Expander 6.3 Online Documentation

Introduction

EXPANDER (EXpression Analyzer and DisplayER) is a java-based tool for ar expression data. It is capable of (1) preprocessing (2) visualizing (3) clustering and (5) performing downstream analysis of clusters and biclusters such

enrichment and promoter analysis (i.e. analysis of gene groups for enrichment factor binding sites in their promoters).

EXPANDER incorporates several conventional gene expression analysis a custom ones that have been developed in the computational genomics gro University, and provides them with an easy-to-operate user interface.

EXPANDER versions are available for Windows OS and for Linux/Unix O the pre-installation of the Java Runtime Environment (JRE) 5.0 or later (Expan first version that fully supports java 1.7). The Java Runtime Environment can b http://java.sun.com/javase/downloads/index.jsp.

The CEL file preprocessing and the newly added SAM filter utilities require the p of one of the recent versions of R, a free software environment for statistical com graphics. For installation instructions, please refer to R External Application secti

Starting EXPANDER

Double click on the **Expander.bat** file, which is located under the Expander direc (alternatively, in Linux, open a Terminal window, cd into the Expander directory, command: './Expander.bat').

When running on Linux/Unix OS, make sure that you have rwx permissions for th directory and for the directory in which your data is located. Also make sure that permissions for all *.exe files that are under your Expander directory.

Upon running the program, the main menu bar appears:

Input Data

Expander operates on the following types of data:

a) **Gene expression data** – For most of EXPANDER's steps for analysis of gene data, the technique used for obtaining the expression estimates doesn't make a Whatever technique (e.g., **expression arrays, RNA-Seq**) was used, the input ex should be summarized in a matrix (tab-delimited txt file; see File Formats sectior rows correspond to probes/genes and columns – to samples.

Values can be either relative intensities data, expected as log 2 (R/G) values dat microarrays) OR absolute intensities data, expected as positive expression level density oligonucleotide data). Oligonucleotide data can be loaded with/without de Affymetrix data can also be loaded from CEL files (If R is installed).

When analyzing **RNA-Seq** data, one way to obtain gene expression matrix is to (http://tophat.cbcb.umd.edu/tutorial.html) to align the sequenced reads to the rel genome, and then use Cufflinks (http://cufflinks.cbcb.umd.edu/howitworks.html) (http://www-huber.embl.de/users/anders/HTSeq/doc/count.html#count) to obtair transcript) expression estimates from TopHat output.

If one wishes to perform functional analysis or promoter analysis, an **ID convers** be loaded along with the data file. The conversion file maps each probe ID (first data file into a corresponding conventional gene ID that is used in the GO ann fingerprint files that are supplied with EXPANDER. The conversion file can be middle of the session too, by Data >> Load Conversion File.

b) **Similarity data** – a pre-calculated similarity matrix

c) **Gene groups data** – contains predefined groups of genes. In this c conventional gene IDs that are used by EXPANDER in the GO annotation and files are expected.

For details regarding the Gene ID convention that is used for each organisr Supplied Files section.

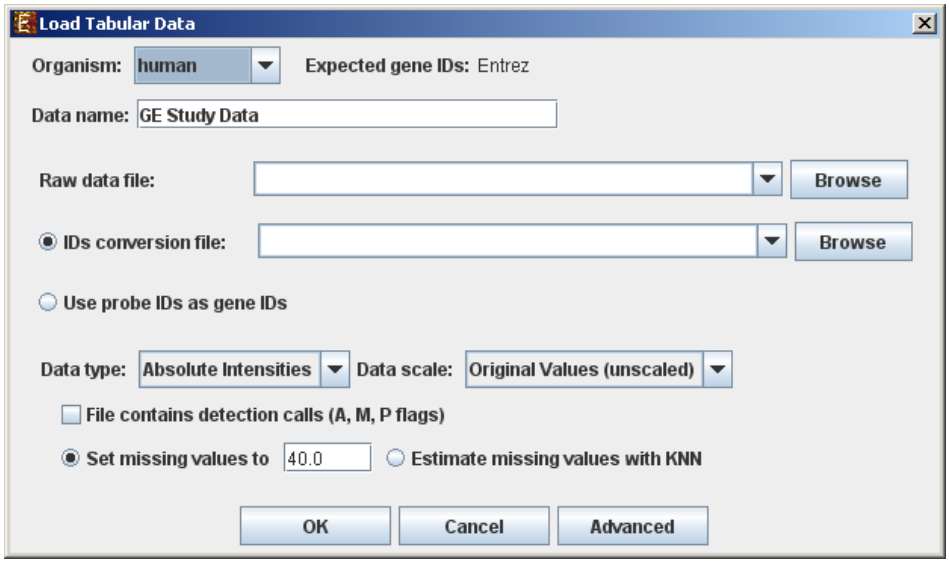
For details regarding the data files formats see the File Formats section.

Loading gene expression data:

Tabular Data File

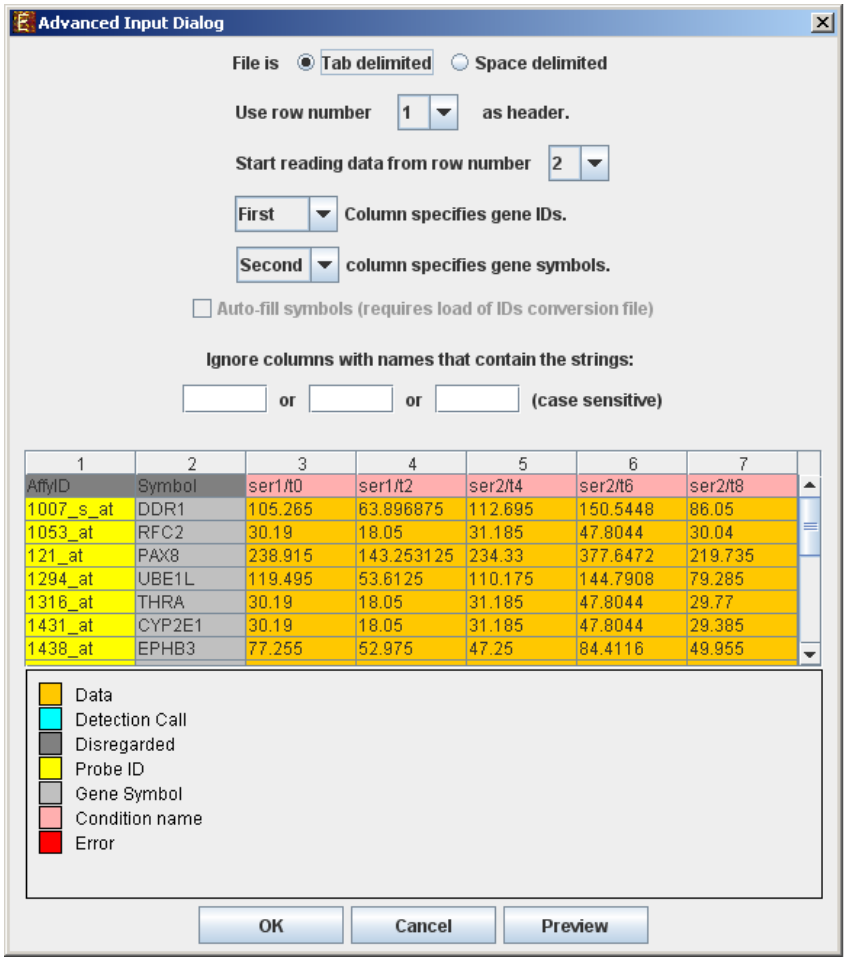
To load tabular expression data, select: *File >> New Session*. From the si *Expression Data >> Tabular Data File*.

When selecting *Tabular Data File,* the following dialog box will appear:



Data type and scale are to be determined according to the input file. If the file cc values, these values will be estimated upon loading the data either by settin arbitrary value (if the 'Set missing value to ____' option is selected) or by utilizir Nearest Neighbors) method (if the 'Estimate missing values with KNN' option the file contains Affymetrix detection calls data, the relevant check box must be may change / erase the default floor value, to which all entries that are below th set (this option is available only for absolute intensities data).

Advanced Input Dialog: Upon pressing the 'Advanced' button after filling the ' field, an 'Advanced Input Dialog' appears. This dialog box can be used in order data load of files that are not in the required format. The first few rows and colur are displayed in a table, demonstrating the way the data is read by the prograı the current input values.



CEL Files

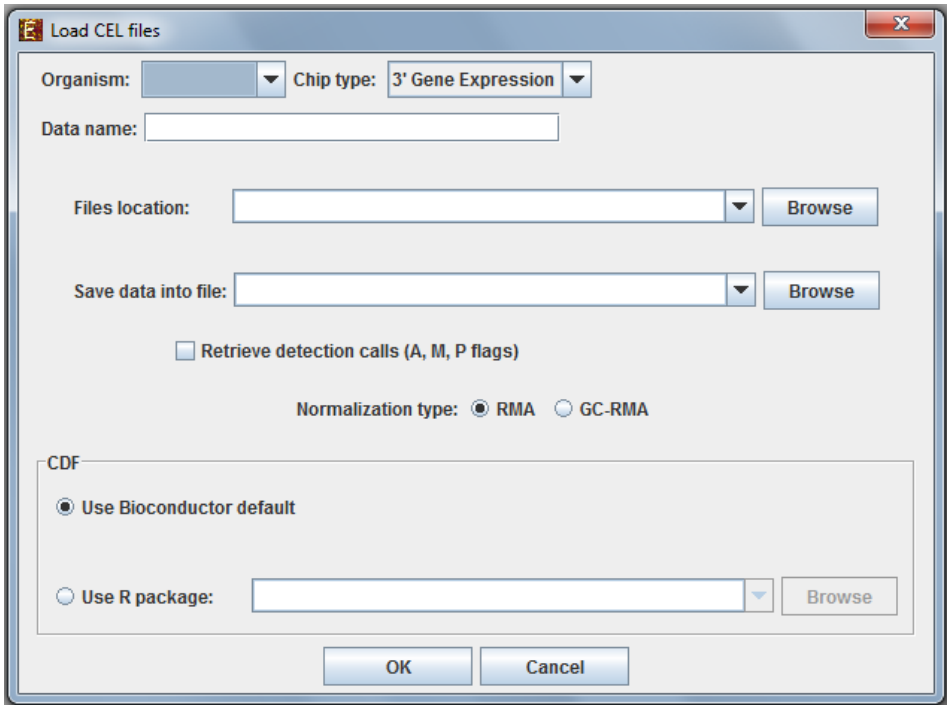
To load expression data from CEL files, select: *File >> New Session*. From the s
Expression Data >> CEL Files.

The load of CEL requires installation of R software (see R External Application
with specific packages, as detailed below. An open internet connection is also r
operation.

Expander supports CEL files of three chip types:

1. **3' Gene Expression** - requires Bioconductor “affy” package
2. **Whole-Transcript Gene Expression** (Gene 1.0 chips) – requires the prior i
cdf package for the used chip (see links below).
3. **Alternative Splicing** (Exon 1.0 chips) requires the prior installation of a cdf p
used chip (see links below). * Please note that we estimate the overall expr
transcript, not exon-by-exon. Therefore, this becomes 'gene data' rather th
splicing data'.

When selecting *CEL Files,* the following dialog box will appear:



Please choose the relevant organism and chip type. Then browse to the folder
files are located (*Files location*), and choose where to save the expression file
the CEL files preprocessing.

Preprocessing and normalization method: The default method in Expander is F
for 3' gene expression arrays, you may select GC-RMA instead (taking into
content bias). Before using GC-RMA, please make sure you have the "gcrm
installed (see R External Application section).

CDF environment choice: You may use the default Bioconductor CDF enviro
chips or browse to an alternative CDF package which you have already insta
whole transcript and alternative splicing chips (for which there is no default Bio
environment), you will need to supply an alternative CDF package (see links bel

Note: GC-RMA requires the probe sequence information of the chip. If you de
the default Bioconductor CDF environment, and have GC-RMA as the preproce
you must have the suitable probe package installed in addition to the CDF altern

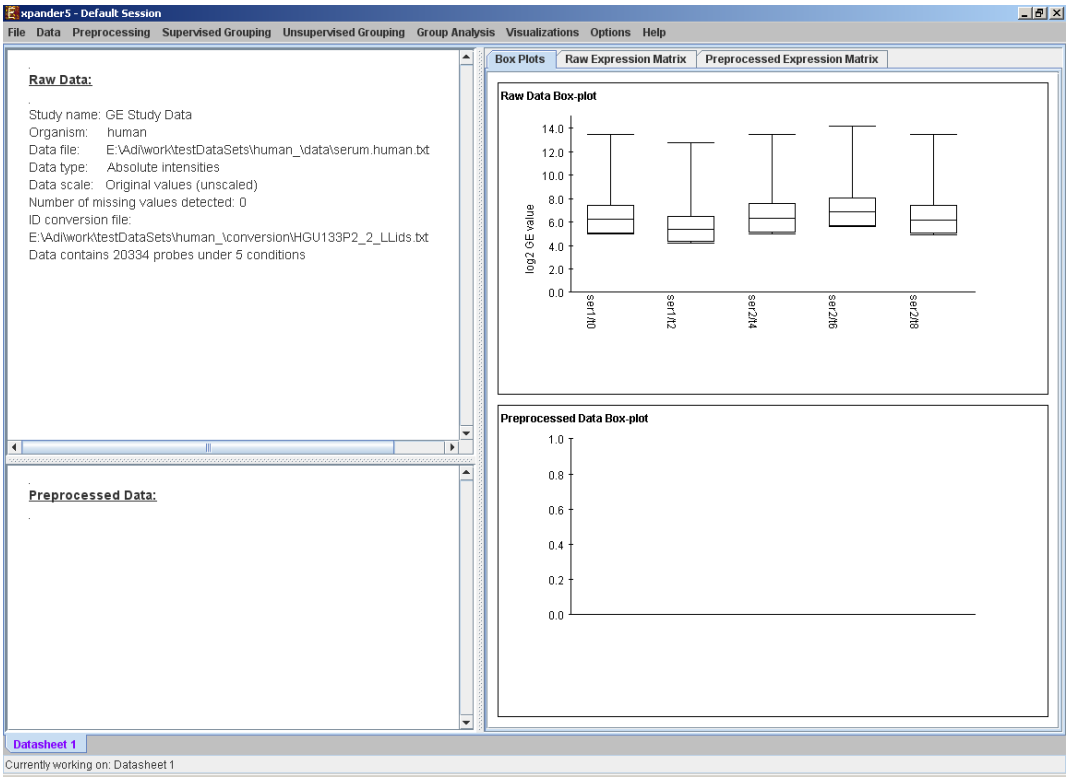
Link for downloading CDF environment packages (for 2nd option):

http://www.bioconductor.org/packages/release/data/annotation/

If Expander cannot find your R software, a window will appear, asking you
location. Please browse to the location of your R software. In Windows, R.exe fil
located in the 'bin' folder of R software. In Linux, you may type 'which R' in the c
find R path. If you have a few versions of R installed, please make sure to point
version in which the Bioconductor "affy" package has been installed.

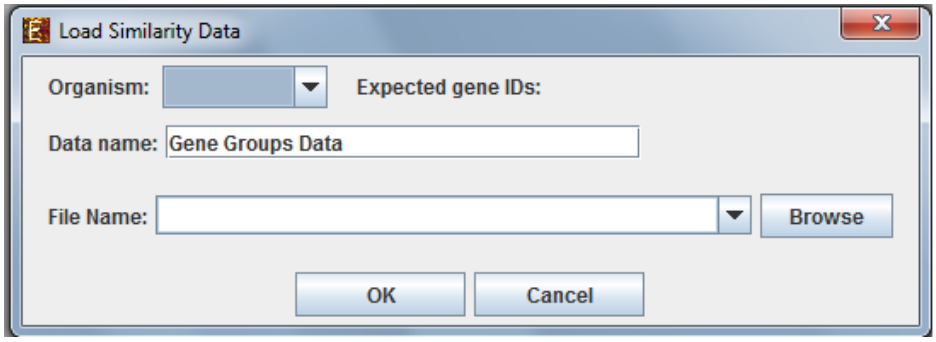
Once the CEL files preprocessing is done, a corresponding tabular data file is g
'Load Study' dialog will appear, as in loading Tabular Data.

After loading a gene expression data set, a ‘Session Data’ display tab is add
window (see example below). It contains information regarding the raw data
chart, and an expression matrix visualization of the raw data. If detection calls e
file, their statistics for each probe appear in 3 columns in the heat maps (expres
in a scale between 0 and 1, corresponding to the relative part of each of the det
M and A). The detection calls statistics for each condition are displayed in a
two tables (one for the raw data and another for the preprocessed data) and a
percent.



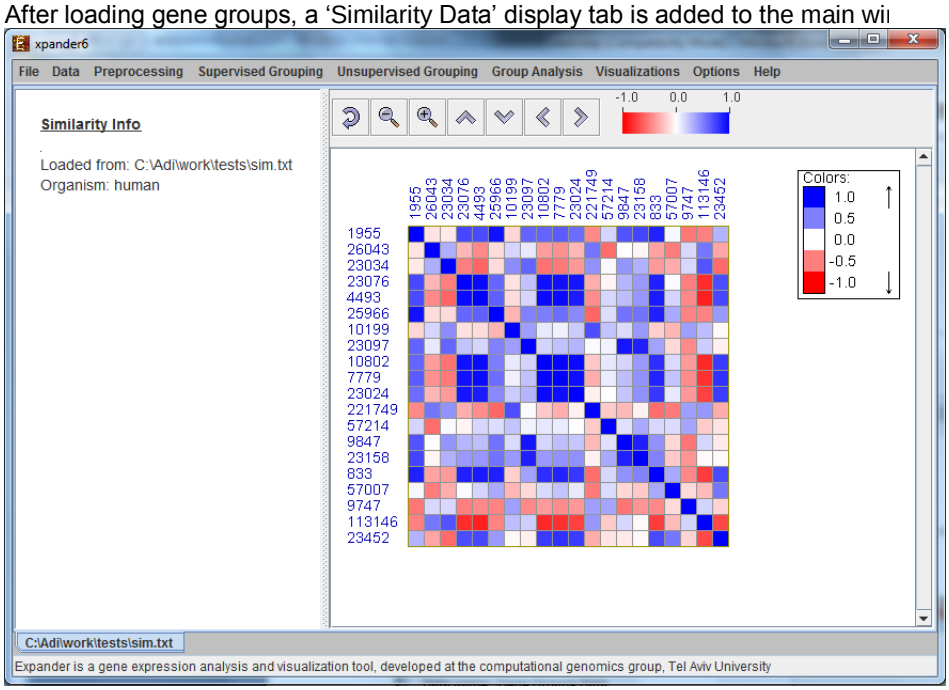
Working on similarity data no associated expression data

To start working on similarity data (no expression data associated) select File>>
Session>> Similarity Data...
The following dialog box will appear:



For details regarding the data files formats see the File Formats section.

After loading gene groups, a 'Similarity Data' display tab is added to the main wi

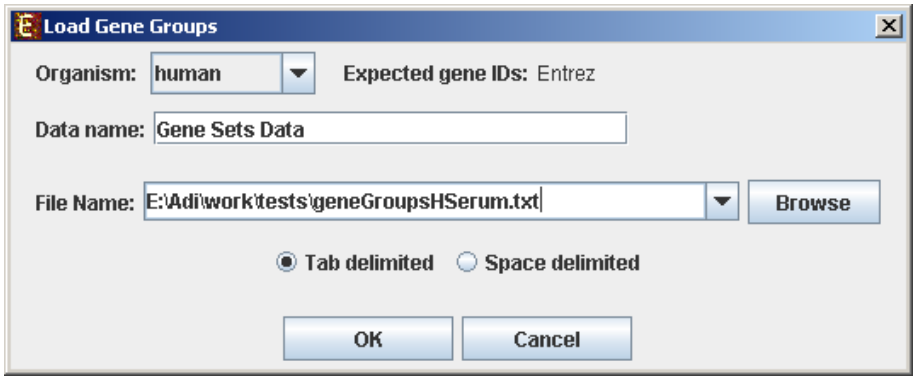


Currently similarity data can only be clustered using the Hierarchical clustering p
selecting *Unsupervised Grouping>>Hierarchical Clustering>>Cluster...* The resul
be used to generate groups (for further details see Hierarchical Clustering).

Working on Gene Groups with no associated expression data

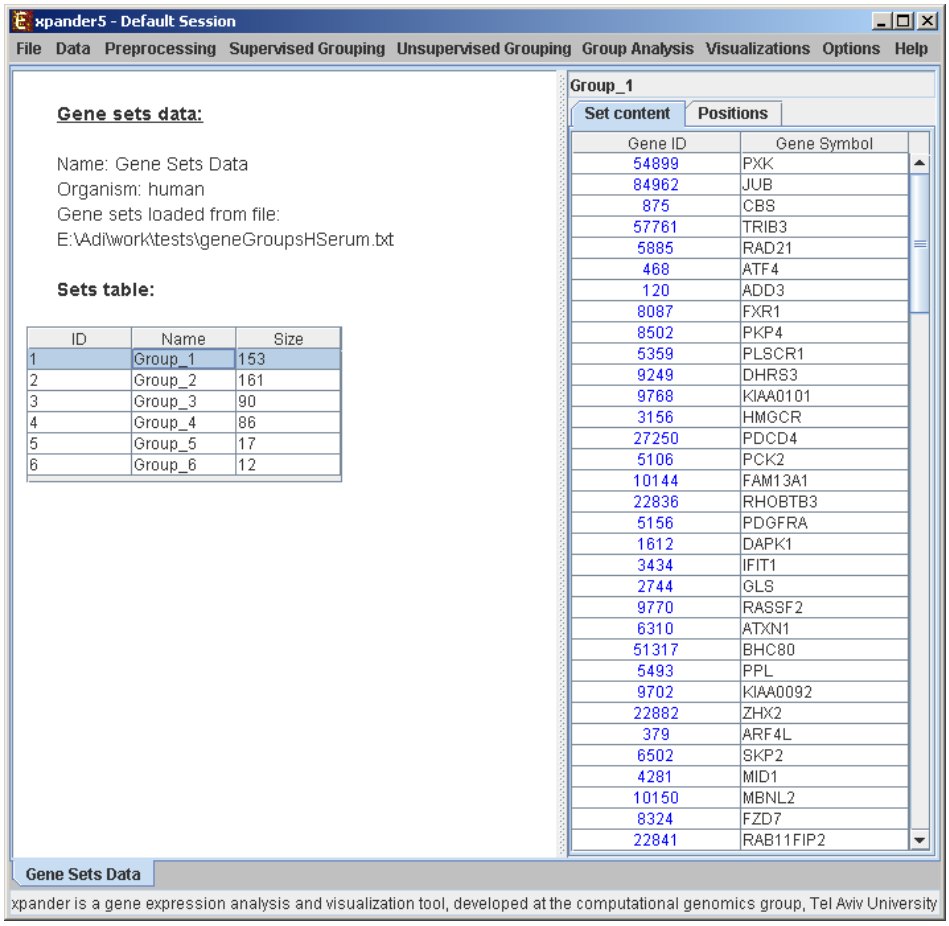
To start working on gene groups (no expression data associated) select *File>*
From the submenu select *Gene Groups*.

The following dialog box will appear:



For details regarding the data files formats see the File Formats section.

After loading gene groups, a 'Session Data' display tab is added to the mai
example below). It contains information regarding the data file, and a table
different groups (serial number, name and size). Group names can be modified
corresponding cell in the table. Upon clicking on a row in the table, the corres
pane appears on the right. It contains a list of the genes in the group and
chromosomal positions. If a network file has been loaded (via *Data>>Load Net*
graph, induced by the group is displayed as well.



Preprocessing GE Data

The following preprocessing operations can be performed using EXPANDER:

1) **Flooring** (*Preprocessing >> Floor Data*): setting all expression values that are b threshold (set by the user) into that threshold. This can be done either by setting itself, or by setting the percentile that should be used as floor value.

2) **Merging conditions** (*Preprocessing >> Merge conditions*): merging a selected s profiles (columns) in the dataset into one profile, in which each entry holds the of the merged entries.

3) **Merging probes according to gene ID** (*Preprocessing >> Merge Probes* automatically shrinks the matrix so that all rows of probes from the same gene a one average row, identified by the corresponding gene ID.

4) **Normalization**: required in order to remove systematic variation, i.e. variatic reasons other than biological differences between RNA samples. Expa normalization only for absolute intensities data, since it is assumed that the rel data (e.g. cDNA microarrays) is already normalized, as it is input after perfo (log2R/G).

Normalization can be performed using the following schemes:

a) **Quantile normalization** (*Preprocessing >> Normalization >> Quantile*), in w data is used.

b) **Non-linear baseline normalization** (*Preprocessing >> Normalization > Baseline*), which uses a baseline array (can be selected by the user). In normalization function is calculated using pseudo Loess regression of the M vs. The subset of genes that are used to evaluate the normalization function car genes' (recommended when most genes in the dataset are expected to expressed) or a 'rank invariant set' of genes (recommended when there can be of differentially expressed genes).

For more details regarding the normalization schemes see the References sectio

5) **Condition filtration**: the conditions used in the analysis can be manually filtere *Preprocessing >> Filter Conditions*. This will bring up a dialog box in which the the required conditions from a list.

6) **Gene (probe) filtration**: can be performed in order to filter out some of expressed genes, and perform downstream analysis on a smaller informative genes.

Probe filtration can be performed using the following schemes:

a) **t-Test** (*Preprocessing >> Filter Probes >> t-Test*): When using this method, o demonstrate differential expression between two condition subsets are selected.

b) **SAM -** Significance Analysis of Microarray (*Preprocessing >> Filter Pro* selects probes that demonstrate differential expression between conditions sub choose 2 or more subsets (multi-class tests are supported). This method uses p get an 'empirical' estimate for the FDR of the reported differential genes (for References section). Before using SAM, please make sure you have **R softwa the "samr" package** installed (see R External Application section).

c) **Fold Change** (*Preprocessing >> Filter Probes >> Fold Change*): when usi only genes that are over/under expressed by at least n fold in at least k arrays and k are determined by the user). The fold change can be calculated in re selected baseline array (b) the minimal expression value of the gene OR (c) value when working on relative intensities (depending on the user's selection).

d) **Variation** (*Preprocessing >> Filter Probes >> Variation*): In this method, the genes are selected (k is determined by the user). Variance is used to measu relative intensities data, and Coefficient of Variation is used to measure variati intensities data.

e) **Detection calls** (*Preprocessing >> Filter Probes >> Detection calls*): i probes/genes are filtered according to the number of expression signals for whic call is 'P' (Present). It can only be operated if the data file contains detection info

f) **Load Probe Subset** (*Preprocessing >> Filter Probes >> Load Probe Subs* set is loaded from an external txt file (for details regarding the format pleas Formats section).

7) **Divide by Base** (*Preprocessing >> Divide by Base*) – Divides each entry in a pr by the corresponding entry in the profile of a selected base condition. This can conditions or for subsets of the conditions.

8) **Log data** (*Preprocessing >> >> Log Data*) – Performs log2 operation on each en

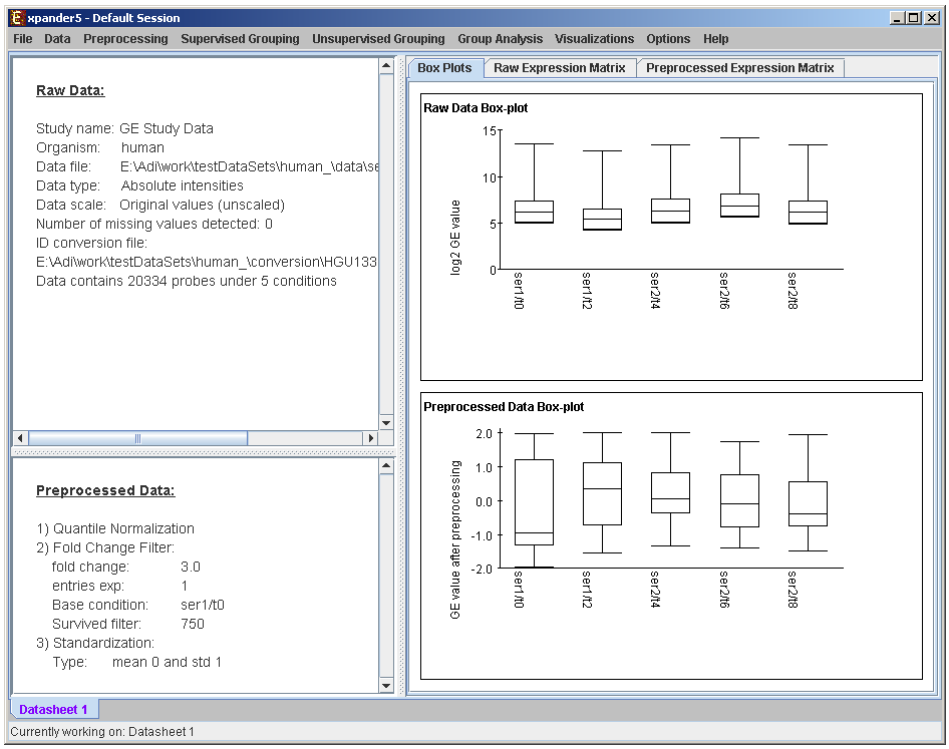
9) **Standardization**: When expression values between different genes are very general expression patterns are similar (high Pearson Correlation values), we w see this similarity when looking on a pattern display. Since the absolute values are different, a manipulation is required, in order to view the patterns on the sa manipulation is called standardization.

