Genatomy Tutorial

What is Genatomy?

Genatomy is a visualization tool for biological data (gene expression, genotypes, growth curves, copy number variation and more), that can be used to analyze the data mathematically and to study the biological aspects of the data and the results.

Genatomy is developed by Bioinformaticians for Bioinformaticians. It "understands" biological data such as gene names, chromosomal location and species. The development team maintains a database for several species that contains full genome information, GO categories, gene sets and more. It can also perform many tasks widely used by Bioinformaticians, such as gene sets enrichments, clustering, GSEA[1] and SAM[2].

This tutorial explains the basic steps of loading, visualizing, analyzing and interpreting microarray data and other types of data. It does not, however, cover all features of Genatomy. We refer you to our user manual for more features and information about file formats. The tutorial is based on *S. cerevisiae* microarray and genotype data[3] which is available for download as a zip file at http://www.c2b2.columbia.edu/danapeerlab/html/Genatomy/example.zip.

We will first show how to load microarray and genome information data. We then explain how to cluster the data using different algorithms, how to load gene sets and run hypergeometric enrichment. We also explain the features that Genatomy includes which allow biological interpretation of these results, and then we show how to perform linkage analysis with Genatomy. To conclude, we explain how to share your results with your collaborators and how to export your results to various formats.

Genatomy was (and still is) developed in Prof. Dana Pe'er's Lab at Columbia University. If you use it for your publication, please cite *Litvin et al. , PNAS 2009*. We appreciate any comment, suggestion and (even) bug reports. Please email us at: genatomy@gmail.com

1. Running Genatomy

Genatomy is a java based application, allowing it to run on Windows, MacOS and Unix. It best runs on Java 1.6, but can also run on Java 1.5. It is advisable that you make sure that you have Java 1.6, and if necessary, install the latest version as explained at http://www.c2b2.columbia.edu/danapeerlab/html/Genatomy/java.pdf.

To run Genatomy, just double click on the icon of *Genatomy.jar*. Genatomy checks for updates with every run, and notifies you when update is available. To update, just ask Genatomy to download itself.

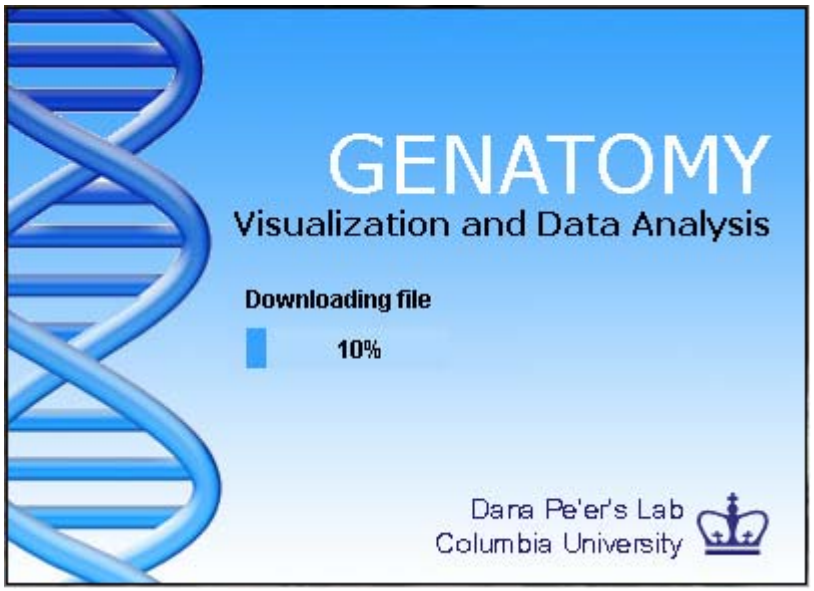

Figure 1 – Genatomy updates itself

2. Creating a project with microarray data

Creating the project

For our first project, we will create a project with S. *cerevisiae* microarray data from "expression.tab" file available in our example zip file at http://www.c2b2.columbia.edu/danapeerlab/html/Genatomy/example.zip. Genatomy does not load raw cell files, and only accepts processed data after conversation of probe reads to gene expression values.

To create a new project, go to the menu File->New. The following wizard will be shown:

