User Manual

GIPro v1.0

Wodak Lab
Research Institute

SickKids
THE HOSPITAL FOR SICK CHILDREN

http://wodaklab.org | support@wodaklab.org

Table of contents

Introduction **3**
1. **Getting started** **3**
   1.1. Installation 3
   1.2. Initiation 3
   1.3. File input 3
   1.4. Parameters 6
2. Displaying and visualizing data **8**
   2.1. The protein complex network 8
   2.2. Expanding protein complex views 8
   2.3. Search and display 10
   2.4. Heat-maps 11
   2.5. Pearson correlation coefficient 12
   2.6. Histograms 13
3. Export information **14**
   3.1. Complex enrichment information 15
   3.2. Pearson correlation matrices/tables 15
   3.3. Heat-maps 15
4. Technical details **16**
   4.1. Cutoff analysis 16
   4.2. Enrichment analysis 17
   4.3. Hierarchical clustering 18
   4.4. Pearson correlation coefficient 18
5. Shortcuts and tips **19**

Introduction

The advent of high-throughput technologies in proteomics and genetic screening has yielded hundreds of protein complexes and millions of genetic interactions, respectively, in the budding yeast. However, analysis of large-scale datasets continues to be overwhelmingly challenging for biologists despite recent progresses in systems biology. For example, how can one find out which complexes are enriched with aggravating/alleviating genetic interactions? How can one use genetic interactions to determine the functional relationships between two complexes? How can one visually inspect the distribution of genetic interactions and physical interactions among proteins within one or more complexes simultaneously and make sense out of it?

The GIPro Plugin for Cytoscape is developed to analyze large-scale quantitative genetic interaction data statistically to identify functional relationships between genes and between protein complexes, and displays results in Cytoscape. Data tables, heat-maps and histograms are optionally generated for further analysis.

As an example of application of the GIPro plugin, yeast COG complex (Conserved Oligomeric Golgi complex) is found to be enriched with both aggravating and alleviating interactions in the enrichment analysis by integrating genome-wide genetic interaction data with protein complexes. A detailed inspection of the genetic interactions within this complex indicate that it may consist two modules (consisting of COG1 to COG4 and COG5 to COG8, respectively), as aggravating interactions exist between these two modules while alleviating interactions occur within modules. This modular decomposition of COG is in good agreement with existing morphological and biochemical evidence. This example demonstrates that a detailed analysis of GIs can reveal fine functional differences between modules of the same complex.

