Correlation Analysis in Agilent GeneSpring and Mass Profiler Professional

"Is there a relationship between variable X and variable Y? This is a central question for the data analyst. The answer comes from examination of the correlation between the two."

Chen, P; Popovich, P. "Correlation: Parametric and Nonparametric Measures". Sage Publications, 2002.

Technical Overview

Authors

Pritha Aggarwal, Durairaj Renu, and Pramila Tata

Strand Life Sciences

Bangalore, India

Michael Rosenberg

Agilent Technologies, Inc.

Santa Clara, California, USA

Introduction

Correlation analysis allows identification of coregulated molecules such as genes and metabolites as well as identification of relationships between the samples in a study. Introduced in the GeneSpring/Mass Profiler Professional (MPP) 13.0 platform, the correlation framework is supported on most of the datasets generated using high-throughput omics platforms such as Microarray, Mass Spectrometry, and Next Generation Sequencing. The framework supports pair-wise correlations measured using a single technology platform and cross-technology measurements between two different platforms.

This Technical Overview describes the details of the correlation framework supported in GeneSpring/MPP 13.0.

Agilent Technologies

Key Correlation Concepts

Correlation analysis is one of the most widely used statistical techniques. Correlation measures the strength and directionality of the linear relationship between two quantitative variables. (http://en.wikipedia.org/wiki/Correlation_and_dependence).

- When high scores of variable X tend to accompany high scores of variable Y, the two variables are said to be positively correlated. When low scores of X tend to accompany high scores of Y, the two variables are said to be negatively correlated or anticorrelated.
- Correlation values range from –1 to +1. If two variables are positively correlated, the value of their correlation coefficient is close to +1; then +1 itself indicates perfect correlation. Similarly, if the two variables are negatively correlated, their correlation coefficient approaches –1. A low correlation coefficient indicates that there is no significant dependency between the two variables.
- Correlation measures only the degree of linear association between two variables. It does not imply a cause-and-effect relationship between the variables.
- For a straightforward linear regression, goodness of fit equals squared Pearson correlation coefficient (R^2 and R in Figure 1). Therefore, correlation can be thought of as a measure of deviation of empirical data from the linear regression fit.
- At least three data points are required to perform correlation analysis; by definition, for any two data points correlation coefficient is +1, –1, or undefined.
- There are many methods for computing correlation, but the most commonly used are Pearson and Spearman correlation coefficients. The Pearson correlation coefficient (http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient) is broadly used for finding the strength of linear relationships within normally distributed random variables.

  The Spearman correlation coefficient (http://en.wikipedia.org/wiki/Spearman's_rank_correlation_coefficient) is a rank-based algorithm, and is more tolerant to the outliers in datasets than Pearson. Both Pearson and Spearman metrics are available in the GeneSpring/MPP Correlation Framework.
- In biological systems, positive correlation is observed between transcriptional activators and their target genes; inhibitors such as miRNA and their mRNA targets are negatively correlated.

Data Point	varX	varY
1	0.81	8.17
2	0.47	5.72
3	0.26	1.78
4	0.50	3.53
5	0.18	1.26
6	0.81	8.53
7	0.18	2.46
8	0.53	6.16
9	0.16	2.41
10	0.99	10.98
11	0.27	0.91
12	0.60	5.62
13	0.32	4.29
14	0.74	7.25
15	0.06	2.35
16	0.95	8.65
17	0.22	3.18
18	0.90	10.39
19	0.10	0.29
20	0.73	7.01
21	0.10	1.06
22	0.45	5.76
23	0.19	2.67
24	0.67	7.14
	Pearson CC	0.9545

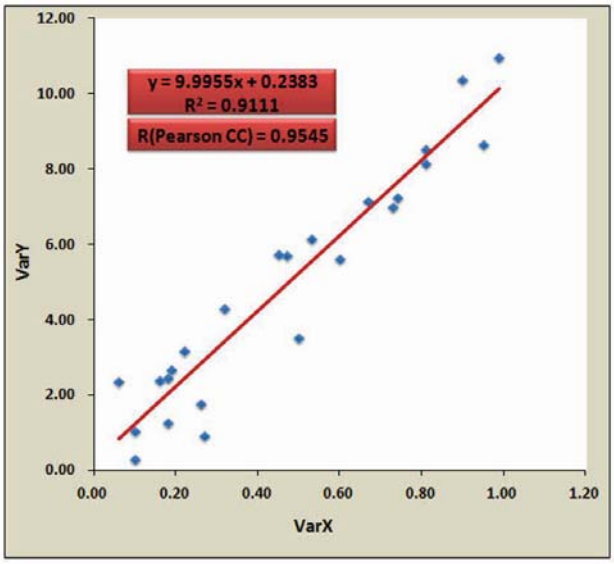


Figure 1. Pearson correlation coefficient (CC) between variables X and Y. The scatter plot of the empirical data with regression line displays the positive correlation between X and Y.