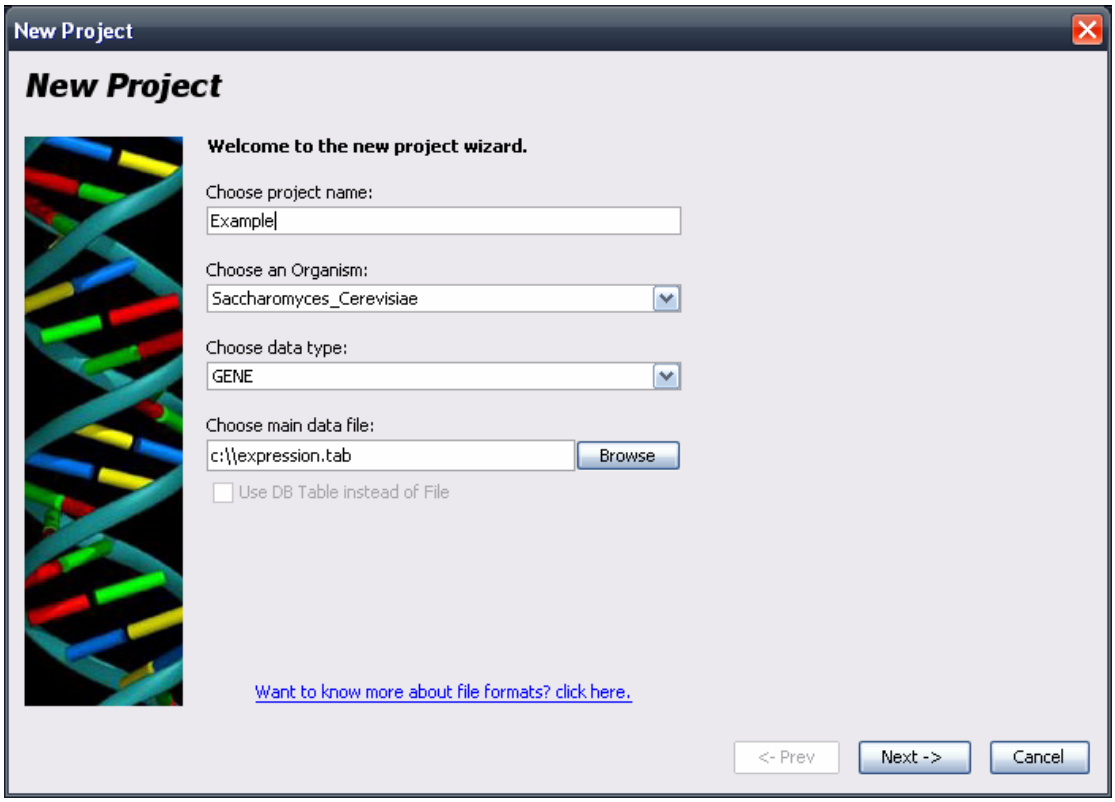

Figure 2 – New project form

Please give the project a name, choose the right organism (*Saccharomyces cerevisiae* in this case) and locate the file "expression.tab" which you extracted from the zip file. Click on "next->" and choose the full genome information file "SGD_features.tab". Click on "Finish". Since you probably do not have the genome information file on your computer just yet, Genatomy will ask your permission to download it from our DB. The file will be downloaded and Genatomy will create the new project.

First look at Genatomy

A new window will open and Genatomy should now look similar to Figure 3. The main window is divided into two main areas:

1. The Data area – capturing most of the window area. This area is also divided into several regions. Currently you can see the gene list (on the right side), the sample names list (at the top), and the expression panel (the rest).

2. The Properties area on the left contains user properties such as colors and size. It displays the information attributed to each region of the data area.

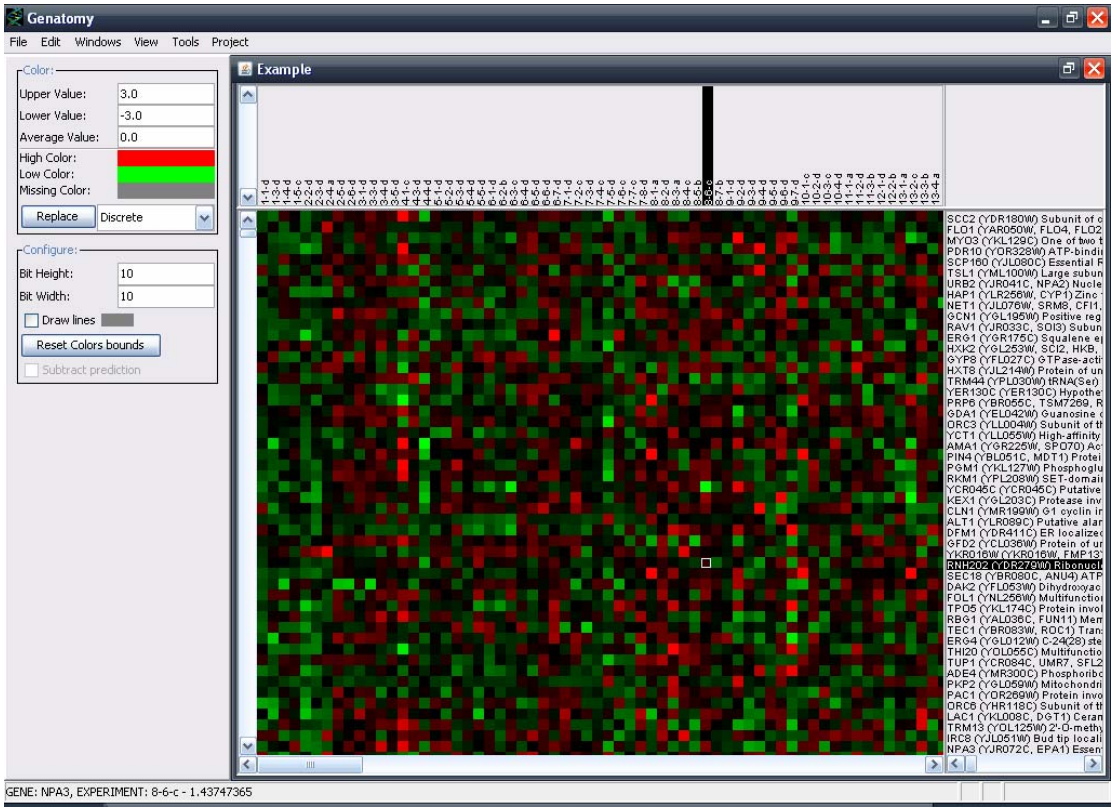

Figure 3 – Expression displayed in Genatomy

Click on the main expression area (the Red-Green area) to display its properties in the properties area on the left. Try to change the colors, size and other visual properties. Notice that you can see and change the properties of the other area (gene and sample name list) by clicking on them.

Saving the project

To save the project go to the menu File->Save as and choose the location and file name. After saving it, the project will appear at the "Recent Projects" list on the menu.

Please Note: the project file DOES NOT contain the data itself, and only saves a reference to the data files.

Tips and Tricks

1. To find a gene in the list press on Control+F (or go to the "Find" menu) and type your searching criteria. The matched string will be highlighted in Red.