Getting started

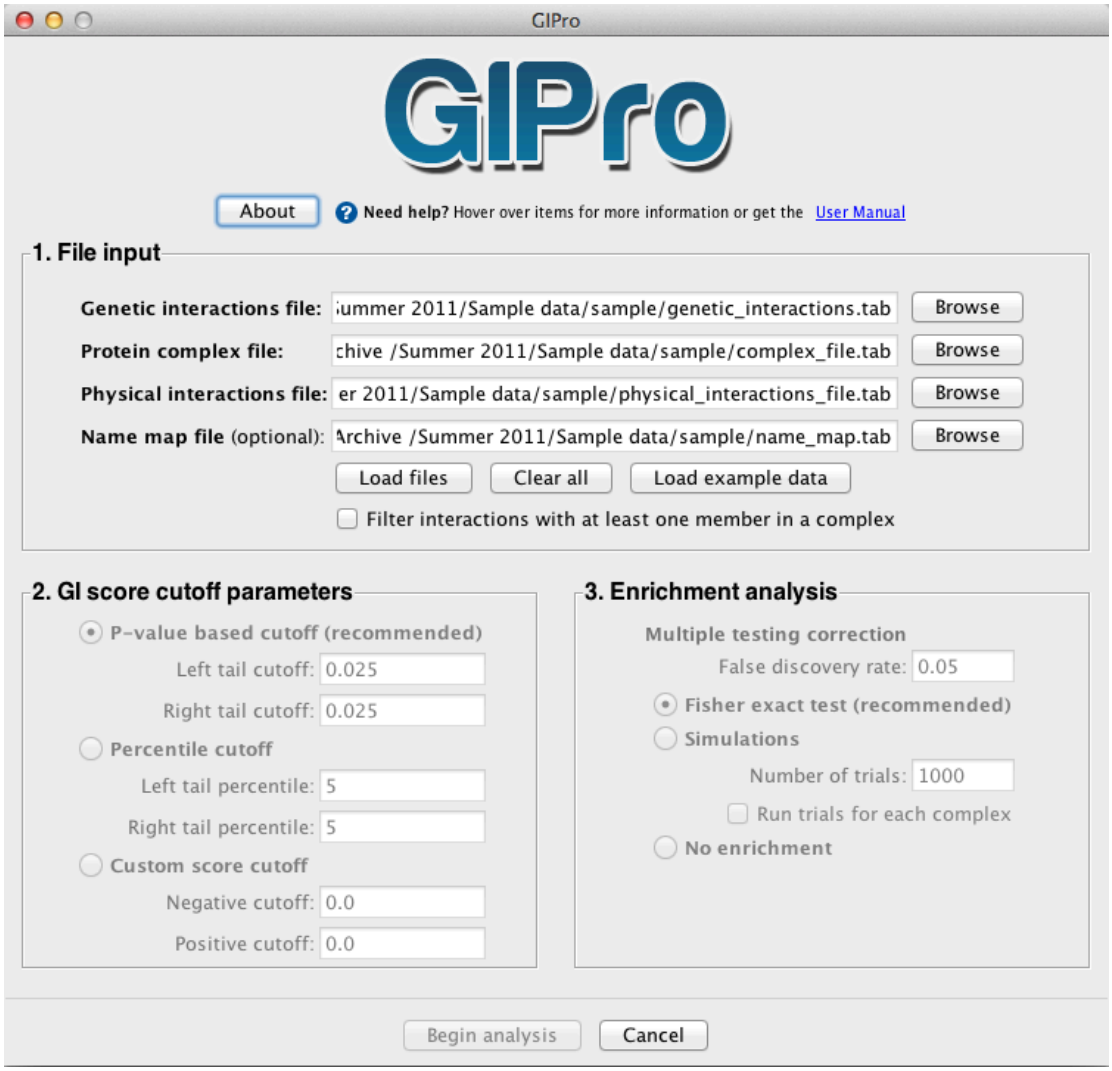
1.1 Installation

Place the **GIProPlugin.jar** into the Cytoscape plugins folder to install. The plug-in is automatically initialized when Cytoscape starts. Java's default memory should be increased with vm arguments when using this plug-in, unless small datasets are being used. See http://cytoscape.wodaklab.org/wiki/How_to_increase_memory_for_Cytoscape for more information on how to increase Cytoscape memory

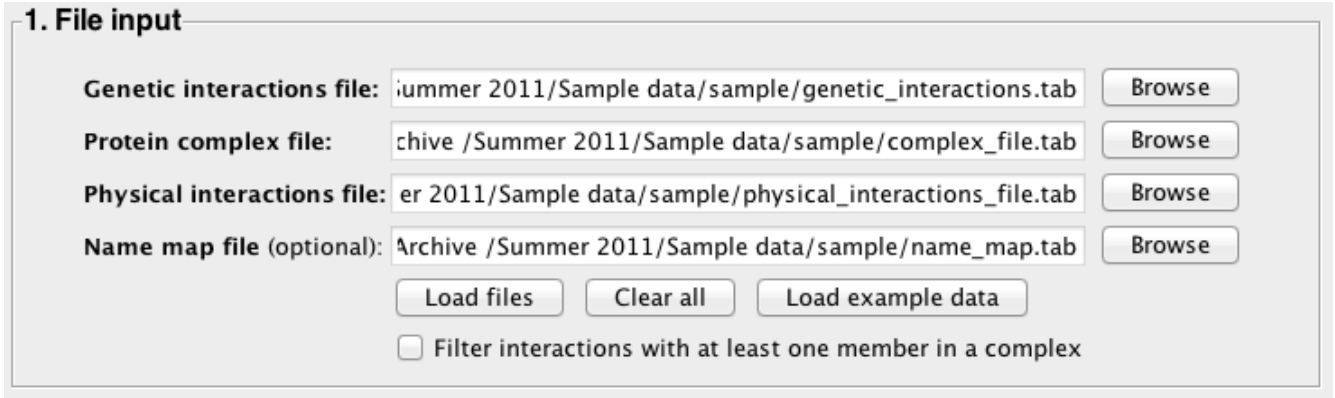
1.2 Initiation

After Cytoscape starts, click **Plugins -> GIPro** and a Wizard will prompt the user for input files and other parameters.

Fig 1. Input screen used for specifying data files and parameters used by the plugin



1.3 File input



There are three required, and one optional files used for this plugin. The purpose of each file is described below:

1) **Functional relations file:** This file is used to specify genetic interaction scores between pairs of genes. Optionally, a p-value can be included for each relation. In this case, relationships that do not meet the p-value cutoff score of 0.05 are ignored.

Delimited by: tab
Header: none
Columns:

1. Gene A
2. Gene B
3. Score
4. P-value (optional)

Sample file (without p-value)**:**

YBL075C	YDL133W	0.00087	
YDL032W	YBL022C	0.01625	
YER054C	YBL012C	0.02615	

Sample file (with p-values)**:**

YBL075C	YDL133W	0.00087	0.012
YDL032W	YBL022C	0.01625	0.005
YER054C	YBL012C	0.02615	0.072

2) **Protein complex file:** This file defines protein compositions for complexes, by listing complex IDs and protein ORF names that belong together.

Delimited by: tab
Header: yes – 2 column header
Columns:

1. Complex ID/name
2. ORF in complex

Sample file:

Complex-name	Gene-name
TRAPP complex	YER054C
TRAPP complex	YDL033C
Rpd3l complex	YBL022C

3) **Physical interactions file:** This file contains physical interaction scores for pairs of proteins.

Delimited by: tab
Header: none
Columns:

1. ORF A
2. ORF B
3. Score

Sample file:

YBL075C	YDL133W	0.00087
YDL032W	YBL022C	0.01625
YER054C	YBL012C	0.02615

4) **Name map file** (optional): This file contains a mapping between ORF names of proteins to gene names.

   Delimited by: tab
   Header: yes – 2 column header
   Columns:
   1. ORF name
   2. Gene name

   Sample file:

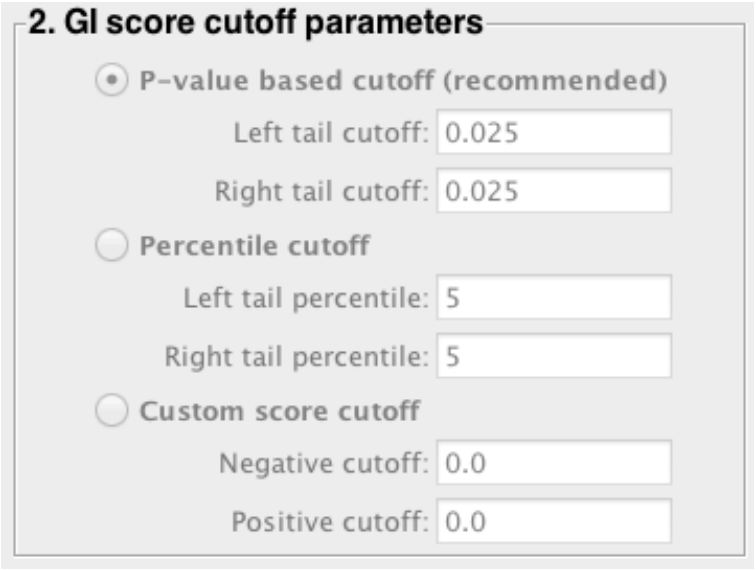
ORF-name	Gene-name
YDL033C	SLM3
YER054C	GIP2

This information can also be found via the plugin. Roll over the file text-fields or **"Browse"** button for more information.

