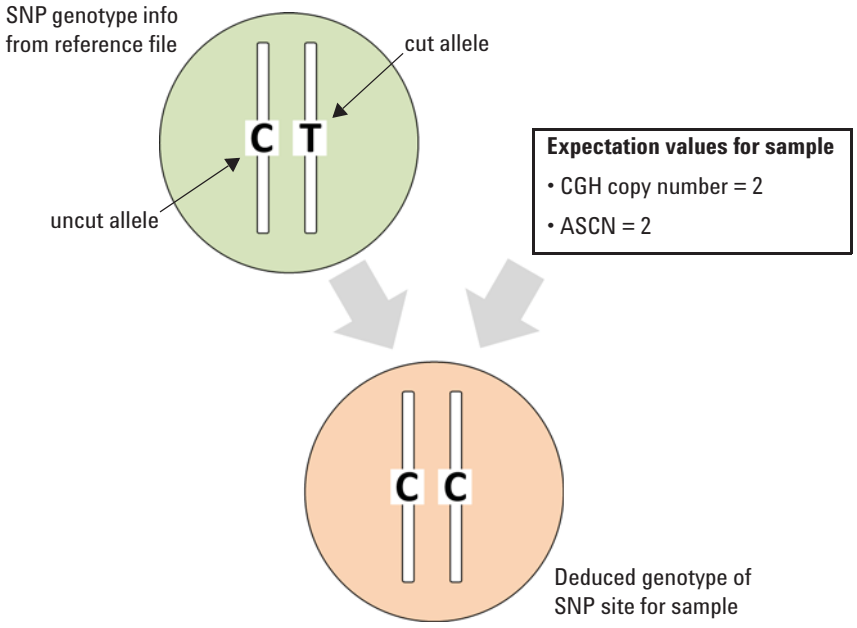


Figure 219 Reference genotype and deduced sample genotype at a specific SNP site. The CGH copy number of 2 indicates a total of two alleles in the sample at the SNP site, and the ASCN of 2 indicates that two alleles are the "uncut" C allele.

Example 2 As in the previous example, the example that appears in Figure 220 shows a genotype of CT for the reference SNP site, and the allele that remains uncut is C. For the sample, the CGH copy number is 4, and the uncut allele copy number (ASCN) is 2. The sample thus has two C alleles. Because SNPs are typically bi-allelic, the other two alleles must be the T allele.

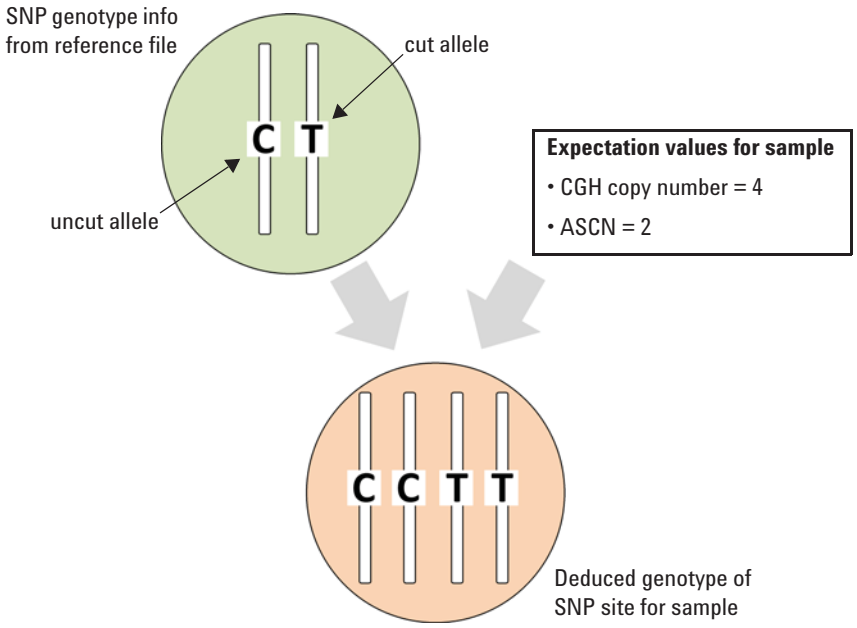


Figure 220 Reference genotype and deduced sample genotype at another specific SNP site. The CGH copy number of 4 indicates a total of four alleles in the sample at the SNP site, and the ASCN of 2 indicates that two of the alleles in the sample are the "uncut" C allele.

NOTE The new dbSNP databases record a small number of SNPs with more than two different alleles. The Agilent assay still works for these sites, with some reinterpretation.

You can inspect or use the reported genotypes in several ways:

- The reported genotypes of SNP sites for the selected experiment appear in the Genotype column in the SNPs tab of Tab View. See "Tab View" on page 233.
- You can create a SNP Genotype Report that contains a summary of analysis parameters and list(s) of reported SNP genotypes. See "To generate a SNP Genotype Report" on page 167.
- You can create a custom genotype reference file that you can import into Agilent Genomic Workbench to use as a reference for future SNP analyses. See "To generate a genotype reference file" on page 164.

LOH (Loss or lack of heterozygosity) algorithm

For the SNP sites that are interrogated by Agilent SNP probes on a CGH+SNP array, the LOH algorithm identifies genomic regions with a statistically significant scarcity of heterozygous SNP calls to discover regions of copy-neutral LOH and uniparental disomy (UPD). Identification of such regions can be used to inform constitutional cytogenetic studies.

The algorithm uses total and allele-specific copy numbers to label each SNP site as homozygous or heterozygous. It then uses a binomial probability distribution to report regions that contain an unusually high fraction of homozygous SNPs.

When you use the LOH algorithm, you must apply the Centralization algorithm. GC Correction is recommended. See "Centralization Algorithm" on page 452 and "GC Correction Algorithm" on page 455. You must also apply the ASCN algorithm, since the LOH algorithm requires ASCN calls as input. See "ASCN (SNP CN) – Allele-specific copy number detection algorithm" on page 534.

The algorithm reports LOH for the regions that are also detected as deletions in standard CGH analysis (see Figure 221). Amplified regions, while not detected as LOH, are visible in the results of CGH aberration analysis, and can also be seen in the SNP CN pane (see Figure 222). Regions of copy-neutral LOH, such as those that arise from UPD and parental consanguinity, are detected by the LOH algorithm, but are not detected by CGH analysis. See Figure 223 and Figure 224.

