

Identification of biological themes in microarray data from a mouse heart development time series using GeneSifter™

VizX Labs, LLC
Seattle, WA 98119

Abstract

Oligonucleotide microarrays were used to study gene expression during mouse heart development. Six timepoints from the E12 stage of embryonic development through 1 year old were examined with the Affymetrix® GeneChip® Mouse Genome U74A array. Data analysis using GeneSifter was performed to identify differentially regulated genes and distinct patterns of gene expression. The goal of this study is to determine the biological significance of the gene lists derived using these methods. ANOVA analysis, followed by adjustment of the raw p-values using the false discovery rate method of Benjamini and Hochberg, identified 1513 genes that showed significant differential regulation across the time series. PAM (partitioning around medoids) clustering separated these genes into two groups: one cluster contained genes showing down-regulation during the time series, the other cluster had genes showing up-regulation. Z-score analysis of the gene ontologies associated with the genes in each cluster identified distinct biological themes associated with each cluster: genes involved in the mitotic cell cycle were significantly overrepresented in the down-regulated cluster, while genes involved in immune response and energy pathways were significantly overrepresented in the up-regulated cluster.

Introduction

Microarrays offer the possibility of measuring the expression levels of tens of thousands of genes in parallel. Recently, several array manufactures have introduced whole genome chips for several model organisms. Advances in microarray technology, sample preparation, and the availability of skilled core facilities, have made the generation of high quality data routine in many settings. However, converting this data into meaningful biological knowledge still remains the primary challenge for many researchers. The steps required for this conversion can be broken down into three discrete sets of tasks: identification of significantly regulated genes, identification of global patterns of gene expression, and the determination of the biological meaning of both individual genes and groups of genes. The first two steps result in a gene list, while the third step deals with understanding the biology behind these gene lists. If experiments are performed with an adequate number of replicates, there are several standard comparative statistical tests available which will generate a list of genes whose differential expression is statistically significant (Draghici, 2002). Unsupervised clustering methods can then be used to identify patterns of gene expression among the differentially regulated genes. Although these steps are critical in the evaluation of microarray data, determining the biological significance of the genes identified is perhaps the most important step in the analysis of this type of data. This analysis is often done on a gene-by-gene basis using annotation supplied by genomic databases. Several challenges exist with this approach. First, a reliable source of gene annotation must be available. If the analysis is

performed manually, it can be extremely time-consuming unless the number of genes in the gene list is limited in some way. However, by analyzing only a small subset of the data, the larger biological context is lost. Automation of microarray data analysis promises to overcome these difficulties and provide high-quality, meaningful information in a short period of time.

GeneSifter combines statistical analysis tools with methods for gene expression pattern recognition. It then allows users to rapidly identify the significance of individual genes, as well as biological themes in sets of genes first identified by the statistical and pattern recognition methods. GeneSifter uses Gene Ontology (GO) Reports and z-scores to summarize the biological processes represented in a particular list. Z-scores can then be used to identify GO terms that are significantly over- or under-represented in this list. This allows for rapid characterization of the broad biological themes affected in a particular experiment. This approach can be efficiently used for gene lists containing thousands of genes. This study outlines the use of these methods to identify biological themes associated with discrete expression patterns in a mouse heart development time series.

The Data

The data used in this analysis was obtained from the CardioGenomics PGA (Genomics of Cardiovascular Development, Adaptation, and Remodeling. NHLBI Program for Genomic Applications, Harvard Medical School. URL: <http://www.cardiogenomics.org> [May 2004]).

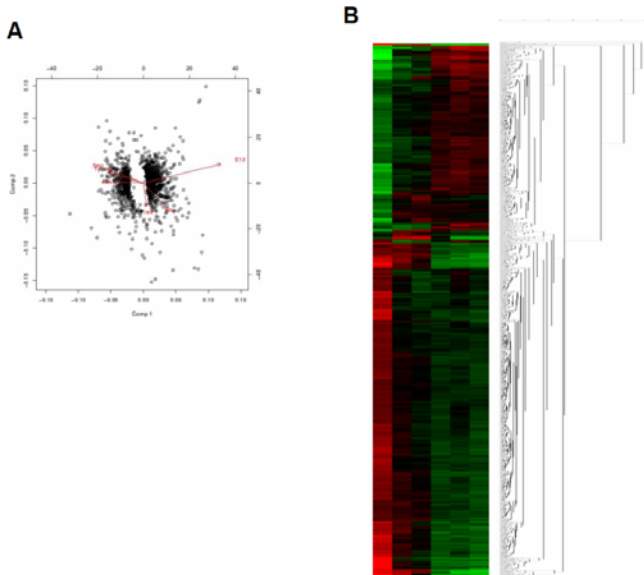


Figure 1: Visualization of filtered gene list. A) PCA analysis of expression data for 1513 significantly regulated genes. B) Hierarchical clustering of expression data for 1513 significantly regulated genes.

