Agilent GeneSpring/MPP Metadata Analysis Framework

Technical Overview

Authors

Srikanthi R., Pritha Aggarwal, Durairaj R., Maria Kammerer, and Pramila Tata
Strand Life Sciences
Bangalore, India

Michael Rosenberg
Agilent Technologies, Inc.
Santa Clara, California, USA

Introduction

Clustering analysis is an efficient way to group the samples and conditions in a dataset into subsets based on the similarity of their abundance profiles. Sample clustering has been broadly used for inferring disease subtypes and for patient stratification. Used in this context, hierarchical clustering can be a very important analysis tool for revealing the molecular mechanisms underlying biological function. New in GeneSpring/MPP 13, the metadata analysis framework allows researchers to visualize the abundance profiles of samples alongside metadata such as administrative, physiological, or technology related information. The metadata visualization framework allows researchers to reveal tacit dependencies between characteristics of the subjects or samples and their gene, metabolite, or protein expression profiles.

Agilent Technologies

Sample Metadata

The sample attributes and associated parameters contribute to metadata. The metadata of a biological sample can be divided into one of the following categories:

1. **Administrative**—who/when/where/how collected the sample
2. **Physiological attributes of the subject**—tumor-normal, blood-biopsy, drug-placebo, cell type
3. **Technology or experiment design**—TNM* staging, treatment time, drug dosage, batch, QC parameters

*TNM staging: a method for classifying malignant tumors based upon tumor size, number of lymph nodes involved, and distant metastasis.

Each of the metadata types can be numerical or categorical, as well as discrete or continuous. Table 1 lists the type of plots supported in GeneSpring for different categories of metadata. The GeneSpring metadata framework supports all types of attributes. In GeneSpring, the researcher can now align experimental metadata alongside the samples in the clustering heatmap either as bar charts, scatter plots, metadata heatmaps, or label plots. Association between various vital, pathological, and molecular parameters and sample clusters allows researchers to identify new relationships between expression patterns and phenotypes.

One example of such analysis is illustrated in Figure 1 using a dataset from The Cancer Genome Atlas (TCGA)[1]. Gene expression data from 220 samples was clustered, and the metadata was used for biological data analysis. For example, TCGA provides information about the main pathways deregulated in these samples. This information for all the samples can be depicted, in a concise manner, using the metadata bar chart functionality, whereby every sample bearing mutations in the WNT pathway is indicated with a green bar (Figure 1, panel A), and mutations in the TGF-beta pathway with a yellow bar (Figure 1, panel B).

A quick glance at Figure 1 shows that a larger number of samples have mutations in WNT pathway as opposed to TGF beta pathway. Details about other implicated pathways can be added in a similar manner. Findings from familial history (for example, number of first degree relatives affected by the same condition, Figure 1, panel C) and survival time (after diagnosis, Figure 1, panel D) are represented as scatter and profile plots, respectively. Information related to the lymph nodes, such as number of examined nodes and number of nodes with a pathological spread is shown using a metadata heatmap (Figure 1, panel E). Other parameters that could be shown here include the mutation rates of individual samples, presence of recurrent mutations, and methylation status of specific genes (Figure 1, panel F), which enable researchers to observe the relationship between expression of target genes, mutation rates, and subanatomical location of the condition.

Table 1. Plots supported in GeneSpring for different categories of sample metadata.

Plot	Numeric attributes	Categorical attributes
Heatmap	Yes	Yes
Scatter Plot	Yes	No
Profile Plot	Yes	No
Bar Chart	Yes	No
Label Plot	Yes	Yes