You may also check the **"Filter interactions with at least on member in a complex"** check box. This is useful if a large relations file is being loaded. If this box is checked, only those relations where at least one of the genes belongs to a complex will be loaded

1.4 Parameters

Cutoff calculation parameters: The solid positive and negative cutoffs are used to determine if an interaction is positive or negative (see technical details section for more information). There are three ways to specify them:



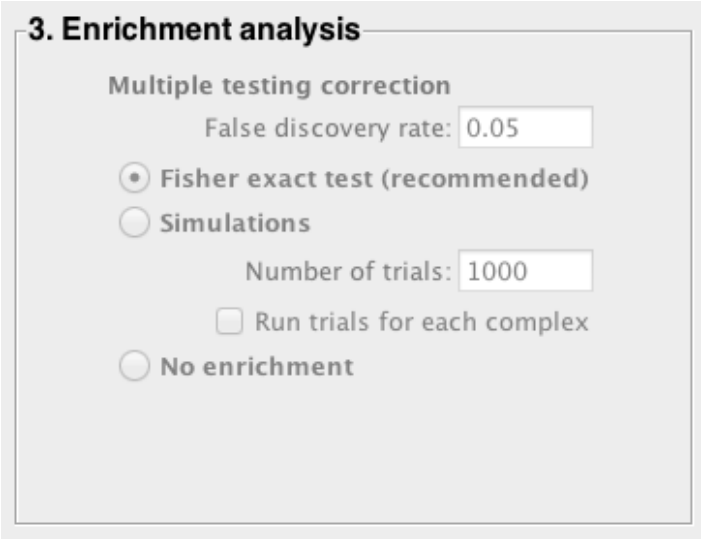
1) **P-value based cutoffs:** by specifying a p-value, an algorithm is run to determine the Gaussian positive and negative solid cutoff values.
2) **Percentile cutoffs:** specifying a value such as 10 percent gives a positive cutoff value at the 90th percentile and a negative cutoff value at the 10th percentile of the relation scores. Note that the

value must be non-negative.

3) **Custom score cutoffs:** a user can enter their own positive and negative cutoff values, which can either, be positive or negative values themselves (as long as the positive cutoff is larger than the negative).

As different values are entered, the custom positive and negative score cutoffs are updated accordingly. These are the actual cutoffs that will be used, and have a direct impact on how the program is run. Refer to the technical details section for details.
Note: with the use of large data sets, the positive and negative cutoffs may take a few seconds to update.

Enrichment analysis parameters:



1) **Multiple testing corrections:** A false discovery rate is entered. This value is used to compute the p-value cutoff used to filter the outputted data. A smaller FDR value will generally result in a smaller p-value. (See technical details for more information)

Simulation Parameters:

2) **Number of Trials:** the number of simulation trials performed for each complex. See the technical details section for more information. The recommended default is 1000. Note: This value must be a positive integer.
3) **Trial for Each Complex:** This check box should be checked if the user would like simulations to be re-run for complexes with the same number of interactions, otherwise they share a distribution. Usually this box should remain unchecked, unless the “number of trials” parameter is small.

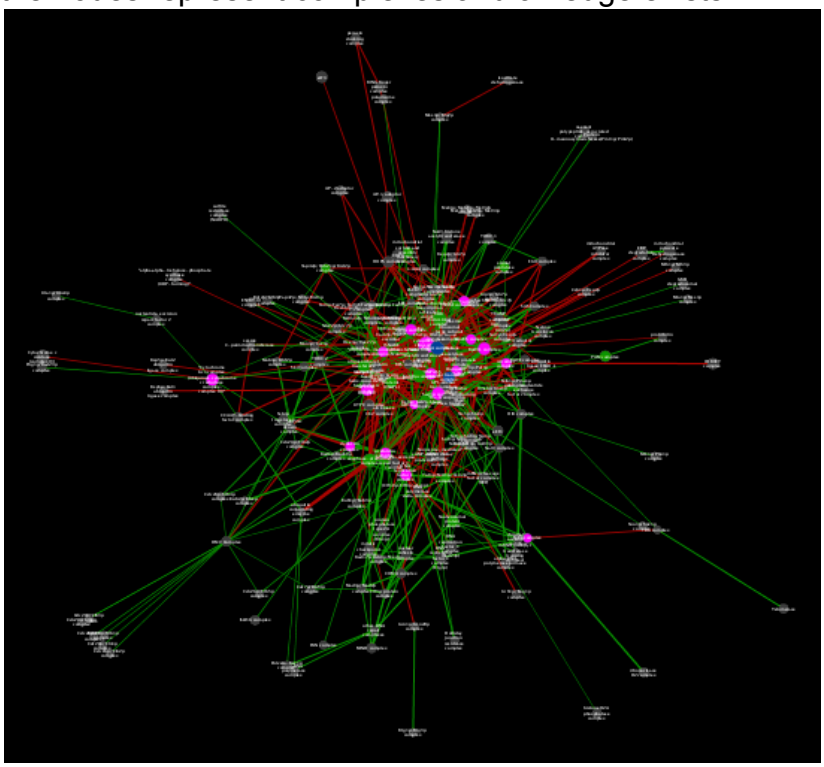
To begin computations, click **“Begin analysis”**. Computation progress and details will be displayed in progress dialog and when finished a protein-complex network will be created in Cytoscape.