The data set was created to study the complex regulatory mechanisms underlying cardiac development. The data are intended for use as a benchmark for other studies involving cardiac development and aging. In addition to allowing the study of gene expression in development and aging of the FVB strain, this data provides a common reference for comparisons with other existing transgenic or knockout models.

The design of this study includes both embryonic and adult ventricular tissue from wildtype FVB mice at the following times:

- Embryonic day 12.5
- Neonatal day 1
- 1 week of age
- 4 weeks of age
- 5 months of age
- 1 year of age

Three samples were examined for each timepoint. For the embryonic stage, three hearts were pooled for each experiment. Two hearts were pooled at the neonatal stage. Adult hearts were analyzed individually. For adult mice at 5 months and one year of age, hearts were collected from male mice. The samples were then hybridized to Affymetrix® GeneChip® Mouse Genome U74A array.

CEL files were obtained for each experiment and were uploaded into GeneSifter using the Advanced Upload tool. Expression values were derived from the probe level data using Robust Multi-array Average, or RMA (Irizarry, et al., 2003). A project was created containing experiments from all timepoints. The three replicates were combined for each timepoint and no additional normalization was performed.

The data was saved using the “Data already normalized” option, as the RMA-derived values were log₂-transformed.

Statistical Analysis

The Affymetrix® GeneChip® Mouse Genome U74A array contains ~10,500 probe sets representing ~9000 different transcripts. Prior to visualization, the project was filtered to identify a subset of genes that were significantly differentially regulated during the development time series. ANOVA was performed on the entire dataset ($p < 0.05$). Correction for multiple testing was performed using the method of Benjamini and Hochberg (Benjamini and Hochberg, 1995) to derive a false discovery rate estimate from the raw p-values. A false discovery rate of 5% was used as a cutoff for statistical significance. This produced a set of 5500 genes. This list was further filtered by setting a fold-change cut-off of 2, relative to the first timepoint. This reduced the size of the list to 1513 genes. This set of filtered genes was saved as a sub-project and then subjected to further analysis.

Ontology	List	Array	z-score
mitotic cell cycle	46	134	5.47
DNA metabolism	61	206	4.99
main pathways of carbohydrate metabolism	21	51	4.66
nuclear mRNA splicing, via spliceosome	17	38	4.6
mRNA metabolism	25	67	4.5
DNA packaging	24	64	4.44
nucleocytoplasmic transport	16	36	4.43
DNA replication	23	64	4.1
RNA splicing	17	42	4.1
cell organization and biogenesis	84	340	4
mRNA processing	22	63	3.85
energy derivation by oxidation of organic compounds	23	68	3.76
energy pathways	23	70	3.6
establishment and/or maintenance of chromatin architecture	19	57	3.34
electron transport	37	136	3.26
hexose catabolism	12	31	3.26
monosaccharide catabolism	12	31	3.26
alcohol catabolism	12	31	3.26
glucose catabolism	12	31	3.26
chromosome organization and biogenesis (sensu Eukarya)	21	67	3.18
glycolysis	10	25	3.1
metabolism	482	2621	3.08
muscle development	18	56	3.07
RNA processing	33	122	3.04
muscle contraction	11	30	2.9
protein targeting	22	76	2.83
nuclear organization and biogenesis	21	72	2.81
organic acid metabolism	42	170	2.78
carboxylic acid metabolism	42	170	2.78
protein folding	16	51	2.78
coenzyme and prosthetic group metabolism	21	73	2.73
chromatin assembly/disassembly	10	28	2.67
ribosome biogenesis and assembly	12	36	2.65
carbohydrate catabolism	12	36	2.65
ribosome biogenesis	12	36	2.65
cell growth and/or maintenance	289	1524	2.64
cytoplasm organization and biogenesis	56	244	2.6
alcohol metabolism	25	93	2.6
fatty acid metabolism	18	62	2.57
cell growth	11	35	2.31

Figure 2: Z-core report for filtered gene list. 40 over-represented biological process ontologies associated with 1513 significantly regulated genes.