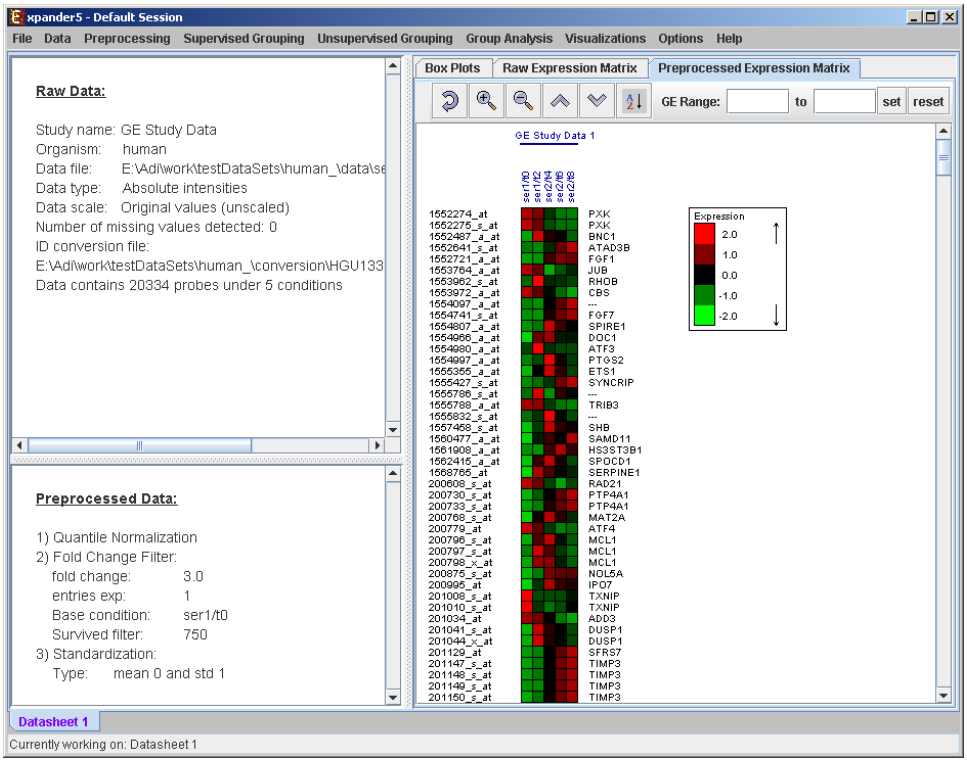
Standardization can be performed using the following schemes:

a) **Mean 0 and Variance 1** (*Preprocessing >> Standardization >> Mean 0 and* normalizes each expression pattern to have a mean of 0 and a variance of 1. appropriate in most cases when working on genes.

b) **Fixed norm** (*Preprocessing >> Standardization >> Fixed Norm*) - no expression pattern to have a fixed norm i.e. expression levels are divided by th expression vector (the root of sum of squares of that vector). This method is ap different mean values or variances are expected for different patterns (e.g. wh conditions and expecting larger variance in later phases of a response.

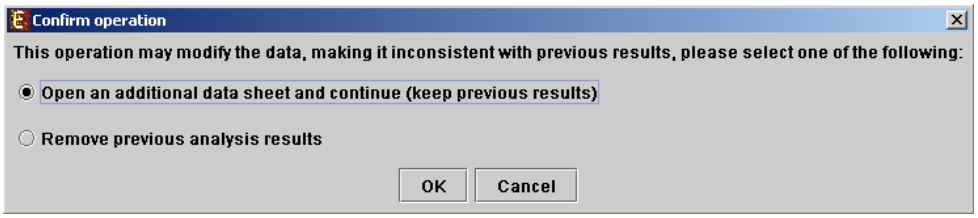
After performing a preprocessing operation, the information regarding the operat the 'Preprocessed Data' section in the 'Session Data' tab. In addition, the 'Prep box plot' and 'Preprocessed Expression Matrix' are automatically updated accor values in the data.





Upon selecting *Preprocessing* >> *Undo* the data is changed to be as it was b recent preprocessing operation was performed, and the corresponding informat from the 'Preprocessed Data' section. The 'Preprocessed Data box plot' and Expression Matrix' are automatically updated accordingly.

All the above operations can be performed before running further analysis or generating displays. When attempting to perform further preprocessing o analysis results and visualizations have been generated, the following dialog box



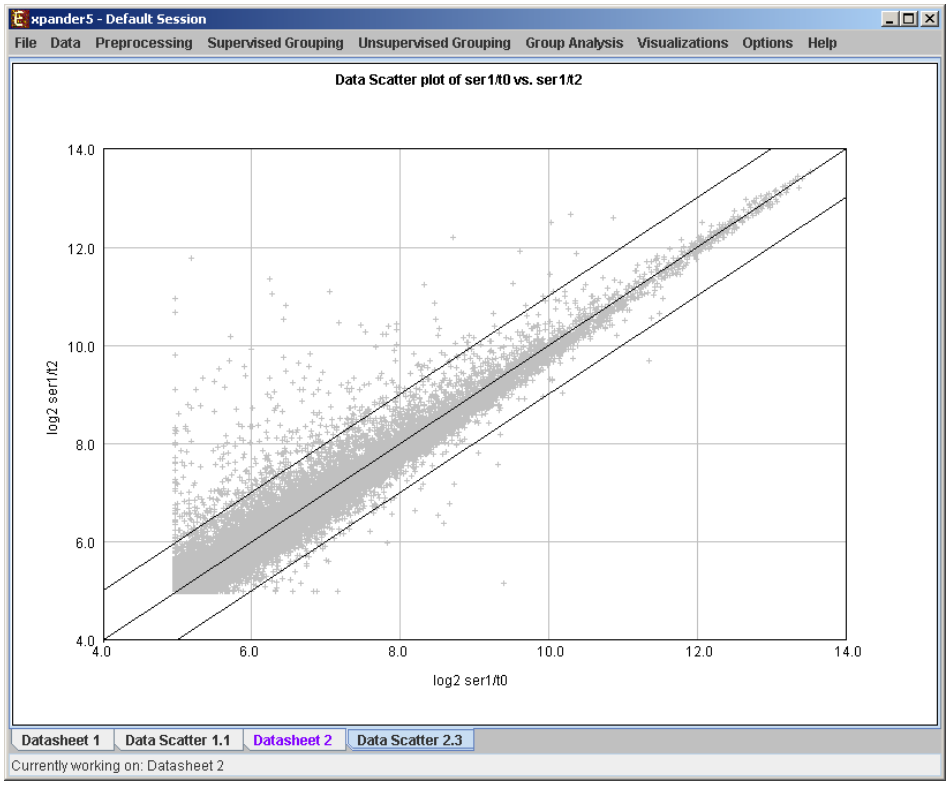
Upon choosing to open an additional data sheet, a new data set view tab called is added to the main frame. The title of this tab is highlighted (colored in purple), it is now the active data sheet (i.e. all further operations refer to this data she data sheet is automatically changed according to the selected (front) visualizatio

Preprocessed gene expression data can be saved to a file at any time *Preprocessing* >> *Save Preprocessed Data*. The data is written in the same for input GE data.

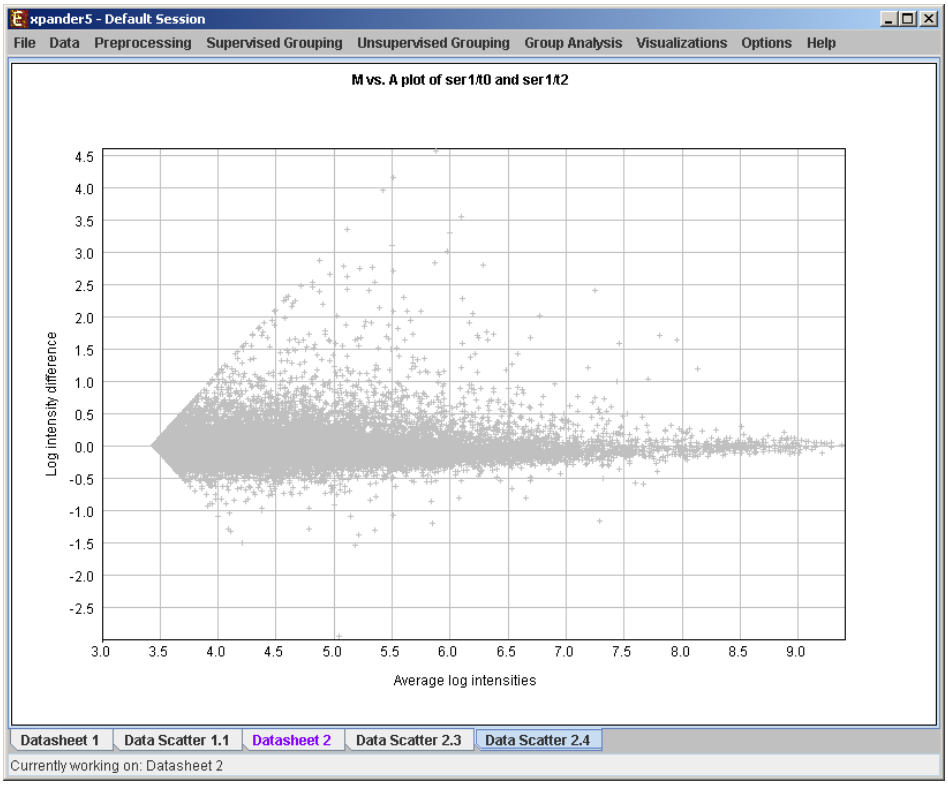
Viewing Data Plots

Expander provides two types of scatter plots visualizations that can be *Visualization >> Scatter Plots...*

Simple plot - Displays a scatter plot of two arrays (selected by the user), in whi (xi,yi) represents the expression value (log expression for un-logged data) of the array vs. the other. For normalized data, points should be located around the y= on the scatter plot).



M vs. A plot (available only for absolute intensities data) - Displays a scatte each point (Ai,Mi) represents the log intensity difference of the i th probe in (selected by the user) vs. the average log value of these intensities.



Differential Expression Analysis

The goal in this analysis is to detect groups of genes that demonstrate differential exp between two/more condition groups.

a) **t-Test** (*Supervised Grouping >> Differential Expression >> t-Test*): W method, genes can be assigned into one of two groups (up-regulated and dc depending on the definitions of t-test parameters.

b) **SAM -** Significance Analysis of Microarray (*Supervised Grouping Expression >> SAM*): this method detects probes that demonstrate differen between conditions subsets. You may choose 2 or more subsets (multi-c supported). The probes are then assigned into two groups (up-regulated and d if 2 condition groups are tested or into one group of differentially expressed o uses permutations to get an 'empirical' estimate for the FDR of the reported dif (for details see the References section). Before using SAM, please make su **software along with the "samr" package** installed (see R External Application

After performing differential expression grouping analysis, a solution visualizatic to the main window. It contains the following views:

Information regarding the algorithm, number of groups (can be either 1 or 2), grouped elements (non-differential), and numerical measures of the groups quali

a) Overall average homogeneity - calculated as the average value of similarity element and the center of the group to which it has been assigned, weighted a size of the group.

b) Overall average separation – calculated as the average similarity between m different groups, weighted according to their sizes.

c) Groups table - contains the number, name (label), size and homogeneity of e

Mean Patterns of the groups with error bars (±1 STD).

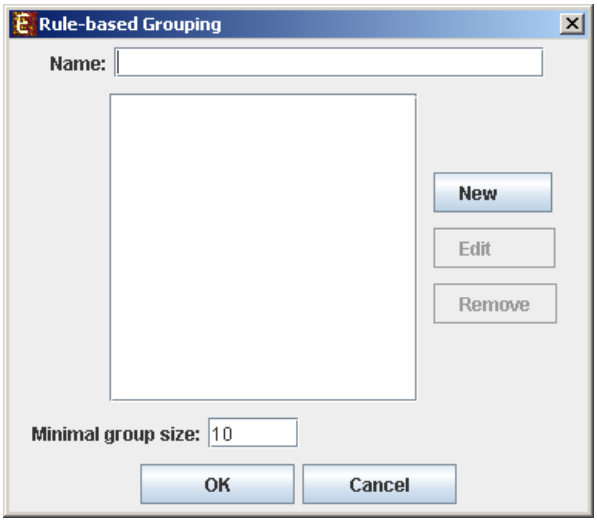
Upon selecting a group, the corresponding pane is displayed on the right. It cc probes, p-values/q-values, fold-change, probe patterns, expression matrix (hea chromosomal locations of the genes. Similarity matrices for probes within the clu for conditions are also displayed in this tab, if the relevant options in the displ selected (see the Settings section). If a network file has been loaded (vi Network), the sub-graph, induced by the cluster is also displayed in the group pa

In order to allow comparison between groups and patterns, the displayed expre are automatically standardized to have mean = 0 and STD = 1.

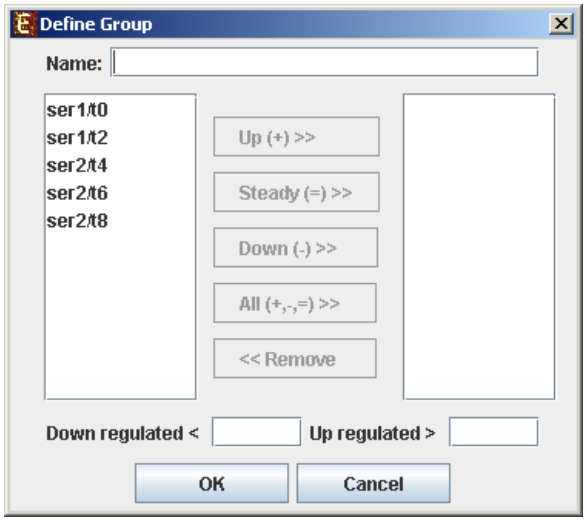
A differential expression solution can be saved using the *File >> Export to tex* solution tab is selected), OR by using the *File>>Save All* option, which will exp within a session to text and image files. A differential expression solution ca using the *Supervised Grouping >> Differential Expression >> Load Solution* optic

Defining a group according to a rule

This can be done by selecting *Supervised Grouping >> Rule-based Grouping.* The fc dialog box will appear:



Upon pressing the “New” button, the following dialog box will appear, to allow definin rule:



In the dialog box, name the new group and select the conditions of interest. For each define weather the expression level should be up-regulated, down-regulated or steac the up-regulation threshold and the down-regulation threshold). These thresholds sh defined. A condition can also be added by pressing the “All” button. In this case a sep will be defined for each of the options of that condition (i.e. a definition of a group usi button can result in more than one group). The visualization for this operation is simil clustering results visualization (described below).

Defining a group according to similarity to a selected prob

This can be done by selecting *Supervised Grouping >> Group by Pattern Similarity. A* box allows setting the similarity measure (Pearson correlation, Spearman correlation distance) and reference probe ID as well as the expected group size. The visualizatic operation is similar to the clustering results visualization (described below).

Clustering GE Data

The goal of clustering is to partition the genes into distinct sets such that assigned to the same cluster should have similar expression patterns, while gen different clusters should have non-similar expression Usually there is no one solution that is the ‘true’ mathematical solution for this good clustering solution should have two merits:

(1) High homogeneity (average similarity between genes from the same cluster).

(2) High separation (average distance/dissimilarity between genes from different

After operating one of the clustering algorithms a clustering results view appe contains information about the solution and its quality including the method a that were used to obtain it, number of clusters, number of singletons (probes assigned to any cluster), overall homogeneity and separation, as well as homogeneity of each cluster. This summary can be used to compare different sc

In order to apply a clustering algorithm to the data, select the required algo *Unsupervised Grouping >> Clustering* menu (options are: **KMeans**, **CLICK**, **S** also use the agglomerative hierarchical clustering algorithm by extracting a pa

existing hierarchical tree, by selecting *Unsupervised Grouping >> Hierarchica Generate Groups* (For details about building such a tree, please go to Hierarchic

Currently similarity data can only be clustered using the Hierarchical clustering p selecting *Unsupervised Grouping>>Hierarchical Clustering>>Cluster...* The resul be used to generate groups (for further details see Hierarchical Clustering).

An existing clustering solution can be loaded from a file by selecting *Unsupen >> Clustering >>Load Solution* (For details regarding the clustering solution file the File Formats section). The **CLICK** algorithm is not designed to find clusters of 15 probes, so it might fail in clustering small datasets.

Fill the required input data in the algorithm input dialog box and press th The parameters required for each method are as follows:

Algorithm	Required parameters
KMeans	Expected number of clusters.
SOM	Grid width, grid length (width*length >= number of clusters) and number of iterations.
CLICK	Homogeneity value (0-1): allows the user control over the homogeneity of the resulting clustering, i.e. the average similarity between elements in the same cluster. This parameter serves as a threshold in various steps in the algorithm, including the definition of cluster kernels, singleton adoptions and kernel merging. The default value for this parameter is the estimated homogeneity of the true clustering. The higher the value assigned to this parameter the tighter the resulting clusters.
Hierarchical tree partition	Distance threshold (if extracting by distance): 0-1 the minimal tree distance that is reqiored for two nodes to be assigned to the same group • It is also possible to partition the tree according to manual node selection that is performed on the hierarchical view (see Hierarchical Clustering).

Details about the algorithms can be obtained through the relevant articles in t section.

After clustering is performed, a clustering solution visualization tab is adde window. It contains the following views:

Information regarding the clustering algorithm, number of clusters, number elements (singletons), and numerical measures of the clustering quality, includin

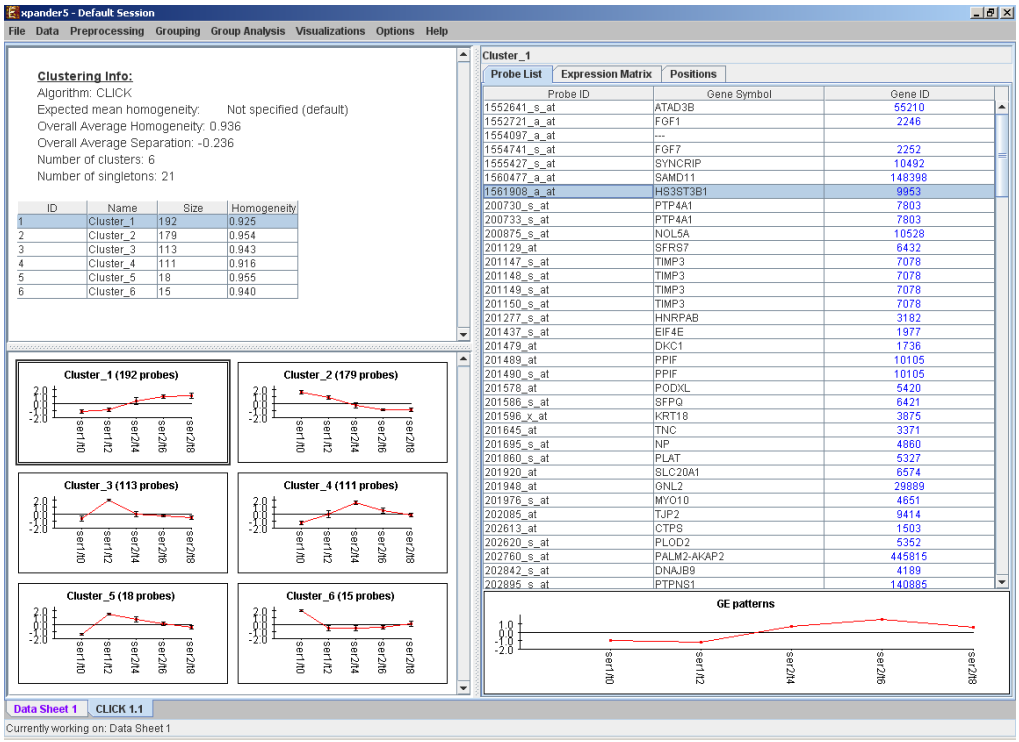
d) Overall average homogeneity - calculated as the average value of similarity element and the center of the cluster to which it has been assigned, weighted a size of the cluster.

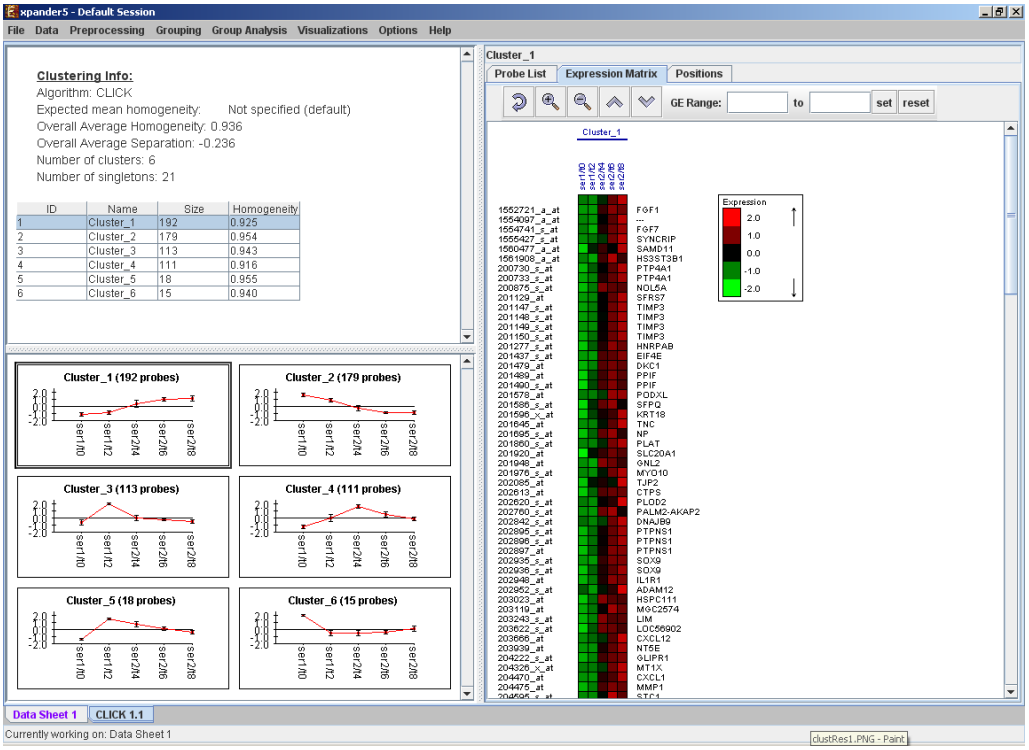
e) Overall average separation – calculated as the average similarity between m different clusters, weighted according to their sizes.

f) Clusters table - contains the number, name (label), size and homogeneity c
The name of a cluster can be changed by editing the corresponding cell in the ta

Mean Patterns of all clusters with error bars (±1 STD).

Upon selecting a cluster (from the clusters table or from the mean patte
corresponding cluster pane is displayed on the right. It contains a list of probes,
expression matrix (heat map) and the chromosomal locations of the genes. Sim
for probes within the cluster as well as for conditions are also displayed in
relevant options in the display settings are selected (see the Settings section). I
has been loaded (via Data>>Load Network), the sub-graph, induced by the
displayed in the cluster pane.





After performing enrichment analysis (for details see the Enrichment Anal enrichment has been detected in the selected cluster, the corresponding analysis information are added to the single cluster view.

In order to allow comparison between groups and patterns, the displayed expre are automatically standardized to have mean = 0 and STD = 1.

A clustering solution can be saved using the *File >> Export to text* op corresponding clustering view as the selected tab) OR by using the *File>>S* which will export all solutions within a session to text and image files. A clusteri be reloaded using the *Unsupervised Grouping >> Clustering >> Load Solution*.

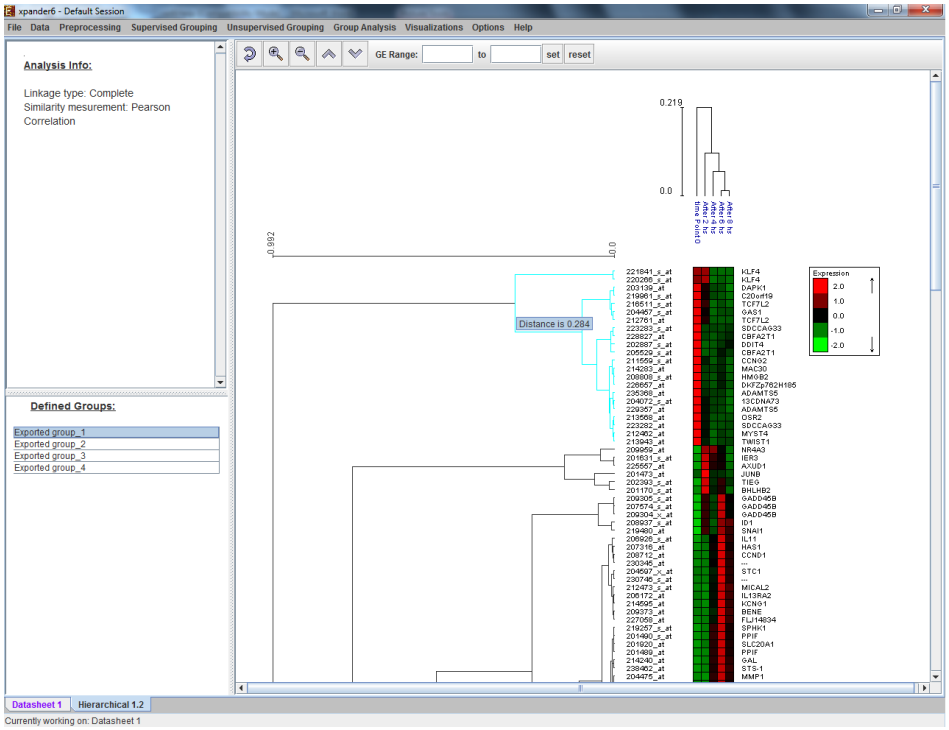
Hierarchical Clustering and Visualization

This tool uses the agglomerative algorithm to calculate a dendrogram tree for patterns (probe patterns) and/or profiles (condition profiles). The type of linka which the distance between a new node and the rest of the nodes is calculate algorithm can be set via an input dialog (for details regarding the algorithm References section). Note that it does not generate a partition of the probes t distance measurement used in the algorithm is (1-Pearson Correlation)/2.

To perform hierarchical clustering, select *Unsupervised Grouping >> Hierarch* Upon selecting this option, a dialog box appears in which the 'linkage type' para the algorithm can be set. After pressing 'OK', the algorithm will be operated bot patterns and on the condition profiles.

The resulting trees are displayed next to an expression matrix so that the prob
vertically on the left and the condition tree appears horizontally above the ma
next to each tree indicates the range of distance values between vectors corres
leaves. The tool tip indicates the distance value corresponding to the cursor l
tree.

If condition attributes file has been loaded for the analyzed dataset, a matrix re
these attributes will be displayed below the expression matri



Upon clicking on the vertical tree, a corresponding sub tree is highlighted (select
be defined as a group by right clicking on the same location and selecting the “E
option from the right click menu. The sub tree is then added as a group of the bo
of the display.

Upon selecting one of the groups that have been previously defined and added t
the bottom left panel, the corresponding sub tree is selected.

A previously selected sub tree can be removed from the list by right clicking on tl
corresponding group in the bottom left panel and selecting remove group.

Manually selected groups can then be defined as a grouping solution by selectin
Unsupervised Grouping>>Hierarchical Clustering>>Generated Groups>> From S
subtrees.