After the complex network is displayed in Cytoscape, if the user decides to change the parameters, click **“Adjust parameters”** on the bottom left of the Cytoscape panel to bring up the “update parameters” panel to enter new parameters; then click **“Apply changes”** to restart the analysis. The current network will be destroyed and a new complex network will appear.

Displaying and visualizing data

2.1 The protein complex network

In the generated complex network, the nodes represent complexes and an edge exists between two complexes if and only if there exists a significant enrichment of positive and/or negative interactions.



Nodes:

- **Size:** scaled according to the number of genes within the complex, using a logarithmic scale.
- **Color:** represents whether the complex is enriched with genetic interactions. A **magenta** complex means the complex is enriched with positive interactions, **green** represents enrichment with negative interactions and **blue** represents an enrichment of both positive and negative genetic interactions. Complexes have a default **dark-grey** color if they are not enriched with either type of interaction.

Edges: edges exist between a complex pair if their between complex interactions are positively or negatively enriched

- **Color:** a **red** edge represents enrichment with positive interactions, while **green** represents enrichment with negative interactions. Double edges are created between complex pairs who's between complex interactions are enriched with both positive and negative interactions.
- **Thickness:** proportional to the significance of the enrichment p-value. The more significant, the thicker are the edges.

2.2 Expanding protein complex views

By clicking the 

button in the left panel, a network of interacting genes contained in the selected complexes is created. Genes belong to a complex are laid out in a circle, and the relative position of the circle is determined by its location in the original complex network. Some of the network properties are listed below:

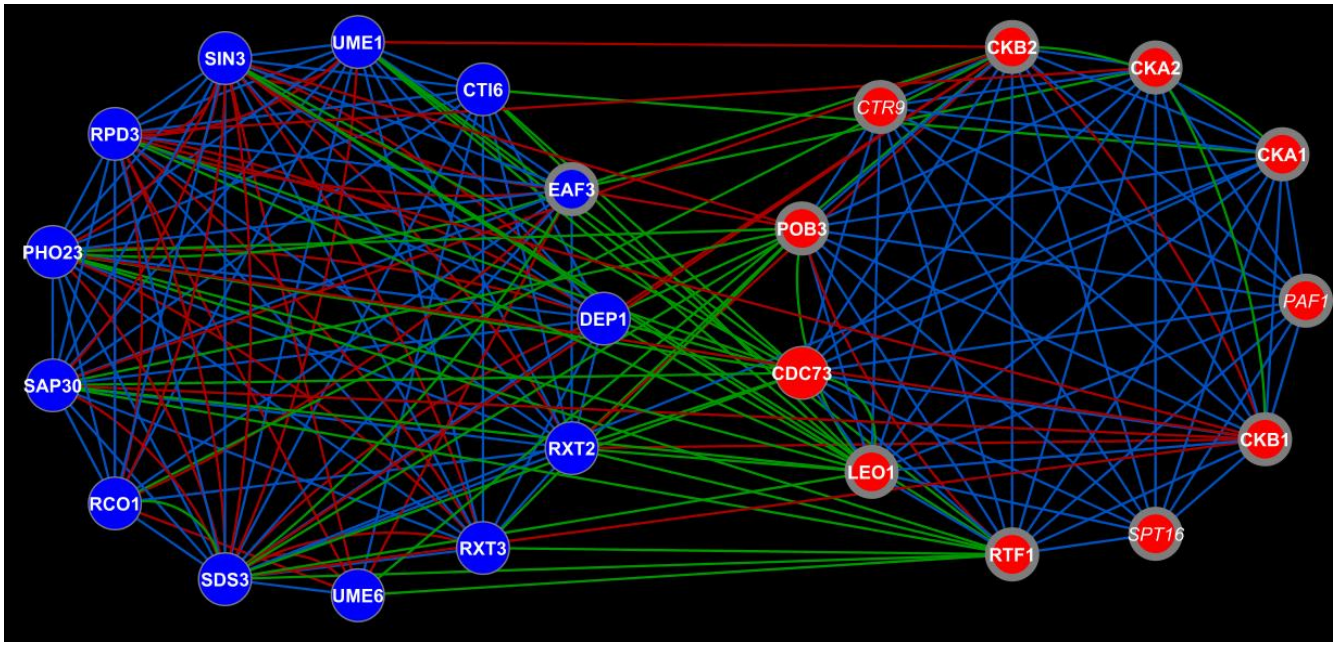
Nodes:

- **Node color:** represents the complex that the gene belongs to. Nodes with thick border indicate that the proteins are shared by multiple complexes.
- **Node label:** is in ***italic*** font if no genetic interactions exist in the provided data for that gene or in **normal** font otherwise.