2. Right click on a gene name will send you to the official website for that gene.
3. Changing related properties together, such as width and height, is possible by changing one of them and pressing on CTRL+Enter.
4. The gene, sample and expression value that the mouse points at are displayed in the message bar at the bottom of the window.

3. Clustering the data

As a first approximation for the underlining network that created the data we see, we can use clustering. We will first divide the data into modules using k-means clustering, and then use hierarchical clustering to sort the modules.

k-means

To cluster the data, go to Project->Cluster->K-Means and choose 20 as the number of initial clusters (first row, see figure 4).

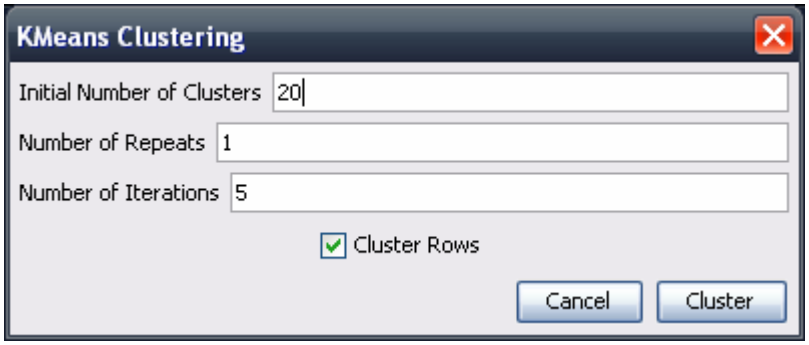

Figure 4 – k-means clustering configuration

After a few seconds the run will be completed and you will be asked to save the results into a file. Once the results are saved, the filter panel (see figure 5) will appear inside the properties area at the left side of the window.

A word on conventions – A **module** defines a set of genes and samples with or without a regulatory program. A **filter** is a group of modules, usually defined by the file from which the modules were loaded.

The panel is divided into 3 part – at the top located a box with all loaded filters; the area at the middle which occupies most of the panel contains a list of modules inside

the selected filter; and the bottom area contains navigation and other configuration button.

By choosing one of the modules, the main display will change and now contains only the genes of the selected module.

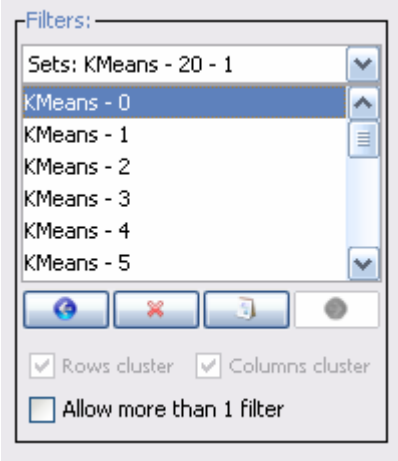

Figure 5 – Filters panel

Hierarchical Clustering

As you can probably notice, the modules are more or less coherent, but the columns (samples) are not sorted in any rational order. To fix that, we will cluster the columns of each of the modules independently.

Go to "Project->Cluster->Hierarchical". The hierarchical clustering configuration form will appear (figure 6). Select to cluster all modules of the k-means filter, select to cluster using the Euclidean metric, and unselect the "cluster rows" checkbox in order to cluster the columns. Run the algorithm.

Now the modules are sorted and the signals become clearer.

To see the dendrogram, select "View->Horizontal Clustering View" from the menu. You can use the dendrogram to zoom-in and focus only at a subset of the columns. The zoom-out and other navigation options will appear, as usual, at the properties area on the leaf side of the window.

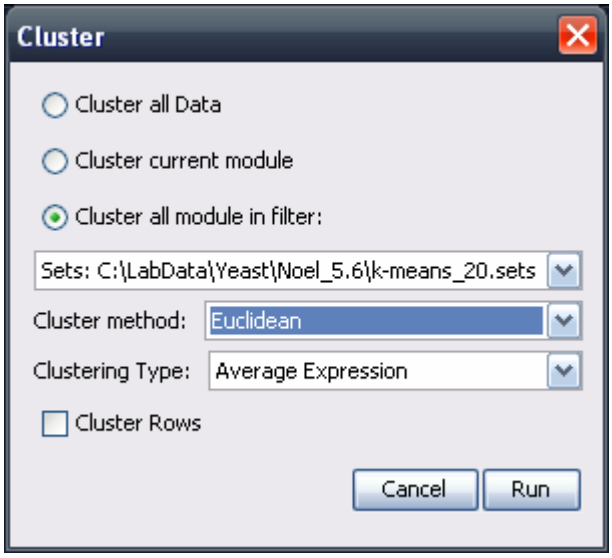

Figure 6 – Hierarchical clustering

Tips and Tricks

1. Use the navigation buttons – the buttons with the arrow icons – in the filter panel (figure 5) to navigate between recently visited modules.
2. To write remarks or to change the name of a module, use the "Name and Notes" button in the filter panel (second button from the right).
3. Sort the modules by name or size by right-clicking on the module list in the filter panel (figure 5).

4. Loading genesets and running enrichment

Genesets

One of the most important and helpful tools of a bioinformatician is comparison to other published genesets. Using Genatomy, you can load, visualize, analyze and compare genesets. We maintain a database of such genesets for several organisms, including the widely used Gene Ontology (GO) database.

To manage the project's genesets go to "Project->Attribute Manager". The form (figure 7) is divided into 3 parts: The upper left is a list of available and loaded genesets (or attributes tables); the upper right is a list of loaded sample attributes (we will use these in section 6); and the bottom is the list of loaded filter files. Notice that the filter that we created earlier using k-means is listed there.