Biclustering GE Data

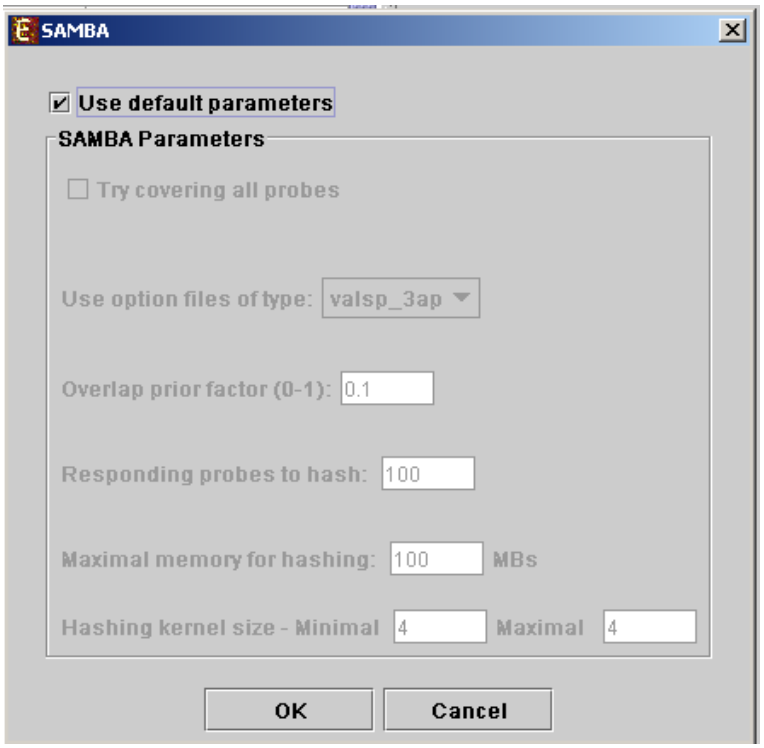
Biclustering is clustering of both genes and conditions of the data into subgrou necessarily disjoint. It enables the user to detect genes that are co-regula subgroup of the conditions, and does not force genes to belong exclusively to o useful when working on datasets which contain a large number of conditions.

Expander incorporates two Biclustering algorithms: ISA (Iterative Signature A SAMBA algorithm (for details see the References section). Before using ISA sure you have **R software along with the "eisa" package** installed (se Application section).

Field Code Changed

In order to apply the ISA algorithm to the data select *Unsupervised Clustering>>ISA*. This operation does not require parameter input.

In order to apply the SAMBA algorithm to the data select *Unsupervised Clustering>>SAMBA*. The following dialog box will appear:



It enables the configuration of some of the parameters for the algorithm. The specifies the different parameters that can be set via this dialog box:

Field	Description

<table>
<tr><td>Use default parameters</td><td>When checked, biclustering parameters (described below) are set automatically (this option is recommended unless the user is familiar with the parameters).</td></tr>
<tr><td>Option files type</td><td>The user can select one out of 6 options. The following table describes the advantages and disadvantages of each option:
<table>
<tr><th>Option name</th><th>fast performance</th><th>less memory required:</th><th>Flexible</th><th>Robust- can handle normalization problems and non gene-expression data</th></tr>
<tr><td>valsp_1</td><td>+</td><td>+</td><td>-</td><td>-</td></tr>
<tr><td>valsp_2</td><td>0</td><td>0</td><td>0</td><td>-</td></tr>
<tr><td>valsp_3</td><td>-</td><td>-</td><td>+</td><td>-</td></tr>
<tr><td>valsp_1ap</td><td>+</td><td>+</td><td>-</td><td>+</td></tr>
<tr><td>valsp_2ap</td><td>0</td><td>0</td><td>0</td><td>+</td></tr>
<tr><td>valsp_3ap</td><td>-</td><td>-</td><td>+</td><td>+</td></tr>
</table>
We recommend the valsp_3ap option (set as default), since it is very flexible, and produces good results also for data that was not normalized properly or for non gene-expression data.</td></tr>
<tr><td>Always cover all genes</td><td>When checked, the solution will cover each gene at least once (each gene will be included in one or more biclusters).</td></tr>
<tr><td>Always cover all conditions</td><td>When checked, the solution will cover each condition at least once (each condition will be included in one or more biclusters). Un checking this option will cause a reduction in the number of biclusters, and the algorithm will run faster.</td></tr>
<tr><td>Overlap prior factor</td><td>Can take values between 0 and 1, describes extent of overlap that is permitted between two different biclusters in the same solution. The higher this parameter is, the more strict the algorithm will be regarding adding a new bicluster (will require less overlap between the new bicluster and the existing ones).</td></tr>
<tr><td>Number of responding genes to hash</td><td>Can take values between 1 and the number of genes in the dataset. Default value is set to 100 (recommended unless data set size < 100). Has impact over the hashing stage in the algorithm.</td></tr>
<tr><td>Maximum hash size (in MB)</td><td>Described the maximum memory size that can be used for the hashing part of the algorithm (the whole algorithm will take up about twice this size of memory).</td></tr>
<tr><td>Maximum hash size</td><td>This parameter determines the number of condition kernel options that are tested and scored in the hashing stage. It can take values from 1 to 7. The default value is 4. In datasets with many conditions raising this number will significantly increase the algorithm run time (may also produce better results).</td></tr>
</table>

Minimum hash size	This parameter determines the minimal size of condition kernel in the hashing stage. It can take values from 1 to 7 and must be <= Maximum hash size. The default value is 4.

Upon clicking ‘OK’ in the dialog box, the SAMBA algorithm is operated on the da

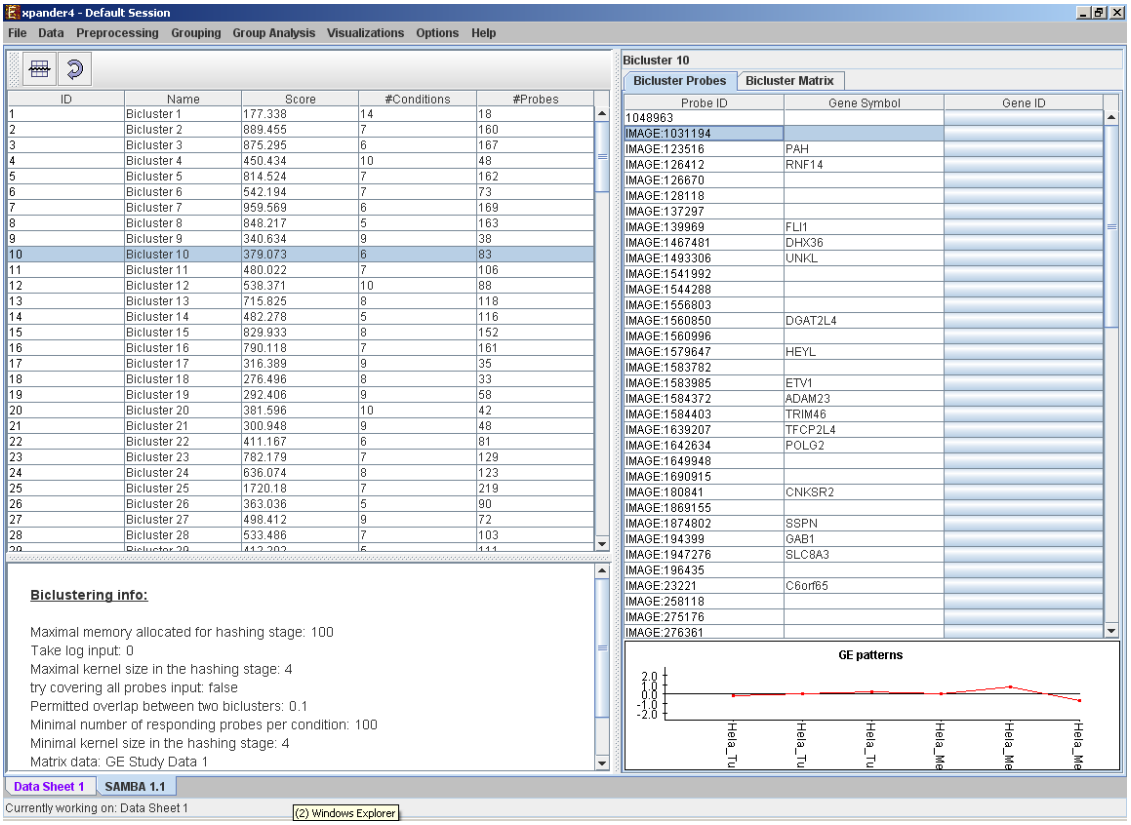
After biclustering is performed a biclustering solution visualization tab is adde window. It contains the following views:

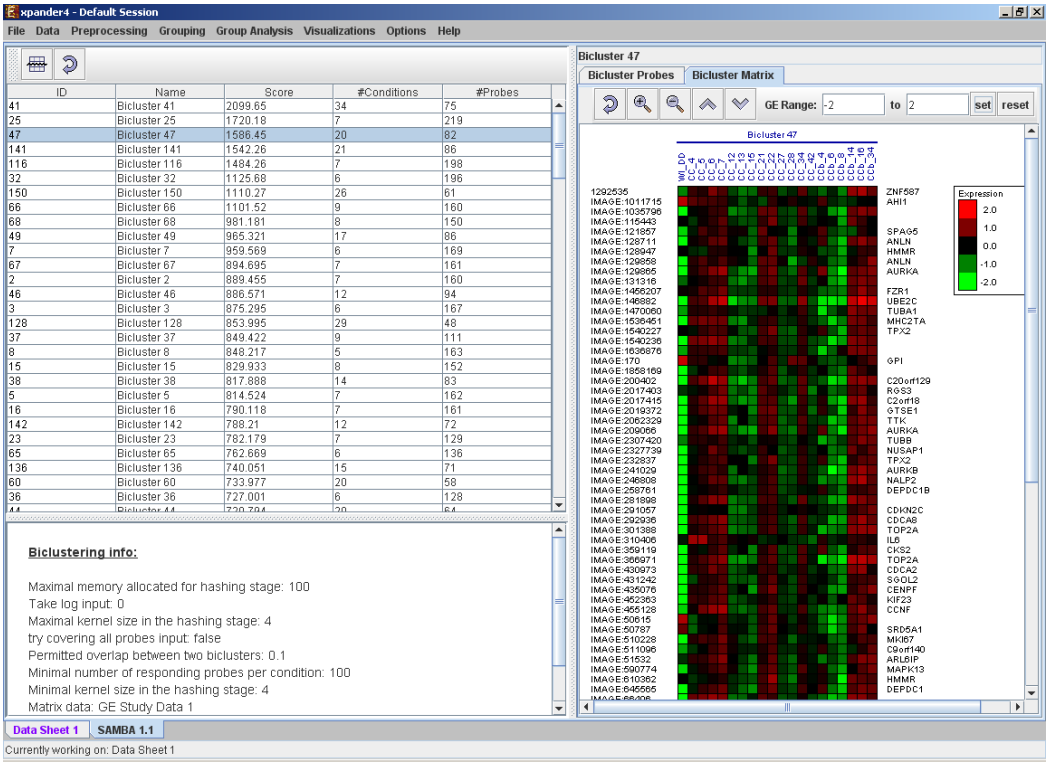
a) Information regarding the biclustering algorithm, and number of resulting biclu

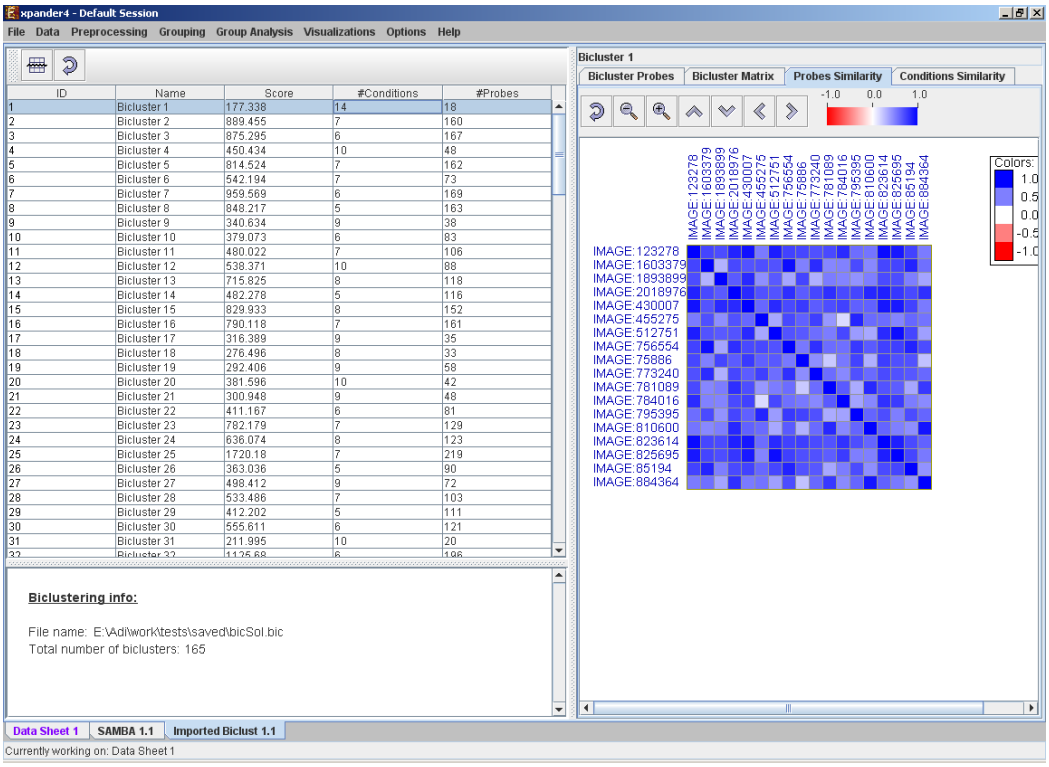
g) Biclusters table – contains the following information for each bicluster: serial r score, number of probes genes and number of conditions. The name of a biclust changed by editing the corresponding cell in the table. The score is given by the algorithm and is size-dependent, thus, it is not recommended to use it to compar two biclusters of different sizes. The table can be filtered to display a subset of th

by clicking on the ‘Filter’ () button in the toolbar. Filtering can be performed a Score, number of probes and number of conditions.

Upon selecting a bicluster (from the biclusters table), the corresponding pane i the right. It contains a list of probes, probe patterns, expression matrix (heat chromosomal locations of the genes. Similarity matrices for probes within the clu for conditions are also displayed in this tab, if the relevant options in the displ selected (see the Settings section). If a network file has been loaded (vi *Network*), the sub-graph, induced by the cluster is also displayed in the cluster p


xpander4 - Default Session
File Data Preprocessing Grouping Group Analysis Visualizations Options Help
ID Name Score #Conditions #Probes
1 Bicluster 1 177.338 14 18
2 Bicluster 2 889.455 7 160
3 Bicluster 3 875.295 6 167
4 Bicluster 4 450.434 10 48
5 Bicluster 5 814.524 7 162
6 Bicluster 6 542.194 7 73
7 Bicluster 7 959.569 6 169
8 Bicluster 8 848.217 5 163
9 Bicluster 9 340.634 9 38
10 Bicluster 10 379.073 6 83
11 Bicluster 11 480.022 7 106
12 Bicluster 12 538.371 10 88
13 Bicluster 13 715.825 8 118
14 Bicluster 14 482.278 5 116
15 Bicluster 15 829.933 8 152
16 Bicluster 16 790.118 7 161
17 Bicluster 17 316.389 9 35
18 Bicluster 18 276.496 8 33
19 Bicluster 19 292.406 9 58
20 Bicluster 20 381.596 10 42
21 Bicluster 21 300.948 9 48
22 Bicluster 22 411.167 6 81
23 Bicluster 23 782.179 7 129
24 Bicluster 24 636.074 8 123
25 Bicluster 25 1720.18 7 219
26 Bicluster 26 363.036 5 90
27 Bicluster 27 498.412 9 72
28 Bicluster 28 533.486 7 103
Biclustering info:
Maximal memory allocated for hashing stage: 100
Take log input: 0
Maximal kernel size in the hashing stage: 4
try covering all probes input: false
Permitted overlap between two biclusters: 0.1
Minimal number of responding probes per condition: 100
Minimal kernel size in the hashing stage: 4
Matrix data: GE Study Data 1
Bicluster 10
Bicluster Probes
Bicluster Matrix
Probe ID Gene Symbol Gene ID
1048963
IMAGE:1031194
IMAGE:123516 PAH
IMAGE:126412 RNF14
IMAGE:126670
IMAGE:128118
IMAGE:137297
IMAGE:139969 FLI1
IMAGE:1467481 DHX36
IMAGE:1493306 UNKL
IMAGE:1541992
IMAGE:1544288
IMAGE:1556803
IMAGE:1560850 DGAT2L4
IMAGE:1560996
IMAGE:1579647 HEYL
IMAGE:1583782
IMAGE:1583985 ETV1
IMAGE:1584372 ADAM23
IMAGE:1584403 TRIM46
IMAGE:1639207 TFCP2L4
IMAGE:1642634 POLG2
IMAGE:1649948
IMAGE:1690915
IMAGE:180841 CNKSR2
IMAGE:1869155
IMAGE:1874802 SSPN
IMAGE:194399 GAB1
IMAGE:1947276 SLC8A3
IMAGE:196435
IMAGE:23221 C6orf65
IMAGE:258118
IMAGE:275176
IMAGE:276361
GE patterns
2.0
1.0
0.0
-1.0
-2.0
Hela_Tu
Hela_Tu
Hela_Tu
Hela_Me
Hela_Me
Hela_Me
Data Sheet 1
SAMBA 1.1
Currently working on: Data Sheet 1
(2) Windows Explorer


xpander4 - Default Session
File Data Preprocessing Grouping Group Analysis Visualizations Options Help
ID Name Score #Conditions #Probes
41 Bicluster 41 2099.65 34 75
25 Bicluster 25 1720.18 7 219
47 Bicluster 47 1586.45 20 82
141 Bicluster 141 1542.26 21 86
116 Bicluster 116 1484.26 7 198
32 Bicluster 32 1125.68 6 196
150 Bicluster 150 1110.27 26 61
66 Bicluster 66 1101.52 9 160
68 Bicluster 68 981.181 8 150
49 Bicluster 49 965.321 17 86
7 Bicluster 7 959.569 6 169
67 Bicluster 67 894.695 7 161
2 Bicluster 2 889.455 7 160
46 Bicluster 46 886.571 12 94
3 Bicluster 3 875.295 6 167
128 Bicluster 128 853.995 29 48
37 Bicluster 37 849.422 9 111
8 Bicluster 8 848.217 5 163
15 Bicluster 15 829.933 8 152
38 Bicluster 38 817.888 14 83
5 Bicluster 5 814.524 7 162
16 Bicluster 16 790.118 7 161
142 Bicluster 142 788.21 12 72
23 Bicluster 23 782.179 7 129
65 Bicluster 65 762.669 6 136
136 Bicluster 136 740.051 15 71
60 Bicluster 60 733.977 20 58
36 Bicluster 36 727.001 6 128
Biclustering info:
Maximal memory allocated for hashing stage: 100
Take log input: 0
Maximal kernel size in the hashing stage: 4
try covering all probes input: false
Permitted overlap between two biclusters: 0.1
Minimal number of responding probes per condition: 100
Minimal kernel size in the hashing stage: 4
Matrix data: GE Study Data 1
Bicluster 47
Bicluster Probes
Bicluster Matrix
GE Range: -2 to 2 set reset
Bicluster 47
Expression
2.0
1.0
0.0
-1.0
-2.0
Data Sheet 1
SAMBA 1.1
Currently working on: Data Sheet 1


xpander4 - Default Session
File Data Preprocessing Grouping Group Analysis Visualizations Options Help
ID Name Score #Conditions #Probes
1 Bicluster 1 177.338 14 18
2 Bicluster 2 889.455 7 160
3 Bicluster 3 875.295 6 167
4 Bicluster 4 450.434 10 48
5 Bicluster 5 814.524 7 162
6 Bicluster 6 542.194 7 73
7 Bicluster 7 959.569 6 169
8 Bicluster 8 848.217 5 163
9 Bicluster 9 340.634 9 38
10 Bicluster 10 379.073 6 83
11 Bicluster 11 480.022 7 106
12 Bicluster 12 538.371 10 88
13 Bicluster 13 715.825 8 118
14 Bicluster 14 482.278 5 116
15 Bicluster 15 829.933 8 152
16 Bicluster 16 790.118 7 161
17 Bicluster 17 316.389 9 35
18 Bicluster 18 276.496 8 33
19 Bicluster 19 292.406 9 58
20 Bicluster 20 381.596 10 42
21 Bicluster 21 300.948 9 48
22 Bicluster 22 411.167 6 81
23 Bicluster 23 782.179 7 129
24 Bicluster 24 636.074 8 123
25 Bicluster 25 1720.18 7 219
26 Bicluster 26 363.036 5 90
27 Bicluster 27 498.412 9 72
28 Bicluster 28 533.486 7 103
29 Bicluster 29 412.202 5 111
30 Bicluster 30 555.611 6 121
31 Bicluster 31 211.995 10 20
Biclustering info:
File name: E:\Adi\work\tests\saved\bicSol.bic
Total number of biclusters: 165
Bicluster 1
Bicluster Probes
Bicluster Matrix
Probes Similarity
Conditions Similarity
-1.0 0.0 1.0
Colors:
1.0
0.5
0.0
-0.5
-1.0
IMAGE:123278
IMAGE:1603379
IMAGE:1893899
IMAGE:2018976
IMAGE:430007
IMAGE:455275
IMAGE:512751
IMAGE:756554
IMAGE:75886
IMAGE:773240
IMAGE:781089
IMAGE:784016
IMAGE:795395
IMAGE:810600
IMAGE:823614
IMAGE:825695
IMAGE:85194
IMAGE:884364
Data Sheet 1
SAMBA 1.1
Imported Biclust 1.1
Currently working on: Data Sheet 1

After performing enrichment analysis (for details see the Enrichment Analysis T enrichment has been detected in the selected bicluster, the corresponding analysis information are added to the single bicluster view, and a column is expression matrix display for each enrichment class, stating for each probe, whe to that class.

A biclustering solution can be saved using the *File >> Export to text* op corresponding biclustering view as the selected tab) OR by using the *File>>S* which will export all solutions within a session to text and image files. A biclus can be reloaded by selecting *Unsupervised Grouping >> Bi-Clustering >> Load* format of the solution file, please refer to the File Formats section:

Network Based Grouping of GE Data

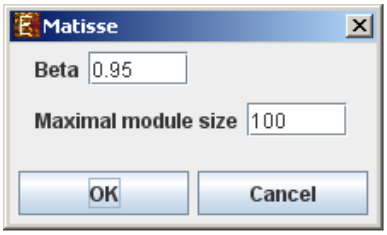
The goal here is to detect groups of genes that demonstrate similar expressio are also highly connected in a given interactions network.

In order to operate these tools, an interactions network in .SIF format needs to b can be done either by selecting *Data>>Load Network…* or via the dialog boxes o

In order to perform network based grouping Expander incorporates two algor and Degas (for details see the References section). The DEGAS algorithm is the expression dataset compares two groups of heterogeneous samples (as studies). The groups detected by these tools are referred to as "modules" an also genes that exist in the network, but are not present in the filtered GE data "Back nodes").

To use the more advanced, stand-alone versions of MATISSE and DEGA flexibility), please refer to the Matisse home page.

In order to apply the Matisse algorithm to the data select *Unsupervised Grou >>Matisse*. The following dialog box will appear:

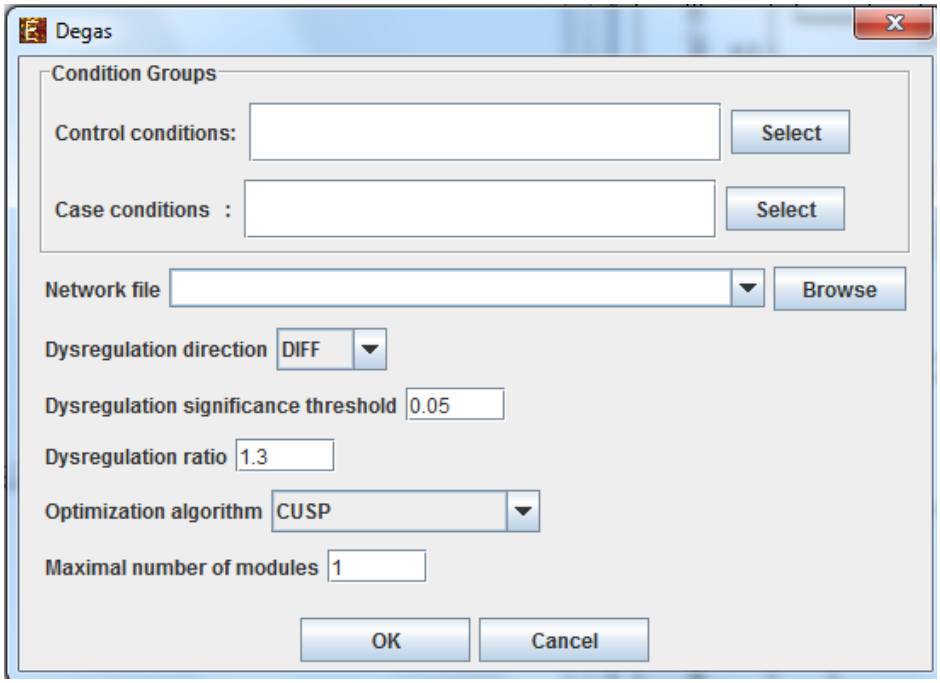


It enables the configuration of some of the parameters for the algorithm:

Field	Description
Beta	The fraction of gene pairs that are expected to be strongly co-expressed in each module
Maximal module size	The maximum size for a detected module.

Upon clicking 'OK' in the dialog box, the Matisse algorithm is operated on the da

In order to apply the Degas algorithm to the data select *Supervised Grou*
>>Degas. The following dialog box will appear:

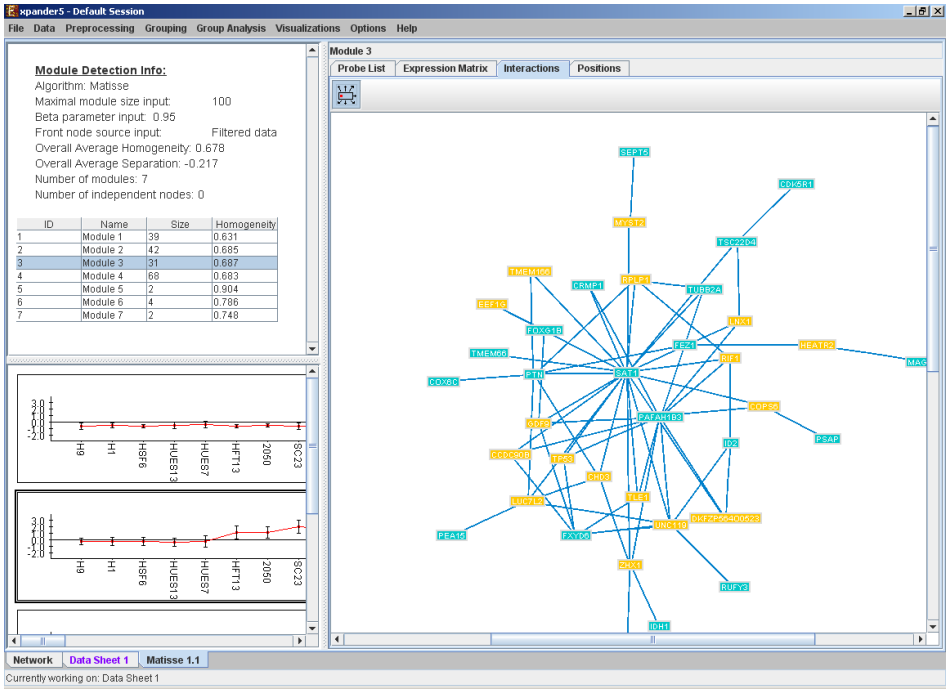


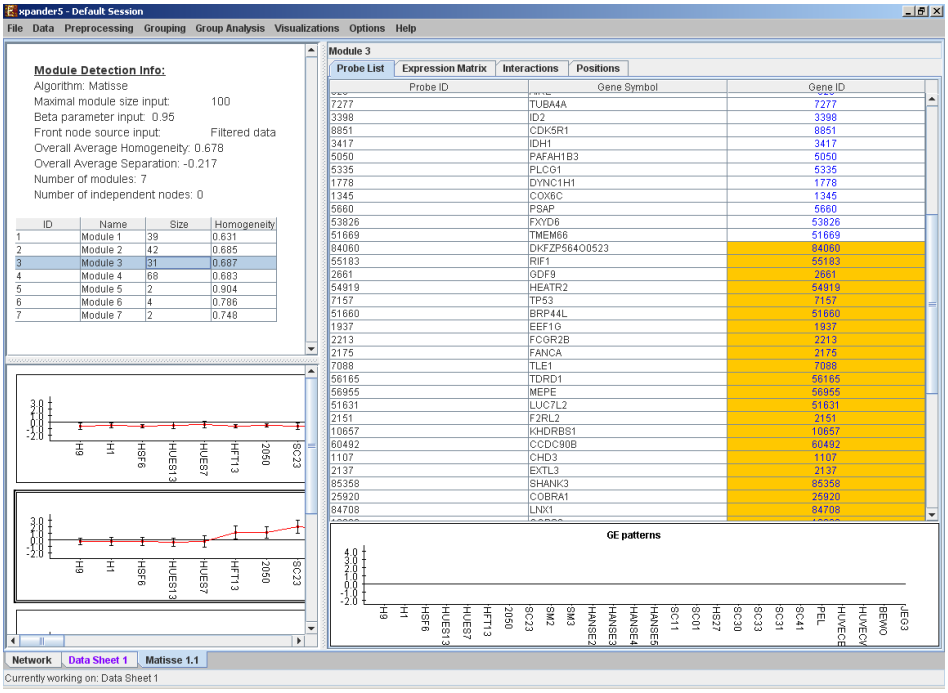
It enables the configuration of some of the parameters for the algorithm:

Field	Description
Case conditions	The case conditions
Control conditions	The control conditions
Dysregulation direction	This parameter will determine which direction of dysregulation will be sought (up/down-regulation/both).
Dysregulation significance threshold	This threshold will be used to identify which genes are differentially expressed in each 'case' sample compared to the controls
Dysregulation ratio	The minimal threshold for the ratio between the gene expression in any of the case conditions and the average expression in the control conditions. Above this threshold a case condition is designated as dysregulated.
Optimization algorithm	The algorithm used to identify dysregulated pathways (DPs). See the DEGAS manuscript for details. CUSP is the recommended option
Maximal number of modules	After DEGAS identifies a significant DP, it removes it from the input data and attempts to identify additional DPs. This parameter specifies the total number of DPs that will be sought.