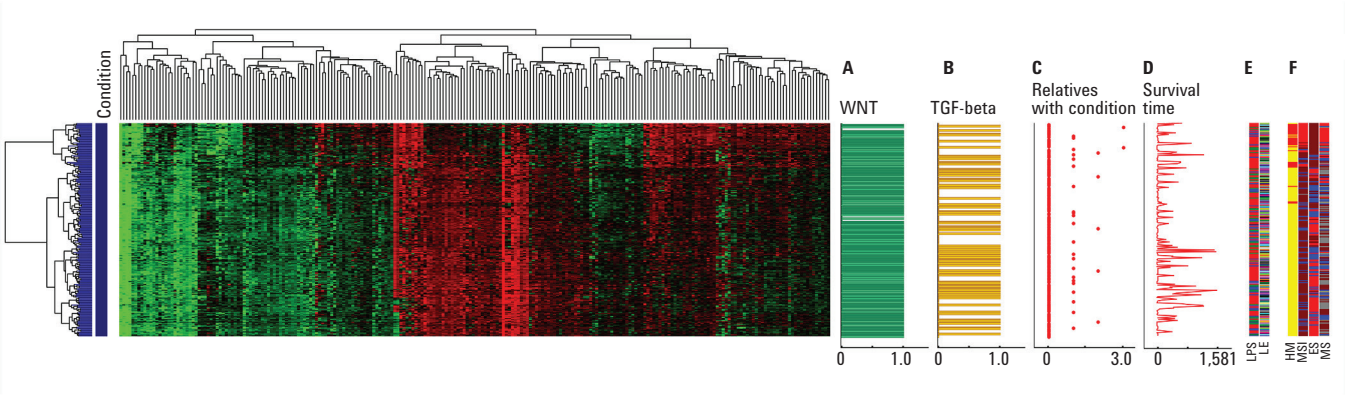

Figure 1. Two-dimensional hierarchical clustering of 220 samples from TCGA with the visual alignment of metadata.
(Labels in panels E and F: LPS – Lymphnode pathologic spread, LE – Lymphnode examined, HM-Hypermutated, MSI – Microsatellite Instability Status, ES – Expression subtype, MS – Methylation subtype)

Import and Visualization of Sample Metadata

Sample metadata is imported in a spreadsheet format using the Experimental Grouping wizard. Once it is imported, a user can launch heatmap, bar chart, profile, scatter, or label plots, and configure appropriate metadata for each plot as shown in Figure 2.

Pathological Assessment and Biological Inferences Using Sample Metadata Plots

The metadata framework is a valuable tool for drawing biological inferences. GeneSpring allows researchers to align metadata with the clustered or unclustered heatmap and to sort the heatmap in the order of any metadata attribute. An ordered heatmap reveals the underlying data patterns, as illustrated by Figure 3. Figure 3A shows the gene expression patterns of samples that exhibited 2-fold or greater fold-change after treatment (GSE21974[2]). Red and green bar charts show the lesion sizes

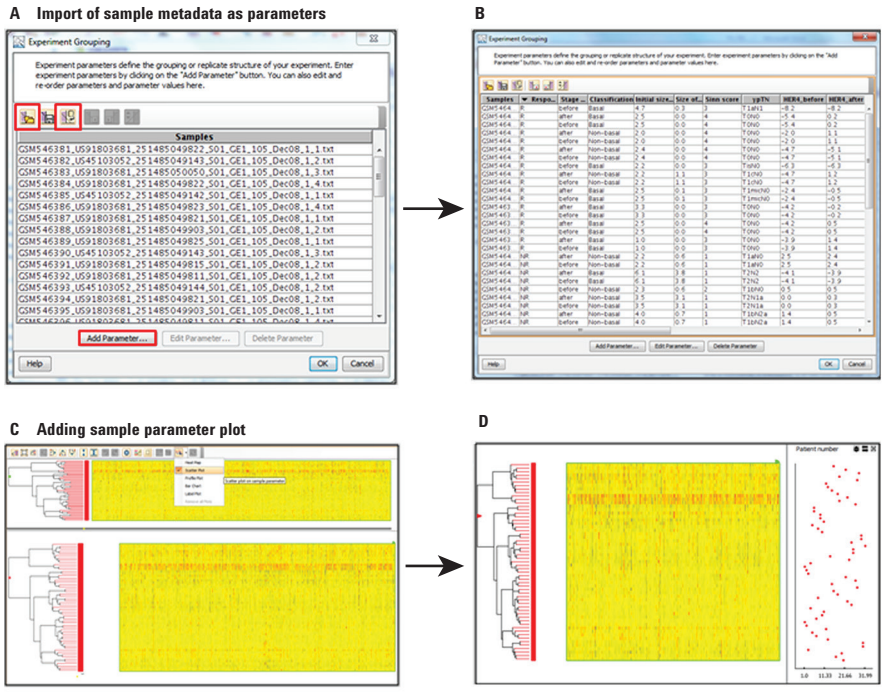

Figure 2. Import of the sample metadata. A) In the Experiment Grouping wizard, sample metadata can be added from a file containing the grouping information, by importing attributes of existing samples, or manually using the **Add Parameter** functionality. B) Imported sample parameters seen in experimental grouping. C) Launching a cluster tree and adding sample parameter plots. D) Metadata viewed as a scatter plot.