Entity-Entity Correlation

Correlation analysis is performed in a pair-wise manner to identify and observe dependency between abundance levels of any pair of biological entities. An entity in GeneSpring/MPP is a gene, metabolite, protein, or a probe in an expression array.

The option to perform correlation analysis is included in the Workflow Browser of the experiment. A correlation analysis can be performed on entities within a single experiment or across two different experiments. For cross-experiment correlation analysis, a Multi-Omics Analysis (MOA) experiment should be created in GeneSpring using the two experiments whose entities would be selected for correlation. Table 1 summarizes the types of GeneSpring experiments that support correlation analysis.

Correlation analysis on entities from a single experiment and from two experiments are explained in Figures 2 and 3 respectively.

Table 1. Analysis types in GeneSpring available for correlation analysis.

Analysis type	Within single experiment	Across two experiments
mRNA expression	Yes	Yes
Exon expression	Yes	Yes
miRNA	Yes	Yes
RTPCR	Yes	Yes
DNA-Seq	No	No
RNA-Seq	No	Yes*
smallRNA-Seq	No	Yes*
Metabolomics	Yes	Yes
Proteomics	Yes	Yes

* To perform correlation in RNA-Seq and smallRNA-Seq experiments, data from Strand NGS v2.1 software (http://www.strand-ngs.com/) should be imported into GeneSpring.

Inputs for Correlation Analysis

The output of the correlation analysis performed within a single or a MOA experiment is determined by the entity list, interpretation, and type of correlation coefficient selected as inputs for the analysis.

Entity list: Pair-wise correlations between the entities in a chosen entity list(s) are calculated. GeneSpring/MPP allows selecting an entity list from the active experiment, or two different entity lists from the two experiments in a MOA experiment.

Interpretation: As any other analysis in GeneSpring, if averaged interpretation is chosen, the mean intensity value for each entity across the replicates will be used for analysis. In the nonaveraged interpretation, the intensity value for the entity in each sample will be used for the analysis.

Supported types of correlation: In GeneSpring 13.0, the Correlation Analysis Framework supports Pearson and Spearman correlation coefficients; other types will be added in future releases. GeneSpring requires a minimum of three valid data points to calculate correlation between a pair of entities. If a nonaveraged interpretation is used, a minimum of three samples have to be included as part of the interpretation. In an averaged interpretation, a minimum of three conditions are required.

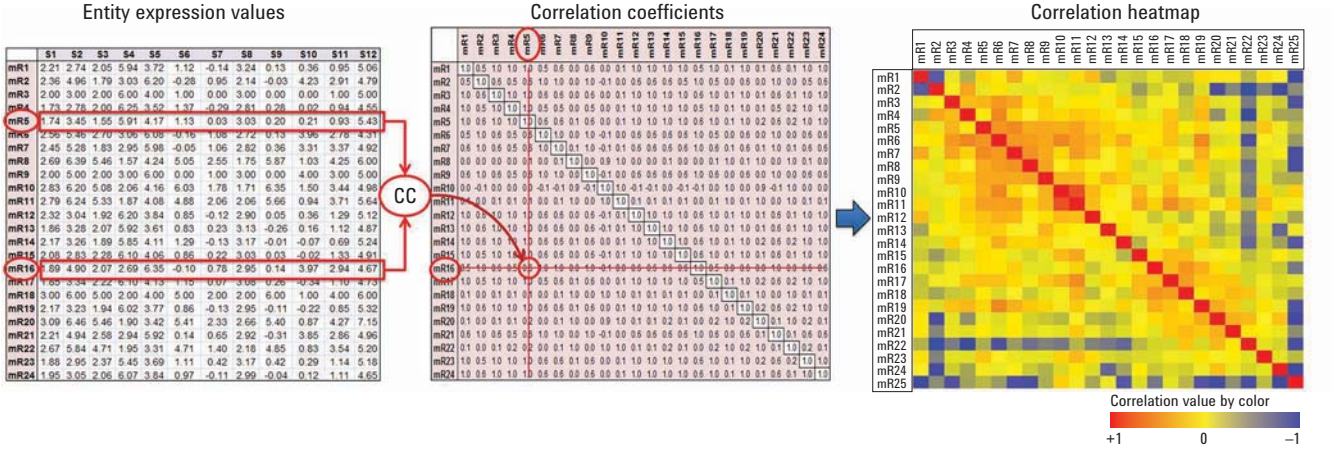


Figure 2. Calculation of correlation coefficients between entities of a single experiment. A table of expression values is created based upon the normalized intensities (abundance) of selected entities in a set of samples. Pair-wise correlation coefficients (Pearson or Spearman) are computed between each pair of rows in the expression table and represented as a heatmap.