Visualization and examination of filtered gene list

The filtered set was then visualized using hierarchical clustering and principal component analysis (PCA). PCA suggested the existence of two discrete clusters of genes (figure 1). The data was also visualized using hierarchical clustering. This method suggested two groups of genes: those whose expression decreased relative to the embryonic samples during development, and those whose expression

increased during development. Several subsets were apparent within each general set, however, the dominant patterns appeared to be increasing and decreasing over time.

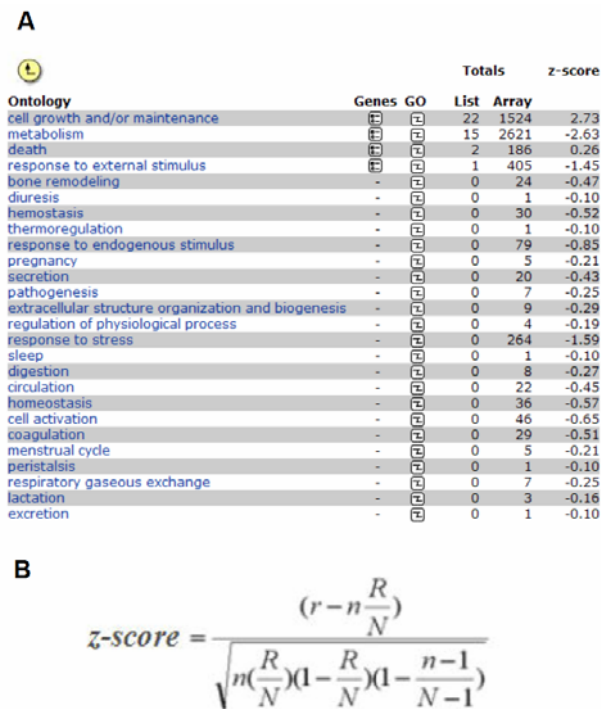


Figure 3: Ontology report A) B) Z-score calculation used to determine whether ontologies are over or under represented in a gene list.

Ontologies and the z-score report

Annotation is available for each of the genes in the filtered dataset. This annotation includes GO terms provided by LocusLink. The GO terms associated with all of the genes in the list can also be summarized with an Ontology Report. The Ontology Report provides information for each GO term in the Biological Process, Molecular Function, or Cellular Component categories. The report includes several pieces of information for each term. The total number of genes with that ontology in the list is displayed, as well as the total number of genes with that ontology on the whole array. This information can be used to calculate a z-score for each ontology term (Doniger, et al., 2003). The z-score indicates whether the specific GO term occurs more or less frequently than expected. Extreme positive numbers (greater than 2) indicate that the term occurs more frequently than expected, while an extreme negative number (less than -2) indicates that the term occurs less frequently than expected. A z-score report can also be generated for each gene list. This report lists all of the ontologies for a particular category with a z-score greater than 2 or less than -2 (figure 3).

A z-score report was generated for the genes in the filtered dataset (figure 2). This report was limited to only the over-represented ontologies in the biological process category. Ontologies with significant z-scores included cell cycle,

energy pathways. A list of the 40 largest z-scores is presented in figure 2

Identification of expression patterns

The z-score report identified specific biological themes in the filtered dataset. Hierarchical clustering and PCA analysis suggested the presence of at least two distinct patterns of gene expression in the time series. It could not be determined whether the ontologies identified in the z-score report are associated with a particular pattern of gene expression for the dataset as a whole. Therefore, we decided to separate the genes in the filtered set based on expression pattern to determine whether specific ontologies segregate with the separate patterns of gene expression.