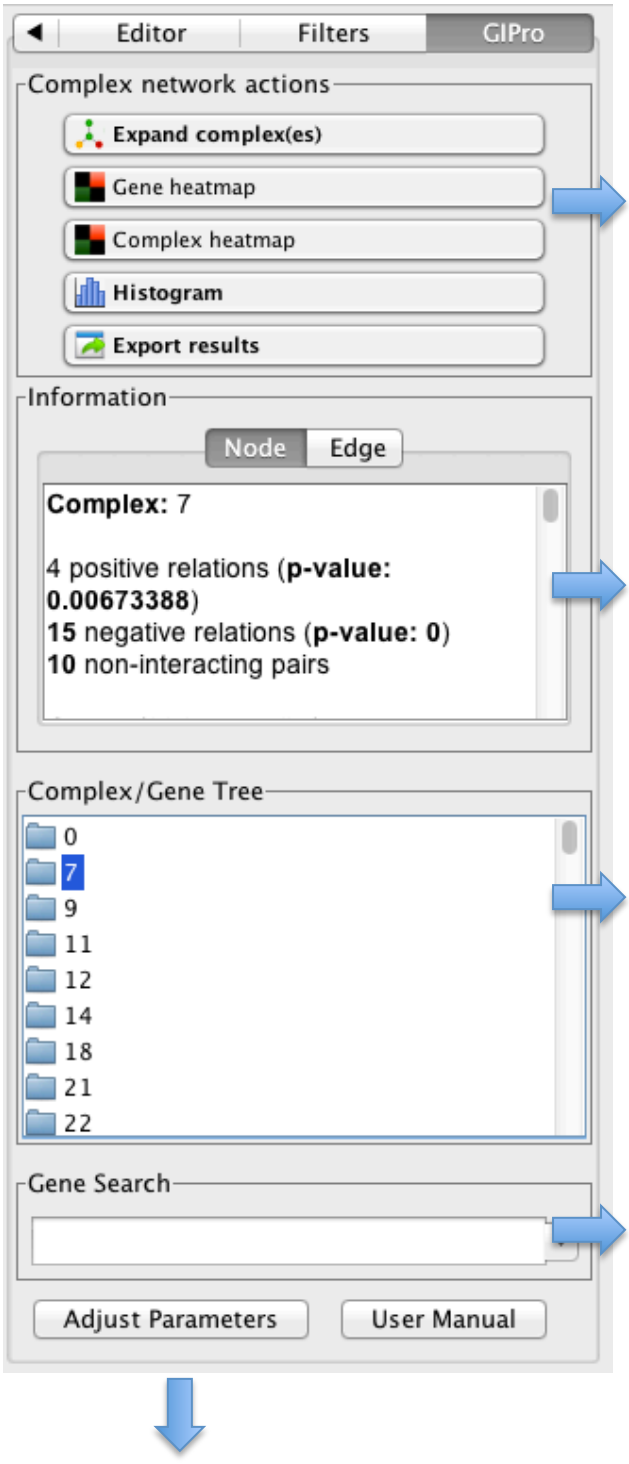
Edges:

- **Color:** the type of interaction occurring between genes. **Red** edges represent positive genetic interactions, **green** edges represent negative genetic interactions, and **blue** edges represent physical interactions. Thickness of **red** and **green** edges is proportional to the score of genetic interactions.

Note: with more than 12 complexes expanded, you may start to see duplicated colors.



2.3 Search and display



Action buttons: used to perform actions on the current network

Nodes and edges information pane: displays information of any node or edge upon its selection

Tree pane: allows navigation of complexes and their genes, will be highlighted to reflected gene or complex selections.

Gene search pane: used to search for a gene in a complex or expanded network. The corresponding complex and gene node (if available) will be selected when searched.

Adjust parameters: Modify your cutoff and enrichment parameters here

2.4 Heat-maps

Heat-maps allow visual representation of the interactions between genes or complexes. There are multiple types of heat-maps that can be generated:

Complex network heat-maps:

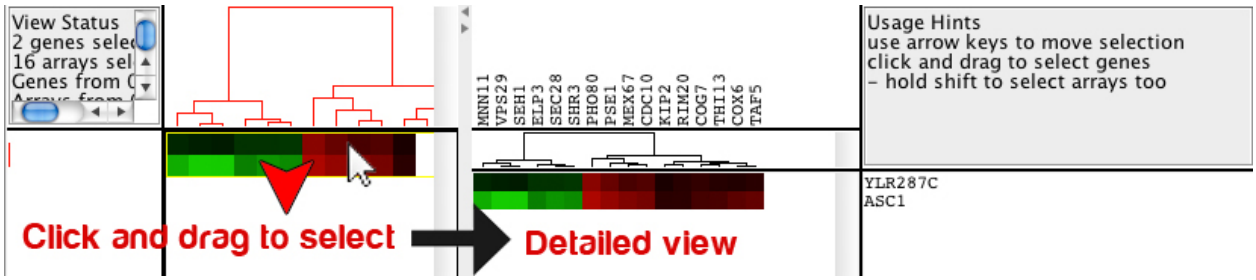
- **Gene heat-map:** select **two or more** complexes and click the Gene heatmap button in the panel, to generate a heat-map of individual interactions between genes in the complexes. The labels beside gene names indicate which complex the genes belong to. The gradient of the color indicate the magnitude of genetic interactions.
- **Complex heat-map:** select **two or more** complexes and click the Complex heatmap button in the panel to generate two heat-maps, one with the average positive score between complex pairs and similarly with the average negative score in the other. The gradient of the color indicate the magnitude of enrichment. Please go to TreeView Menu to change Settings -> Pixel settings -> Contrast to adjust the contrast in order to display the gradient.
-

Expanded view heat-maps:

- **Raw interaction data:** Select **one or more** genes and click Display raw interaction data in the panel to generate a heat-map of interactions between the selected gene(s) and all other loaded genes.

- **Sign patterns:** Select **two or more** genes and click Display sign patterns in the panel to display two heat-maps showing common interactions for the selected genes with same signs (all positive or all negative) in one heat-map and alternating-signs (positive-negative or negative-positive) in the other. **Tip:** to add thresholds for same- and alternating- sign interactions **hold down Ctrl** while clicking.
  Note: array genes that are not interacting are not displayed.