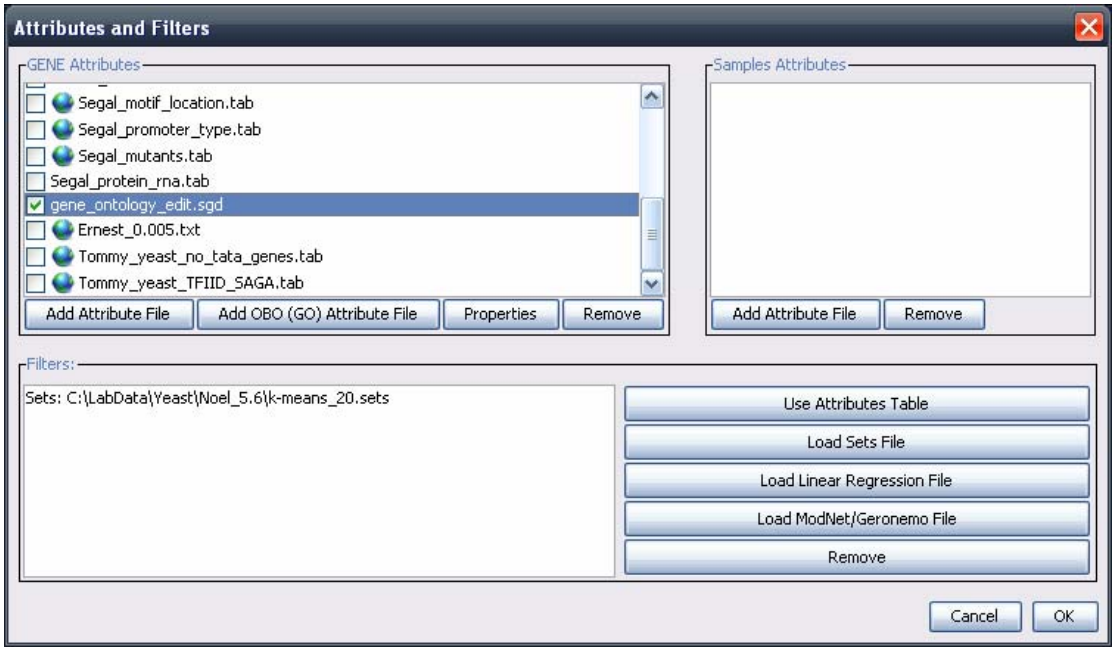

Figure 7 – Attributes and Filter management form

Download and load the S. cerevisiae GO file by selecting "gene_ontology.sgd" from the list. If the file is not currently on your computer, Genatomy will automatically download it.

To view the gene attributes go to "View->Gene Attributes" menu. The attributes, seen as blue and white and another random color squares, will appear to the right of the expression panel.

A word on GO structure – GO contains not only annotations for genes, but also an inheritance structure for the terms (or attributes). Genatomy automatically annotated every gene with all terms upstream to its direct annotations. The colors in the attributes panel indicate the direct and in-direct annotations. You can cancel the automatic aggregation in the GO properties form (see Tips and Tricks #1).

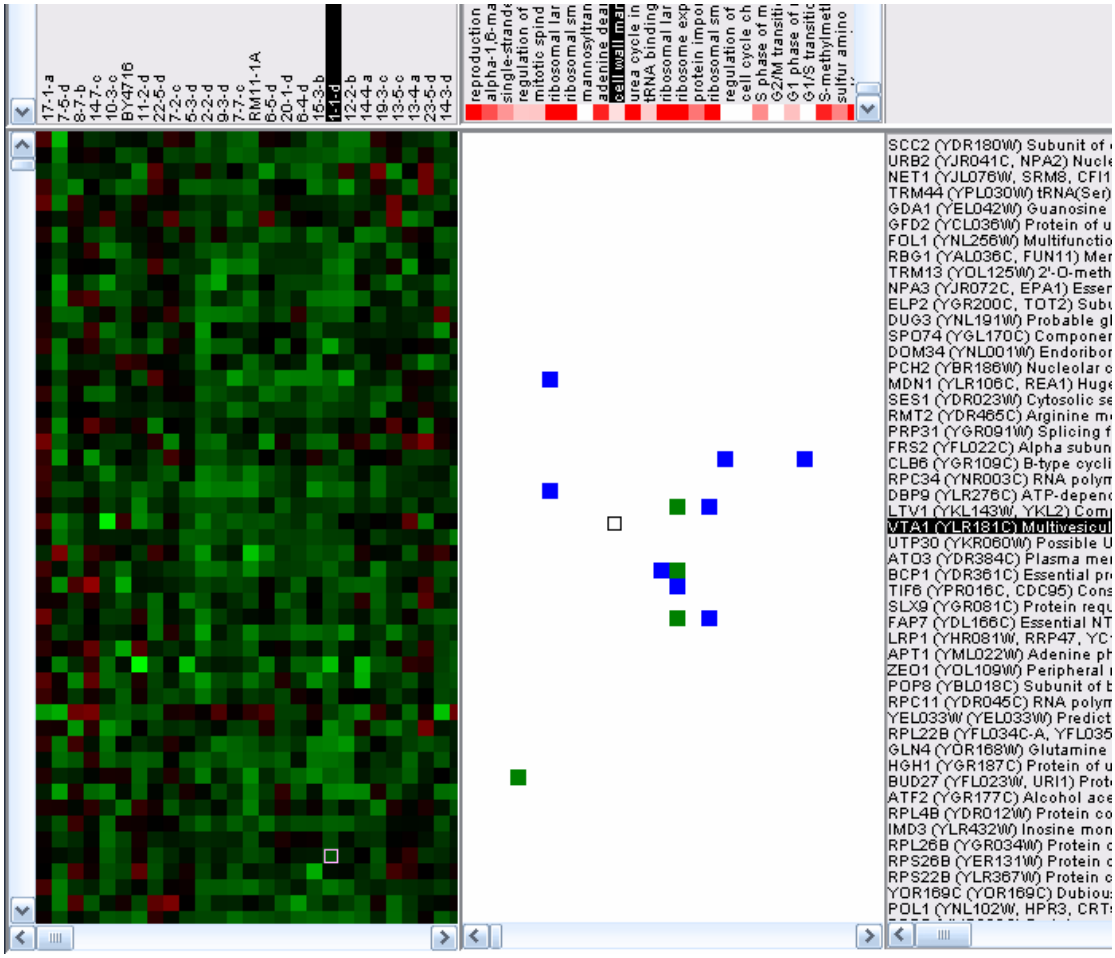

Figure 8 – The gene attribute panel with white, blue and green squares.