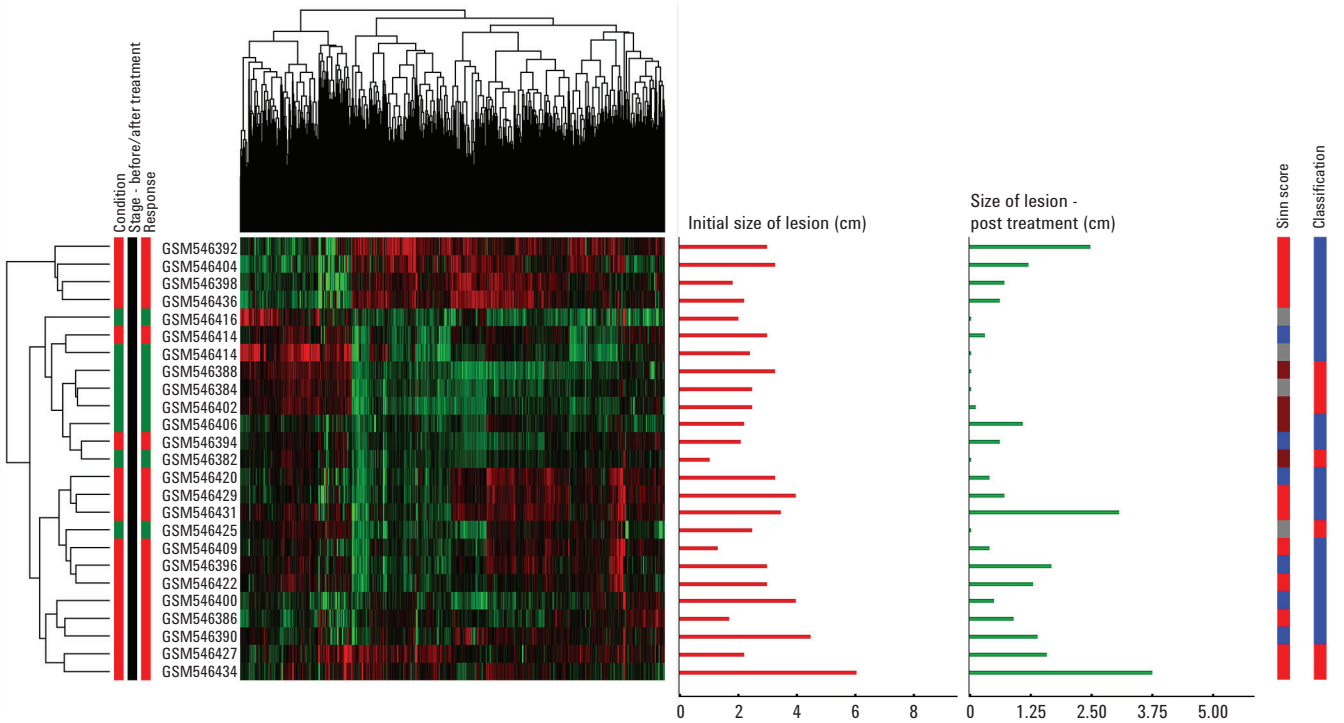

Figure 3A. Lesion sizes before and after treatment are shown by red and green bars while the assigned Sinn scores and classification are depicted by a heatmap.

for individual samples before and after treatment respectively. It is clear from this view that the gene expression pattern shows significant but incomplete correlation with the desired outcome (regression), possibly due to other contributing variables. Incorporating the semiquantitative regression measure, the Sinn score[3]*, and the molecular subtype of the condition allows a user to further observe the relationship between the subtype (basal, nonbasal) and a sample's responsiveness to treatment. The extent of size regression post treatment for each Sinn score and its relationship to the differential gene expression pattern can be highlighted by sorting the heatmap as illustrated in Figure 3B. Thus, the difference between responders and nonresponders becomes more evident. Staging information, the ypTN[4], can be used as label plots to augment the information gleaned from Sinn scores.

In a multifactorial study design such as the present case, expression profiles of samples are governed by many parameters. In these cases, the ability to supplement gene expression clustering with additional metadata and the ability to sort the clustered view using different parameters facilitates exploration of the underlying patterns in the expression profiles of the samples. These, in turn, can be used to answer critical questions, such as if the molecular subtype of a sample influences its response to treatment.