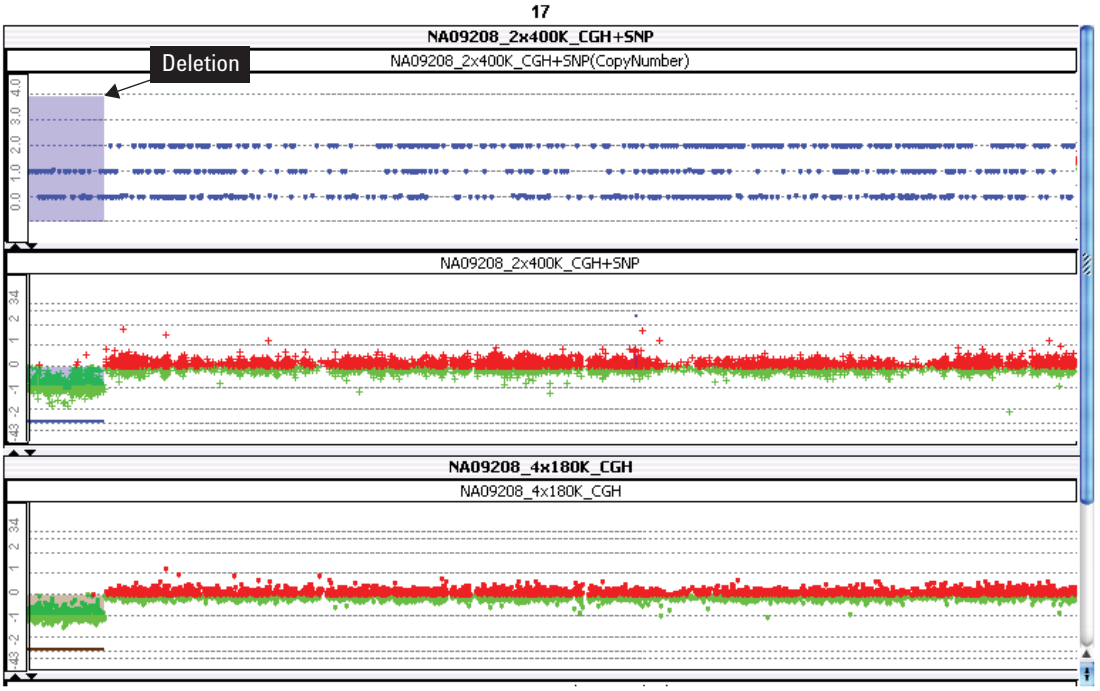


Figure 221 Hemizygous deletion in chromosome 17. **Upper pane** – Deletion identified by the LOH algorithm (purple shaded area) and ASCN calls for SNP probes (blue dots). In the region of the deletion, note the significant absence of probes that correspond to a copy number of 2. **Middle pane** – Log_2 ratio data for the CGH+SNP array. The same deletion is detected by aberration analysis of the CGH probes on the array. **Lower pane** – Log_2 ratio data for the same sample on a standard CGH array. Again, the same deletion is detected by CGH aberration analysis.

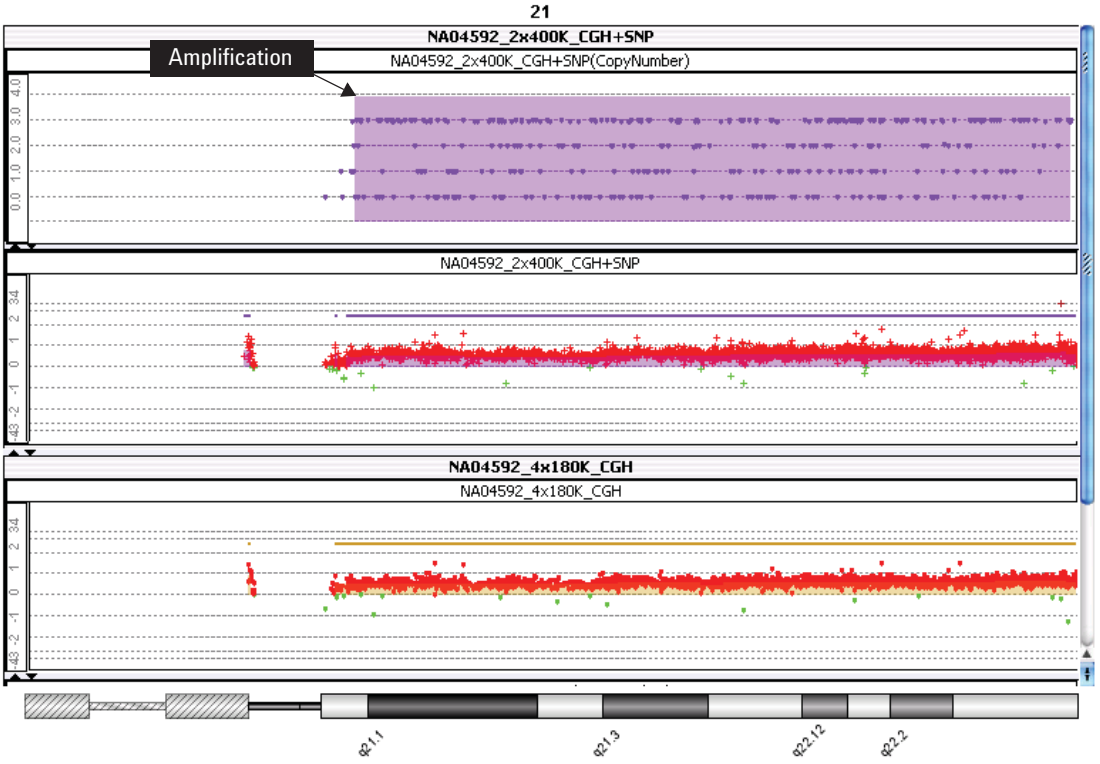


Figure 222 Amplification (trisomy) on human chromosome 21 (purple shaded area). **Upper pane** – ASCN calls for SNP probes (purple dots). Note the significant band of probes that corresponds to a copy number of 3. **Middle pane** – Log_2 ratio data from the CGH+SNP array. The same amplification is detected by aberration analysis of the CGH probes on the array. **Lower pane** – Log_2 ratio data for the same sample on a standard CGH array. Again, the same amplification is detected by CGH aberration analysis.