Upon clicking ‘OK’ in the dialog box, the Matisse algorithm is operated on the da

After running network-based clustering, the solution is displayed in a new tab, v
to the main window. The view is similar to the clustering results display. In the
nodes (genes that appear In the network, but not in the GE data) are marked in y





After performing enrichment analysis (for details see the Enrichment Analysis T enrichment has been detected in the selected module, the corresponding analysis information are added to the single module view, and a column is expression matrix display for each enrichment class, stating for each probe, whe to that class.

A network-based grouping solution can be saved using the *File >> Export to te* the corresponding grouping view as the selected tab) OR by using the *File>>S* which will export all solutions within a session to text and image files. A grouping solution can be reloaded using the *Unsupervised Grouping >> Ne Solution* option or via *Supervised Grouping >> Network >> Load Solution*. For solution file, please refer to the File Formats section:

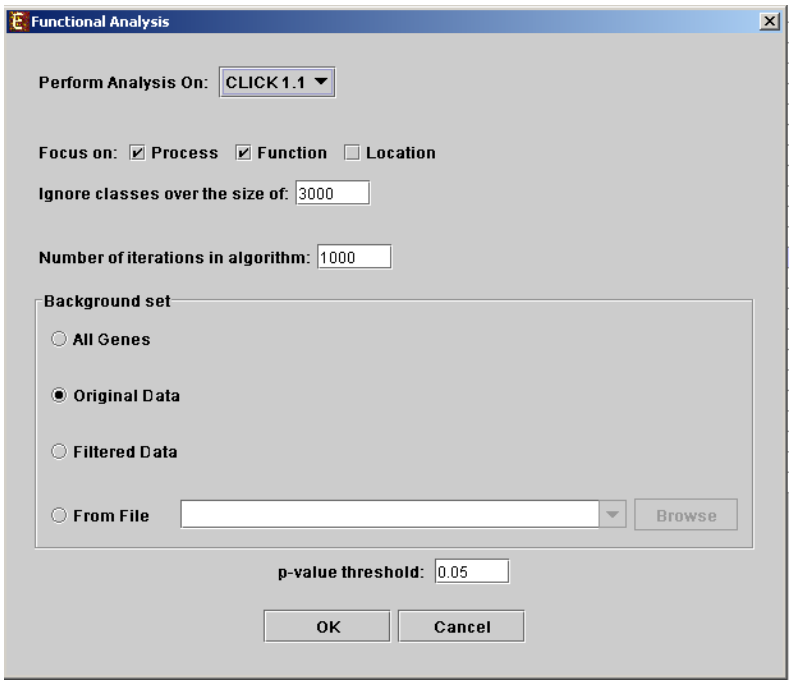
Group Enrichments Analysis Tools

The following analysis can be performed on gene sets, clusters, biclusters, modules, similarity based groups, or the filtered dataset (the analyzed set of set). Before operating any of the enrichment analysis operation (not includin enrichment analysis"), the data files for the relevant organism should be Download can be done by selecting *Help >> Download Data for Organism*. U new session, automatic data download will be suggested if Expander did not relevant organism.

Functional Analysis

This tool performs basic statistical analysis on the distribution of functions of ger cluster. The functions of the genes are determined according to annotation file can be downloaded from the EXPANDER download page (see the Supplied Fil perform this analysis, Expander utilizes the TANGO software, which performs hy enrichment tests and corrects for multiple testing by bootstrapping and estimatin p-value distribution for the evaluated sets.

Before operating functional analysis the annotation files for the relevant orgar downloaded from the download page (more details at introduction of Group An To perform the analysis, select *Enrichment Analysis >> Functional Analysis >>* following dialog box will appear:



The following table specifies the different parameters that can be set via th

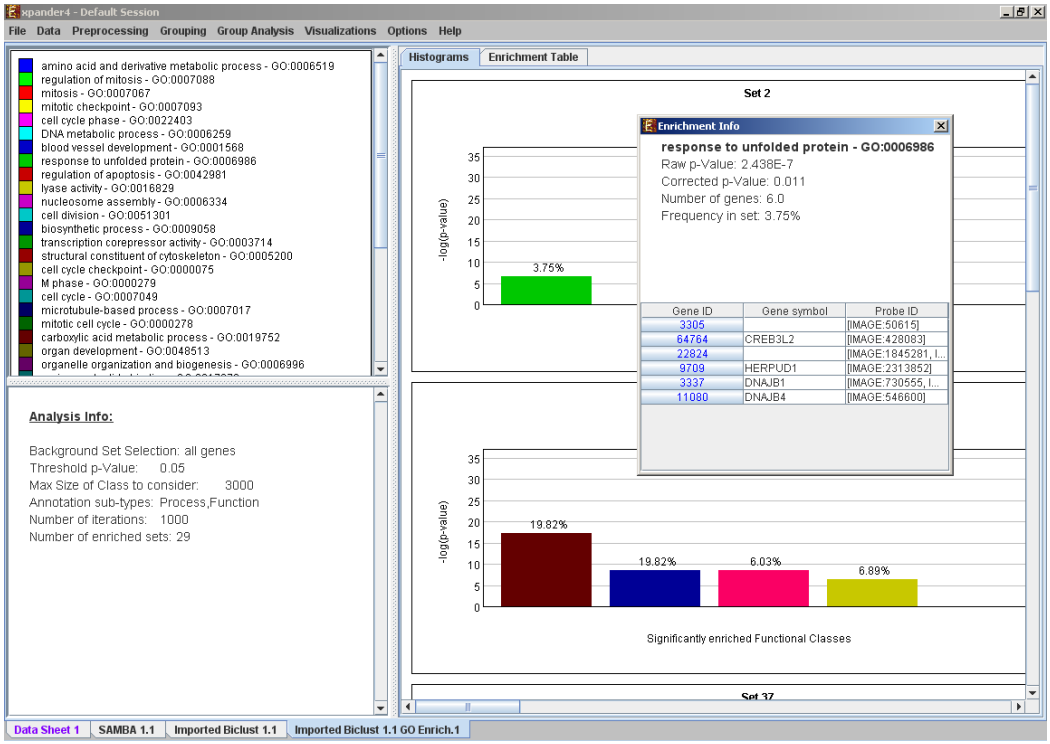
Field	Description
Perform analysis on	The grouping solution on which the analysis will be performed.
Include back nodes	Include genes that are part of the module' but not included in the GE data (Relevant only if the analysis is performed on modules, detected by network based algorithm)
Focus on	Can be used to select annotation subtypes that are of interest (Process, Function and Location). And the analysis will focus on these types only.
Ignore classes over the size of	This parameter states the level in the GO tree at which annotations are too general (class size

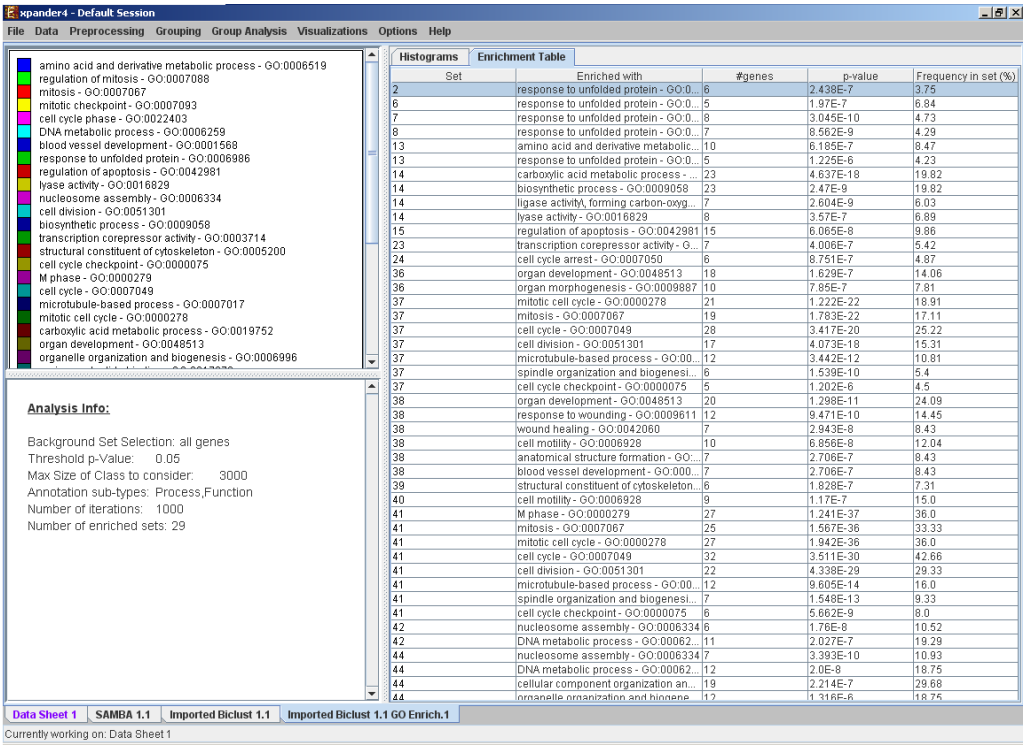
	indicates how general it is) and are thus no longer interesting.
Number of iterations in algorithm	The number of random sampling performed by the algorithm. Increasing this parameter, will increase runtime and will provide higher resolution on corrected p-Values. I.e., corrected p-Values will range between 1/<#iterations> and 1.
Background set	Determines the set of genes that will be used as background in the analysis. Options are: all genes (of the relevant organism), original input data, filtered data or background set from file (see the Files Format section for details regarding the format of an external background set).
Corrected p-value threshold	A functional class will be considered significantly enriched in a cluster/bicluster if its corrected p-value is lower than this threshold. The value in this field should be at least 1/1000, since the TANGO algorithm performs 1000 bootstraps in order to estimate the corrected p-value.

Upon clicking 'OK' in the dialog box, the TANGO algorithm is operated.

After functional analysis is performed a functional analysis solution visualization
the main window. It contains general information regarding the analysis, a
holding all detected enrichments (set ID, functional class, p-value, etc.) and a r
panel along with a color index (mapping each color to a corresponding functio
multi-histogram panel contains one histogram for each probe/gene set/g
enrichment has been detected. Each histogram contains a column for each si
frequent than would be expected by random) functional class. The definitio
depends on the user's selection of threshold p-value i.e., a functional class
significantly enriched in a cluster/bicluster if its corrected p-value is lower th
threshold p-value.

The height of the column is proportional to the significance of this enrichment
log(raw p-value)). The frequency in set (frequency of genes of a functional c
examined set, in %) is written on top of the column. Upon clicking on a column,
displayed containing the class name, raw p-value, corrected p-value, and a list
the cluster/bi-cluster that belong to the class. Upon clicking on one of the gene I
a relevant web page with information regarding this gene is displayed. The
shows the cluster number, size and homogeneity.


xpander4 - Default Session
File Data Preprocessing Grouping Group Analysis Visualizations Options Help
amino acid and derivative metabolic process - GO:0006519
regulation of mitosis - GO:0007088
mitosis - GO:0007067
mitotic checkpoint - GO:0007093
cell cycle phase - GO:0022403
DNA metabolic process - GO:0006259
blood vessel development - GO:0001568
response to unfolded protein - GO:0006986
regulation of apoptosis - GO:0042981
lyase activity - GO:0016829
nucleosome assembly - GO:0006334
cell division - GO:0051301
biosynthetic process - GO:0009058
transcription corepressor activity - GO:0003714
structural constituent of cytoskeleton - GO:0005200
cell cycle checkpoint - GO:0000075
M phase - GO:0000279
cell cycle - GO:0007049
microtubule-based process - GO:0007017
mitotic cell cycle - GO:0000278
carboxylic acid metabolic process - GO:0019752
organ development - GO:0048513
organelle organization and biogenesis - GO:0006996
Analysis Info:
Background Set Selection: all genes
Threshold p-Value: 0.05
Max Size of Class to consider: 3000
Annotation sub-types: Process,Function
Number of iterations: 1000
Number of enriched sets: 29
Histograms
Enrichment Table
Set 2
Enrichment Info
response to unfolded protein - GO:0006986
Raw p-Value: 2.438E-7
Corrected p-Value: 0.011
Number of genes: 6.0
Frequency in set: 3.75%
Gene ID Gene symbol Probe ID
3305 [IMAGE:50615]
64764 CREB3L2 [IMAGE:428083]
22824 [IMAGE:1845281, I...
9709 HERPUD1 [IMAGE:2313852]
3337 DNAJB1 [IMAGE:730555, I...
11080 DNAJB4 [IMAGE:546600]
-log(p-value)
3.75%
19.82%
19.82%
6.03%
6.89%
Significantly enriched Functional Classes
Set 37
Data Sheet 1
SAMBA 1.1
Imported Biclust 1.1
Imported Biclust 1.1 GO Enrich.1


xpander4 - Default Session
File Data Preprocessing Grouping Group Analysis Visualizations Options Help
amino acid and derivative metabolic process - GO:0006519
regulation of mitosis - GO:0007088
mitosis - GO:0007067
mitotic checkpoint - GO:0007093
cell cycle phase - GO:0022403
DNA metabolic process - GO:0006259
blood vessel development - GO:0001568
response to unfolded protein - GO:0006986
regulation of apoptosis - GO:0042981
lyase activity - GO:0016829
nucleosome assembly - GO:0006334
cell division - GO:0051301
biosynthetic process - GO:0009058
transcription corepressor activity - GO:0003714
structural constituent of cytoskeleton - GO:0005200
cell cycle checkpoint - GO:0000075
M phase - GO:0000279
cell cycle - GO:0007049
microtubule-based process - GO:0007017
mitotic cell cycle - GO:0000278
carboxylic acid metabolic process - GO:0019752
organ development - GO:0048513
organelle organization and biogenesis - GO:0006996
Analysis Info:
Background Set Selection: all genes
Threshold p-Value: 0.05
Max Size of Class to consider: 3000
Annotation sub-types: Process,Function
Number of iterations: 1000
Number of enriched sets: 29
Histograms
Enrichment Table
Set | Enriched with | #genes | p-value | Frequency in set (%)
2 | response to unfolded protein - GO:0... | 6 | 2.438E-7 | 3.75
6 | response to unfolded protein - GO:0... | 5 | 1.97E-7 | 6.84
7 | response to unfolded protein - GO:0... | 8 | 3.045E-10 | 4.73
8 | response to unfolded protein - GO:0... | 7 | 8.562E-9 | 4.29
13 | amino acid and derivative metabolic... | 10 | 6.185E-7 | 8.47
13 | response to unfolded protein - GO:0... | 5 | 1.225E-6 | 4.23
14 | carboxylic acid metabolic process - ... | 23 | 4.637E-18 | 19.82
14 | biosynthetic process - GO:0009058 | 23 | 2.47E-9 | 19.82
14 | ligase activity\, forming carbon-oxyg... | 7 | 2.604E-9 | 6.03
14 | lyase activity - GO:0016829 | 8 | 3.57E-7 | 6.89
15 | regulation of apoptosis - GO:0042981 | 15 | 6.065E-8 | 9.86
23 | transcription corepressor activity - G... | 7 | 4.006E-7 | 5.42
24 | cell cycle arrest - GO:0007050 | 6 | 8.751E-7 | 4.87
36 | organ development - GO:0048513 | 18 | 1.629E-7 | 14.06
36 | organ morphogenesis - GO:0009887 | 10 | 7.85E-7 | 7.81
37 | mitotic cell cycle - GO:0000278 | 21 | 1.222E-22 | 18.91
37 | mitosis - GO:0007067 | 19 | 1.783E-22 | 17.11
37 | cell cycle - GO:0007049 | 28 | 3.417E-20 | 25.22
37 | cell division - GO:0051301 | 17 | 4.073E-18 | 15.31
37 | microtubule-based process - GO:00... | 12 | 3.442E-12 | 10.81
37 | spindle organization and biogenesi... | 6 | 1.539E-10 | 5.4
37 | cell cycle checkpoint - GO:0000075 | 5 | 1.202E-6 | 4.5
38 | organ development - GO:0048513 | 20 | 1.298E-11 | 24.09
38 | response to wounding - GO:0009611 | 12 | 9.471E-10 | 14.45
38 | wound healing - GO:0042060 | 7 | 2.943E-8 | 8.43
38 | cell motility - GO:0006928 | 10 | 6.856E-8 | 12.04
38 | anatomical structure formation - GO:... | 7 | 2.706E-7 | 8.43
38 | blood vessel development - GO:000... | 7 | 2.706E-7 | 8.43
39 | structural constituent of cytoskeleton... | 6 | 1.828E-7 | 7.31
40 | cell motility - GO:0006928 | 9 | 1.17E-7 | 15.0
41 | M phase - GO:0000279 | 27 | 1.241E-37 | 36.0
41 | mitosis - GO:0007067 | 25 | 1.567E-36 | 33.33
41 | mitotic cell cycle - GO:0000278 | 27 | 1.942E-36 | 36.0
41 | cell cycle - GO:0007049 | 32 | 3.511E-30 | 42.66
41 | cell division - GO:0051301 | 22 | 4.338E-29 | 29.33
41 | microtubule-based process - GO:00... | 12 | 9.605E-14 | 16.0
41 | spindle organization and biogenesi... | 7 | 1.548E-13 | 9.33
41 | cell cycle checkpoint - GO:0000075 | 6 | 5.662E-9 | 8.0
42 | nucleosome assembly - GO:0006334 | 6 | 1.76E-8 | 10.52
42 | DNA metabolic process - GO:00062... | 11 | 2.027E-7 | 19.29
44 | nucleosome assembly - GO:0006334 | 7 | 3.393E-10 | 10.93
44 | DNA metabolic process - GO:00062... | 12 | 2.0E-8 | 18.75
44 | cellular component organization an... | 19 | 2.214E-7 | 29.68
Data Sheet 1
SAMBA 1.1
Imported Biclust 1.1
Imported Biclust 1.1 GO Enrich.1
Currently working on: Data Sheet 1

Annotation files are currently supplied with EXPANDER for yeast, human, r zebrafish, c-elegans, Arabidopsis, chicken and E. coli, and are updated on a re more information, refer to the Supplied Files section).

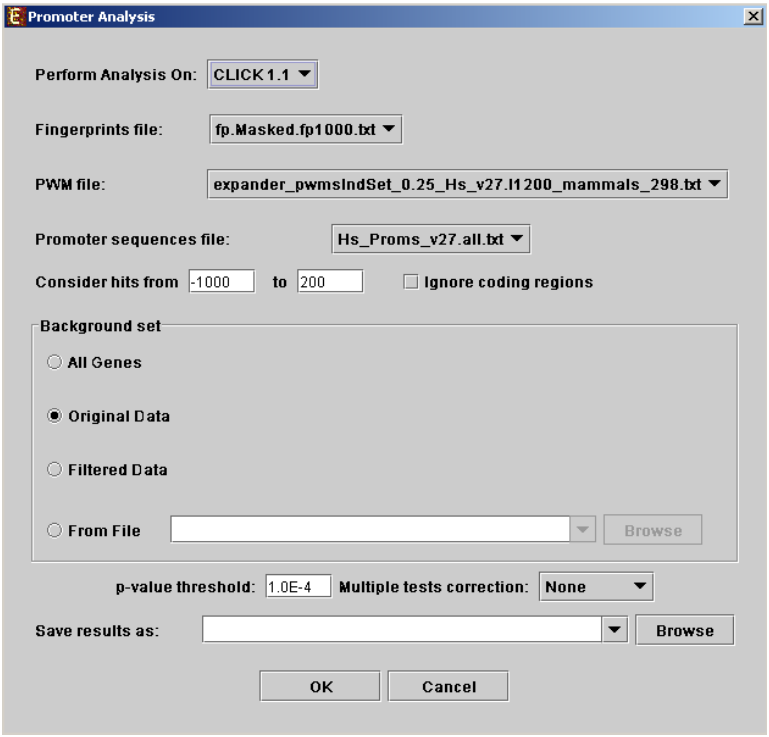
The results of this analysis can be exported to a text file by selecting *File>>Expo* the corresponding view is the selected tab OR by using the *File>>Save All* optio export all solutions within a session to text and image files.

Promoter Analysis

PRIMA

This tool identifies TFs whose binding sites are significantly over-represented ir promoters (i.e. cluster or bicluster). To perform this analysis Expander utiliz (PRomoter Integration in Microarray Analysis) software which performs a statisti the distribution of transcription factor motifs in the promoters of genes within e bicluster. To achieve this, PRIMA uses preprocessed TF fingerprint files, downloaded from the EXPANDER download-page (see the Supplied Files se updated on a regular basis. For details regarding the PRIMA software see t section.

Before operating promoter analysis, the TF fingerprint file for the relevant orgar downloaded from the download page (more details at introduction of Enrich Tools). To perform the analysis, select *Enrichment Analysis >> Promoter Analy* The following dialog box will appear:



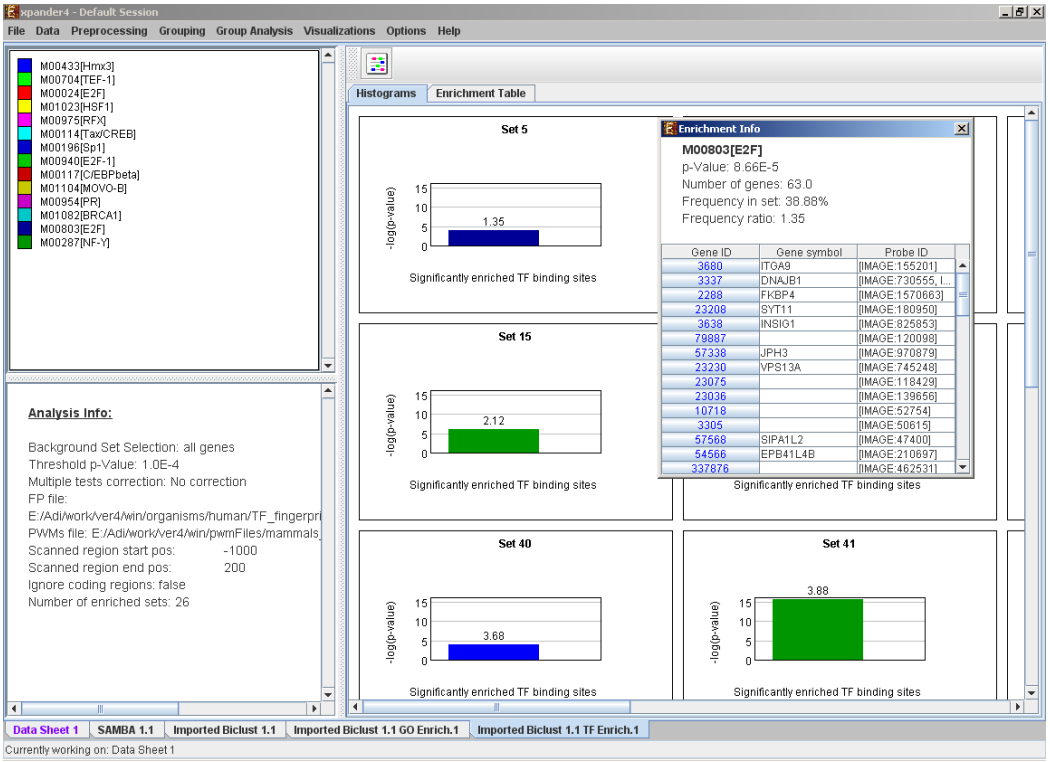
The following table specifies the different parameters that can be set via this dial

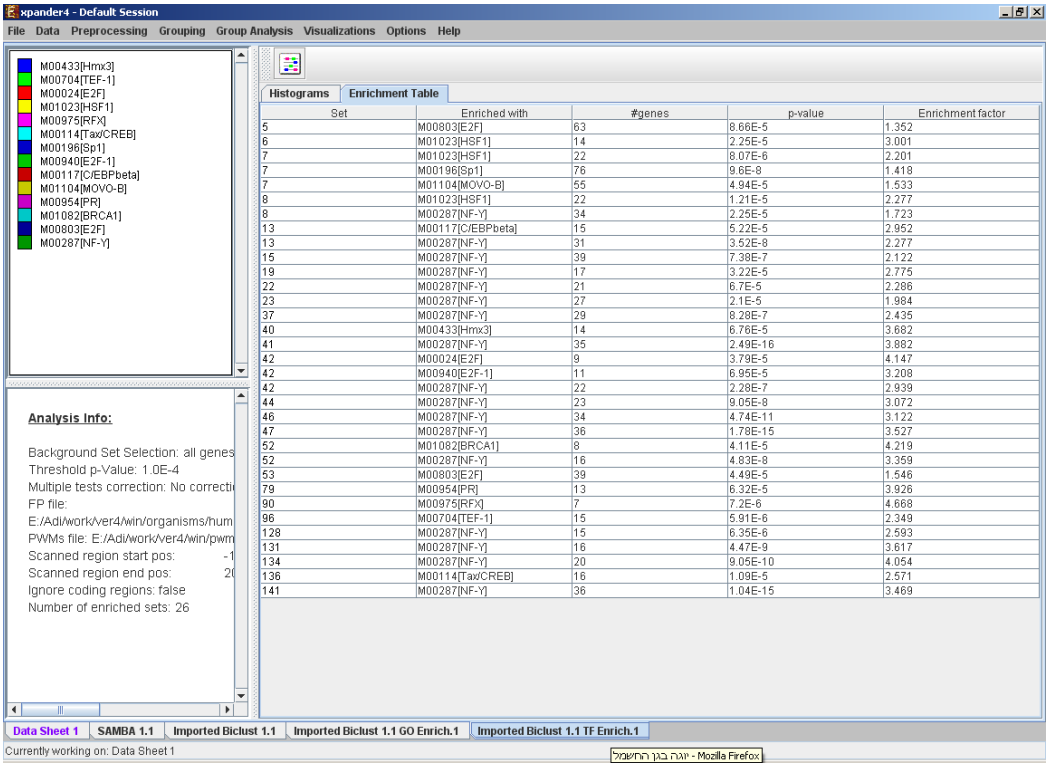
Field	Description
Perform analysis on	The grouping solution on which the analysis will be performed.
Include back nodes	Include genes that are part of the module' but not included in the GE data (Relevant only if the analysis is performed on modules, detected by network based algorithm)
Fingerprints file	Automatically set according to the selection of the organism.
PWM file	Automatically set according to the selection of the organism.
Promoter sequences file	Contains the gene sequences that are used for the TF binding sites display. Automatically set according to the selection of the organism.
Hits range	Determines which regions of the gene are to be analyzed. The possible range depends on the investigated organism (i.e. on the information provided in the TF fingerprint files), and is specified in the Supplied Files section.
Background set	Determines the set of genes that will be used as background in the analysis. Options are: all genes (of the relevant organism), original input data, filtered data or background set from file (see the Files Format section for details regarding the format of an external background set).
Threshold p-value	A TF's binding site will be considered significantly enriched in a cluster if its corrected p-value is lower than this threshold.
Multiple tests correction	Can be set to Bonferroni or None (when set to Bonferroni the corrected p-values are the ones that are compared to the threshold p-value).
Save results as	When filled, the program results are saved in stated txt file.

