

Partition clustering identifies four distinct patterns of gene expression in genome-wide analysis of *Drosophila* immune response

Abstract

Gene expression profiles in adult fruit flies subjected to microbial infection were monitored using the Affymetrix® GeneChip® *Drosophila* Genome Array (De Gregario et al., 2001). CEL files were obtained from the authors' website and RMA was used to derive expression measurements from the probe level data. Of the more than 13,500 genes interrogated, 624 were differentially expressed (at least 1.5 fold-change and less than a 5% false discovery rate) following infection. Hierarchical clustering and PCA suggested that at least 4 distinct patterns of gene expression were present in this filtered dataset. PAM and was used to partition the 624 genes into 4 discrete cluster. The biological process ontologies associated with the genes in each cluster were also analyzed using a z-score report. This analysis showed that distinct biological themes were associated with each of the 4 patterns of gene expression.

Introduction

In this study GeneSifter was used to analyze microarray data generated from a *Drosophila* immune response time series. This analysis process can be broken down into three discrete 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. GeneSifter uses Gene Ontology (GO) Reports and z-scores to summarize the biological processes represented in a particular gene 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 study outlines the use of these methods to identify biological themes associated with discrete expression patterns in an immune response time series.

The Data

Gene expression profiles in adult fruit flies subjected to microbial infection were monitored using the Affymetrix® GeneChip® *Drosophila* Genome Array (De Gregario et al., 2001). Samples were prepared from uninfected adult flies (0hr) and from flies injured by pricking with a bacteria laden needle. The following time points were examined in this study:

1.5 hours after infection (1.5hr)

3 hours after infection (3hr)

6 hours after infection (6hr)

12 hours after infection (12hr)

Three biological replicates were prepared for each time point. These samples were hybridized to the Affymetrix GeneChip *Drosophila* Genome array. CEL files were obtained for each sample from the authors' website. The CEL files were loaded into GeneSifter and RMA was used to derive an expression measurement for each probe set.

Filtering and visualization

The Affymetrix® GeneChip® *Drosophila* Genome array contains ~14,000 probe sets representing ~13,500 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.

A 1.5 fold-change cutoff was first applied and then ANOVA was performed on the 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 filtering produced a set of 624 genes. This set of filtered genes was saved as a sub-project and then subjected to further analysis. Hierarchical clustering and PCA were used to visualize these 624 genes. Both methods identified several distinct patterns of gene expression within the 624 differentially expressed genes (figure 1).

Partition clustering

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. Application of the silhouette method (Kaufman

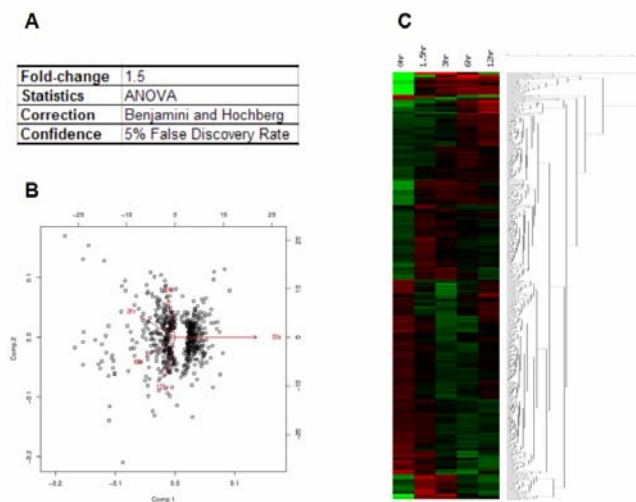


Figure 1: Filtering and visualization. A) Parameters used for filtering times series. These parameters produced a list of 624 differentially expressed genes. B) PCA analysis of expression data for 1513 significantly regulated genes. C) Hierarchical clustering of expression data for 624 differentially expressed genes.

and Rousseeuw, 1990) indicated that $k=4$ gave the best grouping of the filtered data. Figure two summarizes the expression patterns associated with each cluster.