Hypergeometric Enrichment

Now that we create modules using k-means, and loaded genesets (GO annotations), we can calculate hypergeometric p-values. Generally, a hypergeometric p-value is the probability that the overlap of a module and a geneset (or attribute) can occur by random selection of the groups.

To run hypergeometric calculation, go to "Project->Enrichment->Gene Attributes". A configuration window will be shown (figure 9). Choose the k-means filter from the right list, and the GO attributes from the left list and click on "Calculate".

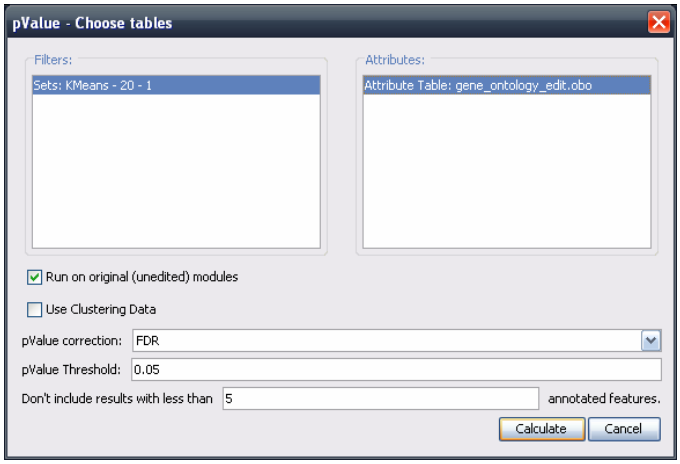

Figure 9 – Hypergeometric configuration window

The run takes a few seconds, and the results are displayed as a table in a new window (figure 10).

PValue - gene_ontology_edit.obo - Sets: KMeans - 20 - 1 - Example

Options

Attribute	DataSet	pValue	Module Ann...	Total Annot...	Module Size	Total Number	qValue
ribonucleoprotein complex biogenesis ...	KMeans - 2	3.725949e-162	219	240	638	3991	6.799856e-159
ribosome biogenesis and assembly	KMeans - 2	1.268209e-159	214	233	638	3991	1.157241e-156
mitochondrion	KMeans - 4	1.988781e-123	191	793	205	3991	1.209842e-120
mitochondrial part	KMeans - 4	1.793820e-118	155	420	205	3991	8.184303e-116
non-membrane-bounded organelle	KMeans - 2	1.280595e-115	311	602	638	3991	3.895142e-113
intracellular non-membrane-bounded ...	KMeans - 2	1.280595e-115	311	602	638	3991	3.895142e-113
rRNA metabolic process	KMeans - 2	2.075281e-99	151	176	638	3991	5.410555e-97
ribonucleoprotein complex	KMeans - 2	7.675861e-99	247	437	638	3991	1.751056e-96
rRNA processing	KMeans - 2	1.534115e-98	149	173	638	3991	3.110844e-96
cytosolic part	KMeans - 2	4.019122e-98	154	184	638	3991	7.334897e-96
nucleolus	KMeans - 2	8.157648e-93	134	150	638	3991	1.353428e-90
RNA metabolic process	KMeans - 2	4.622323e-81	227	431	638	3991	7.029783e-79
RNA processing	KMeans - 2	3.601125e-75	180	299	638	3991	5.055426e-73
macromolecule metabolic process	KMeans - 2	2.571677e-73	462	1598	638	3991	3.352364e-71
translation	KMeans - 2	5.959393e-72	189	337	638	3991	7.250595e-70
ribosome	KMeans - 2	3.543686e-68	174	302	638	3991	4.042017e-66
cytosolic large ribosomal subunit	KMeans - 2	2.779574e-66	84	86	638	3991	2.983954e-64
ribosomal subunit	KMeans - 2	5.671073e-65	143	220	638	3991	5.749838e-63
structural constituent of ribosome	KMeans - 2	9.245188e-61	138	217	638	3991	8.880247e-59
cytoplasmic part	KMeans - 4	1.564613e-58	196	1792	205	3991	1.427709e-56
macromolecule biosynthetic process	KMeans - 2	3.506689e-57	191	401	638	3991	3.047479e-55
translation	KMeans - 4	4.299140e-57	100	337	205	3991	3.566332e-55

Figure 10 – Hypergeometric results form

The results are ordered by their significance. The first column is the annotation (or attribute), and the second is the module name. The rest of the columns enclose p-value, group sizes and percentage of overlap.

Double click on a result to jump back to the module display, with all genes annotated for the selected attribute highlighted in Yellow.

Tips and Tricks

1. You can choose to load only part of the GO annotations by clicking on "Properties" in the attributes window. You can filter annotations by evidence codes, or load only terms in certain domains.
2. You can create a filter and modules using an attribute table. In the attributes window choose the desired attribute table and click on "Use Attribute Table" in the filter area.
3. In many cases, when looking at a subset of genes, many of the attributes do not contain even one annotated gene. You can choose to see only attributes with at least one annotated gene by clicking on the "gene attribute" panel (figure 8), and selecting "Hide empty lines" in the properties area.
4. To add and remove columns in the enrichment results form (figure 10), click on "Options->Column Chooser" menu item.
5. Notice the red-white squares in the panel displaying the attribute names (figure 8). These specify hypergeometric p-values calculated automatically for every displayed module.