After promoter analysis is performed, a promoter analysis solution visualization
the main window. It contains general information regarding the analysis, a
holding all detected enrichments (set ID, TF binding site, p-value, etc.) and a r
panel along with a color index (mapping each color to a corresponding TF bin
multi-histogram panel contains one histogram for each probe/gene set/gr
enrichment has been detected. Each histogram contains a column for each sig
frequent than would be expected by random) TF binding site. The definition
depends on the user's selection of threshold p-value. i.e., a TF binding site
significantly enriched in a cluster/bicluster if its corrected p-value is lower th
threshold p-value.

The height of a column is proportional to the significance of this enrichment log(p-value)), and the frequency ratio (frequency in set divided by frequency in written on top of the column. Upon clicking on a column, a dialog box is displaye

TF accession number in TRANSFAC DB [TF name], p-value, % of covered cluster, relative frequency (frequency in cluster divided by frequency in backgro list of the genes in the cluster which contain the motif in their promoters. Upon of the gene Ids in the table, a relevant web page with information regardin displayed. The display tool tip shows the cluster number, size and homogeneity.

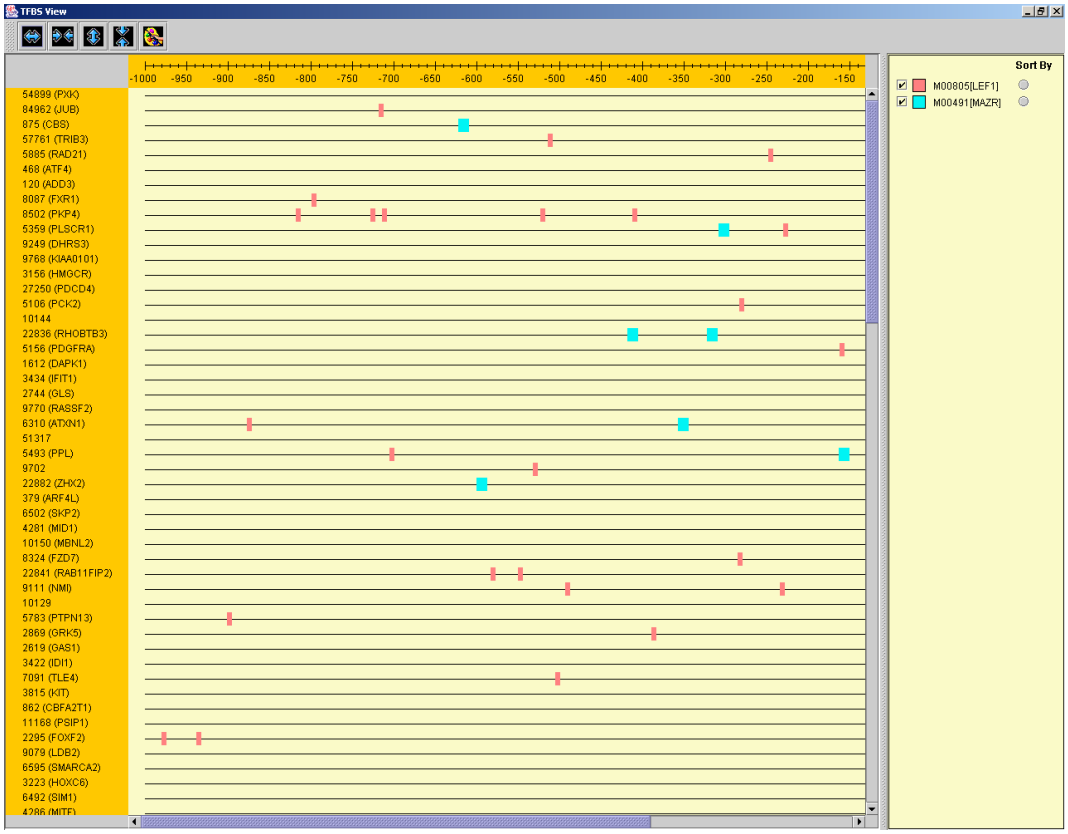




Set	Enriched with	#genes	p-value	Enrichment factor
5	M00803[E2F]	63	8.66E-5	1.352
6	M01023[HSF1]	14	2.25E-5	3.001
7	M01023[HSF1]	22	8.07E-6	2.201
7	M00196[Sp1]	76	9.6E-8	1.418
7	M01104[MOVO-B]	55	4.94E-5	1.533
8	M01023[HSF1]	22	1.21E-5	2.277
8	M00287[NF-Y]	34	2.25E-5	1.723
13	M00117[C/EBPbeta]	15	5.22E-5	2.952
13	M00287[NF-Y]	31	3.52E-8	2.277
15	M00287[NF-Y]	39	7.38E-7	2.122
19	M00287[NF-Y]	17	3.22E-5	2.775
22	M00287[NF-Y]	21	6.7E-5	2.286
23	M00287[NF-Y]	27	2.1E-5	1.984
37	M00287[NF-Y]	29	8.28E-7	2.435
40	M00433[Hmx3]	14	6.76E-5	3.682
41	M00287[NF-Y]	35	2.49E-16	3.882
42	M00024[E2F]	9	3.79E-5	4.147
42	M00940[E2F-1]	11	6.95E-5	3.208
42	M00287[NF-Y]	22	2.28E-7	2.939
44	M00287[NF-Y]	23	9.05E-8	3.072
46	M00287[NF-Y]	34	4.74E-11	3.122
47	M00287[NF-Y]	36	1.78E-15	3.527
52	M01082[BRCA1]	8	4.11E-5	4.219
52	M00287[NF-Y]	16	4.83E-8	3.359
53	M00803[E2F]	39	4.49E-5	1.546
79	M00954[PR]	13	6.32E-5	3.926
90	M00975[RFX]	7	7.2E-6	4.668
96	M00704[TEF-1]	15	5.91E-6	2.349
128	M00287[NF-Y]	15	6.35E-6	2.593
131	M00287[NF-Y]	16	4.47E-9	3.617
134	M00287[NF-Y]	20	9.05E-10	4.054
136	M00114[Tax/CREB]	16	1.09E-5	2.571
141	M00287[NF-Y]	36	1.04E-15	3.469

After performing promoter analysis, TF binding sites can be viewed by selecti

Analysis >> Promoter Analysis >> View Binding Sites OR by pressing the toolba
After selecting the gene group (cluster/bi-cluster etc.) to be viewed, a sep
displayed, containing a line to represent each of the genes in the group,
rectangle, to represent each binding site. A color index appears on the right,
color to the corresponding TF (PWM). A check box next to each of the entri
index allows hiding any of the PWMs, and a radio button next to each of the entı
index allows sorting the genes in the display according to the number
corresponding TF. The toolbar contains tools for vertical and horizontal zooming
file had been selected via the promoter analysis input dialog, the actual se
displayed when the zoom factor (scale) allows it.

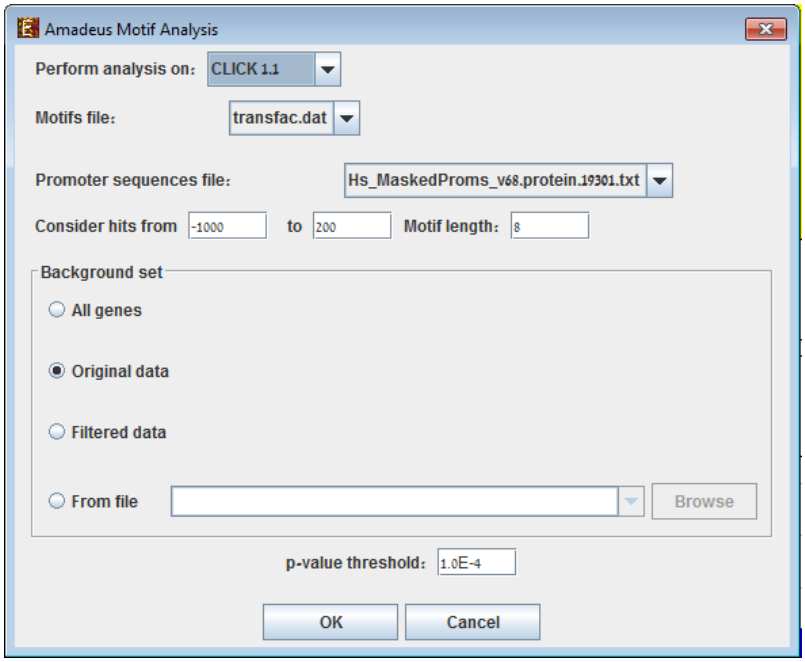


AMADEUS

Another option for performing promoter analysis, is finding enriched *de nov* AMADEUS integrated software in Expander. *Amadeus* is a software for discovery. It seeks for motifs which are enriched in a target set of genes co background set. These sets can be extended to any kind of sequences (e.g., Ch

In order to perform motifs enrichments analysis, select *Enrichment Analysis Analysis >> AMADEUS*.

The following dialog box will appear:



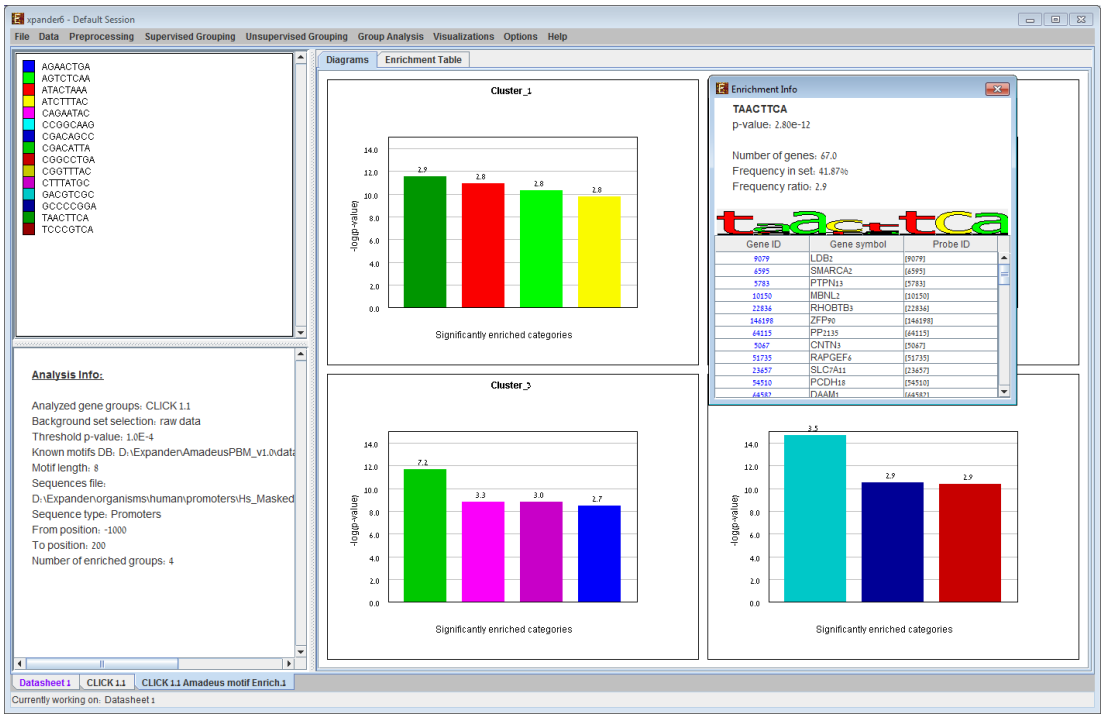
The following table specifies the different parameters that can be set via this dial

Field	Description
Perform analysis on	The grouping solution on which the analysis will be performed.
Motifs file	A matrix table of known motifs in TRANSFAC format. The automatic default is transfac.dat, a public release of TRANSFAC.
Promoter sequences file	Contains the promoter sequences in fasta format. Automatically set according to the selection of the organism. Can be set to any set of sequences (e.g., ChIP-seq peaks).
Motif Length	The length of the motif to be searched for.
Hits range	Determines which sections of the sequences are analyzed. The range depends on the organism (i.e. the average length of a promoter sequence, on the information provided in the TF fingerprint files), and is specified in the Supplied Files section. It can be set manually.
Background set	Determines the set of genes, whose promoter sequences will be used as background in the analysis. Options are: all genes (of the relevant organism), original input data, filtered data or background set from file (see the Files Format section for details regarding the format of an external background set).

P-value threshold	A motif will be considered significantly enriched in a cluster if its corrected p-value is lower than this threshold.

After AMDEUS analysis is performed, an Amadeus motif solution visualization t
the main window. It contains general information regarding the analysis, a
holding all detected enrichments (set ID, Motif binding site, p-value, etc.) and a r
panel along with a color index (mapping each color to a corresponding Motif bir
multi-histogram panel contains one histogram for each probe/gene set/gr
enrichment has been detected. Each histogram contains a column for each sig
frequent than would be expected in the background set) motif binding site. Th
depends on the user's selection of p-value threshold. i.e., a motif binding site
significantly enriched in a cluster/bicluster if its corrected p-value is lower than
value threshold.

The height of a column is proportional to the significance of this enrichment
log(p-value)), and the frequency ratio (frequency in the target set divided by fr
background set) is written on top of the column. Upon clicking on a column, a
displayed containing:



motif binding site, p-value, % of promoters containing the motif in the cluster, rela
(frequency in the cluster divided by frequency in background set), motif logo
position weight matrix and a list of the genes in the cluster which contain the
promoters. Upon clicking on one of the gene ids in the table, a relevant v
information regarding this gene is displayed. The display tool tip shows the c
size and homogeneity.

TF motif fingerprint files and promoter sequence files are currently supplied wi for yeast, human, mouse, rat, fly, zebrafish, c-elegans, arabidopsis and ch updated on a regular basis (for more information, refer to the Supplied Files sect

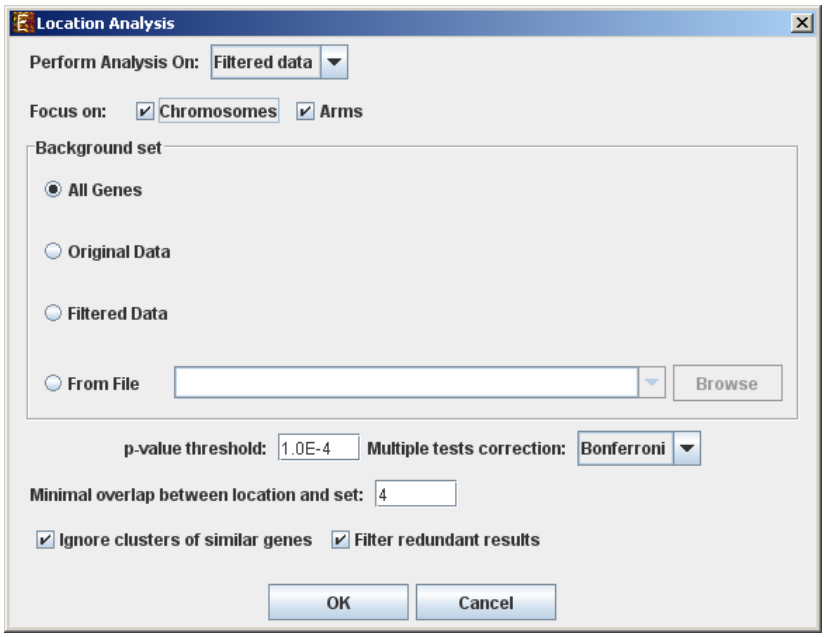
The results of this analysis can be exported to a text by selecting *File>>Export to* corresponding view is the selected tab.

Location Enrichment Analysis

This tool performs basic statistical analysis on the distribution of chromosom genes within each group. The locations of the genes are specified in organisr files, which can be downloaded from the EXPANDER download-page (see the section).

Before operating location analysis, the location data for the relevant organi downloaded from the download page (more details at introduction of Enrich Tools). In this analysis, hyper-geometric enrichment tests are performed, and be (if requested) corrected for multiple testing using the Bonferroni correction.

To perform the analysis, select *Enrichment Analysis >> Location Analys Enrichment*. The following dialog box will appear:



The following table specifies the different parameters that can be set via th

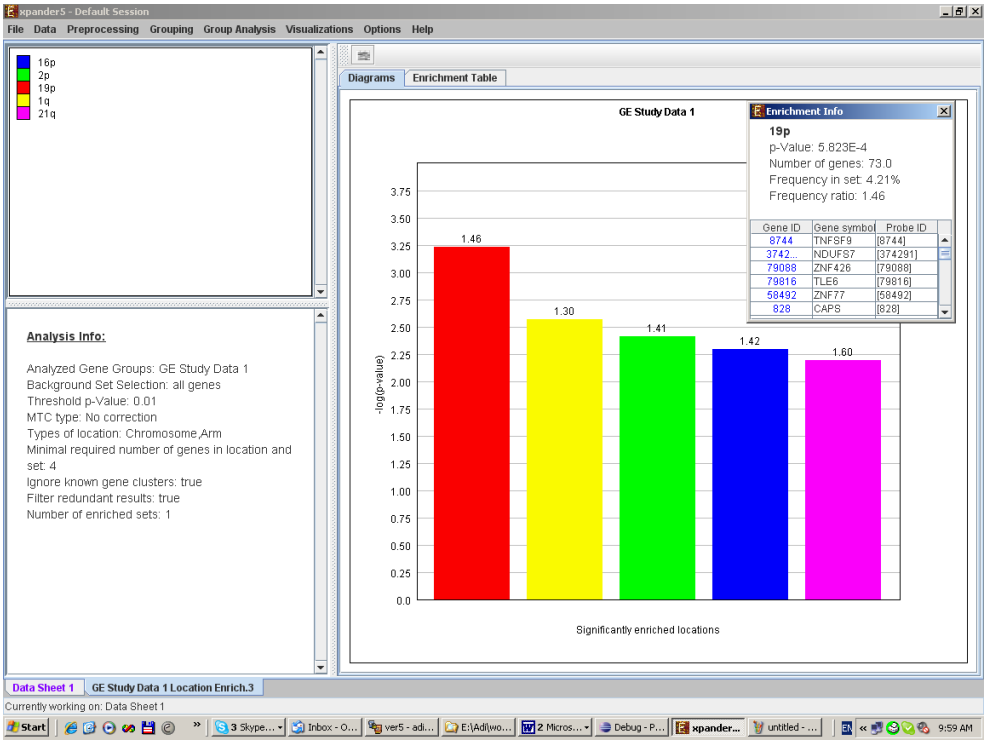
Field	Description
Perform analysis on	The grouping solution on which the analysis will be performed.
Include back nodes	Include genes that are part of the module' but not included in the GE data (Relevant only if the analysis is performed on modules, detected by

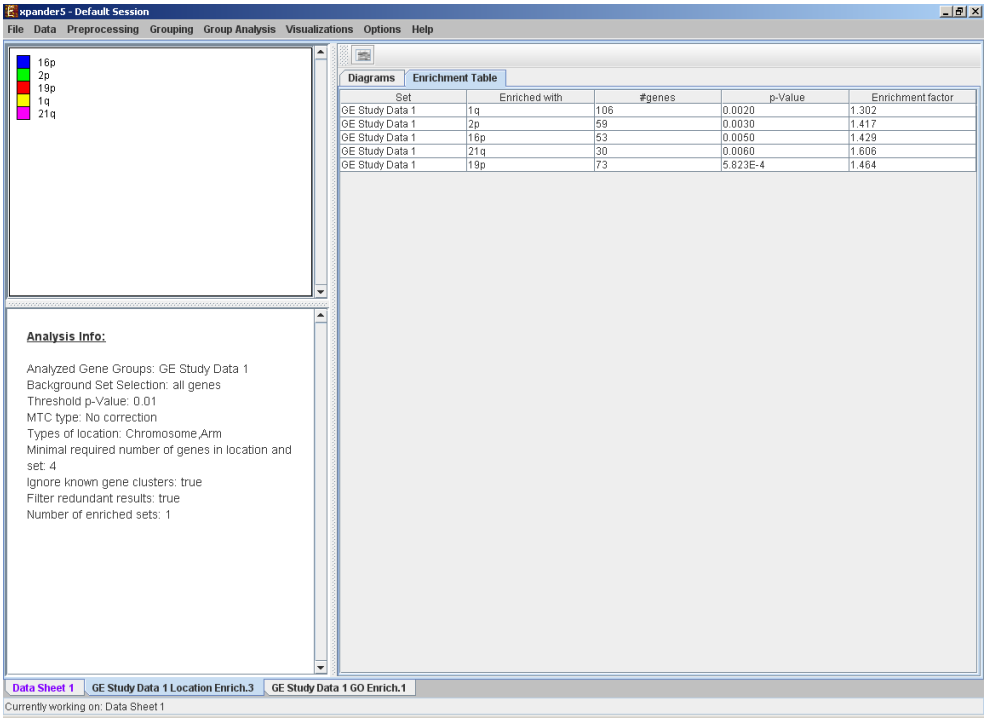
	network based algorithm)
Focus on (Chromosomes, Arms*, Bands*)	Location types to perform analysis on.
Background set	Determines the set of genes that will be used as background in the analysis. Options are: all genes (of the relevant organism), original input data, filtered data or background set from file (see the Files Format section for details regarding the format of an external background set).
p-value threshold	A category/attribute will be considered significantly enriched in a cluster/bicluster if its corrected p-value is lower than this threshold.
Multiple tests correction	Can be set to Bonferroni or None (when set to Bonferroni the corrected p-values are the ones that are compared to the threshold p-value).
Minimal overlap between category and set	The minimal number of genes from a group (cluster/bi-cluster/module etc.) expected to be categorized/attributed by an attribute in order for its enrichment to be accepted.
Ignore clusters of similar genes*	If selected, genes from known homology clusters are not included in the analysis.
Filter redundant results	If selected, the results are filtered, so that out of two enrichments of overlapping areas in the same group, only one is selected (the most significant one).

* If relevant data exists

After the analysis is performed an enrichment analysis solution visualization tab main window. It contains general information regarding the analysis, a sort-abl all detected enrichments (set ID, enrichment category, p-value, etc.) and a r panel along with a color index (mapping each color to a corresponding locati histogram panel contains one histogram for each probe/gene group in which e been detected. Each histogram contains a column for each significant (more would be expected by random) location. The definition of significant depends selection of threshold p-value i.e., a category is considered significantly cluster/bicluster if its corrected p-value is lower than the preset threshold p-value

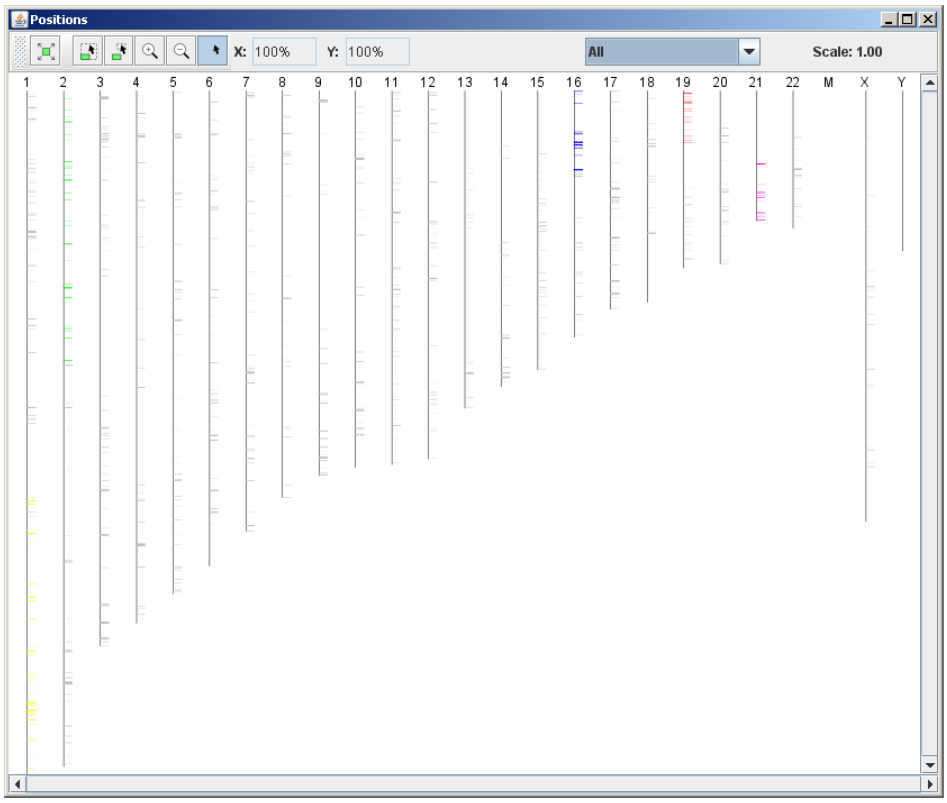
The height of the column is proportional to the significance of this enrichment log(raw p-value)), and the frequency ratio (frequency in set divided by background) is written on top of the column. Upon clicking on a column, a displayed containing the location, corrected p-value, and a list of the genes in are mapped to this location. Upon clicking on one of the gene Ids in the table, page with information regarding this gene is displayed.





After performing location enrichment analysis, the locations can be viewe
Enrichment Analysis >> *Location Analysis* >> *View Locations* OR by pressi
button ( ). After selecting the gene group (cluster/bi-cluster etc.) to be view

frame is displayed, containing an image of all chromosomes on which the p genes in the group are marked. If the gene is located on an area that was i enriched in that group, its position is marked in the same color to this area t results histogram.