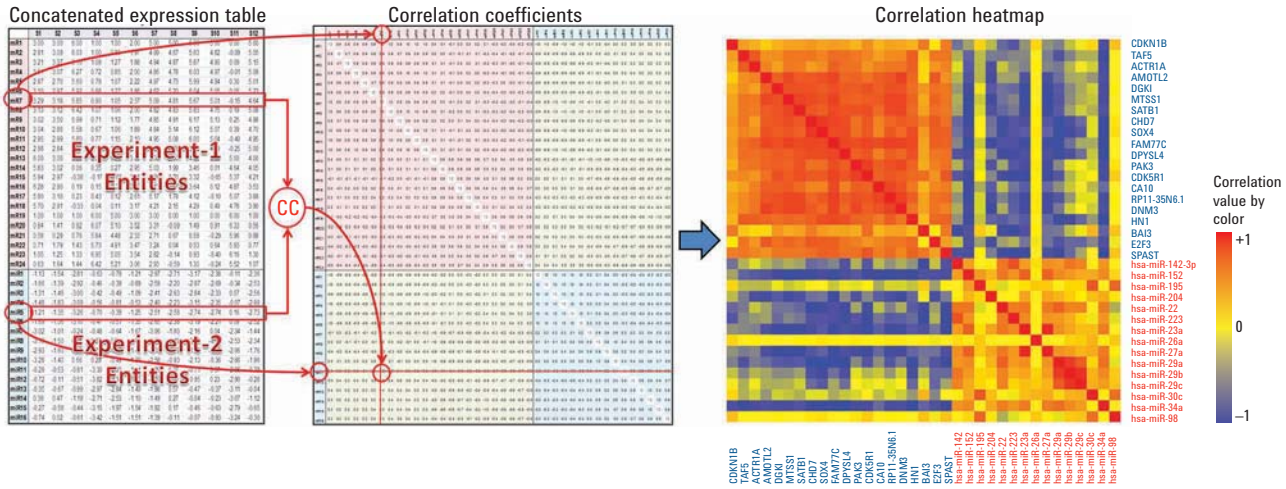


Figure 3. Pair-wise correlation calculation between entities of two different experiments. Similarly to single experiment correlation analysis, individual tables of expression values are created for Experiment 1 and Experiment 2. The two expression tables are concatenated based on the sample pairing information provided by the user. For cross-experiment correlation analysis, it is expected that the same or similar biological samples are measured by the two different platforms. If *m* entities are selected from Experiment 1 and *n* entities from Experiment 2, the concatenated heatmap will have *m*+*n* entities on each axis. The calculated correlation coefficient values are represented as a heatmap.

Correlation Heatmaps

The correlation coefficients for selected datasets are visually represented as a heatmap. A correlation heatmap is a convenient tool for understanding the relationship between multiple entities. The color of each cell in the heatmap is defined by the pair-wise correlation coefficient value between the given entity in the X and Y-axes of the heatmap.

The order of entities in a heatmap is the same in both X and Y dimensions. This applies to correlation performed on entities of a single experiment, or between entities of two different experiments as in a MOA experiment. The resulting heatmap is a square diagram with a diagonal representing the correlation of each entity with itself (that is, all values on the diagonal equal +1). A heatmap view of correlation between entities in a single experiment or MOA experiment are shown in Figures 2 and 3 (the right-most panels). In a MOA, the view can be toggled between a heatmap of all entities or a heatmap of cross-experiment entities, as shown in Figure 4.

Data underlying each cell in a correlation heatmap can be examined as a scatter plot. A scatter plot provides a graphical representation of the empirical abundance values that were used to compute the correlation coefficient between a given entity pair. The regression fit and equation in the plot display the direction and strength of the dependency between the pair of entities (Figure 5).