*Sinn scores range from 0–4, with 0 representing lesions showing no regression and four representing responders in whom no viable residual lesions are seen.

Expression of Potential Key Regulators as Metadata Plots

Identification of key regulators of differentially expressed genes is critical to understanding the pathways involved in disease progression. Figure 3C displays one such scenario using HER4, which is thought to be one of the critical genes associated with response to treatment[2]. To answer the question if HER4 can serve as an indicator of positive response to treatment, the user can export the normalized expression values of the probe corresponding to HER4 from samples before and after treatment and re-import them as metadata attributes. Figure 3C shows expression values of HER4 plotted as bar charts. Note the clear alignment between changes in HER4 expression before and after treatment for Sinn score 3 and 4.

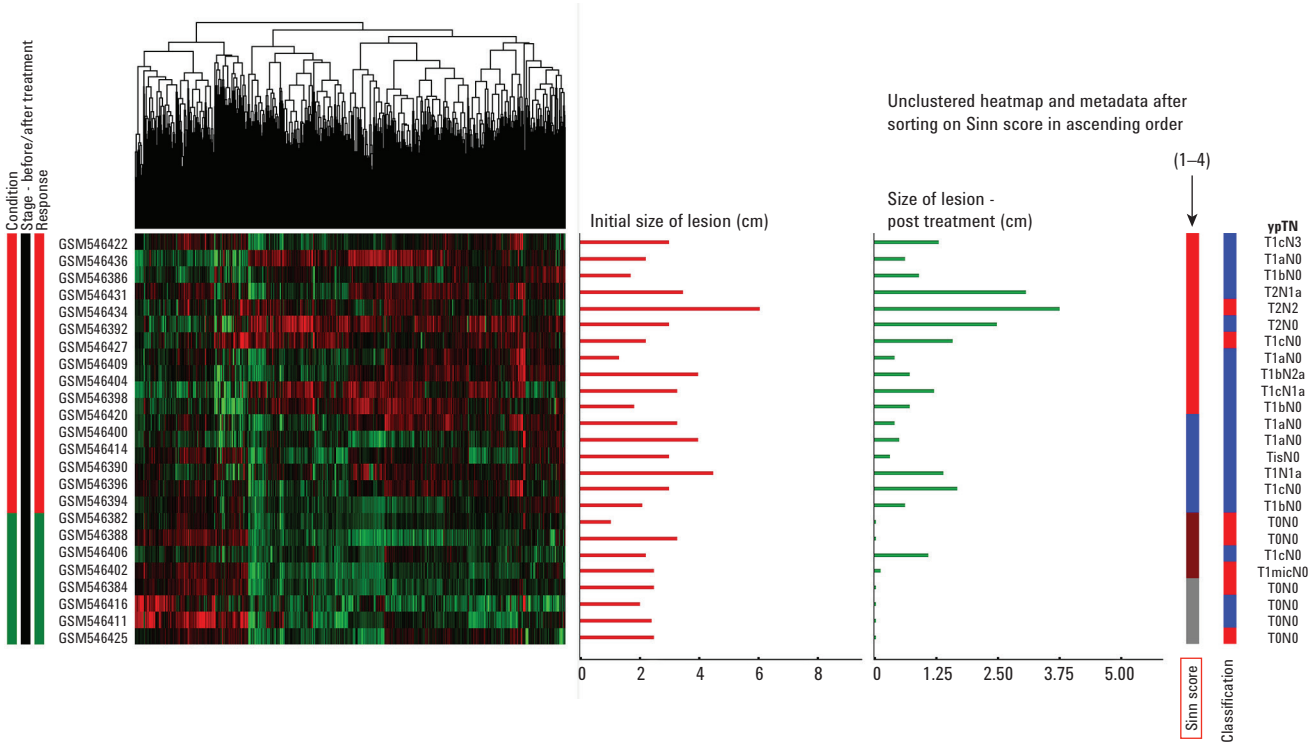

Figure 3B. Heatmap and metadata sorted on Sinn scores. The dendrogram on the left side of the view is removed while the now unclustered sample heatmap and the corresponding metadata in the other metadata plots are reorganized according to the sort order. The brown and grey colors in the Sinn score heatmap correspond to Sinn scores of 3 and 4 (corresponding to non-invasive or no viable lesion residuals). An aberrant sample with lesser lesion size regression and a different gene expression pattern now clearly stands out among samples with a Sinn score of 3 and 4. This sample has a staging of ypT1N0.