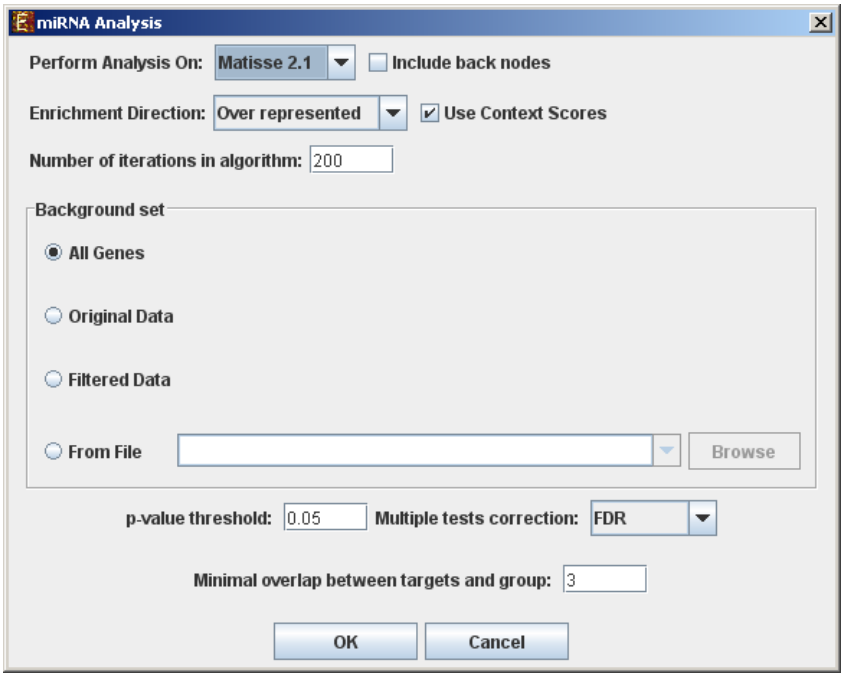
The results of this analysis can be exported to a text file by selecting *File>>Expo* the corresponding view is the selected tab OR by using the *File>>Save All* optio export all solutions within a session to text and image files.

miRNA Targets Enrichment Analysis

This tool performs a statistical analysis on the distribution of miRNA target gene group. The miRNA targets information is supplied in organism-specific data files, downloaded from the EXPANDER download-page (see the Supplied Files sectio analysis, Expander utilizes the FAME algorithm, which performs empirical tests u sampling technique (random permutations) to estimate the empirical p-value dist evaluated groups. This is done while accounting for biases in the 3' UTR sequer

Before operating miRNA enrichment analysis, the location data for the rele should be downloaded from the download page (more details at introduction Analysis Tools). In this analysis, hyper-geometric enrichment tests are perfc results can be (if requested) corrected for multiple testing using the Bonferroni c

To perform the analysis, select *Enrichment Analysis >> miRNA Analysis* > following dialog box will appear:



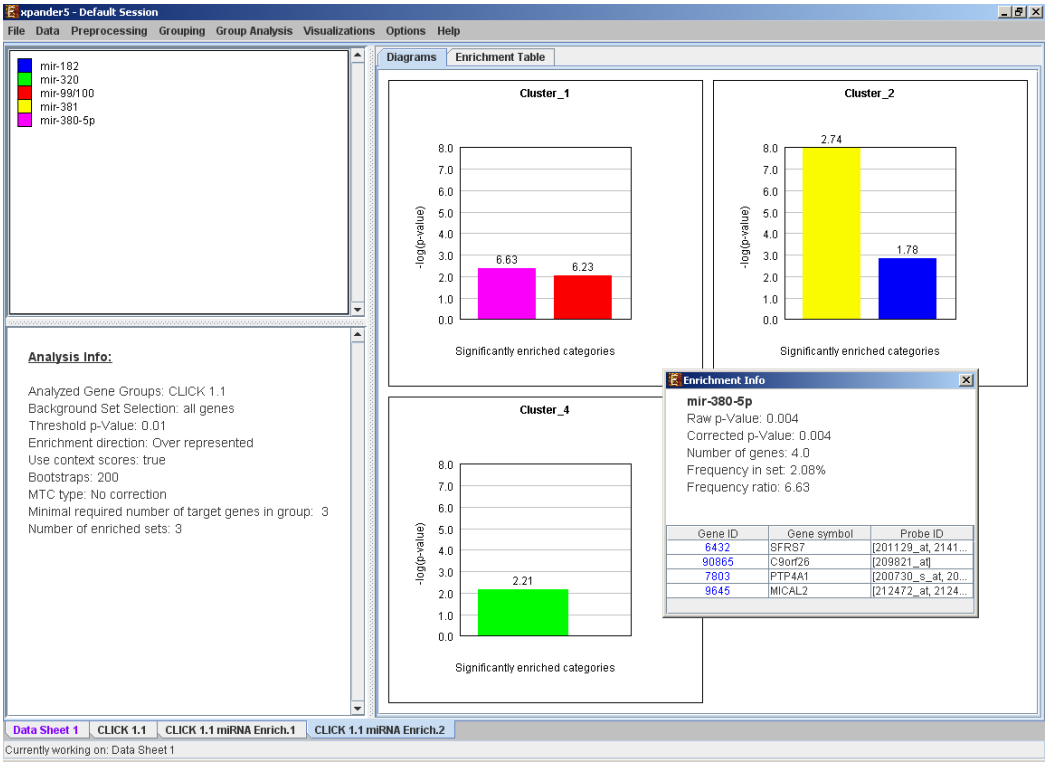
The following table specifies the different parameters that can be set via th

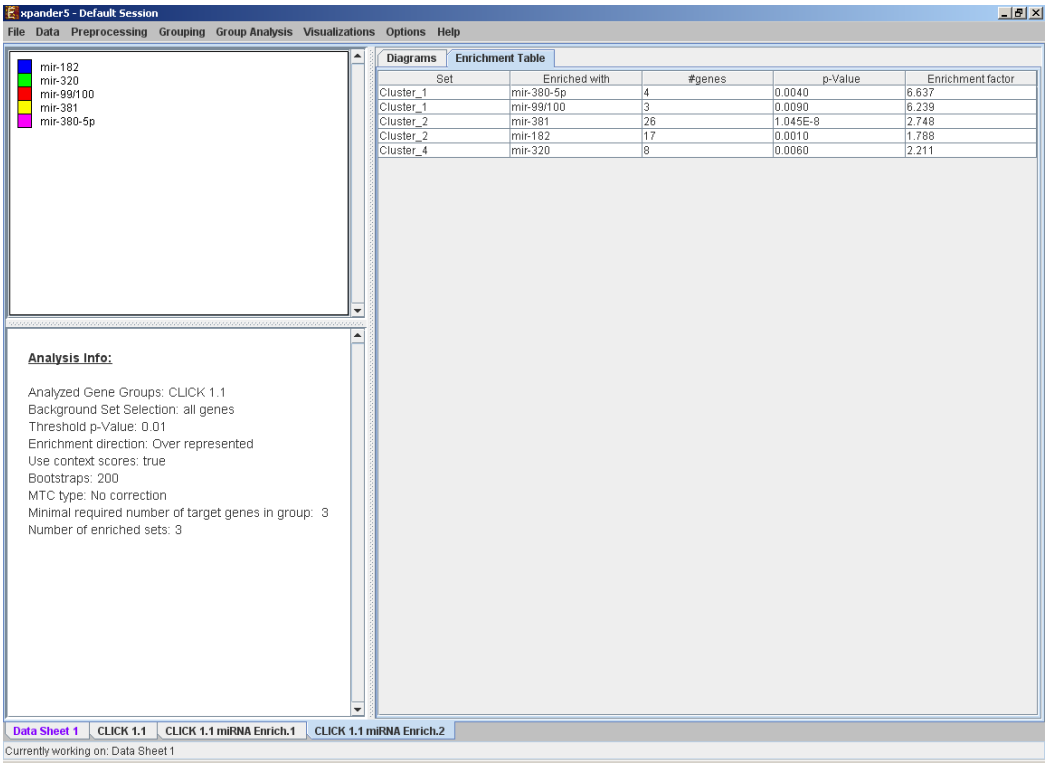
Field	Description
Perform analysis on	The grouping solution on which the analysis will be performed.
Include back nodes	Include genes that are part of the module' but not included in the GE data (Relevant only if the analysis is performed on modules, detected by network based algorithm)
Enrichment Direction	Allows to choose between searching for over-represented targets and searching for under-represented targets.
Use context scores	If context scores are used, FAME will assign a higher weight to miRNA-gene pairs for which at least one target site has a high maximal context score (see References section for further details).
Number of Iterations	The number of random permutations used for the empirical tests.
Background set	Determines the set of genes that will be used as background in the analysis. Options are: all genes (of the relevant organism), original input

	data, filtered data or background set from file (see the Files Format section for details regarding the format of an external background set).
p-value threshold	A category/attribute will be considered significantly enriched in a cluster/bicluster if its corrected p-value is lower than this threshold.
Multiple tests correction	Can be set to FDR, Bonferroni or None (when set to Bonferroni/FDR the corrected p-values are the ones that are compared to the threshold p-value).
Minimal overlap between targets and group	The minimal number of genes from a group (cluster/bi-cluster/module etc.) expected to be categorized/attributed by an attribute in order for its enrichment to be accepted.

After the analysis is performed an enrichment analysis solution visualization tab main window. It contains general information regarding the analysis, a sort-abl all detected enrichments (group name, enriched miRNA target, p-value, etc. histogram panel along with a color index (mapping each color to a correspondin multi-histogram panel contains one histogram for each probe/gene group in wh has been detected. Each histogram contains a column for each significant (mor would be expected by random) miRNA target. The definition of significant d user's selection of threshold p-value i.e., an mRNA target is considered signifi in a group of genes if its corrected p-value is lower than the selected threshold p

The height of the column is proportional to the significance of this enrichment log(raw p-value)), and the frequency ratio (frequency in set divided by background) is written on top of the column. Upon clicking on a column, a displayed containing the miRNA name, corrected p-value, and a list of the gen that are mapped to this location. Upon clicking on one of the gene Ids in the ta web page with information regarding this gene is displayed.


xpander5 - Default Session
File Data Preprocessing Grouping Group Analysis Visualizations Options Help
mir-182
mir-320
mir-99/100
mir-381
mir-380-5p
Diagrams
Enrichment Table
Cluster_1
Cluster_2
Cluster_4
-log(p-value)
Significantly enriched categories
6.63
6.23
2.74
1.78
2.21
Enrichment Info
mir-380-5p
Raw p-Value: 0.004
Corrected p-Value: 0.004
Number of genes: 4.0
Frequency in set: 2.08%
Frequency ratio: 6.63
Gene ID
Gene symbol
Probe ID
6432
SFRS7
[201129_at, 2141...
90865
C9orf26
[209821_at]
7803
PTP4A1
[200730_s_at, 20...
9645
MICAL2
[212472_at, 2124...
Analysis Info:
Analyzed Gene Groups: CLICK 1.1
Background Set Selection: all genes
Threshold p-Value: 0.01
Enrichment direction: Over represented
Use context scores: true
Bootstraps: 200
MTC type: No correction
Minimal required number of target genes in group: 3
Number of enriched sets: 3
Data Sheet 1
CLICK 1.1
CLICK 1.1 miRNA Enrich.1
CLICK 1.1 miRNA Enrich.2
Currently working on: Data Sheet 1

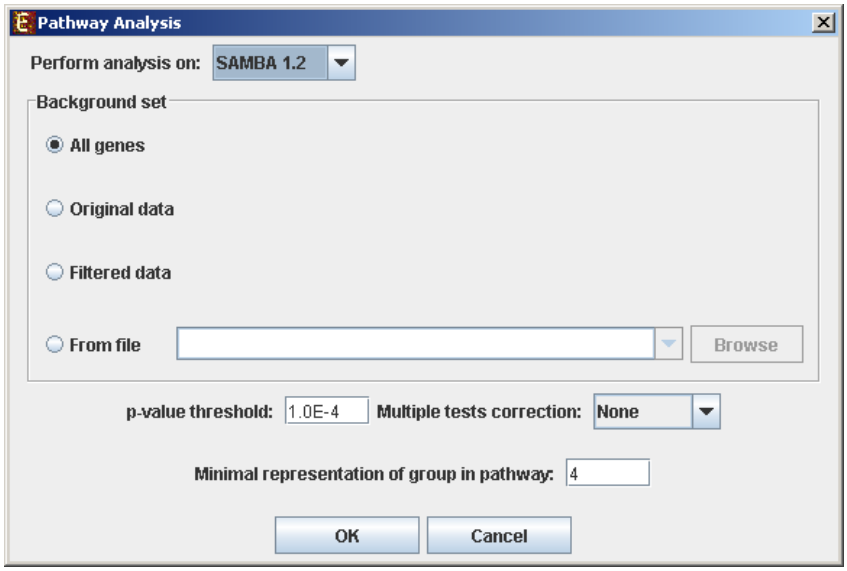

xpander5 - Default Session
File Data Preprocessing Grouping Group Analysis Visualizations Options Help
mir-182
mir-320
mir-99/100
mir-381
mir-380-5p
Diagrams
Enrichment Table
Set
Enriched with
#genes
p-Value
Enrichment factor
Cluster_1
mir-380-5p
4
0.0040
6.637
Cluster_1
mir-99/100
3
0.0090
6.239
Cluster_2
mir-381
26
1.045E-8
2.748
Cluster_2
mir-182
17
0.0010
1.788
Cluster_4
mir-320
8
0.0060
2.211
Analysis Info:
Analyzed Gene Groups: CLICK 1.1
Background Set Selection: all genes
Threshold p-Value: 0.01
Enrichment direction: Over represented
Use context scores: true
Bootstraps: 200
MTC type: No correction
Minimal required number of target genes in group: 3
Number of enriched sets: 3
Data Sheet 1
CLICK 1.1
CLICK 1.1 miRNA Enrich.1
CLICK 1.1 miRNA Enrich.2
Currently working on: Data Sheet 1

The results of this analysis can be exported to a text file by selecting *File>>Expo* the corresponding view is the selected tab. OR by using the *File>>Save All* optio export all solutions within a session to text and image files.

Pathway Enrichment Analysis

This tool performs a statistical analysis on the representation of KEGG and pathway maps within each group. The KEGG and WikiPathways information organism-specific data files, which can be downloaded from the EXPANDER (see the Supplied Files section In this analysis, hyper-geometric enrichm performed, and the results can be (if requested) corrected for multiple tes Bonferroni correction.

To perform the analysis, select *Enrichment Analysis >> Pathway Analysis Enrichment Analysis >> Pathway Analysis >> WikiPathways* . The following appear:



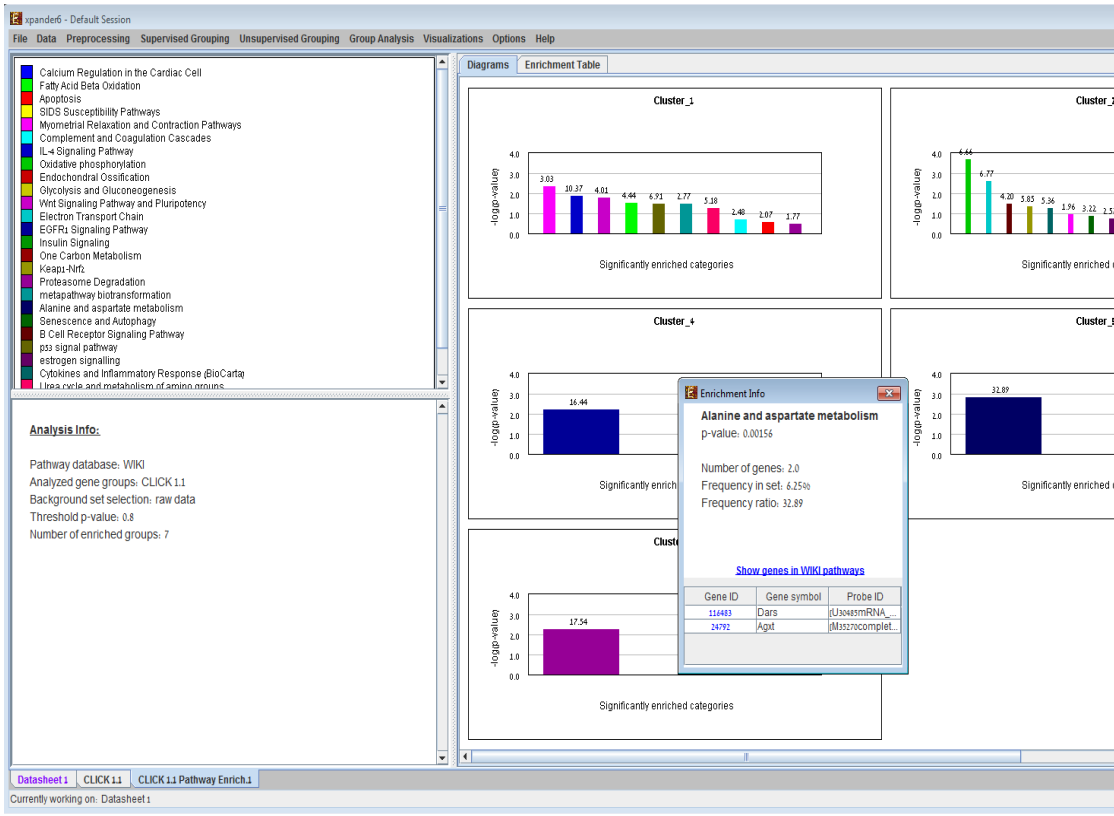
The following table specifies the different parameters that can be set via th

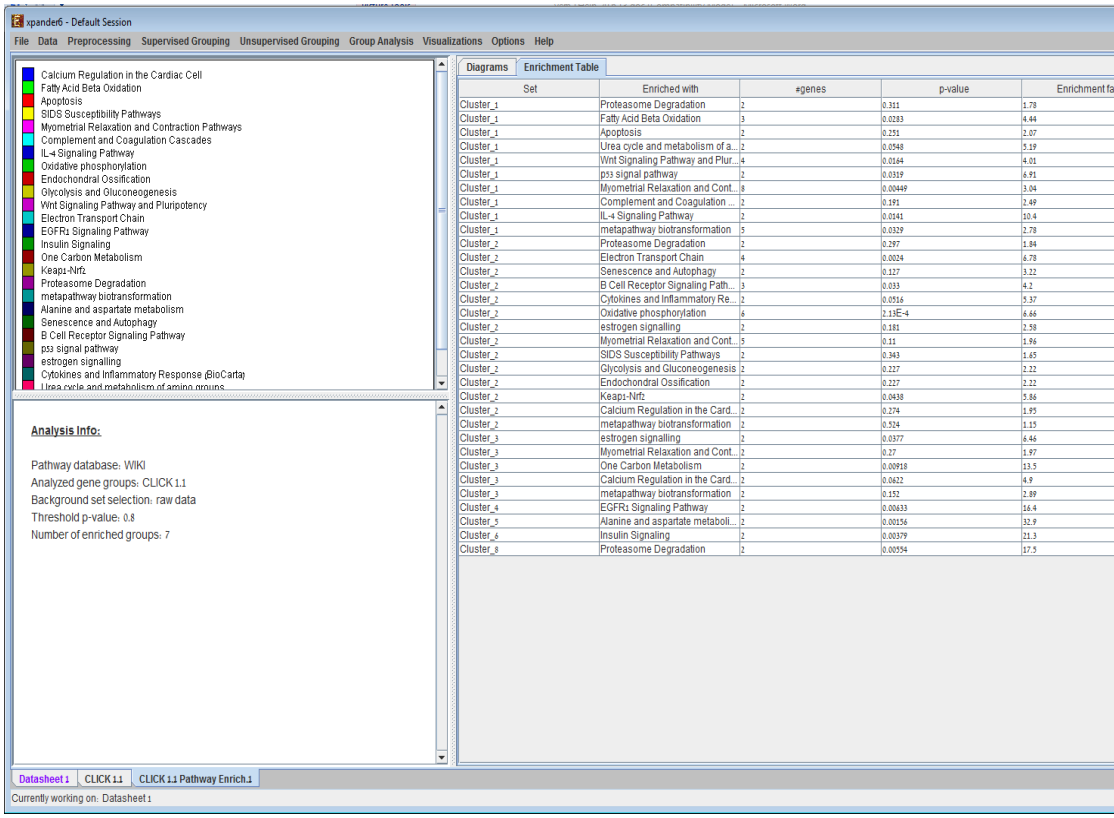
Field	Description
Perform analysis on	The grouping solution on which the analysis will be performed.
Background set	Determines the set of genes that will be used as background in the analysis. Options are: all genes (of the relevant organism), original input data, filtered data or background set from file (see the Files Format section for details regarding the format of an external background set).
p-value threshold	A category/attribute will be considered significantly enriched in the pathway in a

	cluster/bicluster if its corrected p-value is lower than this threshold.
Multiple tests correction	Can be set to Bonferroni or None (when set to Bonferroni the corrected p-values are the ones that are compared to the threshold p-value).
Minimal overlap between category and set	The minimal number of genes from a cluster/bi-cluster expected to be categorized/attributed by an attribute in order for its pathway analysis to be accepted.

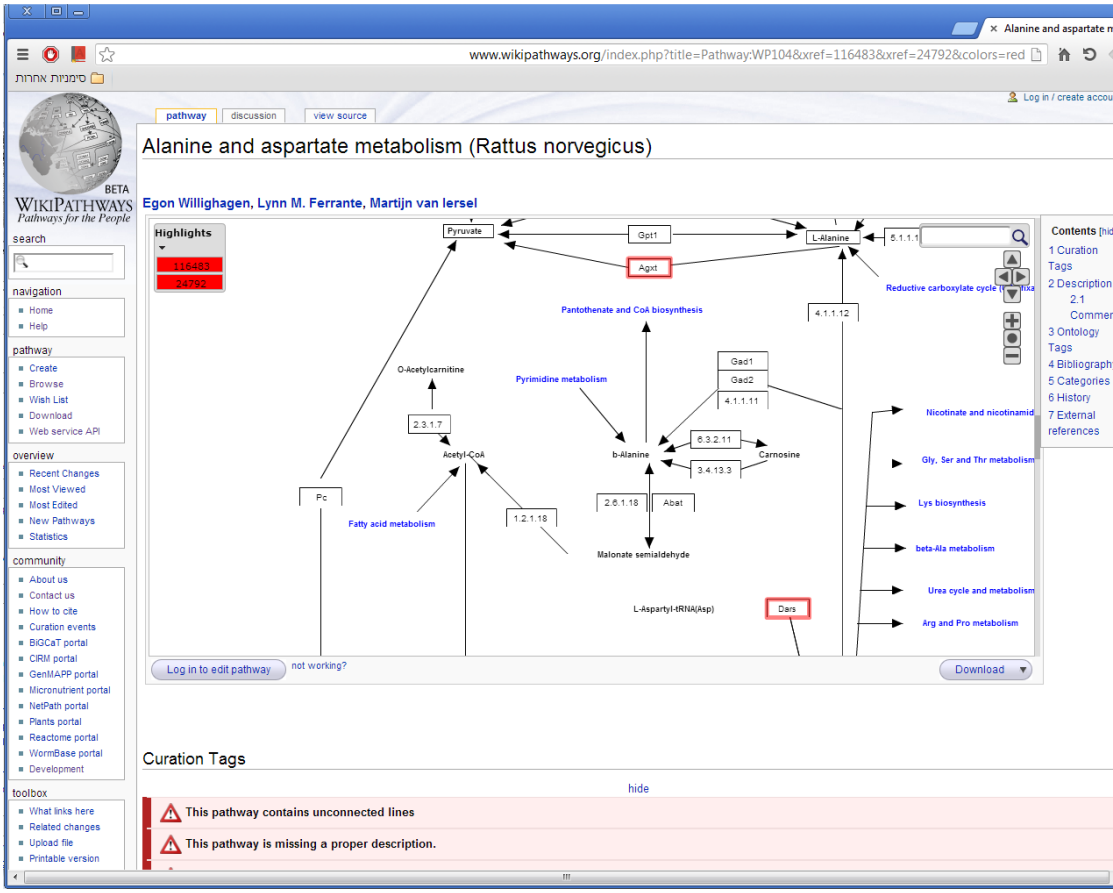
After the analysis was performed a Pathway analysis solution visualization tab is main window. It contains general information about the analysis, a sorted ta all detected pathways (group name, enriched pathway target, p-value, etc histogram panel along with a color index (mapping each color to a correspond The multi-histogram panel contains one histogram for each probe/gene g enrichment has been detected. Each histogram contains a column for each sig frequent than would be expected by random) pathway target. The definition depends on the user's selection of threshold p-value i.e., a pathway target significantly enriched in a group of genes if its corrected p-value is lower tha threshold p-value.

The height of the column is proportional to the significance of this enrichment log(raw p-value)), and the frequency ratio (frequency in set divided by background) is written on top of the column. Upon clicking on a column, a displayed containing the pathway name, corrected p-value, link to the relevant web page, and a list of the genes in the group that are included in the correspon Upon clicking on one of the gene Ids in the table, a relevant web page w regarding this gene is displayed.


xpander6 - Default Session
File Data Preprocessing Supervised Grouping Unsupervised Grouping Group Analysis Visualizations Options Help
Calcium Regulation in the Cardiac Cell
Fatty Acid Beta Oxidation
Apoptosis
SIDS Susceptibility Pathways
Myometrial Relaxation and Contraction Pathways
Complement and Coagulation Cascades
IL-4 Signaling Pathway
Oxidative phosphorylation
Endochondral Ossification
Glycolysis and Gluconeogenesis
Wnt Signaling Pathway and Pluripotency
Electron Transport Chain
EGFR1 Signaling Pathway
Insulin Signaling
One Carbon Metabolism
Keap1-Nrf2
Proteasome Degradation
metapathway biotransformation
Alanine and aspartate metabolism
Senescence and Autophagy
B Cell Receptor Signaling Pathway
p53 signal pathway
estrogen signalling
Cytokines and Inflammatory Response (BioCarta)
Urea cycle and metabolism of amino groups
Analysis Info:
Pathway database: WIKI
Analyzed gene groups: CLICK 1.1
Background set selection: raw data
Threshold p-value: 0.8
Number of enriched groups: 7
Diagrams
Enrichment Table
Cluster_1
-log(p-value)
3.03
10.37
4.01
4.44
6.91
2.77
5.18
2.48
2.07
1.77
Significantly enriched categories
Cluster_4
16.44
Cluster_5
32.89
17.54
Enrichment Info
Alanine and aspartate metabolism
p-value: 0.00156
Number of genes: 2.0
Frequency in set: 6.25%
Frequency ratio: 32.89
Show genes in WIKI pathways
Gene ID
Gene symbol
Probe ID
116483
Dars
(U30483mRNA_...
24792
Agxt
(M35270complet...
Datasheet 1
CLICK 1.1
CLICK 1.1 Pathway Enrich.1
Currently working on: Datasheet 1



Upon clicking on the link to the pathway map web page, the web browser dis
with the relevant genes highlighted in it.

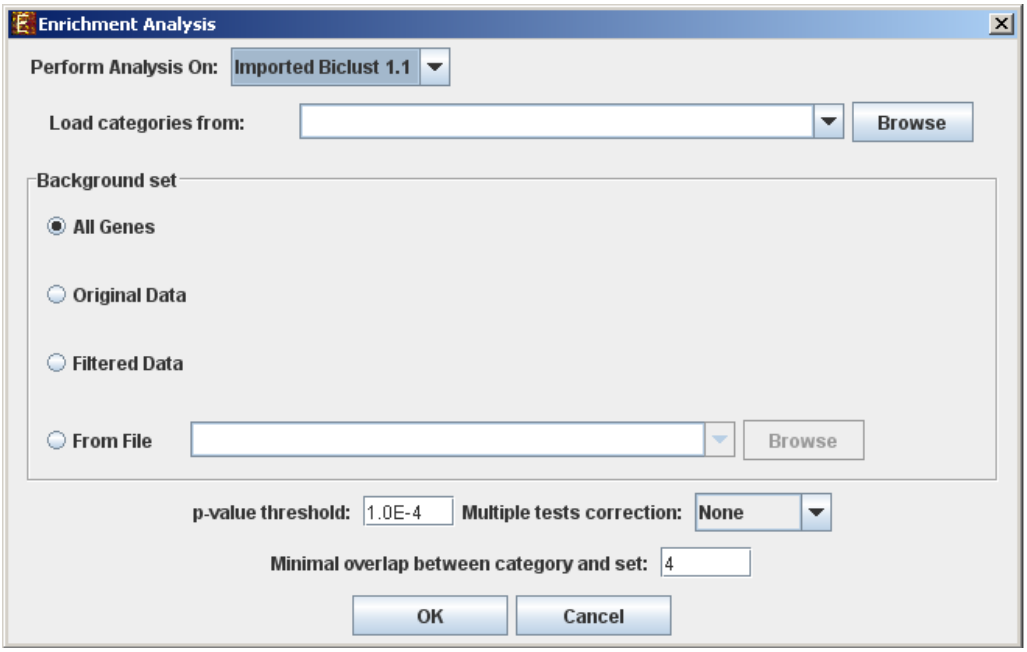


The results of this analysis can be exported to a text file by selecting *File>>Export to* corresponding view is the selected tab. OR by using the *File>>Save All* option, which solutions within a session to text and image files.

General Enrichment Analysis

This tool performs basic statistical analysis on the distribution of categories/attri within each group. The categories/attributes of the genes are to be determined b imported as a text (for details regarding the required format, see the File Form this analysis, hyper-geometric enrichment tests are performed, and the resu requested) corrected for multiple testing using the Bonferroni correction.