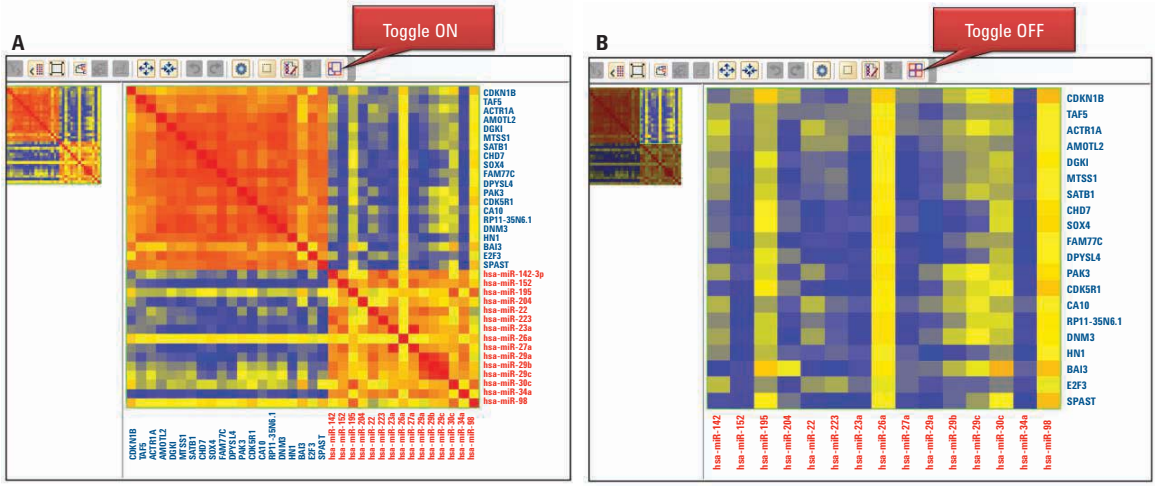


Figure 4. MOA correlation heatmap. The view can be switched between (A) *all input entities* or (B) *cross-experiment only* using the ON/OFF toggle in the menu bar.

Filtering Entities in a Correlation Heatmap

The output of many statistical analyses performed in GeneSpring/MPP is saved as entity list-associated data, for example, *Fold change*, *p-value*, or *Regulation*. In the correlation analysis framework, researchers can use any associated data to filter the correlation heatmap. Filtering is supported both on numerical data, such as *fold change* or *p-value*, and categorical data, for example, *Up/Down Regulation* (Figure 6A). If more than one filter is applied at a given instance, entities passing all of the applied filters are retained in the view (Boolean AND).

Correlation framework supports filtering a heatmap based on any attributes associated with entities, including external attributes. External entity attributes are imported into GeneSpring by uploading an entity list with all associated values from a tab-delimited text or an Excel file. In the example illustrated in Figure 6A, each imported gene (Figure 6B) has a pathway it belongs to as an associated value. Filtering a heatmap by a pathway name (for example, Receptor Tyrosine Kinases (RTK) in Figure 6B) allows a user to explore the relationship between the expression levels of the members of the selected pathway.

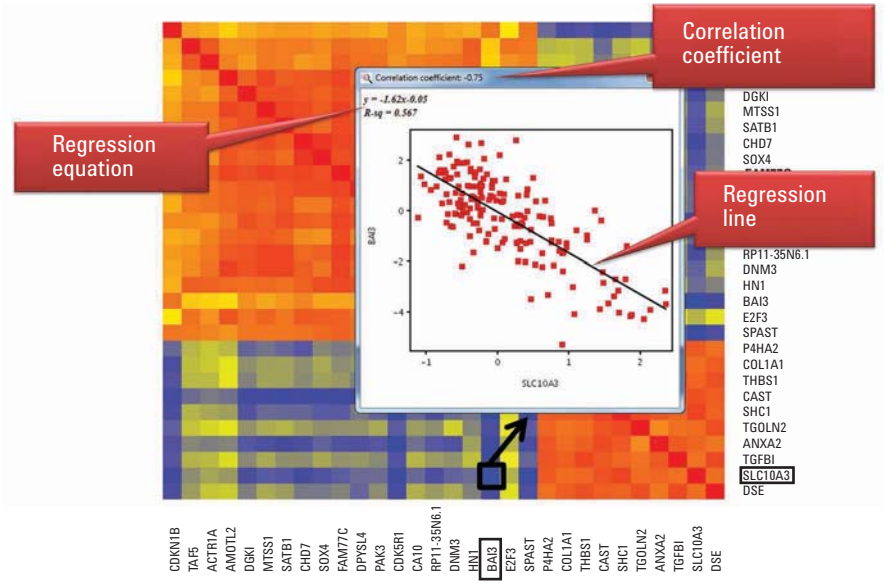


Figure 5. Example of a scatter plot showing negative correlation between the highlighted pair of entities, including empirical data, linear regression fit, and correlation coefficient.

The correlation framework allows saving the entities that passed a filter or filters. In the experiment navigator, the new entity list will appear under the node for correlation analysis where it was generated. For example, in Figure 6, where the heatmap was filtered to display only members of the RTK pathway, correlation between the members of this pathway can be saved for future analysis and reference.

In a MOA correlation, two filtering tabs are provided, one for each experiment.

