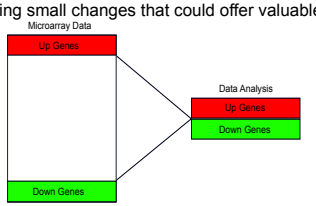
A Gene Group Database

for Systematic Analysis of Microarray Data

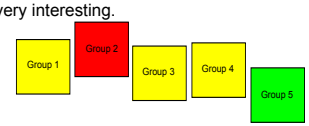
Madelaine Marchin and Chris Seidel
Stowers Institute for Medical Research, Kansas City, Missouri, USA

Introduction

Often when analysing microarray data, it is easy to focus only on the genes that show the most difference between two samples. This artificially crops the center out of the data, in effect ignoring small changes that could offer valuable insight.



Instead, gene group analysis (also known as gene set enrichment or iterative group analysis) aims to take predefined groups of genes and check their overall values in the microarray data. If a group's member genes are showing a change in expression, that group is suddenly very interesting.

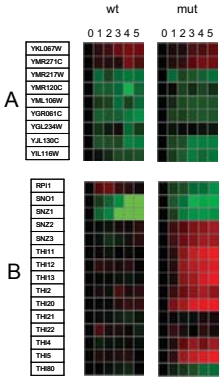


Groups of genes can be created for a number of different reasons. Groups of genes are limited only by the imagination, and could include:

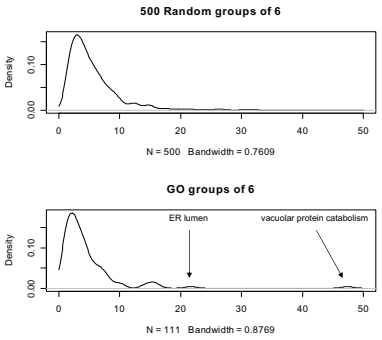
- Gene Ontology groups
- Chromosomal locations
- Pathways
- Groups from experimental results
- genes targeted by a specific microRNA

If you take **every** group of genes that you and other people can think of and examine the behavior of the groups in your data, the results are an easy to analyze list of groups changing in your experiment that can offer surprising insights.

Group Analysis



Two groups of genes are shown above, as measured in a time course of yeast induced for expression of either a wt protein, or a misfolded mutant protein, which causes ER stress. Genes in group A are roughly the same under the two conditions, whereas genes in group B show differential expression in the wt cells relative to the mutant cells.



For any given group of genes, a distance based on expression values can be calculated between two conditions, such as wt and mutant. For the example above, distances between wt ant mutant were calculated for groups containing randomly chosen genes. One can see that groups with random members give a distribution of distances. However when distances are calculated for groups of genes based on Gene Ontology terms, some groups partition themselves far away from the mean. In this case, groups relevant to ER stress distinguish themselves. Alternative methods of group analysis exist but all require a source of groups.

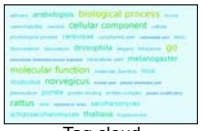
Gene Group Database

Gene group analysis does not work without groups of genes. We built a web database (http://research.stowers-institute.org/microarray/gene_groups) for people to use with group analysis.

We do not limit the type of group users can upload. We encourage users to upload any groups they have developed. Groups might show enrichment in someone else's data set, which could be meaningful to both parties. We originally populated the database with groups from Gene Ontology and L2L (1,2). Here is a screenshot of the download page.

You can search groups to find all groups containing genes of interest. When users upload groups, they are asked to "tag" them with various category terms for easy user-generated categories. On the download page, a tag cloud shows the tags that have the most member groups and allows for easy navigation.

Search options

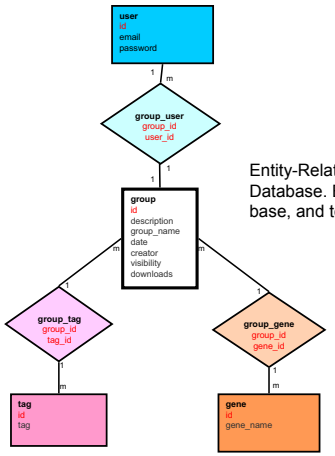


Tag cloud

Users can upload and download groups without registering. Users who would like to keep track of private gene groups may register with a password. Groups can be downloaded in your chosen file format (currently three formats are offered to correspond with iGA, Catmap, and GSEA). A help page explains how to use iGA, Catmap, and Gene Group Database to analyze data.

Behind the scenes

The database back-end is a MySQL database with 7 tables to hold information about groups, users, tags, and genes. The website is written in HTML, PHP and javascript.



Entity-Relationship Diagram for Gene Groups Database. Each shape is a table in the database, and terms in red are keys.

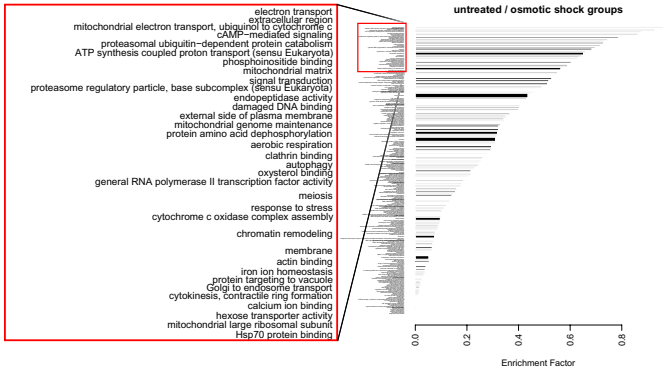
Other databases

Other gene group databases have been created. GSEA has MSigDB (Molecular signatures database), which contains groups of genes, including groups based on chromosomal position, signaling pathways, and experimental gene expression results from literature. MSigDB does not currently allow users to upload new groups or redistribute the database.

L2L is another gene list database. All lists represent experimental gene expression results. Users can contribute their own experimental results. This database has been included in our Gene Group Database.

Our database imposes no limitation on groups, is generally extensible, encourages sharing, and serves multiple group analysis platforms.

Visualization



Shown here are results using groups from our database and iGA on data from Elena Hidalgo, S. pombe untreated vs. Osmotic shock. Shown are the groups that most changed from untreated to osmotic shock.

Summary

-We created a public web database to store gene groups for microarray data analysis.

-We used our gene groups to run group analysis on a data set.

Contact Information
Chris Seidel, Managing Director of Microarray, Stowers Institute
cws@stowers-institute.org
http://research.stowers-institute.org/microarray/gene_groups

Acknowledgements
Thanks to Elena Hidalgo for the S. pombe untreated vs. osmotic shock data and Antony Cooper for the ER stress data.

References

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., & Cherry, J. M. et al. (2000). Nature Genetics, 25(1), 25-29.

Breitling, R., Amtmann, A., & Herzyk, P. (2004). BMC Bioinformatics, 5.

Breslin, T., Ede´n, P., & Krogh, M. (2004). BMC Bioinformatics, 5.

Newman, J. C., & Weiner, A. M. (2005). Genome biology, 6

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., & Gillette, M. A. et al. (2005).