5. Interpreting the results

Now that you have some results, you are probably wondering what exactly lead to these outcomes. For example, you might want to check what the differences between two modules are; how similar a gene expression is to another module; and why only a subset of an annotation belongs to a module while the other annotated gene do not.

The "Module Editor" tool in Genatomy can answer all these question and more.

Open it by clicking on "Tools->Module Editor", and the panel will appear under the gene expression area (figure 11).

The "Module Editor" reflects the current status of a module, and allows you to add or remove genes, based on gene names, attributes or modules.

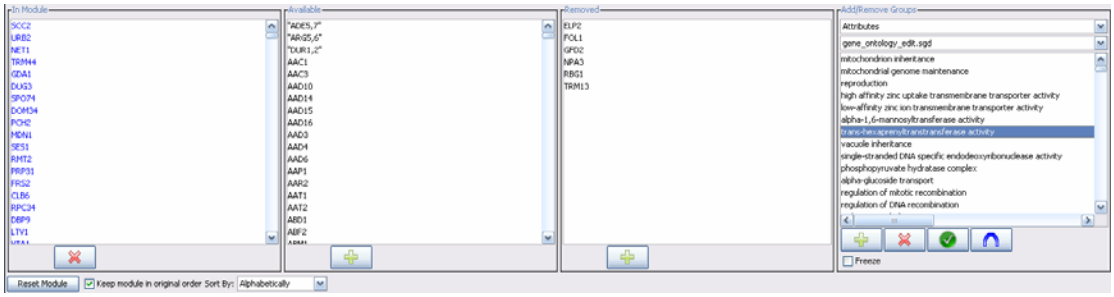

Figure 11 – Module Editor

It is divided into 4 parts. The left area – "In Module" - contains the list of gene currently in the module, the area right to it – "Available" – contains genes that you can add to module, the next area – "Removed" – lists all genes that were originally in the module but removed from it. The right area – "Groups" – allows you to add/remove groups of genes, either attributes or modules.

You can always choose to reset the module and reverse any changes you have made by clicking on "Reset Module".

While the first 3 areas are pretty straightforward, the right most area, the "Groups" area, is a bit more complex and demands an explanation. First, you can choose any geneset loaded in Genatomy by choosing a category – Attributes or Filters, then a File and then the geneset itself.

To see how it works, open the "Module Editor" and go to the enrichment results (figure 10). Double click on any result. As you can see, you are back in the main view, the annotated genes are highlighted, and for the new feature - the annotation is now selected in the "Module Editor".

You can now choose, for example, to check what the other genes annotated for the selected attribute are doing. Just click on the plus (+) button under the group list. You can also choose to remove all annotated genes (X), replace the module with the gene in the selected group (V), or get only the genes that are both in the module and in the selected group (intersection).

Tips and Tricks

1. The genes that were added to the module are listed in black, while the genes from the original module appear in blue.

2. Right click on a gene will reveal to which modules the gene belong.
3. A summary form on a module is available throw "Tools->Module Overview".
4. Once changes were made to the modules, you can run hypergeometric enrichment on the edited modules. Just unmark the relevant checkbox in the hypergeometric configuration form (figure 9).

6. Loading sample attributes and performing association analysis

Sample Attributes

After loading attributes for the genes, we will now load attributes for the samples. Sample attributes can describe several types of information, such as genotypes, experimental conditions and more, and can be binary, discrete or continuous.

Go to the "Attribute Manager" via "Project->Attributes Manager", and under the sample attributes click on "Add Attribute Table". Locate the file from our example zip file named "genotypes.tab" and load it.

To see the sample attributes go to "View->Sample Attributes". You should see the attributes above the expression area and the sample names. Notice that Genatomy identified the attributes as binary and assigned two colors for the two values. You can change these colors by clicking on the sample attributes panel and configuring it in the properties panel.

Association Analysis

Genatomy allows you to perform simple and quick association analysis, finding genes correlated with sample attributes.

Since our sample attributes are binary, we can use t-test together with permutation testing to select significant associations. Genatomy also support other scoring metrics such as Welch's t-test, Anova and ranked Anova.

Go to "Project->Differentially Expressed" and open the association running form (figure 12). Leave the option as is, and click on Run. The run should take about a minute to complete.

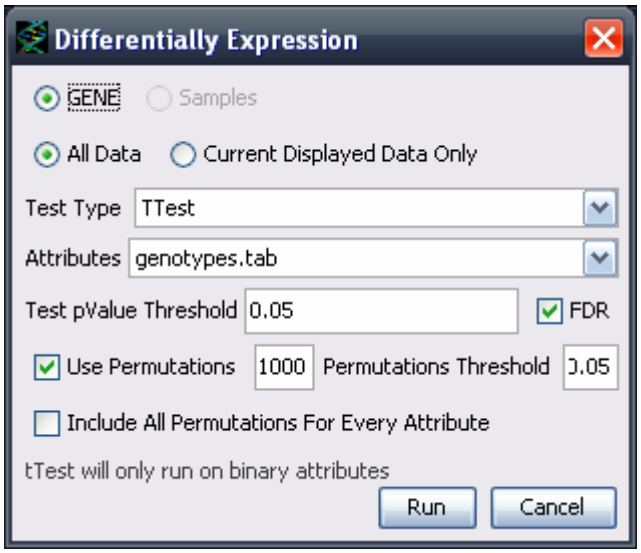

Figure 12 – association running form

The results were added as a new filter, with each sample attribute and its correlated genes creating a module. Choose the filter from the "Filters panel" (figure 5) and look at the resulted modules. To see the sample attribute which created the module, open the "Module Network View" via the menu "View->Module Network View" (figure 13).