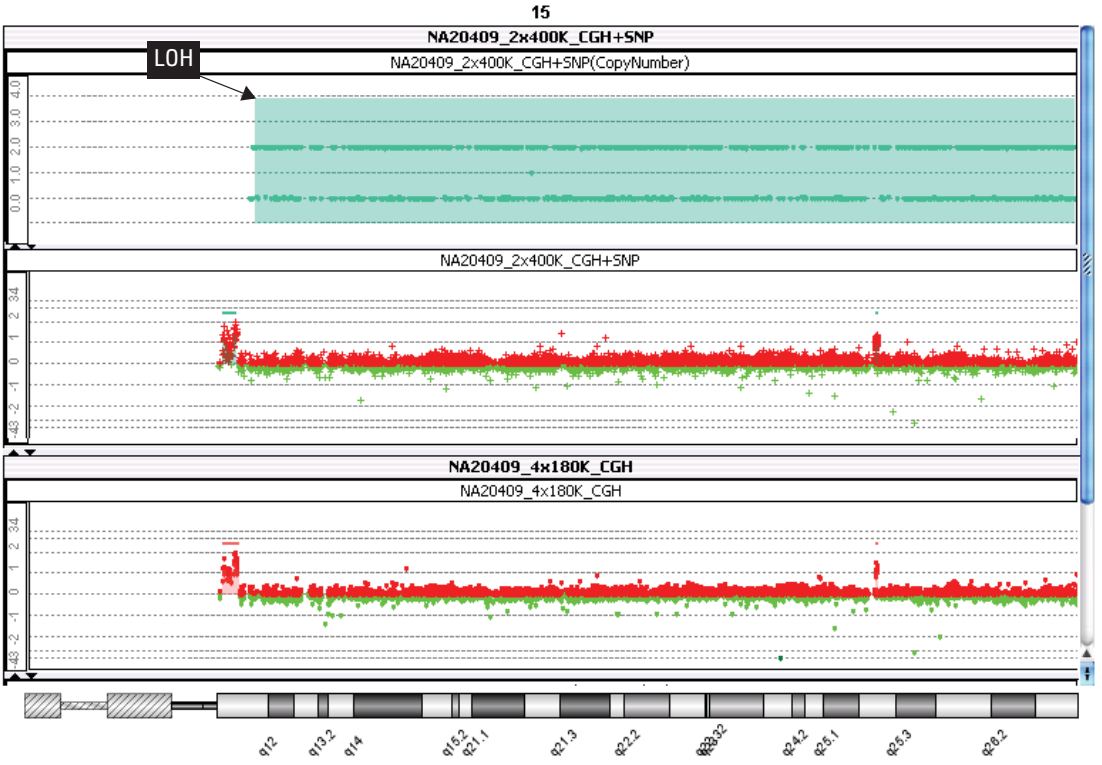


Figure 223 LOH caused by UPD of the entire chromosome 15. **Upper pane** – ASCN calls for SNP probes (green dots). In the region of the detected LOH (green shaded area), note the lack of probes that correspond to a copy number of 1. **Middle pane** – Log_2 ratio data from the CGH+SNP array, with an aberration analysis of the CGH probes on the array. The CGH aberration analysis does not detect the region of copy-neutral LOH caused by the UPD. **Lower pane** – Log_2 ratio data for the same sample from a standard CGH array. Again, the copy-neutral LOH is not detected by standard CGH analysis.

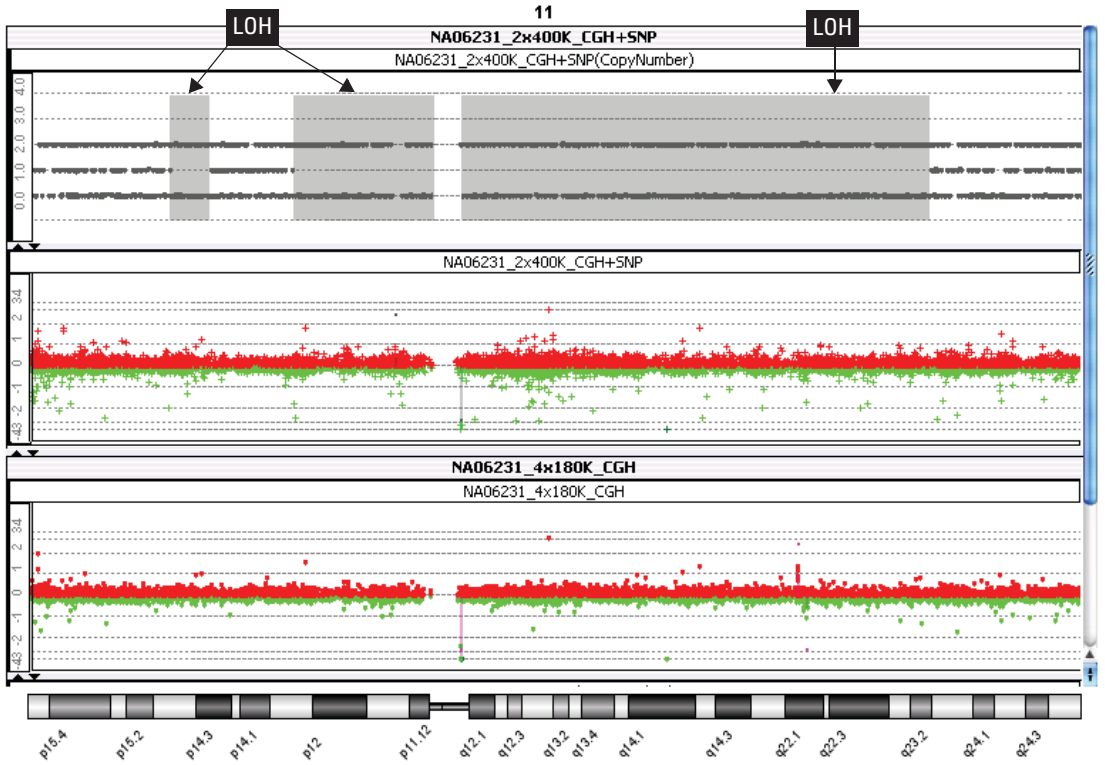


Figure 224 LOH in chromosome 11 caused by parental consanguinity. **Upper pane** – ASCN calls for SNP probes. In the regions of LOH (gray shaded regions), note the lack of probes that correspond to a copy number of 1. **Middle pane** – Log_2 ratio data from the CGH+SNP array, with an aberration analysis of the CGH probes on the array. The CGH aberration analysis does not detect the copy-neutral LOH regions. **Lower pane** – Log_2 ratio data for the same sample from a standard CGH array. Again, the copy-neutral LOH regions are not detected by standard CGH analysis.