In order to examine the two sets of genes separately, k-medoids clustering was performed on the filtered set of genes. The partitioning around medoids (PAM) method was used, and several values for k were chosen, ranging from 2-6. Both PCA and hierarchical clustering suggested the existence of two clusters. Application of the silhouette method (Kaufman and Rousseeuw, 1990) confirmed that k=2 gave the best grouping of the filtered data. k=2 gave an average silhouette width of 0.599 which was the largest value obtained for the values of k examined. Cluster 1 contained 533 genes and showed a trend of up-regulation relative to the embryonic timepoint. Cluster 2 contained 980 genes and showed decreased expression during development. Heatmaps for each cluster were generated, and confirmed the overall expression patterns in each cluster (figure 4).

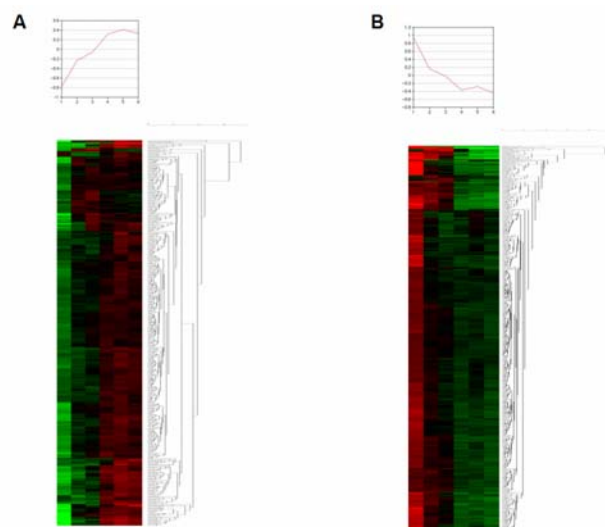


Figure 4: Visualization of clusters identified by PAM. A) Center and hierarchical clustering of genes in cluster 1. B) Center and hierarchical clustering of genes in cluster 2.