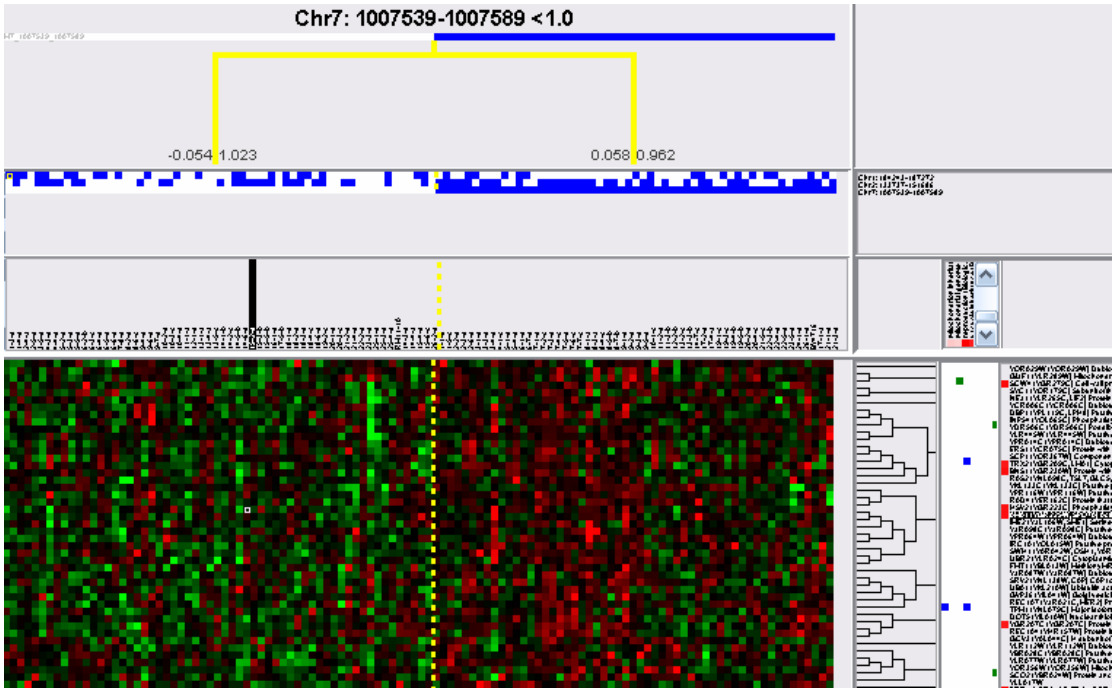

Figure 13 – Association results

As you have probably noticed, the modules contain both up- and down-regulated genes. You can sort the modules using Hierarchical clustering (figure 6).

Genotypes and Genatomy

In this section we loaded genotypes as sample attributes. Genotypes can be anything from a point mutation, via allele types till large genomic regions.

Genatomy can interpret a genotype name and display information on it. For example, right click on a genotype name will open a window listing all genes around its area.

Genatomy also allows you to easily identify CIS regulation. Notice the red square next to the gene names in figure 13. The figure presents a module being regulated by a genomic region in chromosome 7. The red squares represent the distance of these genes from the genomic region.

7. Exporting the results

Sharing a Project

As mentioned above, the project file does not contain the data itself but rather links to the files added to the project. Sharing a project by collecting all the related files can be very annoying.

Genatomy can do it for you. Go to "File->Export->Zip Project" to collect all project files into one zip file. To open the project on another computer, open the zip file first and load the project into Genatomy.

Exporting Figures, Tables and Filters

You can export virtually anything from Genatomy. Among other options, you can export enrichment results to a table, and export the current view to a picture file. Your exporting options are listed for every window under "File->Export".

8. In Brief: other important features

There are many more features not described in this tutorial. We have selected several other cool/important features to briefly describe here. For more information on them and other feature, please read our extended user manual.

Module Networks View

A module network is a module with a tree-based regulatory network. We have already encountered such a module when we created a module based on association analysis (figure 13). Another example of a module network view is shown in figure 14, and here is a short list of features for module networks:

1. Regulatory trees can be constructed from any kind of data type (expression, CNV, genotype) or a mixture of types.
2. You can edit the regulatory trees inside Genatomy by right-clicking on a branch in the tree in order to add, remove and replace branches.
3. You can sort the genes by their matching to the regulatory program and the current module in the "Module Editor".