Biological significance

The biological process ontologies associated with the genes in each cluster were examined using a z-score report. The z-score report identifies ontologies that are significantly over-represented in a gene list. The z-score report identified distinct biological themes associated with each cluster. Figure 3 lists the predominant ontologies over-represented in each cluster. See supplemental material to view a comprehensive list of ontologies identified for each cluster.

Summary

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

Mean silhouette width: 0.390

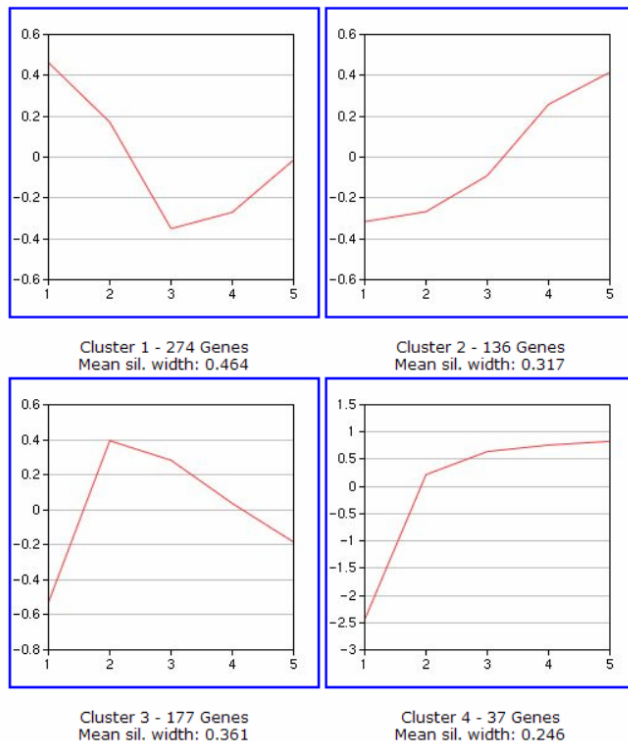


Figure 2: Partition clustering of the filtered gene list. PAM was used to separate 624 differentially expressed genes into 4 groups based on expression pattern. Each line graph summarizes the gene expression pattern for that cluster. The number of genes in each cluster is listed below the graphs. Silhouette widths measure how the genes in each group are clustered and can be used to select the best number of clusters for a set of genes.

Cluster 1

Ontology	List	Array	z-score
carbohydrate metabolism	43	321	11.81
energy pathways	19	88	10.74
lipid metabolism	36	402	7.73
glycolysis	7	25	7.63
proteolysis and peptidolysis	25	479	3.33
protein catabolism	25	483	3.29

Cluster 2

Ontology	List	Array	z-score
SRP-dependent cotranslational protein-membrane targeting, translocation	3	5	11.77
amino acid biosynthesis	4	50	4.29
proteolysis and peptidolysis	13	479	2.98
protein catabolism	13	483	2.94
RNA splicing	4	113	2.19

Cluster 3

Ontology	List	Array	z-score
antifungal humoral response (sensu Protostomia)	5	9	12.67
antifungal polypeptide induction	4	7	11.5
negative regulation of protein-nucleus import	2	2	10.89
Toll signaling pathway	7	24	10.57
defense response	22	407	6.15
amine metabolism	12	205	4.79
dorsal/ventral axis specification	5	48	4.77

Cluster 4

Ontology	List	Array	z-score
antibacterial humoral response (sensu Protostomia)	11	18	39.27
defense response to fungi	5	6	30.94
defense response to Gram-negative bacteria	7	12	30.58
defense response to Gram-positive bacteria	6	10	28.71
defense response to bacteria	11	43	25.21
defense response	16	407	11.17

Figure 3: Distinct biological themes are associated with each cluster. Each table lists highly over-represented biological process ontologies for that cluster.

References

CEL files were obtained from

<http://www.fruitfly.org/expression/immunity/data.shtml>.

Data originally published in -

De Gregorio E, Spellman PT, Rubin GM, Lemaitre B.

Genome-wide analysis of the *Drosophila* immune response by using oligonucleotide microarrays.

Proc Natl Acad Sci U S A. 2001 Oct 23;98(22):12590-5.

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

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

Irizarry, et al. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data.

Biostatistics 4(2): 249-64.