Algorithm

The LOH algorithm uses two main inputs:

- The expectation values for CGH copy number. This value is calculated for every CGH aberration interval and non-aberrant interval on the array.
- The expectation value for the uncut SNP allele copy number (ASCN) for each SNP site under study in the sample genome. These values are the output of the ASCN algorithm. See "ASCN (SNP CN) – Allele-specific copy number detection algorithm" on page 534.

For each SNP site that is interrogated on the array, the algorithm deduces, to the extent possible, whether the given SNP is homozygous or heterozygous. The algorithm scans the SNP sites in the genome and uses an iterative, stepwise expansion and comparison method to develop putative LOH regions. This part of the algorithm is similar to the ADM-1 aberration detection algorithm. See "ADM-1" on page 468. As it develops these regions, the algorithm calculates an LOH score for each region after each expansion. The algorithm reports a final LOH call for genomic regions whose scores exceed a user-defined threshold (default 6). The highest scoring of overlapping LOH regions is reported.

Step 1 Calculate LOH scores

To calculate an LOH score for a given region, the algorithm first assigns each SNP as either homozygous or heterozygous. It assumes a binomial distribution, and calculates the probability that the given set of SNP sites forms an LOH region that is not just a result of random chance. For a region that contains n total SNP sites, x of which are homozygous, the algorithm first calculates the probability that x or fewer of n sites would be homozygous as a result of random chance:

$$F(x, p, n) = \sum_{1}^{x} P(x, p, n) = \frac{n!}{x!(n-x)!} p^{x} (1-p)^{(n-x)}, \quad \textbf{(51)}$$

where F is the cumulative distribution function, P is the probability distribution function for the binomial distribution, and p is the prior probability that a given SNP is heterozygous. p is estimated from the genome-wide fraction:

$$p_{(\text{heterozygous})} = \frac{n_{\text{heterozygous}}}{n_{\text{heterozygous}} + n_{\text{homozygous}}}, \quad (52)$$

where n is the total number of SNP sites in each given state. If the value of p is less than 0.22, the algorithm sets p to 0.22.

Smaller values of F signify a lower likelihood that this event happens as a result of random chance. The LOH score is the negative common logarithm of this probability:

$$\text{LOH score} = -\log_{10}[F(x, p, n)] \quad (53)$$

Step 2 Separate regions if needed

In this step, the algorithm examines each putative LOH region. If a region contains a statistically significant number of consecutive heterozygous SNPs, that region is separated into two smaller regions. Each of the resulting smaller regions is then scored again.

Step 3 Make final LOH calls

As the algorithm develops putative LOH regions, it makes a final LOH call when the score for a given region falls below a user-specified threshold value. You enter the threshold value when you set up LOH analysis. Agilent recommends that you use a value of 6.0 as a starting point for this threshold value for the LOH score. A higher value makes the algorithm more stringent (*i.e.* it will report fewer false positives, but also fewer true positives).

Visualization

When you apply the LOH algorithm in Agilent Genomic Workbench, LOH calls can appear in the Genomic Viewer in Genome View (Figure 225), Chromosome View (Figure 226), and Gene View (Figure 227 and Figure 228). In addition, you can create a SNP Aberration & LOH Report that contains a summary of analysis parameters and a list of all aberrant and LOH intervals. See "To generate a SNP Aberration & LOH Report" on page 168.

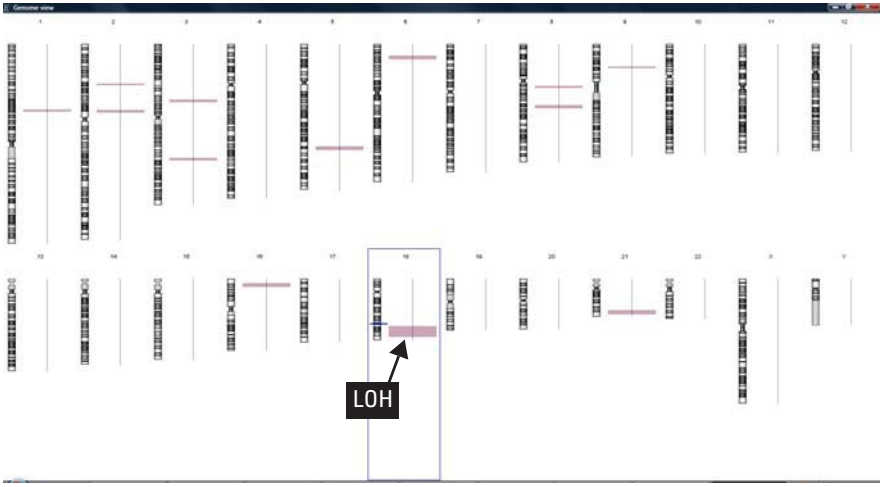


Figure 225 Visualization of LOH regions in Genome View in the Genomic Viewer of Agilent Genomic Workbench. LOH calls appear as shaded areas next to the relevant chromosomes.

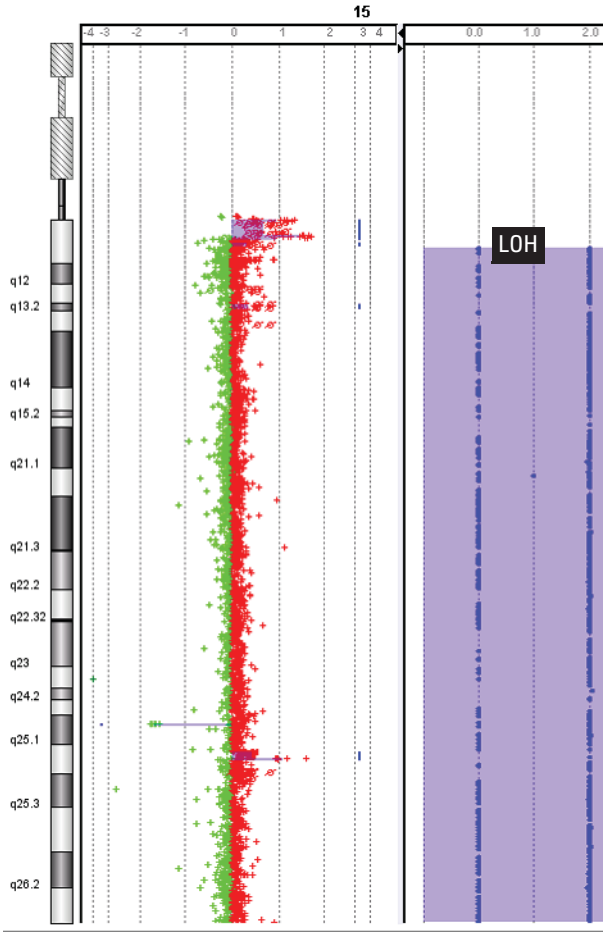


Figure 226 Chromosome View of the Genomic Viewer in Agilent Genomic Workbench. **Left pane** – Log_2 ratio data and aberration calls on chromosome 15 based on data from the CGH probes on a CGH+SNP array. **Right pane** – SNP CN panel, with ASCN calls (blue dots) for SNP sites. An LOH call appears as a large purple shaded area. Note the absence of sites with a copy number of 1.