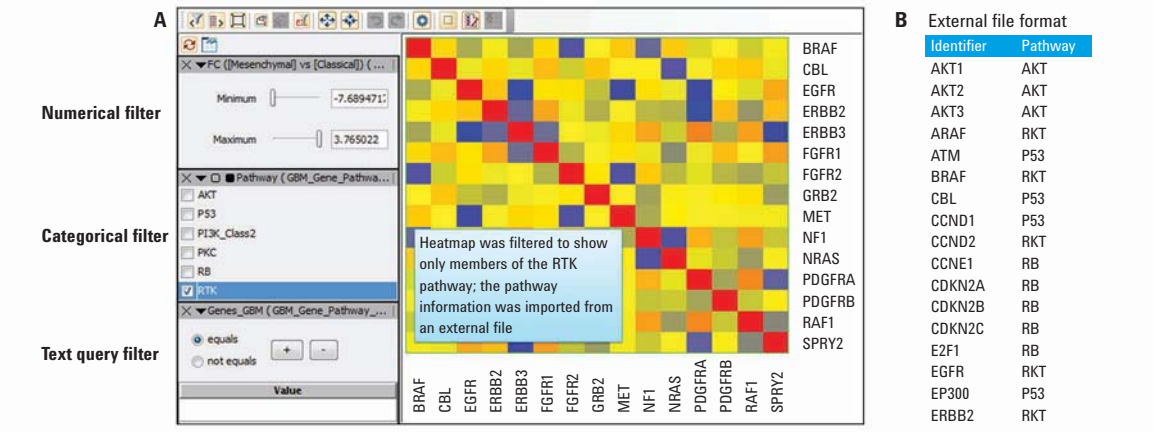


Figure 6. Numerical and categorical filtering. A) In this example, **Fold Change** (FC) is a numerical filter and **Pathway** and **Genes_GBM** are categorical filters. If the number of different categories is less than 30, it will be displayed in the filter panel as checkboxes (see **Categorical filter** in the figure). A text box will be displayed if the number of values exceeds 30 (see **Text Query filter** in the figure). B) Is an example of an external entity list with associated values used for filtering.

Clustering Entities in a Correlation Heatmap

The GeneSpring/MPP Correlation Framework allows hierarchical clustering of entities based on their correlation coefficients rather than their expression values. It is an important functionality because co-expressing entities are likely to share common biological functions and, therefore, a clustering result may suggest a new hypothesis or confirm existing assumptions. The clustering parameters supported by the correlation framework are summarized in Table 2.

In a single experiment, clustering is performed based on the pair-wise correlation coefficients between all entities in a given entity list. In a MOA experiment, clustering is performed based on cross-experiment correlation values only (Figure 7). If a filtering option was applied prior to clustering, in either a single-technology or a MOA experiment, clustering is performed only on those entities that passed the filters. If filtering is applied after clustering, the clustering dendrogram is removed, and the order of entities in the heatmap is reset to default.

Table 2. Summary of parameters supported for hierarchical clustering in GeneSpring.

Parameter	Value
Distance metric	Euclidean, Squared Euclidean, Manhattan (Cityblock), Maximum (Chebychev), Minimum, Differential, Canberra, Harmonic, Pearson's Centered, Pearson's Uncentered (Cosine), Pearson's Centered – Absolute, or Pearson's Uncentered – Absolute
Linkage rule	Average, Centroid, Ward's, Median, Single or Complete

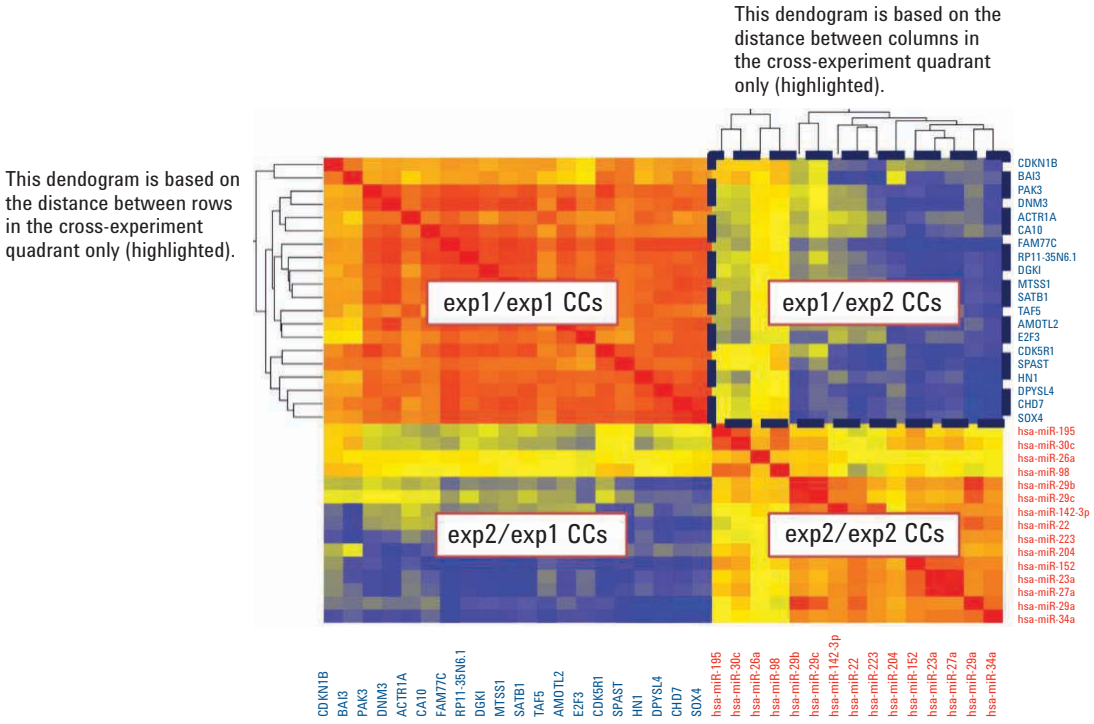


Figure 7. Clustering in MOA uses profile defined by cross experiment correlation values.