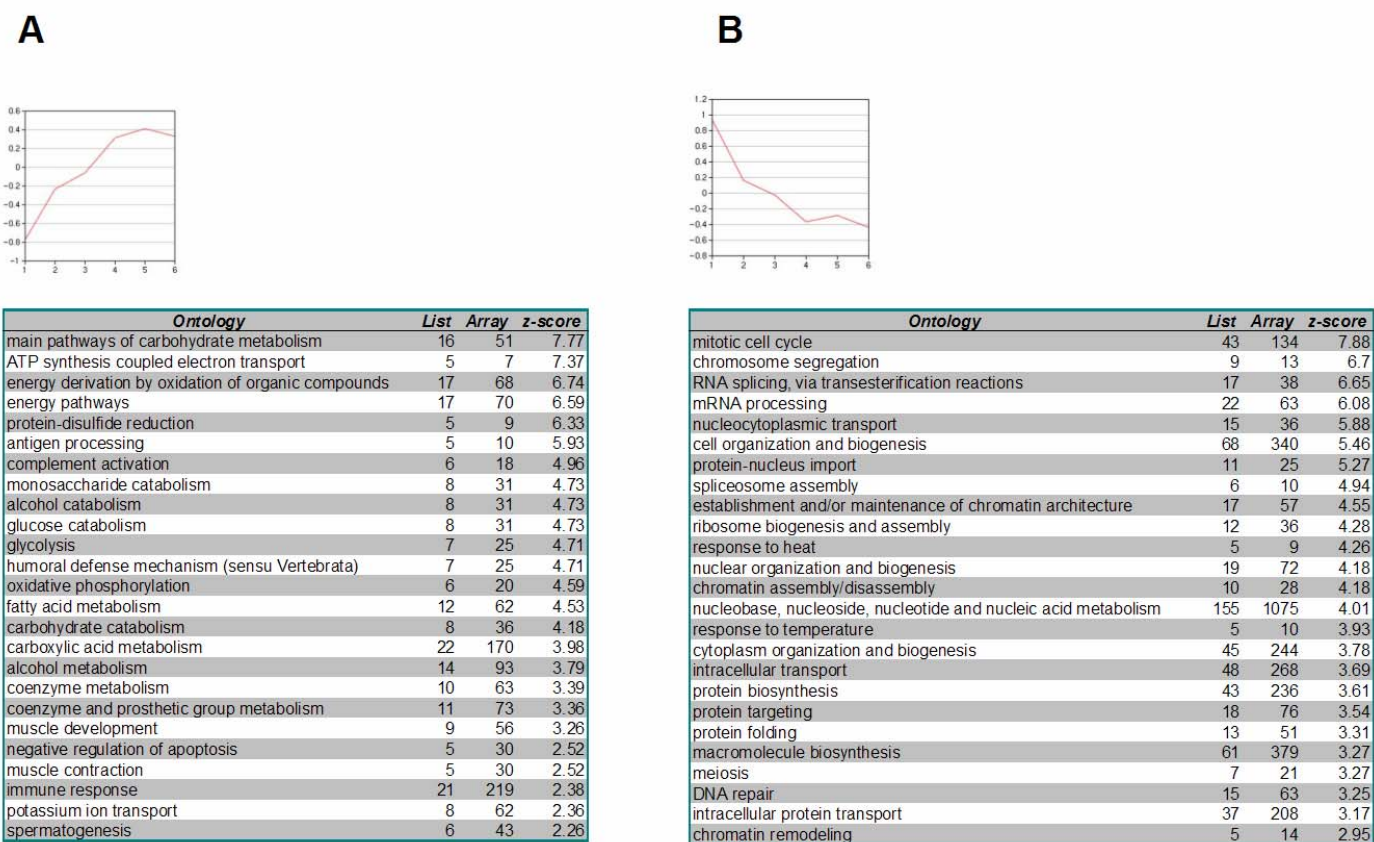


Figure 5: Z-score reports for genes in clusters 1 and 2. A) Top 25 biological process ontologies over-represented in cluster 1. B) Top 25 biological process ontologies over-represented in cluster 2.

Distinct biological themes are associated with different expression patterns

Ontology reports were examined within each cluster (figure 5). The z-score report identified distinct biological themes associated with each cluster. Prominent themes identified in cluster 1 included immune response, energy pathways, and carbohydrate metabolism. Cluster 2 included genes involved in cell cycle, mRNA processing and protein transport. Figure 5 lists the top 25 biological process ontology identified for each cluster. See supplemental material to view a comprehensive list of ontologies identified for each cluster.

Summary and Conclusions

The goal of the study was to identify differentially regulated genes and patterns of gene expression among this set of genes. Additionally, we sought to characterize the biological themes associated with each pattern of gene expression. Expression measurements were derived from the probe level data contained in the CEL files using RMA. We used RMA because it has been shown to reduce noise, especially at lower intensity values, when compared to MAS 5 (Rosati, et al., 2004). Although all analysis presented here used RMA-transformed data, the results were similar when analysis was performed using MAS5-derived expression values (data not shown).

As there were three replicates for each timepoint, ANOVA was chosen for assessing the statistical significance of any changes in gene expression across the time series. Since over 10,000 tests were performed, we chose to adjust the p-value based on the number of tests performed. The method of Benjamini and Hochberg was used to adjust the raw p-value such that the adjusted p-value represented the false discovery rate. This produced a list of 5500 genes that were differentially regulated. We further applied a 2-fold cutoff to produce a list of 1513 genes that had a least a 5% FDR and displayed at least a 2-fold change in expression across the time series, relative to the earliest timepoint. This cutoff was chosen to limit our analysis to only the most highly regulated genes. However, this also excluded many genes that show expression changes that were statistically significant, but of a low magnitude. Further analysis would presumably identify genes that showed a lesser degree of regulation.

In order to visualize the filtered set of 1513 genes, and identify dominant patterns of gene expression, the data was visualized using hierarchical clustering. This analysis revealed two distinct patterns of gene expression – genes down-regulated during development and genes up-regulated during development. Many less dominant patterns were observed within each cluster, but except for a small set of

outliers, most of the difference seemed to be accounted for by an early split in the dendrogram, which divided the data into the two groups described. The data was also visualized using principal component analysis (PCA). When the genes were plotted against the first two principal components, two obvious groups were identifiable.