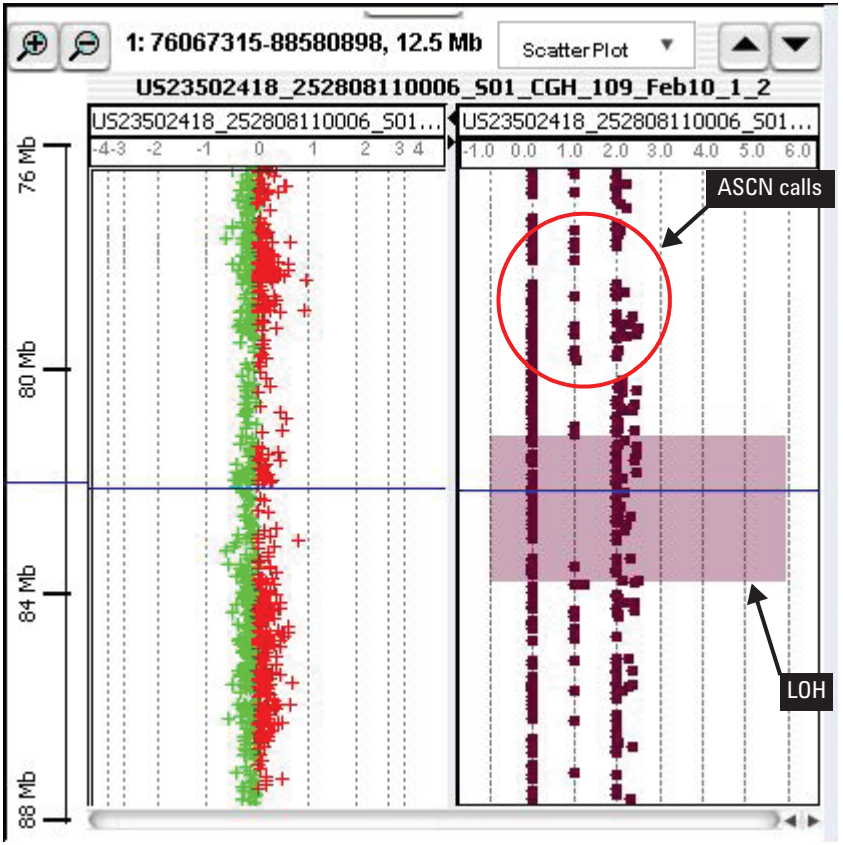


Figure 227 Gene View of the Genomic Viewer in Agilent Genomic Workbench, showing a 12.5 Mb region of chromosome 1. **Left pane** – Log_2 ratio data from a CGH+SNP array. **Right pane** – ASCN calls for SNP sites. LOH call appears as a mauve shaded region. In the LOH region, note the absence of sites with a copy number of 1.

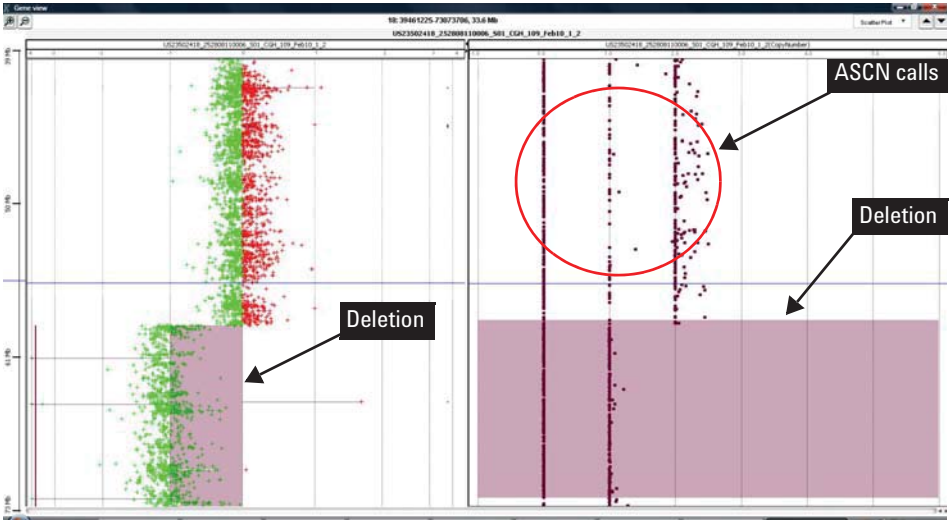


Figure 228 Gene View of the Genomic Viewer in Agilent Genomic Workbench, showing a 33.6 Mb region of chromosome 18. A hemizygous deletion affects part of chromosome 18. **Left pane** – Log_2 ratio data from the CGH+SNP array. Deletion call, based on data from the CGH probes on the array. **Right pane** – SNP CN panel, with ASCN calls for the same region, with the analogous deletion call made by the LOH algorithm. In the aberrant region, note the lack of SNP sites with a copy number of 2.

Appendix

This section contains additional information about statistical algorithms and is useful for understanding algorithm steps in detail.

Enrichment Analysis - the Hypergeometric distribution

The Hypergeometric distribution is a generalization of the binomial distribution. The key difference between the two distributions is that the Hypergeometric distribution samples without replacement and is used to calculate the overabundance, or *enrichment* of a certain event occurring, given the constraints of the total number of matching events possible. In this way, the Hypergeometric distribution addresses a question of conditional probability - the probability that one event occurred, given that another event has occurred within the groups under consideration.

Because of the similarity to the binomial distribution, it is familiar to use terms used to describe binomial events - mainly, the concept of the success or failure of an event to occur. The first step to enrichment analysis using the Hypergeometric distribution is to partition trials into successes and failures based upon a given condition. For example, one might wish to measure the probability of an aberration event occurring (a success), where the probability itself (usually called p) is a measure of the likelihood of observing the number of events under a certain condition compared to observing the same number of events in a random model. Another experimental parameter might then be considered, in order to ascertain whether or not the probability of the number of observed mutations is increased This is called an enrichment analysis, and such an analysis asks whether or not the observed number of mutations is enriched in a certain group. Other factors might be the sex of the subject, the age of the subject, the medical history, and so forth.