Exporting Correlation Coefficients

The correlation framework supports exporting the correlation coefficients for all, filtered, or selected entities. The associated data and annotations can also be exported (Figure 8).

Highlighting and Selecting Entities

The correlation framework in GeneSpring/MPP permits highlighting the subset of entities in a correlation heatmap that matched the selected entity list. Entities matching the selected list are highlighted by a red hash bar located next to the corresponding row and column. By default, unmatched entities are hidden from view by setting the transparency level to 0 (Figure 9). The transparency is customizable and can be changed in the heatmap property dialogue.

Entities of interest in the correlation heatmap can be selected and saved as an entity list of selection. The saved entity list can be used for further analysis in GeneSpring as any other entity list. For example, the entities that make up a cluster of interest can be selected and saved as an entity list. Performing Gene Ontology analysis or Pathway analysis on the saved entity list would identify the biological function or pathway that is enriched in the cluster of interest.

		CC's											COLUMN ANNOTATIONS							
	Identifier	E1_ZEB2	E1_NKX2-2	E1_MTSS1	E1_CRMP1	E1_MYT1	E1_SOX11	E1_DBN1	E2_hsa-miR-10a	E2_hsa-miR-143	E2_hsa-miR-145	E2_hsa-miR-155	E2_p (Corr)	E2_Regulat	E2_FC (abs)	E1_chr	E1_Entrez (	E1_map_locat	E1_Symbol_	E1_UniGene_cluster
CC's	E1_ZEB2	1.000	0.585	0.707	0.526	0.603	0.526	0.507	-0.218	-0.132	-0.100	-0.262				2	9839	2q22	ZEB2	Hs.34871
	E1_NKX2-2	0.585	1.000	0.670	0.681	0.729	0.706	0.614	-0.200	-0.207	-0.169	-0.309				20	4821	0pter-q11.23	NKX2-2	None
	E1_MTSS1	0.707	0.670	1.000	0.595	0.764	0.633	0.578	-0.195	-0.165	-0.127	-0.378				8	9788	8p22	MTSS1	Hs.336994
	E1_CRMP1	0.526	0.681	0.595	1.000	0.735	0.757	0.701	-0.262	-0.223	-0.238	-0.402				4	1400	4p16.1-p15	CRMP1	Hs.135270
	E1_MYT1	0.603	0.729	0.764	0.735	1.000	0.729	0.727	-0.243	-0.119	-0.148	-0.418				20	4661	20q13.33	MYT1	Hs.279562
	E1_SOX11	0.526	0.706	0.633	0.757	0.729	1.000	0.743	-0.100	-0.233	-0.211	-0.252				2	6664	2p25	SOX11	Hs.432638
	E1_DBN1	0.507	0.614	0.578	0.701	0.727	0.743	1.000	-0.132	-0.065	-0.150	-0.228				5	1627	5q35.3	DBN1	Hs.130316
	E2_hsa-miR-10a	-0.218	-0.200	-0.195	-0.262	-0.243	-0.100	-0.132	1.000	0.065	0.015	0.214	1.73E-04	down	1.453309					
	E2_hsa-miR-143	-0.132	-0.207	-0.165	-0.223	-0.119	-0.233	-0.065	0.065	1.000	0.758	0.045	2.11E-02	down	1.32976					
	E2_hsa-miR-145	-0.100	-0.169	-0.127	-0.238	-0.148	-0.211	-0.150	0.015	0.758	1.000	0.091	4.00E-03	down	1.446995					
	E2_hsa-miR-155	-0.262	-0.309	-0.378	-0.402	-0.418	-0.252	-0.228	0.214	0.045	0.091	1.000	9.94E-10	down	2.045588					
ROW ANNOTATIONS	E2_p (Corr)								1.73E-04	2.11E-02	4.00E-03	9.94E-10								
	E2_Regulation								down	down	down	down								
	E2_FC (abs)								1.453309417	1.329760313	1.446995139	2.045587778								
	E1_chromosome	2	20	8	4	20	2	5												
	E1_Entrez Gene	9839	4821	9788	1400	4661	6664	1627												
	E1_map_location	2q22	2ter-q11.23	8p22	4p16.1-p15	20q13.33	2p25	5q35.3												
	E1_Symbol_from_	ZEB2	NKX2-2	MTSS1	CRMP1	MYT1	SOX11	DBN1												
	E1_UniGene_clust	Hs.34871	None	Hs.336994	Hs.135270	s.279562	Hs.432638	s.130316												

Figure 8. View of an output file after exporting correlation coefficients. Along with pair-wise correlation coefficient values, the entity associated data and annotations are exported both for rows and columns. The Correlation Coefficients are shown in peach color, the column annotations in blue, and the row annotations in green.