To further determine whether these groups represented a specific pattern of gene expression, partition clustering was used. PAM was chosen for several reasons. PAM is a k-medoid-based method, which has been shown to be both more robust, and more reproducible, than the popular k-means algorithm (Dudoit and Fridlyand, 2002). Rather than selecting starting centers at random, PAM evaluates all possible starting centers and chooses the “best” centers to start cluster building. This gives consistent results when clustering is repeated. The choice of k for partitioning methods is an important issue and several methods have been suggested for selecting an appropriate k. These include empirical methods with very subjective guidelines, such as selecting a value that maximizes the diversity of clusters, yet minimizes the duplication of patterns. Using a method such as PCA offers a more robust alternative. With this method the user tries to estimate the number of clusters based on the grouping of genes clustered against the first two principal components. Depending on the data, this method can be less subjective than the empirical methods, but when the clusters are not well separated it may be difficult to resolve two adjacent groups. The silhouette method uses a post-hoc measurement to determine the fit of the clustering with the actual data structure. An average silhouette width can be calculated for each value of k examined. This value can then be used in an objective way to choose the most appropriate k, the best being the value that which produces the largest average silhouette width.

With the filtered dataset, PCA analysis showed two distinct groups of genes. Therefore, we chose k=2, although several other values were also examined (data not shown). Application of the silhouette method (with k=2) resulted in the largest average silhouette width (best structure). Larger values of k produced clusters with similar centers, differing only in magnitude. Thus, all three methods suggested that two was the most appropriate value for k. We next attempted to determine the biological functions that might be represented in these two groups of genes.

In addition to these visualizations, a z-score report was created for each group in the filtered dataset. The z-score report ranks GO terms based on their over- or under-representation in the current gene list. In the two clusters of genes previously identified, clear metabolic and signaling pathways were identified. In the first cluster, which demonstrated up-regulation over the time course, we identified genes involved in immune response, energy pathways, and carbohydrate metabolism. In the second cluster, representing down-regulated genes, we identified several different gene families, including cell cycle, mRNA processing, and protein transport.

We have demonstrated an approach to analysis of microarray data which takes into account both statistical and biological parameters to determine the results of an experiment. The application of statistical tests (ANOVA), as well as basic filtering methods based on the fold change in gene expression can eliminate much of the noise (genes not affected by the experimental condition), and generate a useful gene list. Using several different visualization and clustering methods, including PCA, PAM, and silhouettes, we showed that the data contain two major patterns of gene expression: one which was up-regulated over time, and the other which trended towards down-regulation. We found cell cycle genes to be robustly up-regulated, and believe these genes to be critical in the differentiation processes inherent in development. Across the same time course, genes involved in energy metabolism were appropriately down-regulated, as energy use decreased during development. Immune response genes were also down-regulated, on a similar time scale as the energy metabolism genes. These gene families are all known to play a role in mouse ventricular development. In addition, we also found several specific genes, which were not part of a larger pathway, to be differentially regulated. Among these were: thrombin receptor, MARCKS-like protein, and calpain 6. There are undoubtedly other specific genes, as well as other gene families, which are important to these developmental processes. In this report, we have demonstrated that the use of GeneSifter combines statistical analysis, pattern recognition and the determination of biological significance, allowing users to rapidly identify biological themes associated with a pattern of gene expression and to understand the biology of a gene cluster.

Supplemental Material

Please contact us at info@GeneSifter for information about complementary use of GeneSifter for analysis of the FVB heart development time series. Additionally, tutorials for reproducing the analysis presented in this study are also available.

References

- Draghici. 2002. Statistical intelligence: effective analysis of high-density microarray data. *Drug Discovery Today*, 7(11)-suppl.: 55-63.
- Parmigiani, et al. *The Analysis of Gene Expression Data: Methods and Software*. New York: Springer; 2003.
- Xu and Li. A comparison of parametric versus permutation methods with applications to general and temporal microarray gene expression data. *Bioinformatics* 19(10):1284-1289.
- Dudoit, et al. 2003. Multiple hypothesis testing in microarray experiments. *Statistical Science* 18(1): 71-103.

Reiner, et al. 2003. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* 19(3):368-375.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B.*, **57**, 289–300.

Rosati, et al. 2004. Comparison of different probe-level analysis techniques for oligonucleotide microarrays. *Biotechniques* 36(2): 316-322.

Irizarry, et al. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4(2): 249-64.

Li and Wong. 2001. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA* 98(1): 31

Bolstad, et al. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19(2): 185-93.

Kaufman L, Rousseeuw PJ: *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley; 1990. First description of the silhouette method in partition clustering.

Kaufman L, Rousseeuw PJ: *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley; 1990.

Dudoit and Fridlyand. 2002. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology* 3(7):1-21.

Doniger, et al. 2003. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biology* 4:R7.