To demonstrate the derivation of the Hypergeometric distribution, an enrichment analysis such as that used in the Differential Aberration algorithm (see "Differential Aberration Analysis") begins by partitioning a group of samples based upon whether or not an event is observed in that group. For N total subjects, assume that n aberrations are observed within a genomic interval in normal samples (preselected by some criteria),

and therefore N-n aberrations are observed within the same interval in abnormal samples. From these, assume that the total number of all aberrations within a given genomic interval is m, and that therefore are N-m absence of detected aberrations in the same interval. An aberration can be either an amplification (gain) or deletion (loss), so the calculations will have to be repeated to assess enrichment for both scenarios.

The probability of observing y aberrations within healthy subjects for a given interval can therefore be described using the binomial distribution:

$$P_Y(y) = \binom{n}{y} p^y (1-p)^{n-y} \quad \textbf{(54)}$$

The probability of observing z aberrations within abnormal samples is also described using the binomial distribution:

$$P_Z(z) = \binom{N-n}{m-y} p^{m-y} (1-p)^{N-m-(m-y)} \quad \textbf{(55)}$$

And the probability of observing m total aberrations is derived from the binomial distribution:

$$P_M(m) = \binom{N}{m} p^m (1-p)^{N-m} \quad \textbf{(56)}$$

The probability of observing the number of aberrations from healthy subjects within the interval, given the probability of observing the number of aberrations in the total sample for that same interval is given by the conditional probability property:

$$P(Y_y | M_m) = \frac{P(Y_y M_m)}{P(M_m)} \quad \textbf{(57)}$$

Equation 57 allows use of the probability of observing the number of successes and failures in each of the partitions from the samples normalized by the probability of observing this aberration in the population:

$$P(y|m) = \frac{\binom{n}{y} p^y (1-p)^{n-y} \cdot \binom{N-n}{m-y} p^{m-y} (1-p)^{N-n-(m-y)}}{\binom{N}{m} p^m (1-p)^{N-m}} \quad \textbf{(58)}$$

Rearranging Equation 58 yields:

$$P(y|m) = \frac{\binom{n}{y}\binom{N-n}{m-y}p^m(1-p)^{N-m}}{\binom{N}{m}p^m(1-p)^{N-m}} \quad (59)$$

Equation 59 cancels to yield the Hypergeometric distribution, which is a quick way to calculate enrichment independently of the binomial probabilities, and instead entirely upon the binomial coefficients:

$$P(y|m) = \frac{\binom{n}{y}\binom{N-n}{m-y}}{\binom{N}{m}} \quad (60)$$

Because this calculation is a distinct probability, the calculation is usually summed across all number of observations equal or greater than that of the number of aberrations in the interval from the partitioned sample to yield a p-value, which is known as the Hypergeometric tail score (H.G.T.), or simply the probability under the extreme tail of the Hypergeometric distribution:

$$HGT = p-value = P(Y \geq y|m) = \sum_{y}^{m} \frac{\binom{n}{y}\binom{N-n}{m-y}}{\binom{N}{m}} \quad (61)$$

The hypergeometric tail probabilities reflect the probability of observing y or more number of gains or losses in a group in this example - *i.e.* it is a measure of the enrichment of the gains or losses in a group.

Cluster Analysis - Distance-based methods

Genomic Workbench provides two distance-based metrics for clustering. Both are distance measures between two points, identified by NIST as L_m distances or Minkowski distances. The points measured are the coordinates of two vectors (shown here in two dimensions for simplicity):

$$A = (x_1, y_1 \ldots) \quad (62)$$

$$B = (x_2, y_2 \ldots) \quad (63)$$

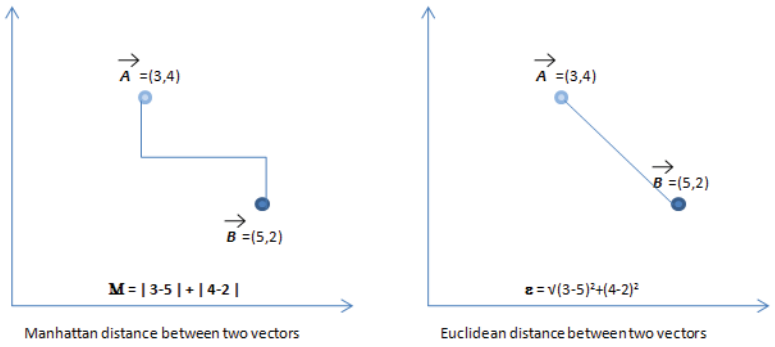


Figure 229 Distance-based measures available in Genomic Workbench.

Manhattan distance

Manhattan distance is also called City Block distance, or Minkowski's *L1* metric. It is the distance in multidimensional space between two points as taken by following each dimensional axis. In two dimensions, this is similar to the distance through a city if restricted to following a grid-like street pattern. The calculation for Manhattan distance between two vector coordinates X and Y is:

$$D_m(X,\ Y) = \sum |x_i - y_i| \quad \textbf{(64)}$$

Euclidean distance

Euclidean distance is also called Minkowski's *L2* metric. It is the distance in multidimensional space between two points connected by a straight line, and is therefore shorter than the Manhattan distance for all cases where the number of dimensions is greater than 1 (i.e., a straight line). The calculation for Euclidean distance between two vector coordinate systems X and Y is:

$$d = |x - y| = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2} \quad \textbf{(65)}$$

For both the Manhattan and Euclidean distance measures, if the distance between two arrays is small they are considered close. However, a key difference between the Manhattan and Euclidean distances is the way in which dimensional distances are treated. Each dimension represents a coordinate of the aberration profile, or vector, from a given sample. If the vector is two-dimensional as in Figure 229, then changing both of the two coordinates results in a change of distance less in Euclidean space than

changing just one of the coordinates, while the change in distance is the same for either case in Manhattan space. Euclidean distance is therefore less resistant to outlying coordinates.

Euclidean distance (Normalized Euclidean)

Distance similarity uses a normalized Euclidean distance where the Euclidean distance is normalized over the dataset. It is equal to 1 – normalized Euclidean distance. In order to normalize over the dataset, the following steps are followed:

1 Calculate the minimal and maximal values of an interval I over the number of samples to be clustered using the interval values.