To perform the analysis, select *Enrichment Analysis >> General Enrichmer Detect Enrichment*. The following dialog box will appear:



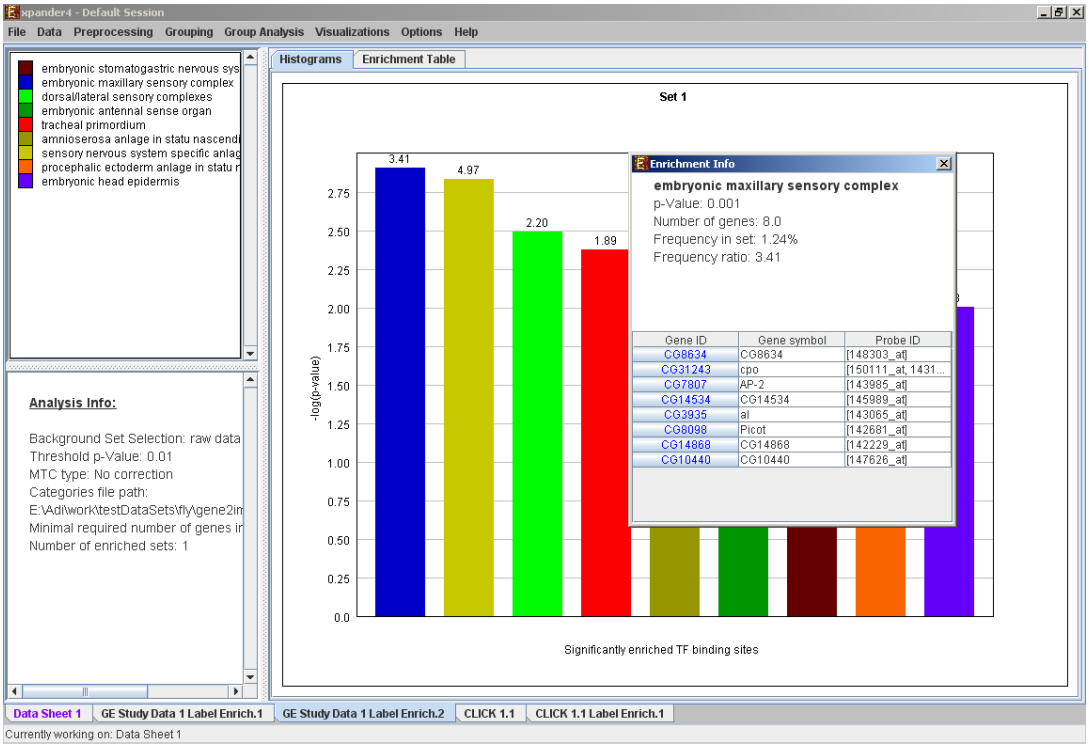
The following table specifies the different parameters that can be set via th

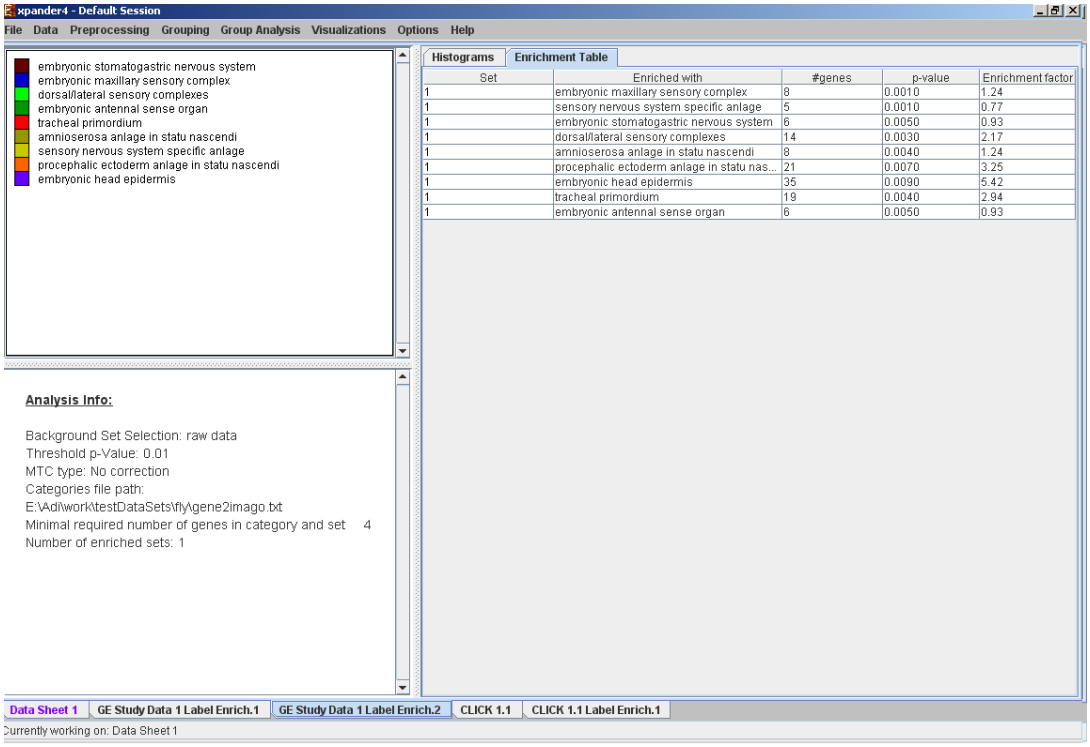
Field	Description
Perform analysis on	The grouping solution on which the analysis will be performed.
Include back nodes	Include genes that are part of the module' but not included in the GE data (Relevant only if the analysis is performed on modules, detected by network based algorithm)
Load categories from	Input field for the file path, holding the gene categories/attributes.
Background set	Determines the set of genes that will be used as background in the analysis. Options are: all genes (of the relevant organism), original input data, filtered data or background set from file (see the Files Format section for details regarding the format of an external background set).
p-value threshold	A category/attribute will be considered significantly enriched in a cluster/bicluster if its corrected p-value is lower than this threshold.
Multiple tests correction	Can be set to Bonferroni or None (when set to Bonferroni the corrected p-values are the ones that are compared to the threshold p-value).
Minimal overlap between category and	The minimal number of genes from a cluster/bi-cluster expected to be categorized/attributed by

set	an attribute in order for its enrichment to be accepted.

After the analysis is performed an enrichment analysis solution visualization tab main window. It contains general information regarding the analysis, a sort-abl all detected enrichments (set ID, enrichment category, p-value, etc.) and a r panel along with a color index (mapping each color to a corresponding categ histogram panel contains one histogram for each probe/gene set/group in wh has been detected. Each histogram contains a column for each significant (mor would be expected by random) category. The definition of significant depends selection of threshold p-value i.e., a category is considered significantly cluster/bicluster if its corrected p-value is lower than the preset threshold p-value

The height of the column is proportional to the significance of this enrichment log(raw p-value)), and the frequency ratio (frequency in set divided by background) is written on top of the column. Upon clicking on a column, a displayed containing the class name, corrected p-value, and a list of the genes ir cluster that belong to the category. Upon clicking on one of the gene Ids in the ta web page with information regarding this gene is displayed. The display tool cluster number, size and homogeneity.



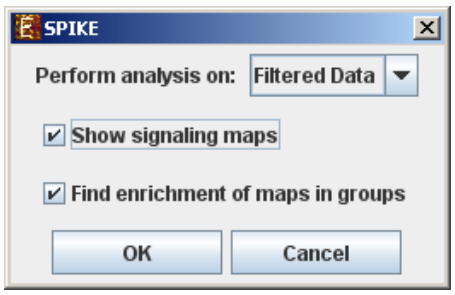


The results of this analysis can be exported to a text file by selecting *File>>Expo* the corresponding view is the selected tab. OR by using the *File>>Save All* optio export all solutions within a session to text and image files.

Network Based Enrichment Analysis

This tool allows browsing through signaling data to view the sub-graphs th induced by the analyzed gene groups. It also enables the user to search f enrichment of these groups in highly curated signaling maps. To perform t Expander interfaces with the SPIKE software and database. For further in regarding the SPIKE software see the References section.

To perform the analysis on one/more of the gene groups defined in Expar clusters, bi-clusters, modules, loaded gene sets or filtered data), select *Er Analysis >> Network >> SPIKE>>Gene Groups*. The following dialog box



The following table specifies the different parameters that can be set via tl box:

Field	Description
Perform analysis on	The grouping solution on which the analysis will be performed.
Show signaling Maps	For each group display regulatory data induced by the genes included in the group.
Find enrichment of maps in groups	For each group, search for signaling maps that are enriched with genes included in the group.

Pressing OK in the dialog box will launch the SPIKE application. When op first time, the launch takes a few minutes, since it has to build a local data this point on, please refer to page 12 in the SPIKE user manual.

SPIKE can also be operated on a sub-group of genes that is derived from enrichment solution in Expander. I.e. a group of genes that has a commor that was found to be enriched by one of the enrichment analysis operatio to operate SPIKE on such a *group, select: Enrichment Analysis>>Networl SPIKE>>Enrichment Derived Sets*.

Gene Set Enrichment Analysis (GSEA)

GSEA (Subramanian et al 2005) considers experiments with genomewide expre from samples belonging to two classes. Genes are ranked based on the correlat their expression and the differential expression between classes distinction or pr the user.
Given an *a priori* defined set of genes *S*, the goal of GSEA is to determin members of *S* are randomly distributed throughout the ranked list of genes (found at the top or bottom. It is expected that sets related to the phenotypic disti to show the latter distribution.

There are two key elements of the GSEA method in Expander:

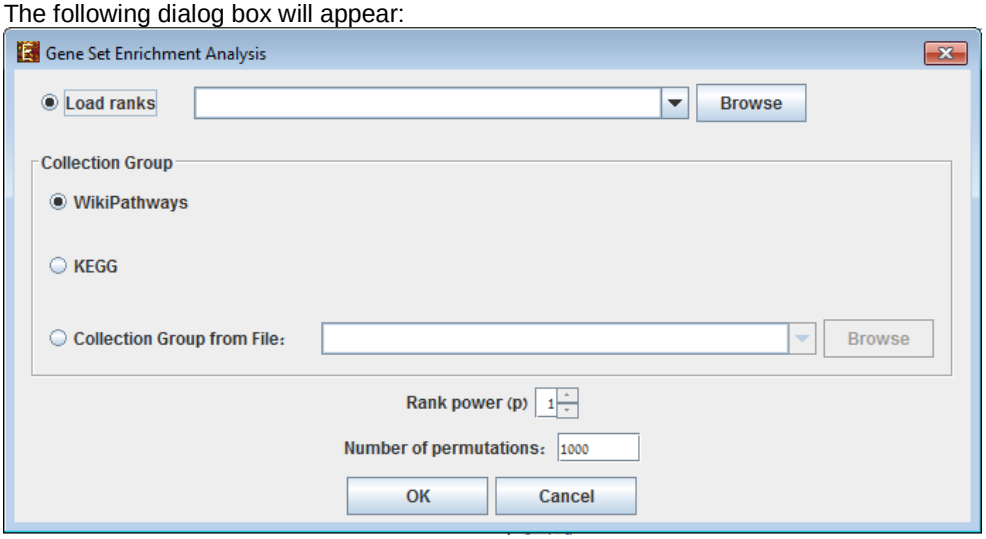
Step 1: Calculation of an Enrichment Score. Enrichment score (*ES*) reflects which a set *S* is overrepresented at the extremes (top or bottom) of the entire ra score is calculated by walking down the list *L*, increasing a running-sum sta encounter a gene in *S* and decreasing it when we encounter genes not in *S.* Th the increment depends on the correlation of the gene with the phenotype. T

score is the maximum deviation from zero encountered in the random walk. It cc
weighted Kolmogorov–Smirnov-like statistic.

Step 2: Estimation of Significance Level of ES. An estimation of the statistic
(nominal *P*-value) of the *ES* is done by using an empirical phenotype-based p
procedure that preserves the complex correlation structure of the gene ex
Specifically, the phenotype labels are permuted again and the *ES* of the ge
permuted data is re-computed, which generates a null distribution for the *E*
provided a pre-ranked list of genes then a random shuffling of the ranked list is
The empirical, nominal *P* value of the observed *ES* is then calculated relati
distribution. Importantly, the permutation of class labels preserves gene-gene cc
thus, provides a more biologically reasonable assessment of significance t
obtained by permuting genes.

There are 2 ways to perform GSEA:
1. GSEA on a pre-Ranked list of Genes without loading gene expression data
2. GSEA on a gene expression data

GSEA on a pre-Ranked list of Genes without loading gene expression data
To perform analysis on a pre-ranked list of genes, select *File->New Session->
Analysis (GSEA).*
The following dialog box will appear:



The following table specifies the different parameters that can be set via tl
box:

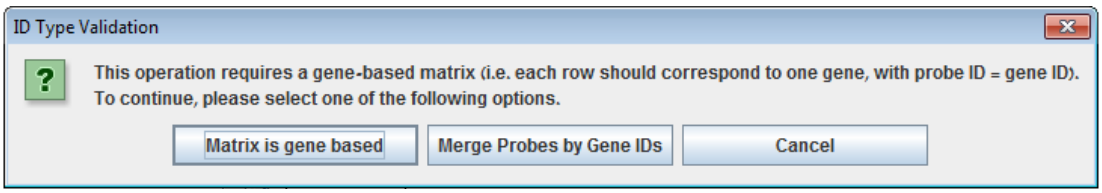
Field	Description
Load ranks	User pre-ranked list of genes file composed of two columns – first with genes and second with values
Collection Group	Can be chosen between: WikiPathways, KEGG or User pre-defined file with gene sets*
Rank Power (p)	If p=0 then ES is reduced to standard Kolmogorov–Smirnov statistic. If p=1then ES is a weighted Kolmogorov–Smirnov-like statistic.

Number of permutations	For estimation of the Significance Level of ES

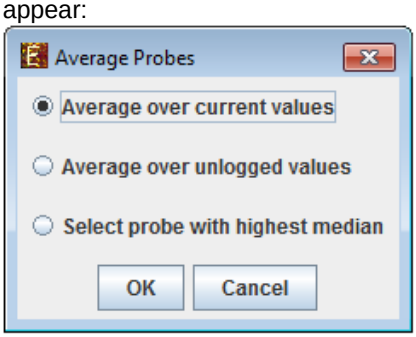
Please refer to Results section in "GSEA on a gene expression data" to interpret

GSEA on a gene expression data

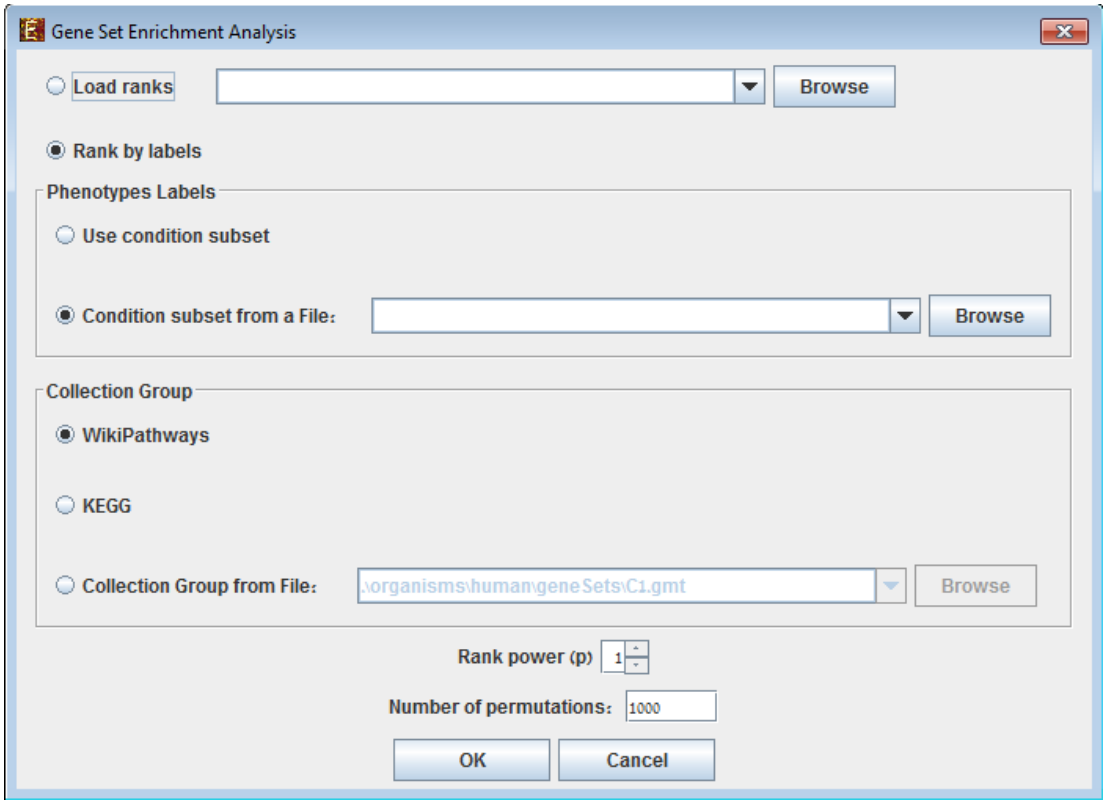
To perform the analysis on the gene expression, select *Enrichment Analysis* >
The following dialog box will appear:



The user can choose between "Matrix is gene based" (i.e each row should cor
gene, with probe ID = gene ID) or "Merge Probes by Gene IDs".
In case the user chose "Merge Probes by Gene IDs", a dialog box titled "Avera
appear:



After choosing the preferred merging option or "Matrix gene is based" in "ID T
dialog box, "Gene Set Enrichment Analysis" dialog box will appear:



The following table specifies the different parameters that can be set via tl box:

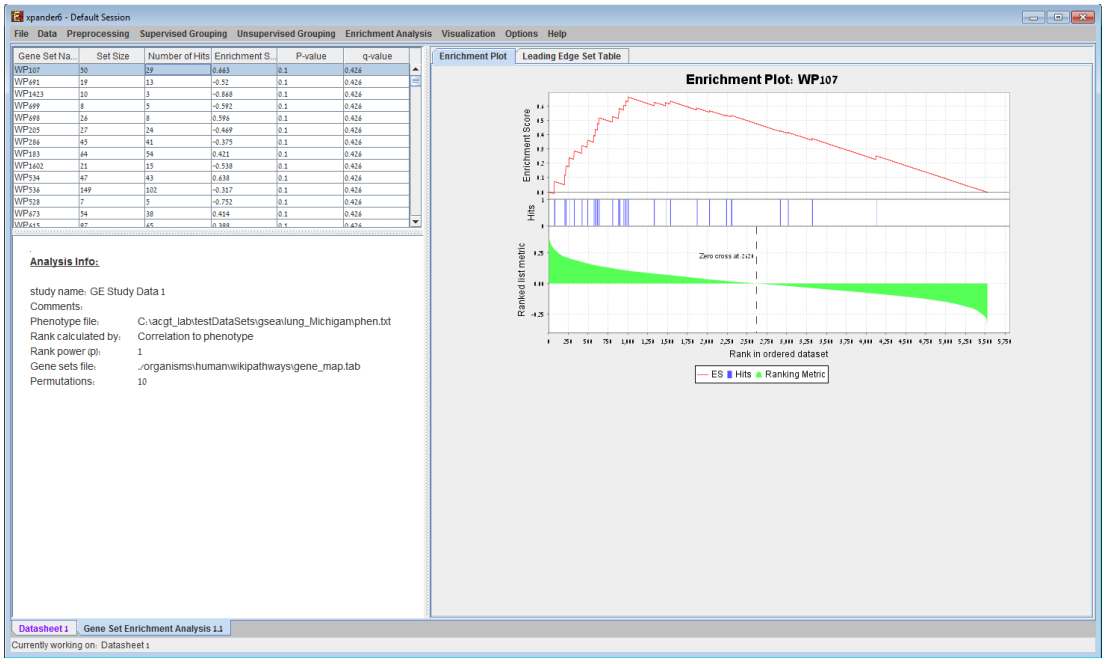
Field	Description
Load ranks	User pre-ranked list of genes file composed of two columns – first with genes and second with values
Use condition subset	Can be used when the matrix is composed of two condition subsets
Condition subset from a File	A file with a single tab delimited row which contains the phenotype for each condition
Collection Group	Can be chosen between: WikiPathways, KEGG or User pre-defined file with gene sets*
Rank Power (p)	If p=0 then ES is reduced to standard Kolmogorov–Smirnov statistic. If p=1then ES is a weighted Kolmogorov–Smirnov-like statistic.
Number of permutations	For estimation of the Significance Level of ES

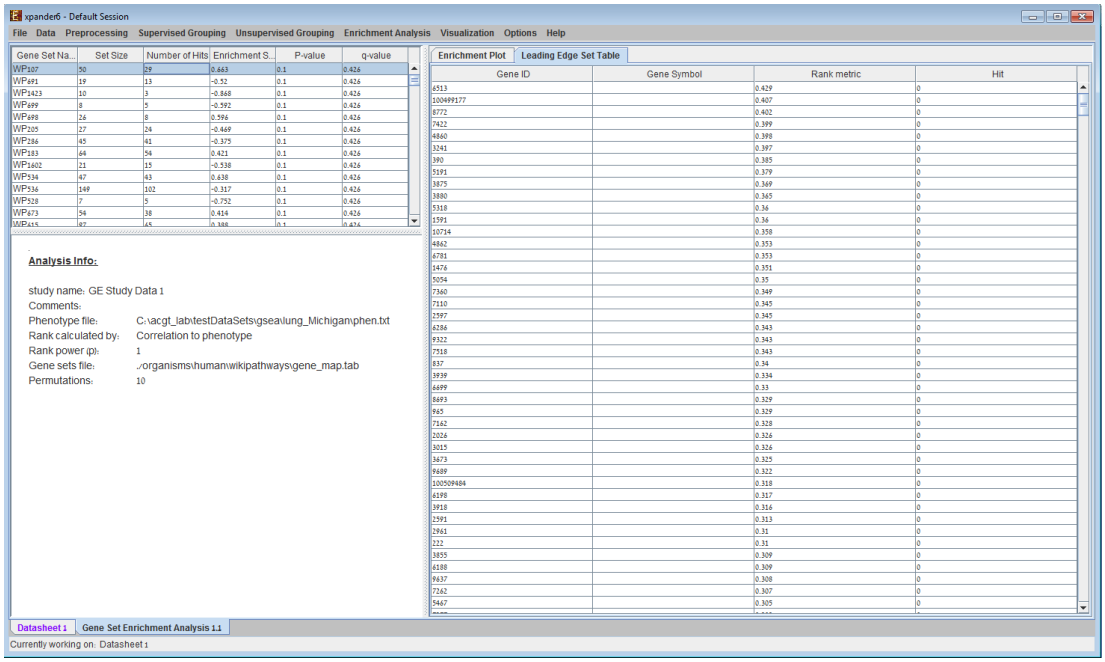
*Collection Group File

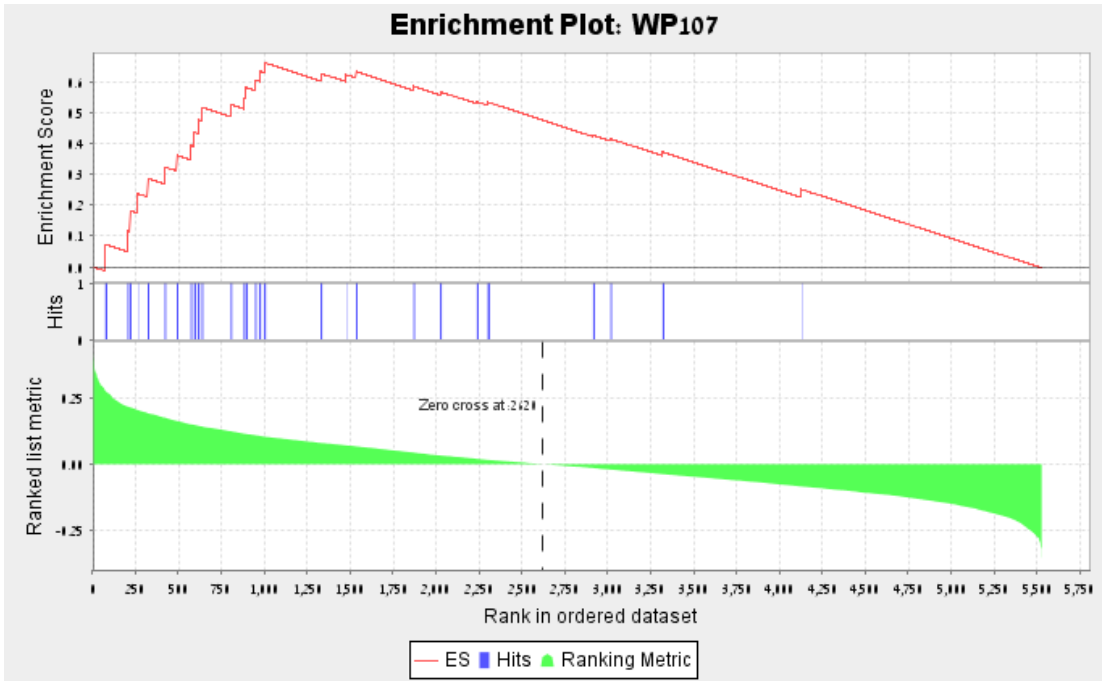
More gene sets are available to download via MSigDB and files can b
"Collection Group File" field in Collection Group.
The MSigDB gene sets are divided into 7 major collections:
C1.gmt – Positional gene sets
C2.gmt – Curated gene sets
C3.gmt – Motif gene sets
C4.gmt – Computational gene sets
C5.gmt – GO gene sets
C6.gmt – Oncogenic signatures
C7.gmt – Immunologic signatures
For further information, please refer to: http://www.broadinstitute.org/gsea/msigd

Results

After the analysis is performed a gene set enrichment analysis solution visu
added to the main window. It contains general information regarding the analy
table holding all gene sets (Gene set name, set size, Number of hits, Enrichi
value and q-value(FDR)), an enrichment plot for each gene set selected in
enrichment plot panel contains a graph of the enrichment score for each gene
list, a bar of hits of the genes in the gene set with the genes in the ranked list ar
metric of the genes, and a tab - leading edge set table that contains Gene ID,
Ranke metric and Hit (if the gene was hit by a gene in the gene set). The le
table contains only genes that appear before the maximum enrichment score.







The enrichment plot can be saved as image file by right clicking on the graph->S

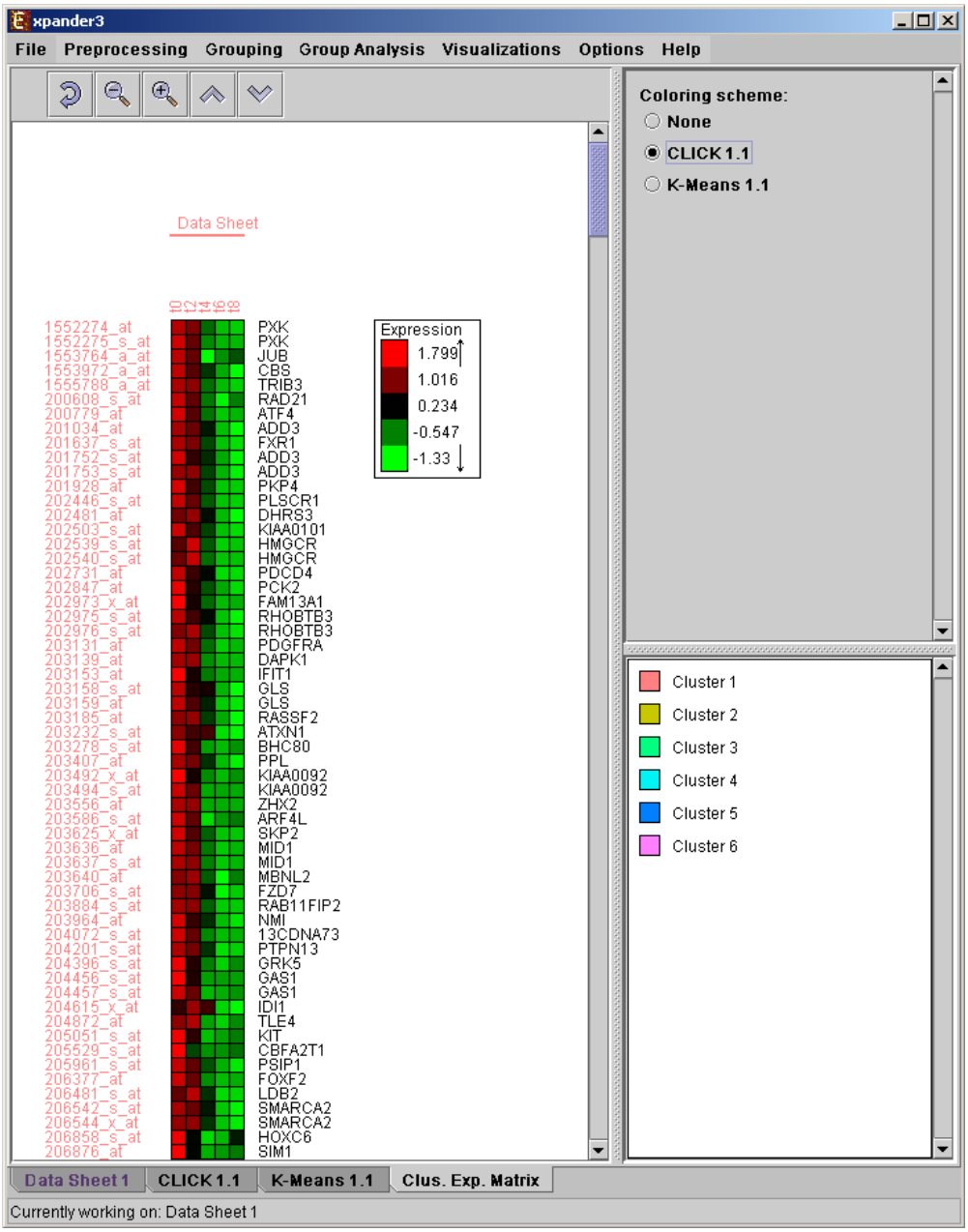
The results in the tables can be saved as image file by selecting *File>>Save A* the corresponding view is the selected tab OR by using the *File>>Save All* op export all solutions within a session to text and image files.