Once the heat-map is generated, click and drag to select one or more rows to be viewed in more detail:



To modify the contrast of the pixels navigate to **Settings -> Pixel settings** and adjust the contrast sidebar.

Note:

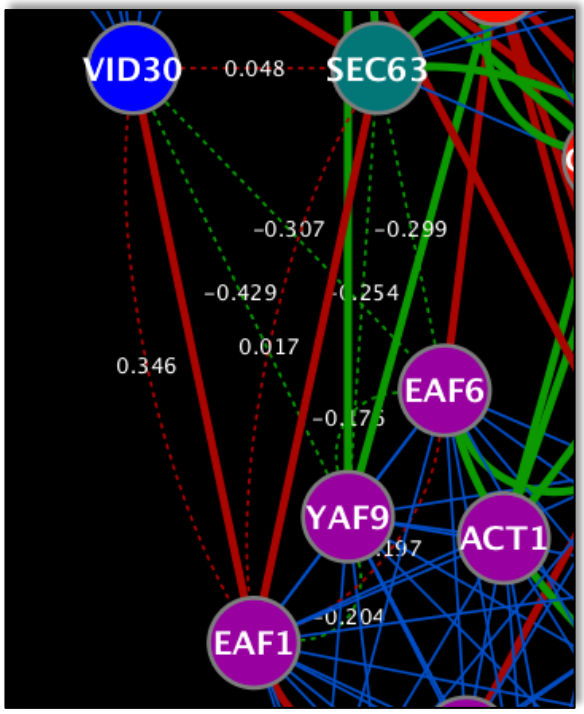
- Hierarchical clustering is applied to heat-maps where needed. For more details see the technical details (Section 4.1)
- All heat-maps are generated using TreeView. For more information on how to use TreeView, visit http://jtreeview.sourceforge.net/.

2.5 Pearson correlation coefficient

Correlation edges:

In order to see how well a pair of genes is correlated, a Pearson correlation coefficient can be generated. In the **expanded network** view, select **two or more** genes and choose 

to add correlation edges for all possible pairs of selected genes to the current subnetwork. These edges are **dashed**, colored **red** if a positive correlation or **green** if a negative correlation and labeled with the Pearson correlation coefficient (*r*). If a correlation edge is not added, there is insufficient data to generate it.

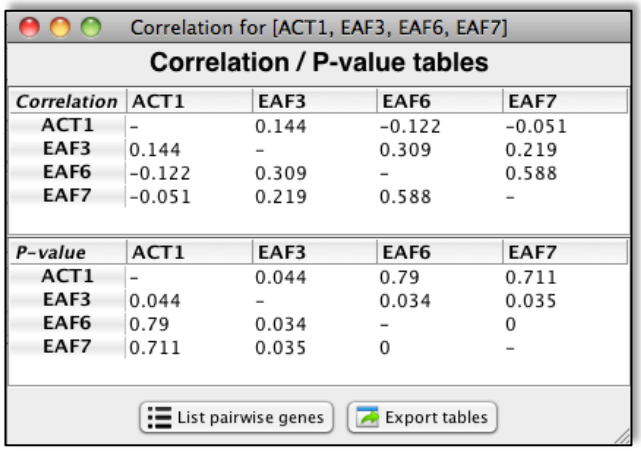


Tip: to add correlation edges above a specific threshold, **hold down Ctrl** while clicking and specify a positive and negative cutoff.

Correlation tables/matrices:

An alternative way of viewing correlation data is through a table or a pairwise matrix. Select **two** or more genes in the expanded network view and click 

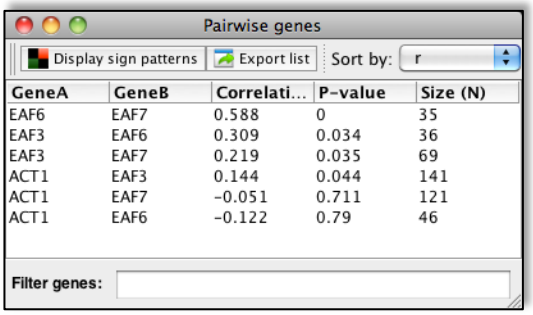
. Two matrices are generated, the top one showing pairwise Pearson correlation coefficients and the bottom one showing pairwise p-values representing the significance of the pairwise correlation. A hyphen **'-'** represents pairs with no data and an asterisk **'*'** represents pairs with insufficient data for a correlation coefficient/p-



value.

To export both matrices, click Export tables and select a save directory. The matrices are saved in a tab-delimited format to the specified directory.

To see a list of pairwise genes, their correlation, p-value and number of interactions used in the calculations *(N)* click List pairwise genes. This list can be sorted by selecting a field from the drop-down menu or filtered using the filter text-field at the bottom of the window. To filter, enter **one or more** gene name (separated by a whitespace) to show pairs containing the entered gene(s).