2 Calculate the middle value of that interval as:

$$I_{mid} = \frac{I_{max} + I_{min}}{2} \quad (66)$$

3 Calculate the range of that interval as:

$$I_{range} = I_{max} - I_{min} \quad (67)$$

4 Calculate the updated score for the interval as:

$$I_{score} = \frac{I_{value} - I_{mid}}{I_{range}} * 0.5 + 1.0 \quad (68)$$

Where 0.5 and 1.0 are the expected middle and range.

5 Calculate 1 – Euclidean distance as described in Equation 65 above using the updated interval score.

NOTE If Aberration Calls are used as the Data Choice for clustering, then Cosine Similarity is automatically calculated instead of Distance Similarity.

Cluster Analysis - Similarity-based methods

Genomic Workbench provides two similarity-based measures for aberration profiles. Both use the multidimensional vector from two samples to find the correlation coefficient called r, which is a relative measure of the strength of the linear relationship between the two samples. The measure r ranges in value from -1 to +1. The closer the value is to |1|, the stronger the relationship. An r value of 0 indicates that the vectors are orthogonal.

Cosine angle

The cosine of the angle between two vectors A and B measures how related the two samples are across all intervals containing aberrations. If the two vectors are the same then this measure is 1. The measure of similarity r measured in the cosine of the angle between two vectors A and B is:

$$r = \frac{(A \cdot B)}{\|A\|\|B\|} \tag{69}$$

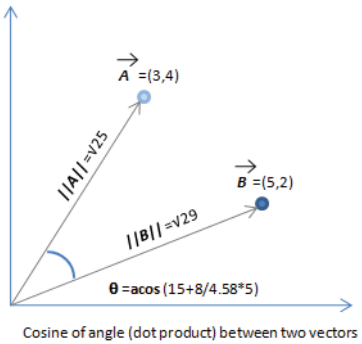


Figure 230 The cosine-angle similarity metric.

Pearson correlation

The Pearson correlation measures the angle of the vectors from two arrays around the mean of the aberration vectors. This mean can be positive or negative depending on whether there is an abundance of amplifications or deletions. Using the Pearson correlation often results in more negative correlations than using the cosine correlation. This is determined by the mean of the vectors, which is used as a baseline. If the mean is largely positive, then aberration losses will result in a more negative correlation. The measure r of similarity by Pearson correlation between two vectors X and Y is:

$$r = \frac{\left(\sum xy - n\bar{x}\bar{y}\right)^2}{\left(\sum x^2 - n\bar{x}^2\right)\left(\sum y^2 - n\bar{y}^2\right)} \tag{70}$$

Cluster Analysis - Linkage methods

When calculating distances between individual samples, the specified similarity metric is used directly. However, when the distance calculation involves a cluster, or group of samples, there is a need to specify how to extend the similarity metric to cover groups. In Genomic Workbench, the following linkage methods can be specified:

- Single linkage
- Average linkage
- Complete linkage

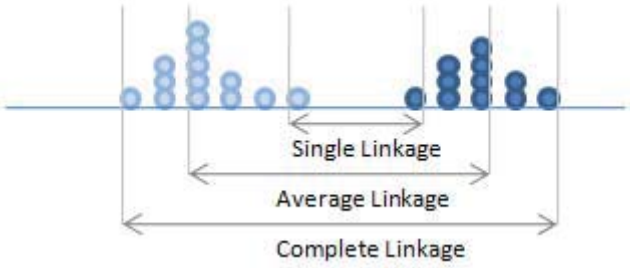


Figure 231 Linkage measures between clusters in Genomic Workbench.

These are described below:

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 connections of a cluster, but in doing so may tend to find spread out chain-like clusters containing dissimilar elements.

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.

Complete linkage The distance between two clusters is calculated as the largest distance (dissimilarity) between any two members in the two clusters. Complete linkage maximizes the maximum within-cluster distance, and so tends to find discrete, compact clusters. However, it may overemphasize small differences between members.

Multiple testing correction – False Discovery Rate in the Discrete Wavelet Transform

False Discovery Rate, or FDR, is a multiple testing correction procedure that allows a percentage of called breakpoints in the data (as identified by the Wavelet Transform) to be called false positives (meaning they are called true breakpoints when really they are detected merely because the data was noisy). This type of adjustment is necessary when making multiple comparisons such as is done with the set of putative breakpoints in the data. The Wavelet Transform creates a list of breakpoints that have associated *Z*-scores, which measure the distance of the transformed signal from a static mean of zero (meaning no breakpoint). These *Z*-scores are simply the number of standard deviations that a point falls away from the mean in a normal distribution. Because of this, they can be easily converted to *p*-values by taking the value beyond the *Z*-score under the normal curve. In Benjamini and Hochberg FDR, the *p*-values are ranked for each putative breakpoint from largest to smallest. Each *p*-value is then multiplied by the total number of breakpoints in that ranked list (let's call that n) and then divided by its rank. If it is less than the filter cutoff, it is deemed significant. For, say, the 3rd p-value in that list, the corrected *p*-value is =*p*-value*(n/3). If the corrected *p*-value is less than the threshold, that breakpoint is kept. The threshold can be set to any value and will allow a portion of the detected breakpoints to be considered

statistically significant. In the HMM algorithm user interface parameters, the "Q" in FDRQ parameter stands for the "q" parameter in the original Haar segmentation implementation. Q-values are a type of Bayesian posterior p-value tied to the rate of minimum false discoveries in the data and is the minimum FDR over which a given statistic can be rejected. The FDRQ in the Haar wavelet is therefore thresholded on false positives. For example, if the threshold is set to 0.3, then a breakpoint is used in construction of the stationary distribution if the multiple testing probability of being a false positive is less than 30%.

The Haar Wavelet and associated FDR user defined setting is all part of the HMM construction phase of the HMM algorithm. For a probabilistic model to be applied, the probabilities of that model must be estimated and optimized. The Haar Wavelet takes a complex signal of the probe log ratios and converts it into a signal in which breakpoints between a region of continuous no change and a region of amplification (for example) are detected. No change and amplification are two "states" in the HMM, or bins into which we want to group the signals. By counting how many times we see such a transition between these states we can set up probabilities for the model to follow (probabilities are just counts divided by the total number of events). This means that if we see many breakpoints, we assign probabilities between the states which allows transitions to occur more quickly – for example, a transition from NC to gain to loss in three probes is very rapid and may not adequately reflect the true biological activity underlying the model. By reducing the number of breakpoints, the model can be made more resistant to changing states quickly. This means that once the model enters a state like amplification it is more likely to stay there. The combination of these model probabilities is called the stationary distribution. Since the FDRQ can increase or reduce the number of breakpoints, it therefore has an effect on how easily the model is allowed to move between states, or alternatively, how likely the model is to remain in one state for a number of probes. Stringent thresholding will reduce the number of spurious transitions between states in the HMM but may result in loss of sensitivity to local changes in the data.