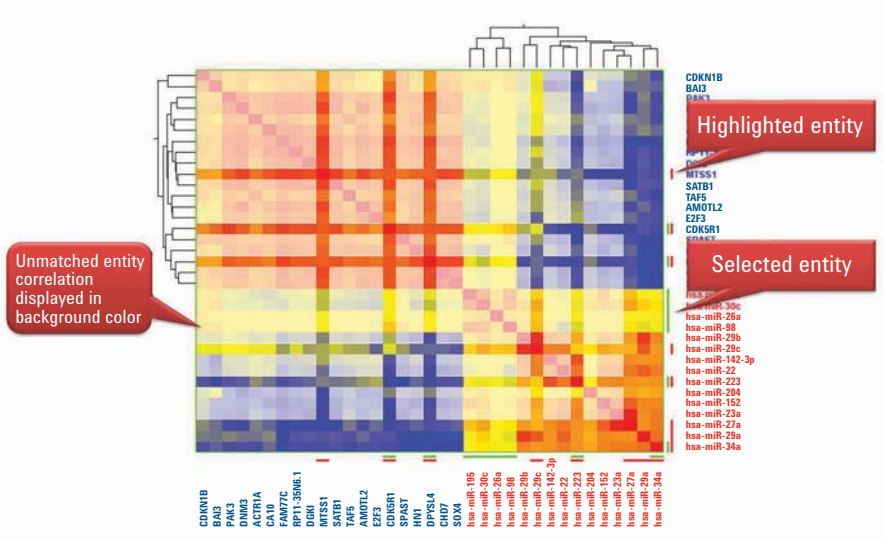


Figure 9. Highlighted entities are shown with a red color hash bar, and selected entities are indicated by a green color hash bar. The highlight and selection colors are customizable and can be changed in the heatmap property dialogue.

Sample-Sample Correlation

In addition to entity-entity correlation, the GeneSpring/MPP Correlation Framework supports pair-wise correlation analysis between biological samples within a given experiment. Sample correlation allows identification of the condition-wise relationships that may exist between the samples in a study.

Sample correlation is performed on the same experiment types that are supported for entity correlation (Table 1). Sample correlation is supported only between the samples in a single experiment; it is not supported for MOA experiments.

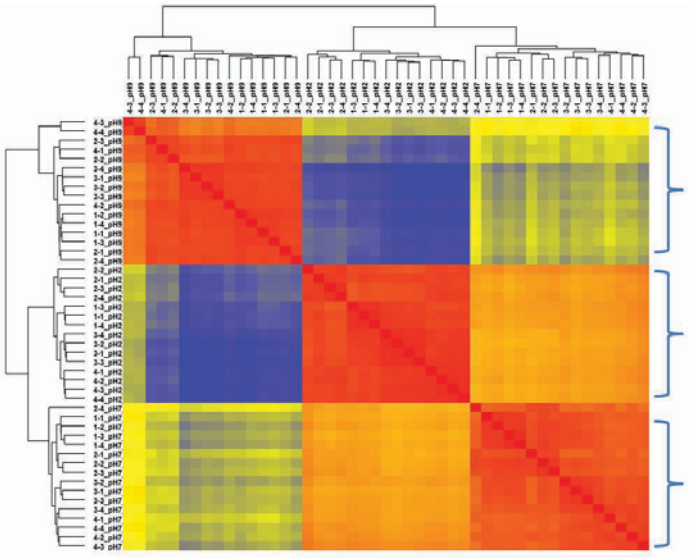
The sample correlation analysis is defined by the entity list, interpretation, and type of correlation coefficient selected as inputs for the analysis. Table 3 depicts an example of a correlation computed between Sample 1 and Sample 2 based on the signal intensity values of Genes 1–12. A selected entity list and interpretation are used to define the samples and entities for correlation analysis.

Sample-sample correlation analysis is sensitive to baseline transformation performed during the experiment creation. The biological inference derived from the analysis will be altered by using baseline transformed data. It is advisable to perform sample-sample correlation on data which has not been baseline transformed.

The output of sample-sample correlation is displayed as a heatmap. Samples in the heatmap can be clustered using hierarchical clustering based on the correlation profiles of the samples (Figure 10).

Table 3. Example calculation of correlation between two given samples.