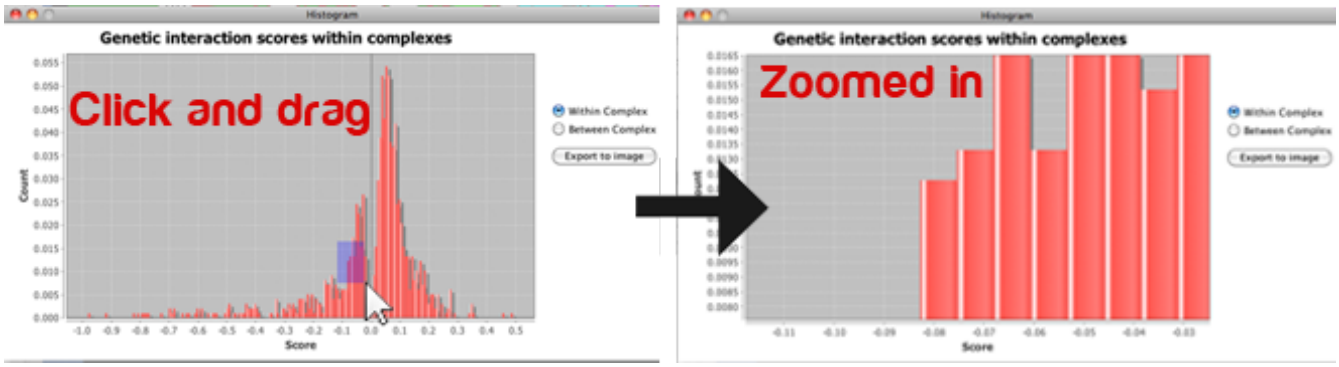


To generate a heat-map of sign patterns between both genes click Display sign patterns (**see section 2.4**). Clicking Export list allows you to export the current list as-is to a tab-delimited text file to the specified directory.

For more details on how the Pearson correlation coefficient is calculated, **see technical details.**

2.6 Histograms

Click the Histogram button in the panel to generate a histogram of all between- and within-complex interaction scores.
Click and drag your mouse to zoom into the histogram.



To export them as an image, click the **"Export to image"** button and specify the save directory.

Export information

3.1 Complex enrichment

Click the  button in the panel to generate files containing information about the enrichment, select the files you wish to output and the **"Export selected complexes only"** checkbox to export results on the selected complexes in the protein complex network. The three files that can be outputted are:

1. **Within-complex enrichment file:** this file contains information on the p-values generated for positive and negative genetic relations for every complex in a spreadsheet format.
   File saved as: givenName_{within}.txt
   Columns
   1) **Name** indicates the name of the complex in question.
   2) **Number of Genes** indicates the number of genes in the complex as stated by the complex file.
   3) **Actual Number of Genes** indicates the number of genes in the complex that are also in the functional relations file.
   4) **Full List** gives a total list of genes.
   5) **Interacting List** gives a list of the genes also in the genetic relations file.
   6) **Interactions** total number of positive, negative and neutral interactions within the complex.
   7) **Pos/Neg/Zero Relation** indicates the number of positive, negative, and neutral relations in the complex.
   8) **Pos/Neg pvalue** is calculated during the statistical analysis for the positive or negative interactions (See technical details for more information)
2. **Between-complex enrichment file:** this file contains information on the distribution and p-values generated for positive and negative genetic relations for every possible complex pair in a spreadsheet format. Only complex pairs with significance are displayed.
   File saved as: givenName_{between}.txt
   Columns:
   1) **Complex1** first complex in the complex pair
   2) **Complex2** second complex in the complex pair
   3) **Number of pairs** the total number of possible pairs between the complexes (e.g. if Complex1 has 3 genes and Complex2 has 4 genes, the total number of pairs is 12).
      Note: If the two complexes share a gene, the gene self loop is not counted. For example, if Complex1 contains genes 1, 2, 3 and Complex2 contains genes 3, 4. The total number of pairs is (3 × 2) - 1 = 5 (an edge is subtracted since gene 3 belongs to both complexes).
   4) **Actual Number of Pairs** number of pairs that have a score in the functional relations file. This is the number used to perform analysis.
   5) **Pos/Neg/Zero Relation** the number of positive, negative, and neutral relations in the complex pair.

6) **Pos/Neg pvalue** is calculated during the statistical analysis for the positive or negative interactions (See technical details for more information).
7) **Significance** Denotes the significance of the complex pair.

3. **Complex enrichment matrix**: this file contains the significance of interactions between complexes. These are displayed in matrix form, with a value of **'1'** representing positive significance, and **'-1'** representing negative significance or a hyphen **'-'** if no significance exists.
   File saved as: givenName_{matrix}.txt

   Sample output file:

complex_name	**complex1**	**complex2**	**complex3**
complex1	-	1	-1
complex2	1	-	1
complex3	-1	1	-

3.2 Pearson correlation matrices/tables

Correlation/P-value matrices

When exporting correlation/p-value matrices, two files are saved
GIPro_correlation_matrix.txt
GIPro_pvalue_matrix.txt
In the following format:

gene_name	**RPL38**	**HHF1**	**MSO1**
RPL38	-	0.823	0.052
HHF1	0.823	-	*
MSO1	0.052	*	-

Ranked pairwise correlation/p-value list

When exporting the pairwise pair-wise list generated**,** one tab-delimited file is saved
GIPro_rank_list.txt
In the following format:

GeneA	**GeneB**	**r**	**p-value**	**N**
RPL38	HHF1	0.823	0.032	56
HHF1	MSO1	0.997	0.001	89
MSO1	RPL38	0.012	0.572	13

3.3 Heat-maps

Export to image

You can export a generated heat-map as an image by selecting **Export-> Export to Image** in the heat-map window. Next to "**Total size:**" adjust the dimension of the heat-map being exported, choose a save path / file name by clicking **"Browse"** and choose an image format from the drop down menu. Finally, click **"Save"** to export the heat-map as an image.

Note: to exclude gene and/or array dendogram from the exported file, deselect the checkboxes **"Gene Tree"** and/or **"Array Tree"**

Similarly, the legend can be exported as an image by selecting **Export-> Export ColorBar** to Image.

Export data matrix

To export a generated heat-map as a data matrix, select **Export -> Save data**. Under **"Field(s) to print"** select **"YORF"** for gene identifiers or "GID" for gene names. Choose the save path / file name by clicking **"Browse"**. Finally, click **"Save"** to export the data matrix

Note: for more details on using TreeView, see http://jtreeview.sourceforge.net/.