Segmentation- Discrete Haar Wavelet Transform and Gaussian mixtures for HMM

A wavelet transform is a method for examining the signal at different levels of resolution *i.e.* it is a multiresolution signal analysis method. It can be crudely likened to looking at the signal characteristics at different zoom levels. The undecimated discrete wavelet transform method proposed by Mallat[5] decomposes the signal into an approximation sub band and a set of detail subbands. The approximation sub-band contains the approximation of the signal at the current level of resolution. The detail subbands contain the higher frequencies of the signals and are made up of wavelet coefficients. In the undecimated version each sub band contains the same number of coefficients given by:

$$w_L[n] = \frac{1}{\sqrt{2^{L+1}}}\left(\sum_{k=n}^{n+(2^L-1)} y[k] - \sum_{k=n-2^L}^{n-1} y[k]\right) \quad (71)$$

The wavelet coefficient is therefore a difference between two averages (this is the same as a running sum in the look ahead algorithm). Wherever a breakpoint occurs in the data, the wavelet coefficient is some value other than zero. If there is no breakpoint the wavelet coefficient is zero for no as it describes a difference between two identical averages. Here L is the level of the sub-band. When (additive) noise is present the coefficient would be close to zero for no breakpoint and would have larger values wherever breakpoints occur for large L.

Therefore if we take local maxima we find occurrences of the breakpoints

$$z_L[k] = localmax(|w_L[n]|) \quad (72)$$

The high amplitude coefficients in $z[k]$ now correspond to the breakpoints, whereas the low amplitude ones correspond to noise. To keep only the high amplitude ones, we apply FDR thresholding.

To apply FDR thresholding, we first sort the z's in descending order and for each z calculate a two-sided p-value from the normal distribution. Then we threshold this p-value by the supplied parameter and keep only the z's that pass this threshold After gathering the thresholded values into a list we unify them keeping in mind the possibility that the same breakpoint may be detected at slightly different position in different sub bands.

We keep only the segments whose z is at least 3 standard deviations. To estimate the standard deviation the Median absolute deviation (MAD) estimator is used. We then group together the probes belonging to each state (*e.g.* amplification, deletion or no change). This is done by considering each interval found above, considering its direction and grouping together probes in all intervals with same direction. A Gaussian distribution is then fitted to each group and the initial estimates of probabilities and distributions is calculated.

References

1 Diskin, Sharon J., Mingyao Li, Cuiping Hou, Shuzhang Yang, Joseph Glessner, Hakon Hakonarson, Maja Bucan, John M. Maris, and Kai Wang 2008. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res.* 36:e126. http://nar.oxfordjournals.org. doi:10.1093/nar/gkn556.

2 Hedenfalk, I. et. al. (2003) Molecular classification of familial non-BRCA1/BRCA2 breast cancer. *Proc. Natl. Acad. Sci. USA,* 2003 Mar 4;100(5):2532-7.

3 Kincaid, R., A. Ben-Dor, and Z. Yakhini. Exploratory visualization of array-based comparative genomic hybridization. *Information Visualization*, 2005, 4:176-190. doi:10.1057/palgrave.ivs.9500095.

4 Lipson, Doron, Yonatan Aumann, Amir Ben-Dor, Nathan Linial, and Zohar Yakhini. Efficient Calculation of Interval Scores for DNA Copy Number Data Analysis. Proceedings of Recomb, 2005.

5 S. G. Mallat. *A Wavelet Tour of Signal Processing.* Academic Press, 1998.

6 Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics.*" **5**, 4, 557-572.

7 E. S. Venkatraman and Adam B. Olshen. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, **23(6)**: 657-663.

8 JV Braun, HG Muller (1998), Statistical methods for DNA sequence segmentation. *Statistical Science,* 13(2):142-162.

9 E. S. Venkatraman and Adam B. Olshen. (2005). DNAcopy: A Package for Analyzing DNA Copy Data. *The Bioconductor Project*: http://www.bioconductor.org.

10 R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. Markov Chains and Hidden Markov Models. *Biological Sequence Analysis: Probabilistic models of proteins and nucleic acids.* Cambridge University Press, 1998. 46-68.

11 A. Jensen and A. la Cour-Harbo. *Ripples in Mathematics: The Discrete Wavelet Transform.* Springer-Verlag Berlin, 2000.

12 Stanton A. Glantz. Correlation and Correlation Coefficients. *Primer of Biostatistics, 5th Edition*. McGraw-Hill, 2002. 262-265.

13 Pierre Baldi and G. Wesley Hatfield. Statistical analysis of array data: Dimensional reduction, clustering, and regulatory regions. *DNA Microarrays and Gene Expression*. Cambridge University Press, 2002. 78-84.

14 Lipson D, Aumann Y, Ben-Dor A, Linial N, Yakhini Z. (2006) Efficient calculation of interval scores for DNA copy number data analysis. *J. Comput Biol*. 2006 Mar;13(2):215-28.

15 Ben-Dor, A., Lipson, D., Tsalenko, A., Reimers, M., Baumbusch, L., Barrett, M., Weinstein, J., Børresen-Dale, A.-L. and Yakhini, Z. (2007). Framework for identifying common aberrations in DNA copy number data. Research in Computational Molecular Biology, Lecture Notes in Computer Science, Springer, Berlin/Heidelberg, Vol. 4453.

16 Rice, John A. (1995) The Hypergeometric Distribution. *Mathematical Statistics and Data Analysis, 2nd Edition*. Duxbury Press, 1995, 39-40.

17 Chaya Ben-Zaken Zilberstein et al., *Lecture Notes in Computer Science: Regulatory Genomics: RECOMB 2004 International Workshop, Revised Selected Papers* in Eskin E. and Workman, C. (eds.), Springer Berlin / Heidelberg 2005(3318). ISBN: 3-540-24456-5

www.agilent.com

In this book

This guide describes how to use the CGH application of Agilent Genomic Workbench Lite Edition to analyze CGH data and to perform comparative genomic hybridization (CGH), copy number variation (CNV), and single nucleotide polymorphism (SNP) studies. It also describes how you can find common aberrant regions, and correlate CGH and expression data.

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

PR7000-0048

© Agilent Technologies, Inc. 2010, 2015

Revision A0, December 2015

G3800-90038

Agilent Technologies