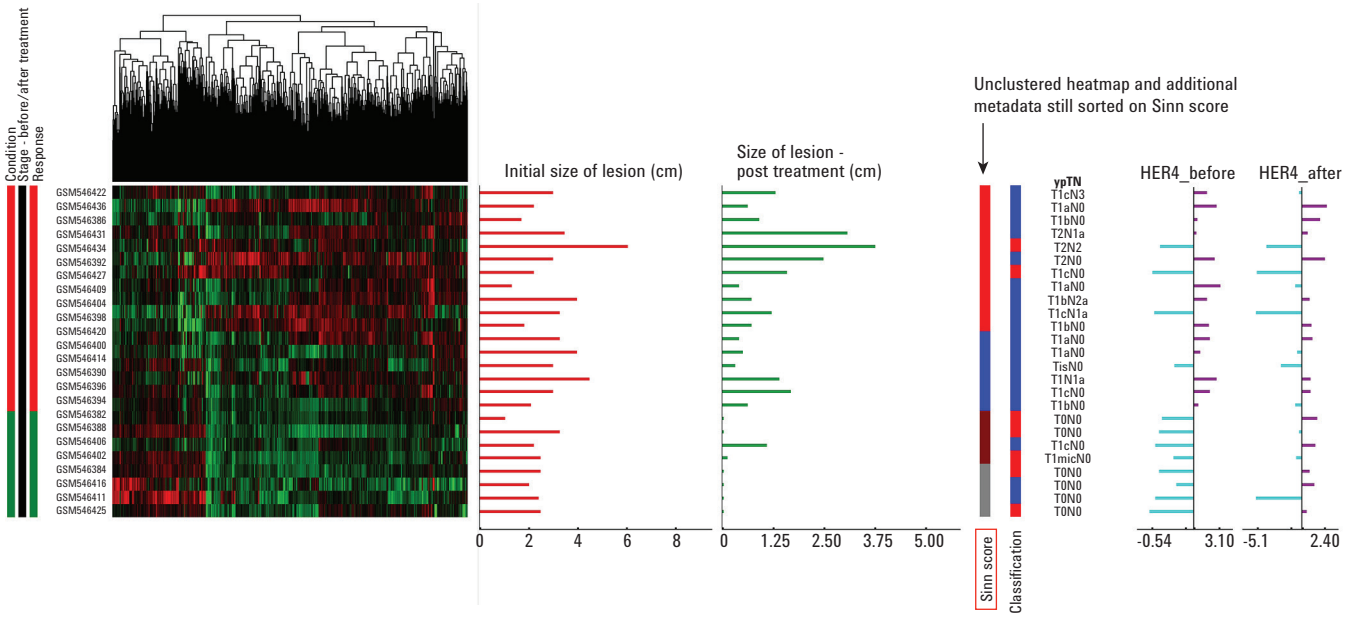

Figure 3C. HER4 shows up-regulation in samples with size regression (Sinn score 3 and 4). Normalized expression values for HER4 (probe A_32_P183765) are shown before and after treatment.

Conclusions

Multiple interdependent factors govern behavior of complex biological systems. The GeneSpring metadata analysis framework provides important visual cues for biological interpretation of the gene, protein, or metabolite abundance patterns. The innovative synchronized views of the expression heatmaps generated from omics data, combined with physiological attributes, help reveal the intrinsic interplay between the parameters, allowing scientists to better understand and interpret complex systems.

References

1. *Nature* 19 July **2012**, *487*, pp 330-337. http://www.ncbi.nlm.nih.gov/pubmed/22810696
2. *Oncol. Rep.* Oct **2011**, *26(4)*, pp 1037-45. http://www.ncbi.nlm.nih.gov/pubmed/21769435
3. *JCO* May 20, **2012**, *vol. 30 no. 15*, pp 1796-1804. http://www.ncbi.nlm.nih.gov/pubmed/22508812
4. Purposes and Principles of Cancer Staging http://www.kliinikum.ee/ho/images/stories/attachments/102_Purpose and Principles of Cancer Staging.pdf.

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

Notes

www.agilent.com/chem

This information is subject to change without notice.

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

PR7000-0098

Published in the USA, January 4, 2016
5991-4984EN

Agilent Technologies