	Sample 1	Sample 2
Gene 1	0.651	1.372
Gene 2	0.818	1.590
Gene 3	0.945	1.716
Gene 4	0.578	0.643
Gene 5	0.464	1.186
Gene 6	0.675	0.947
Gene 7	0.323	0.642
Gene 8	0.304	0.774
Gene 9	0.043	0.783
Gene 10	0.943	1.452
Gene 11	0.908	1.686
Gene 12	0.415	0.808
Pearson	0.822	



The sample-sample correlation heatmap indicates a strong relationship within a pH as compared to *Plasmodium falciparum* infection status.

Tube no.	NRBC	IRBC at 10 %	SLO 250 stock units	Set A	Set B	Set C
1-1	500 µL			pH 2	pH 7	pH 9
1-2	500 µL			pH 2	pH 7	pH 9
1-3	500 µL			pH 2	pH 7	pH 9
1-4	500 µL			pH 2	pH 7	pH 9
2-1	500 µL		10 µL	pH 2	pH 7	pH 9
2-2	500 µL		10 µL	pH 2	pH 7	pH 9
2-3	500 µL		10 µL	pH 2	pH 7	pH 9
2-4	500 µL		10 µL	pH 2	pH 7	pH 9
3-1		500 µL		pH 2	pH 7	pH 9
3-2		500 µL		pH 2	pH 7	pH 9
3-3		500 µL		pH 2	pH 7	pH 9
3-4		500 µL		pH 2	pH 7	pH 9
4-1		500 µL	10 µL	pH 2	pH 7	pH 9
4-2		500 µL	10 µL	pH 2	pH 7	pH 9
4-3		500 µL	10 µL	pH 2	pH 7	pH 9
4-4		500 µL	10 µL	pH 2	pH 7	pH 9

Figure 10. Sample-sample correlation heatmap for a metabolomics study[2]. Clustering on correlation coefficients clearly demonstrates that samples group together based on their pH values rather than infection status (NRBC = Noninfected RBCs, IRBC = Infected RBCs).

References

1. The results shown in the document are in whole or part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/

2. Sana, T. R; *et al.* Global Mass Spectrometry Based Metabolomics Profiling of Erythrocytes Infected with *Plasmodium falciparum*. *PLoS ONE* **2013**, *8(4)*: e60840. doi:10.1371/journal.pone.0060840.

Notes

Notes

Ordering Information

Product number	Product description
Mass Profiler Professional	
G3835AA	Mass Profiler Professional (MPP) Perpetual
G9274AA	Mass Profiler Professional (MPP) Perpetual Upgrade
G3836AA	Pathway Features for MPP Perpetual
G9275AA	Pathway Features for MPP Perpetual Upgrade
G9277AA	Sample Class Predictor (Perpetual). Allows the use of class prediction models generated by MPP with MSD ChemStation or MassHunter
G9281AA	Mass Profiler Pro (MPP) Concurrent License; allows unlimited installations but only one user to access the program at a time
G9282AA	Mass Profiler Pro (MPP) Concurrent License Upgrade; requires previous purchase of G9281AA
GeneSpring	
G5886AA	GeneSpring GX Standard Perpetual Academic + 1 year SMA
G5887AA	GeneSpring GX Standard Perpetual Commercial + 1 year SMA
G5888AA	GeneSpring GX Standard Upgrade - Academic
G5889AA	GeneSpring GX Standard Upgrade - Commercial
G5890AA	GeneSpring GX Concurrent Perpetual Academic + 1 year SMA
G5891AA	GeneSpring GX Concurrent Perpetual Commercial + 1 year SMA
G5892AA	GeneSpring GX Concurrent Perpetual Upgrade - Academic
G5893AA	GeneSpring GX Concurrent Perpetual Upgrade - Commercial
G3784AA	GeneSpring GX Standalone 1 year - Academic
G3782AA	GeneSpring GX Standalone 2 year - Academic
G3780AA	GeneSpring GX Standalone 3 year - Academic
G3783AA	GeneSpring GX Concurrent 1 year - Academic
G3781AA	GeneSpring GX Concurrent 2 year - Academic
G3779AA	GeneSpring GX Concurrent 3 year - Academic
G3778AA	GeneSpring GX Standalone 1 year - Commercial
G3776AA	GeneSpring GX Standalone 2 year - Commercial
G3774AA	GeneSpring GX Standalone 3 year - Commercial
G3777AA	GeneSpring GX Concurrent 1 year - Commercial
G3775AA	GeneSpring GX Concurrent 2 year - Commercial
G3773AA	GeneSpring GX Concurrent 3 year - Commercial

www.agilent.com/chem

For Research Use Only. Not for use in diagnostic procedures.
This information is subject to change without notice.

Published in the USA, March 17, 2016
5991-5165EN
PR7000-0389

Agilent Technologies