Technical details

4.1 Cutoff analysis

A genetic relation is considered positive if it's score exceeds the positive cutoff value, and is considered negative if it is below the negative cutoff. The positive and negative cutoffs are specified in the **"Cutoff Parameters"** panel, and are based on p-value, percentile, or user-specified custom cutoffs.

1) **p-value based cutoff:**

The background scores are approximated as a normal distribution; then identifying true genetic interactions amounts to finding outliers to the background distribution. For each score x_i in the dataset, we determined its probability of belonging to this background distribution by calculating a normalized score z_i as follows:

$$z_i = \frac{x_i - m}{\sigma} = \frac{x_i - m}{\frac{IQR}{IQR_{norm}}}$$

where m is the median and IQR is the inter-quartile range between the first and third quartile. IQRnorm = 1.34898 is the inter-quartile range of the standard normal distribution (with a mean of 0 and a standard deviation of 1). This version of z score is more robust than the conventional z score defined on mean and standard deviation in the sense that median and inter-quartile range are less susceptible to the impact of true interactions which always lie at the tails of the distribution, thus provide a more accurate definition of the background distribution. Each z score corresponds to a p-value in standard normal distribution. Right-tail and left-tail probabilities can be specified separately. Based on the p-value entered, a z score is calculated for each tail, and the genetic interaction score cutoff that corresponding to the z score is determined.

2) **Percentile cutoff:**

The scores are sorted in rising order. Left-tail and right-tail percentiles specify the bottom and top x percent of the entire data, respectively. The value for x can be different for each tail. The percentiles are translated into scores internally by the program.

3) **Custom score cutoff:**

Negative cutoff and positive cutoff will be applied to the scores directly.

Note:

- If a fourth column, containing the interaction p-value, is included in the **Functional relations file**, the p-value must be greater than 0.05 to meet the filtration requirements.
- A complex is only considered if more than half its genes partake in genetic interactions

4.2 Enrichment analysis

In order to determine whether complexes are significantly enriched with positive, negative or both types of interactions, a p-value for the number of positive (given **at least one** positive interaction) and negative relations (given **at least one** positive interaction) in each complex is generated. Based on the **type of enrichment** selected, one of the following is performed:

1) **Fisher Exact Test**
   This statistical method compares the number of positive interactions within a complex, with the number of positive background interactions to generate a right tailed p-value for the complex's positive interactions. Similarly, a right tailed p-value for the complex's negative interactions is generated. **Note:** p-values for positive and negative interactions in a complex are only calculated when there are at least one type of that of interaction.
2) **Simulations**
   The simulation creates distributions for the number of positive (***i***) and negative (***j***) relations within each complex or between a complex pair. The number of possible pairs (***n***) for a given complex is calculated. Given the number specified in the **number of trials** text field (**m**), the algorithm makes ***n*** draws from the **Functional relations file** scores ***m*** times. The number of times the positive relations drawn are greater than or equal to ***i*** is recorded and divided by ***m*** to generate an empirical p-value signifying the likelihood of observing ***i*** positive interactions by chance. A p-value is generated similarly for negative interactions.

   Note: If the "**Run trials for each complex"** check box is unchecked, the algorithm will re-use distributions for complexes with the same number of relations to expedite the algorithm.

   For example, if a complex has 9 proteins and 6 positive interactions and the number of trials specified is 1000, the program randomly draws 36 (9x8/2, divided by 2 to get rid of duplicate edges) relations from the functional relations scores 1000 times, and counts the number of positive interactions each time to generate distributions for the number of positive and negative relations. The number of times that 6 or more positive interactions are observed within the 36 draws is recorded and an empirical p-value is generated by dividing this number into 1000. The same is done for negative interactions.

Once the p-values are generated, a within complex p-value cutoff is calculated which complexes must meet in order to be considered statistically enriched.
This is done using the **false discovery rate** specified in the **enrichment parameters**. The within complex p-values are arranged in increasing order, and the maximum index ***i*** is found, such that for all indices smaller than ***i:***

$$Pvalue_i < \frac{i}{M} FDR$$

where ***M*** is the total number of p-values and ***FDR*** is the specified **false discovery rate**. The corresponding p-value is then the **"within-complex p-value cutoff"**. If the positive and/or negative p-value for a complex is below the p-value cutoff, the complex is considered significantly enriched with that type of interaction.

The procedure above is repeated for between complex interactions, to find significantly enriched complex interaction edges.

4.3 Hierarchical clustering

When generating complex or query heat-maps, hierarchical clustering is used to group genes or complexes into groups or "clusters" such that those within a cluster are closely related to one another. The metric used in the clustering is the **Euclidean distance** and an **average linkage criterion**.
For more information see the homepage of the algorithm at http://function.princeton.edu/WCluster/.

4.4 Pearson correlation coefficient

When generating a Pearson correlation coefficient for a pair of genes, genetic interaction data of both genes is used. Only common interactions between both genes are considered during the calculations. For example, if geneA interacts with geneX, geneY, geneZ and geneB interacts with geneX, geneY, only the pairwise scores of geneX and geneY are used.

Shortcuts

Shortcut	Description
Ctrl + 1	Generate subnetwork using custom list of one or more complex
Ctrl + 2	Generate gene heat-map using custom list of two complexes
Ctrl + 3	Generate complex heat-map using custom list of two or more complexes
Ctrl + 4	Sort complexes in the tree by the number members in the complex in descending order
Ctrl + 5	Sort complexes in the tree by the number members in the complex in ascending order
Ctrl + 6	Sort complexes in the tree by the number members in the complex in alphabetical order
Ctrl down + click	When clicking **"Display sign patterns"** or **"Add correlation edges"**: Allows cutoff to be applied when using both features