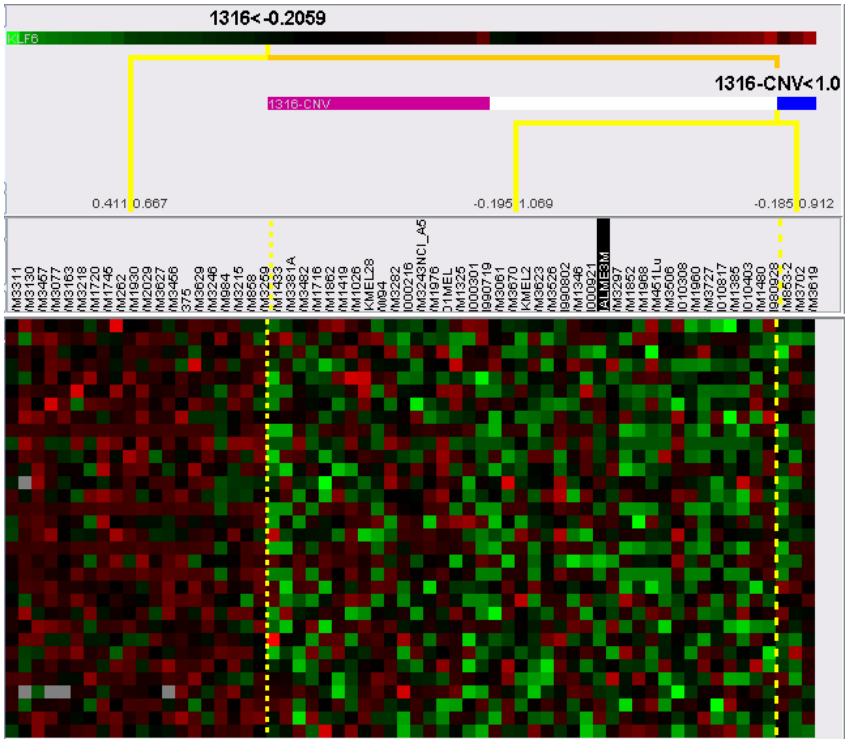

Figure 14 – Regulation tree with mixed data types

Linear Regression

Genatomy can also load, display and change module with regulatory program based on linear regression (see figures 15-16). Genatomy allows you to change the regression coefficients, add/remove features and refit the model, and even add interaction terms between features.

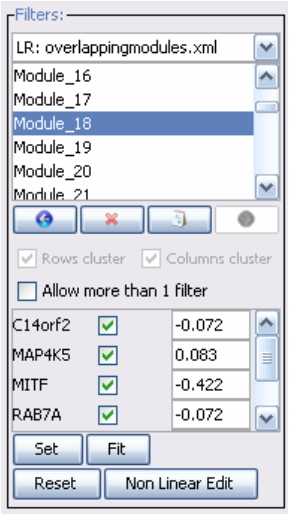

Figure 15 – Control linear regression

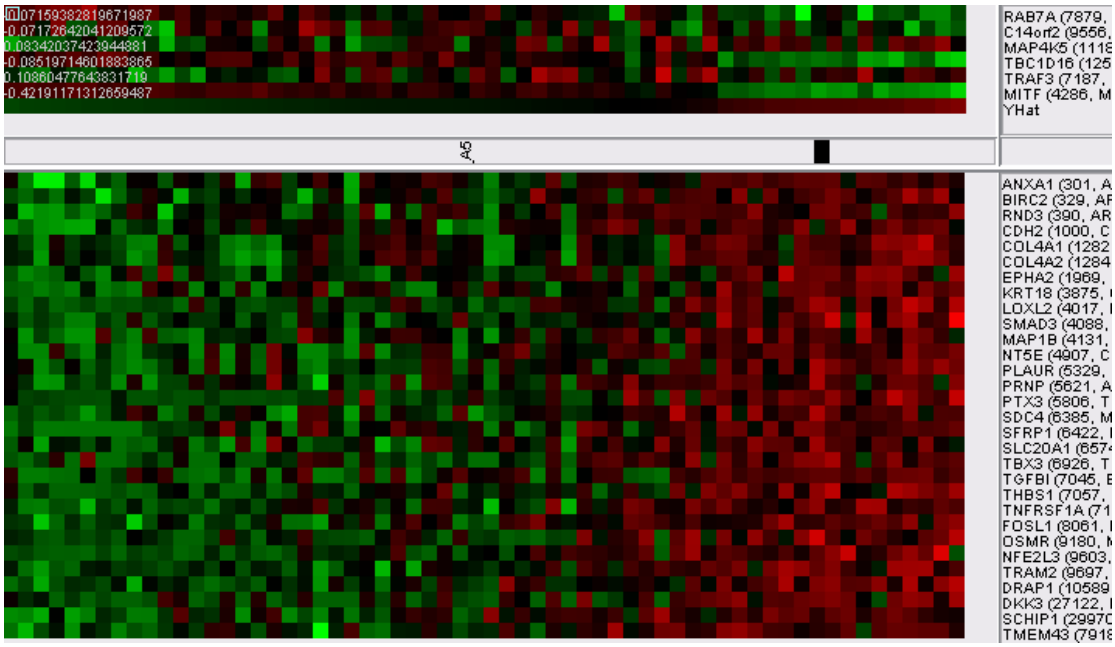

Figure 16 – Linear Regression module view

Coloring Schemes

Genatomy automatically recognizes the data of data loaded and match it a coloring scheme, so continuous data will be colored with gradient, and discrete data will receive a color for each of the values.

Another coloring scheme is "Ranges", which is based on ranges of values that get the same color. For example, values greater than 0 and smaller than 1 will all be painted in Red.

If Genatomy made a mistake and assigned the wrong coloring scheme, or if you would like to change the coloring scheme, click on the panel you want to change the color to, and use the list and the "replace" button in the colors panel in the properties area (figure 17).

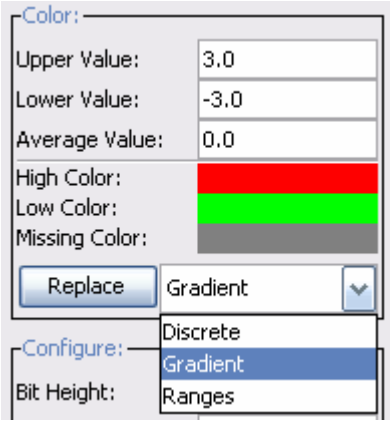

Figure 17 – Changing coloring scheme

References

1. **Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A.* Oct 25 2005;102(43):15545-15550.**
2. **Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A.* Apr 24 2001;98(9):5116-5121.**
3. **Brem RB, Storey JD, Whittle J, Kruglyak L. Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature.* Aug 4 2005;436(7051):701-703.**