Matrix Visualizations

An expression matrix (Heat-map) visualization is integrated in many of EXPANI
This visualization is similar to the red-green matrix representation of Eisen et a
does is to render the gene-expression data on the screen in color, where green i
expression, and red indicates over expression. Color rendering can be configur
in one of the following manners: (a) by setting the range (top and bottom value
values (default values are set according to the data scale, e.g. 40-1000 for no
absolute intensities data) or (b) by setting the percent of values, which are to l
as extreme values from each edge (by default set to 5%). The manner
configuration (i.e. (a) vs. (b)) can be set via the 'Data Matrix View' tab in the 'Di
dialog box, available from *Options* >> *Settings*. The red/green coloring sc
changed to blue/yellow (using *Options* >> *Settings* >> *Display* >> 'Data Matrix Vi

A color scale appears next to the matrix (upper right side). The displayed tool
probe ID and condition title corresponding to the row and column on which
placed, and the expression value in that position. The matrix toolbar contains
out, reset scale (to reset zoom factor), shorten condition title and Elongate condi

Upon selecting *Visualization* >> *Clustered Expression Matrix*, a clustered exp
visualization tab is added to the main window. The probes are ordered in their c
a clustering solution has been previously created, its' name appears next to a
the top right panel. Upon pressing this button, the order of the probes in the di
and probe IDs are colored according to the clusters. The color index at the bott
maps each color to the index of the corresponding cluster.

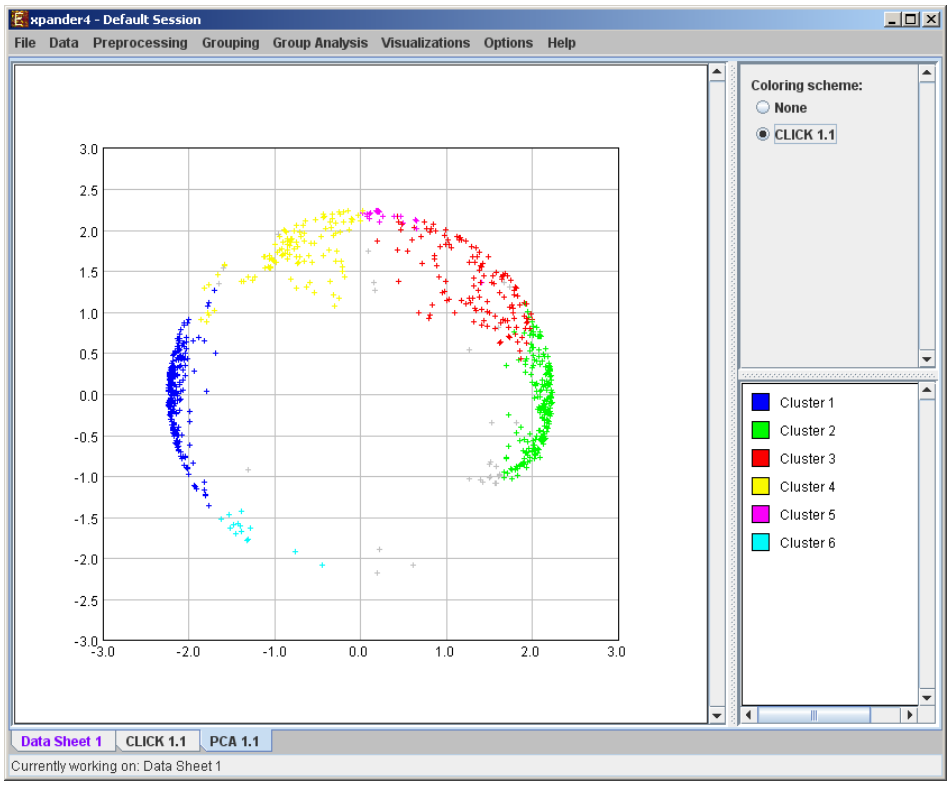


PCA Transformation

This tool transforms the original data from a k (original pattern length) to a space, so that each expression vector is represented by a dot on an XY sca transformation is based on the PCA (Principal Component Analysis) algorithm. tool, select *Visualization >> PCA*.

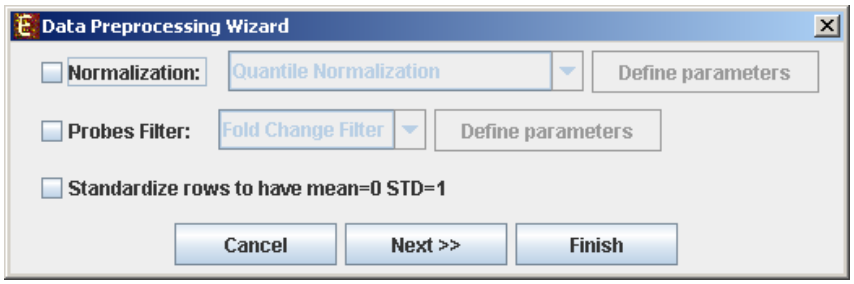
If a clustering solution has been previously created, its' name appears next to a the top right panel. Upon pressing this button, the color of each dot in the di

according to the cluster assignment of the corresponding probe. The color inde
right panel, maps each color to the index of the corresponding cluster.

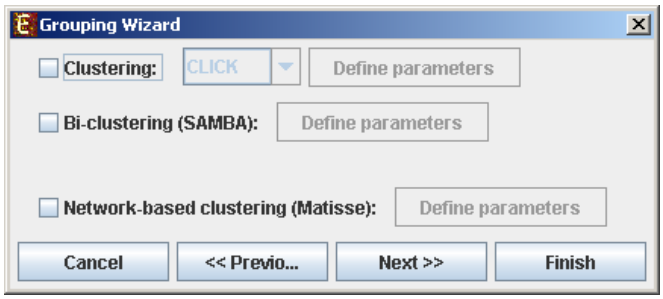


Analysis Wizard

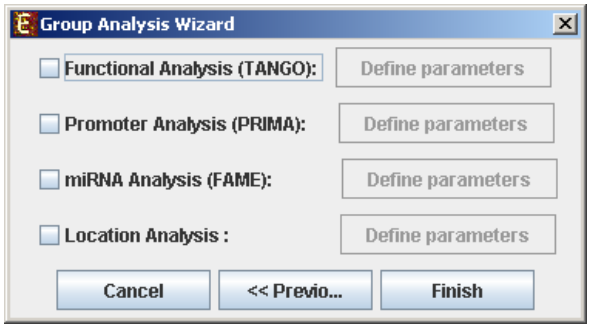
Expander allows performing an automatic analysis on a loaded dataset by
analysis wizard to predefine the analysis stages and parameters.
To use this tool, go to *Data>> Analysis Wizard*.
Upon selecting this option, the following dialog box will appear, allowing tc
required preprocessing operations:



For some of the stages, parameters can be defined by pressing the corres
“Define parameters” button. Upon pressing the “Next>>” button, the follow
box will appear, allowing to define the required grouping operations:



Upon pressing the “Next>>” button, the following dialog box will appear, a
define the required enrichment analysis operations:

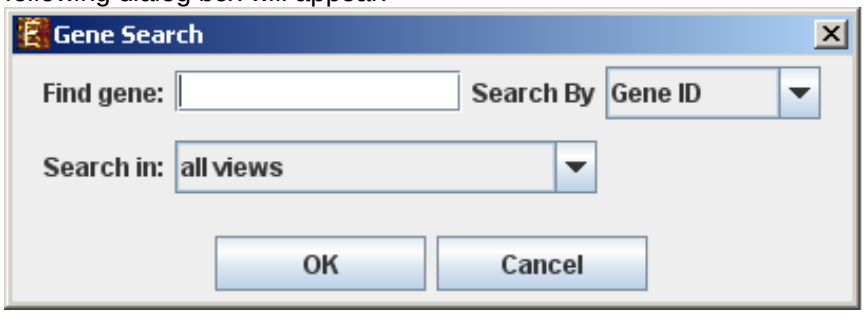


Upon pressing the “Finish” button (in any one of the dialog boxes) the enti
operations defined by the user is performed by Expander, and the corresp
visualizations are generated.

Additional Options

Searching for a gene/probe in the display

A gene can be detected in a display by selecting *Options >> Search Gene*
following dialog box will appear:

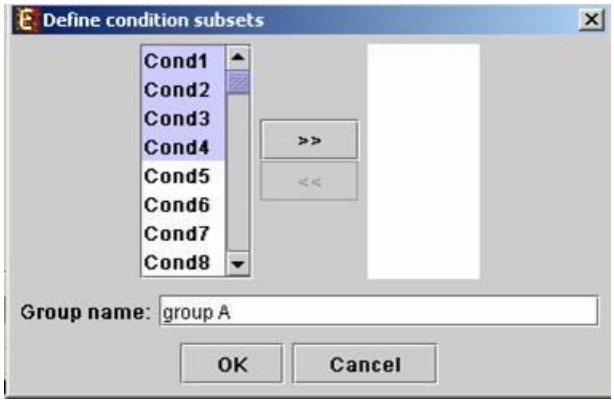


Please type the ID of the gene (can also be symbol/probe ID depending o
selection in the “Search By” combo-box) in the corresponding text box. No

must type the entire name or ID, not part of it. After pressing the “OK” butt
will appear containing text describing the number of items detected in eac
searched views (number of “hits” in each view). In addition, the correspon
elements will be highlighted in all searched displays.

Defining condition subsets

You may group several conditions under a common subset name, by selecting
Condition Subsets. This partition is used for visualization purposes. In the dia
the relevant conditions, type a group name and click on the arrows.



I addition to subset definitions, multiple condition annotations can be loaded u
Data>>Load Condition Attributes. The file should be in a tabular (tab delimited),
correspond to attributes (first column will contain attribute names) and columns correspc
(first row will contain condition labels in the same order as in your expression data)
numeric and/or textual.

Saving and loading sessions

A set of analysis operations performed on one data set can be saved by selectin
Session. It can later be reloaded by selecting *File >> Load Session*. Loadin
saved session will bring up all analysis output and visualizations that had bee
that session, and the user will be able to continue working where he had previou

Closing views

The user can close all open views by selecting *File >> Close All*.

Closing a single view can be performed either by selecting *File>>Close* when th
is selected OR by right clicking on the tab title of the relevant view and selecting
popup menu.

Docking a view into a separate frame

Can be performed either by selecting *Options >> Dock into external frame* wh
view is selected OR by right clicking on the tab title of the relevant view and sele
external frame from the popup menu.

Upon creating the separate frame, the view will be removed from the main
closing the separate frame generated in this manner, the view will be retrieved
window.

Accessing the EXPANDER download page

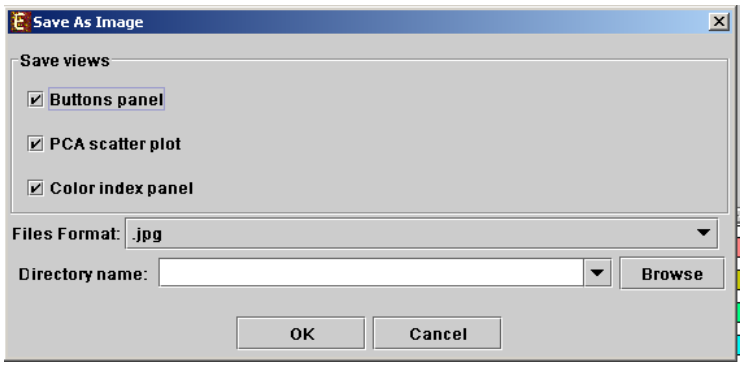
The Expander download page can be accessed directly by selecting *Help* >> *C Page*, while the machine is connected to the Internet.

Printing the display

Each display can be printed by selecting *File* >> *Print* while its tab is selected.

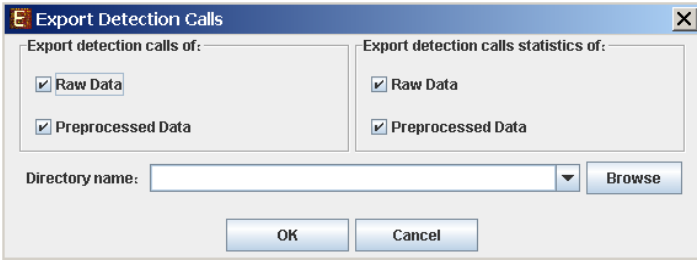
Exporting display into image files

Each display can be exported into image files of type .jpg, .png or .eps (post-s be done by selecting *File* >> *Save As Image*. Upon selecting this option, a dial to the following is displayed. In the dialog box the saved images (sections of th files format, and destination directory name are input.



Exporting detection calls information

The detection calls info of the raw and preprocessed data can be exported in selecting *Data* >> *Export Detection Calls*. Upon selecting this option, the followir displayed. You may export the detection calls and also the statistics of detectior of P, M and A calls per condition), for raw data and for preprocessed data.



File Formats

Expression data file format:

1) Suffix: no limitations.
2) Separating token: tab delimiter.
3) Format:

1st line: contains a string like 'probeId' and a tab delimiter, followed by 'geneSymbol' and a tab delimiter, followed by the names of all conditions sep delimiters. The symbol column is optional – if the file does not contain a symbol specify it in the Advanced Input Dialog box (see Input Data section).

2nd line (**optional**): contains the string '>SERIES', a tab delimiter followed 'SYMBOL ' (if there is a symbol column), a tab delimiter and then all corresponding to the condition (one series assigned for each condition) sep delimiters.

Next lines: Each subsequent line consists of the probe ID (an identifier string th each probe in the chip), followed by a string, which represents the gene full na can be left empty by adding an additional tab delimiter), followed by its expres tab delimited). If the expression file contains missing values, Expander either with a preset value (0 by default), or estimates them using the KNN (K-Near method, depending on the user selection in the data load dialog box.

*For example see files 'expressionData1.txt' and 'expressionData2 Expander/sample_input_files/ directory.

If the data is not in the above format, it may be possible to load it using the 'Ad box, which appears upon pressing the 'Advanced' button in the Expression Da box (see Advanced Input Dialog box in Input Data section).

Expression data with detection calls file format:

1) Suffix: no limitations.
2) Separating token: tab delimiter.
3) Format:

1st line: contains a string like 'probeId' and a tab delimiter, followed by 'geneSymbol' and a tab delimiter, followed by the names of all conditions and de columns alternately, separated by tab delimiters. Each title of condition is follow its detection column.

The symbol column is optional – if the file does not contain a symbol column, p in the Advanced Input Dialog box (see Input Data section).

Next lines: Each subsequent line consists of the probe ID (an identifier string th each probe in the chip), followed by a string, which represents the gene full na can be left empty by adding an additional tab delimiter), followed by its express detection calls values, alternately (all tab delimited). Each expression value is detection value (P, M or A). If the expression file contains missing values, E replaces them with a preset value (0 by default), or estimates them using the K Neighbors) method, depending on the user selection in the data load dialog box.

*For example see files 'expressionWithDetection.txt' in the Expander/sam directory.

If the data is not in the above format, it may be possible to load it using the 'Ad box, which appears upon pressing the 'Advanced' button in the Expression Da box (see Advanced Input Dialog box in Input Data section).

Gene Sets file format:

1) Suffix: no limitations

2) Format: Each line contains a gene ID, a gene symbol (optional) and the nam set (separated by tabs/spaces). The gene IDs are expected to be of the same c

in the GO annotation and TF fingerprint files. For details regarding the Gene that is used for each organism, refer to the Supplied files section.

*For example see file 'geneSetsData1.txt' under the Expander/sample_input_ (see Sample input files for more details).

Probes Filter file format:

Each line contains a single identifier. Identifiers can be probe Ids, gene Ids OR (but not a mixture of these identifier types).

ID conversion file format:

1) Suffix: Currently, there are no limitations regarding the file name suffix.

2) Format: Each line contains the probe id as it appears in the data file, a tab sep corresponding gene ID (e.g. Entrez/Locus-Link ids for mouse and human ge codes for yeast). The second field can be left blank, indicating no conversion for

* It is possible that several probe IDs in the data file will be mapped to the same several ESTs from the same gene).

Clustering files format:

1) Suffix: no limitations.

2) Format: Each line contains the probeID, a tab separator and name/numbe The number 0 is reserved for probes that are left unclustered. The file does not l all probes in the data. If a probe does not appear in the file, it is autom unclustered.

*For example see file 'expressionData1Clustering.sol' (a clustering solution fo expressionData1.txt') under the Expander/sample_input_files/ directory (see Files section for more details).

Biclustering files format:

1) Suffix: `.bic`.
2) Format: the file is composed of two parts, presented here.

Part 1 presents a summary of the biclusters found.

- It begins with the string: `[Bick]` in the first line.
- Following lines contain the bicluster's id followed by its' score, separated by (a line for each bicluster).

Part 2 presents the probesets and the conditions contained in each bicluster.

- It begins with the string: `[Bicd]` in the first line.
- Following lines contain the bicluster id, type of element ('0' for condition, '1' element id (name of condition or probe ID), separated by tab delimiters.

Background set files format:

1) Suffix: no limitation.

2) Format: each line should contain one gene ID. The gene IDs are expected to convention used in the annotation and TF fingerprint files for the organism you (please refer to the Supplied Files section).

Gene annotations/categories files format (for the general analysis):

1) Suffix: no limitation.

2) Format: each line should contain one gene ID and an annotation/category n by a tab delimiter. The gene Ids are expected to be of the same conventic annotation and TF fingerprint files for the organism you are working on (plea Supplied Files section).

Sample Input Files

Several sample files are provided under Expander/sample_input_files/. These fil

expressionData1.txt – A gene expression data file that was generated us microarray technology. This is a partial dataset extracted from a yeast cell generated by Spellman et al 1998 (see the References section). Gene identifiers yeast ORFs, which are the same identifiers used in the annotation and TF finge are supplied with Expander. Thus, no conversion file is required.

ExpressionData2.txt – A gene expression data file that was generated in technology. This dataset was generated in an experiment that was conducted ir on human cells, and has not yet been published. Affymetrix chips of type HG-Fc for this experiment and thus, the HG-Focus conversion file is required for the a downloaded from the download page).

ExpressionData3.txt – taken from Murray JI, Whitfield ML, Trinklein ND, Mye PO, Botstein D: Diverse and specific gene expression responses to stresses in c cells. Mol Biol Cell 2004, 15:2361-2374. A corresponding conversion file (from c Ids) is available at the same directory under the name Data3Conversion.txt.

expressionWithDetection.txt – A gene expression data file with detection calls generated in the Affymetrix technology. This dataset was generated in an experi conducted in our laboratory on human cells. Affymetrix GeneChip HGU133 Plus were used for this experiment.

expressionData1Clustering.sol – A clustering solution that was generated b the dataset in 'expressionData1.txt'.

geneSetsData1.txt – Contains sets of human genes (in Entrez/Locus-Link Ids).

Data3Conversion.txt - A conversion file for expressionData3.txt.

Supplied Files

The following files include gene info files: Gene ID conversion files, GO annotati fingerprint files, promoter sequences, miRNA target scan files, chromosomal po biological pathway files, taken from the KEGG database*. These files should be into "Expander/organisms" directory.

Organism	Size after extraction	Origin of GO annotations	Origin of sequences used for generating TF-fingerprint files	Origin of miRNA targets data files:	Origin of chromosomal location data files	Origin of pathway data files
Human	431MB	NCBI - Oct 2009,GOA@EBI - Nov 2008	Ensembl release 53	TargetScan website version 5	UCSC genome browser Jan 09	KEGG data base November 09, WikiPathways March 2013
Baker's yeast	30.6MB	NCBI - October 2009	SGD database 1.01	-	-	KEGG data base November 09, WikiPathways March 2013
S. pombe	15.7MB	NCBI - November 2008	Sanger GeneDB - October 2008	-	-	KEGG data base November 09, WikiPathways March 2013
Listeria monocytosenes EGD-e	1.74MB	Blast2GO - February 2009	Not available	-	-	KEGG data base November 09

Mouse	338MB	NCBI - Oct 2009	Ensembl release 53	TargetScan website version 5	UCSC genome browser Jan 09	KEGG data base November 09, WikiPathways March 2013
Rat	280MB	NCBI - Oct 2009	Ensembl release 56	-	UCSC genome browser Jan 09	KEGG data base November 09, WikiPathways March 2013
Fly	218MB	NCBI - Oct 2009	Ensembl release 53	TargetScan website version 5	UCSC genome browser Jan 09	KEGG data base November 09, WikiPathways March 2013
C-elegans	278Mb	NCBI - Oct 2009	Ensembl release 56	TargetScan website version 5	UCSC genome browser Jan 09	KEGG data base November 09, WikiPathways March 2013
Arabidopsis	267MB	NCBI - Oct 2009	TAIR - December 2006	-	-	KEGG data base November 09, WikiPathways March 2013
Zebra Fish	254Mb	NCBI - Oct 2009	Ensembl release 53	-	UCSC genome browser Jan 09	KEGG data base Novemb

						er 09, WikiPathways March 2013
Chicken	205Mb	NCBI - Oct 2009	Ensembl release 56	-	-	KEGG data base November 09, WikiPathways March 2013
Tomato	1.7Mb	EBI - March 2009	-	-	-	-
A.Fumigatus	72.5Mb	EBI - Jan 2010	Kevin Verstrepen's lab (via private communication with Nir Osherov)	-	-	KEGG data base November 09
E. coli	1.25MB	EBI - April 2008	Not available yet	-	-	-
Rice	170MB	GO DB Dec 2010	Bioinformatics Core database Washington State University	-	-	KEGG data base November 10
Leishmania	2MB	Zilberstein D. lab Technion - Israel Nov 2011	-	-	-	-

* Users of this product may not download large quantities of KEG

Gene ID conversion files:

Gene ID conversion files for many of the Affymetrix chips can be downlo
Expander download page. The files map each Affymetrix Id into the correspo
Conversion files are generated and added to the download page according to u
you can't find the file you need here, please look it up in the download page, ar
it's not there.

Organism	Chip name
Human	HG-Focus
Human	HGU1332
Human	HG-U95E
Human	HG-U133A
Human	HT_HG-U133A
Human	HG-U133Plus2
Human	Hu-35KsubB
Human	HuGene-1_0-ST
Mouse	MGU74Av2
Mouse	MGU430_2
Mouse	MG430A2
Mouse	MoGene-1_0-ST
Rat	RGU34A
Rat	Rat230_2
Rat	Agilent
C-elegans	C. elegans Genome Chip
Arabidopsis	ATH1
Zebra-Fish	GeneChip Zebrafish Genome Array
Chicken	Affymetrix Chicken Genome Chip
E. coli	Affymetrix E. coli Antisense Genome Array
E. coli	Affymetrix E. coli Genome 2.0 Array

Network files :

Organism	File name	Network origin
Human	Expander.hsa.RualNature05.sif	Towards a proteome-scale map of the human protein-protein interaction network by Rual JF et al. *Nature*. 437(7062):1173-8 (2005)
Human	Expander.hsa.IntAct.s	IntAct database

	if	(http://www.ebi.ac.uk/intact/)
Mouse	Expander.mmu.IntAct.sif	IntAct database (http://www.ebi.ac.uk/intact/)
Rat	Expander.rno.IntAct.sif	IntAct database (http://www.ebi.ac.uk/intact/)
Worm	Expander.cel.SimonisNatMethods08.sif	Empirically controlled mapping of the *Caenorhabditis elegans* protein-protein interactome network by Simonis N. et al. *Nature Methods* 6, 47 - 54 (2009)
Fly	Expander.dme.DroID.sif	DroID database (http://www.droidb.org/)
Yeast	Expander.sce.United.sif	1. High-Quality Binary Protein Interaction Map of the Yeast Interactome Network by Yu et al. *Science* 322(5898):104 – 110 (2008) 2. Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae by Reguly et al. *Journal of Biology* 5(4):11 (2006) 3. Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae by Collins SR et al. *Molecular Cell Proteomics* 6(3):439-50 (2007)
Arabidopsis	Expander.ath.TAIR.sif	TAIR database (http://www.arabidopsis.org/)
E. coli	Expander.eco.Arifuzzaman06.txt	Large-scale identification of protein–protein interaction of Escherichia coli K-12 by Arifuzzaman M et al. *Genome Research* 16(5):686-91. (2006)

Settings

The Settings are accessible from the *Options* menu, and contain *Display* s *External applications* settings.

The **Display** dialog box contains the following tabs:

Clustering Results View – Contains check boxes that configure the follo parameters:

- A common Y-axis scale for all cluster patterns (vs. cluster specific)
- Visible x axis
- Connect all points in a pattern
- Display similarity matrix for probes (using Pearson correlation)
- Display similarity matrix for conditions (using Pearson correlation)

Enrichment Analysis Results View – Contains a check box that configu
the Y-axis scale of all histograms is common OR cluster specific.

Data Matrix View – Allows selection between:

- Range control and extreme values control when rendering expression r
- The red/green coloring scheme can be changed to blue/yellow.

The ***External applications*** dialog box allows specification of the location
executable (required for CEL files loading). In Windows, R.exe file is likely to
the 'bin' folder of R software. In Linux, you may type 'which R' in the command li
path.

R External Application

The CEL file preprocessing and the newly added SAM filter utilities require the
of one of the recent versions of R, a free software environment for statistical
graphics. R can be installed from: http://cran.r-project.org/.

Upon the first time that Expander uses R external application, a window will pop
specify your R software location. Please browse to the location of your R
Windows, R.exe file is likely to be located in the 'bin' folder of R software. In L
type 'which R' in the command line to find R path. If you have a few versions
please make sure to point Expander to a version in which the necessary packa
installed.

You may also specify R location from the menu: *Options >> Settings >> Externa*

To use R utilities, please make sure there are no white spaces in the path of Exp
directory (or the CEL files directory, if loading CEL files). For example, if the nam
Expander folder is 'Expander 5', change it to 'Expander_5'. If Expander is under
Files" it should be moved to another location, because of the space between "Pr
"Files". The R software does not cope well with spaces in the path.
Also, please make sure to have 'write' permission to the Expander\Rscripts direc
loading CEL files, check also that you have 'write' permission to the *Files locatio*
specified in the 'Load CEL Files' dialog box.

After specifying R software location, a window will pop, asking you to approve
automatically installation of R packages when needed.

If you approve automatically installation of R packages then when R utility is use
Expander will automatically install the needed R packages for the used R utility.

If you disapprove automatically installation of R packages then please refe
installation of R packages" section.

Manually installation of R packages

After installing R, please do the following to install the Bioconductor “affy” pac package and the "samr" package:

1. Run R.
2. In the R frame\window type the text: **source("http://bioconductor.org/bi**
3. Press ‘Enter’.
4. In the R frame\window type the text: **biocLite("affy")**
5. Press ‘Enter’.

To install the 'samr' package:

6. In the R frame\window type the text: **install.packages("samr")**
7. R frame\window type the text: **install.packages("impute")**
8. Press ‘Enter’.

To install the ‘eisa’ package:

9. In the R frame\window type the text: **biocLite("eisa")**
10. Press ‘Enter’.

To install the 'gcrma' package:

11. In the R frame\window type the text: **source("http://bioconductor.org/bi**
12. Press ‘Enter’. In the R frame\window type the text:
13. **biocLite("gcrma")**
14. Press ‘Enter’.

You may install only one of the packages, depending on what you wish to use (t "samr", follow instructions number 1, 6 and 7).

FAQ

Linux/Unix problems

Click clustering, Samba bi-clustering, Tango and Prima Promoter analysis algor running on Linux/Unix.

CEL Files Loading Problems

How do I install R and the Bioconductor "affy" package?

Loading of CEL files or performing SAM filter continue for ever

Loading of CEL files fail.

Clustering

When I try to run Biclustering on my data I get a failure notice.

How can I save the clustering expression patterns charts?

Grouping Analysis (functional and promoter analysis)

Saving sessions

Others

Click clustering, Samba bi-clustering, Tango and Prima Promoter analysis algor
running on Linux/Unix.

Answer: Make sure that you have write permission in the Expander directory,
permissions on the files: click.exe, samba.exe, annot_sets.exe and analyzeFi
which are under the Expander directory. If the problem still occurs,
expanderLog.txt and search for the text: “libstdc++.so.5”. If this text appears
message indicating it has not been found), please contact your system adr
report this problem (this is a system problem). If you do not have a system adr
fail to install this library, please contact us (expander@cs.tau.ac.il) and we will tr

How do I install R and the Bioconductor "affy" package?

Answer: please refer to R External Application section.

Loading of CEL files or performing SAM filter continue forever.

Answer: If the operation continues forever (the 'processing, please wait' windov
please check if there is a folder with a space in its name somewhere in the pa
(or the CEL files) directory. For example, if the name of the Expander folder i
change it to 'Expander_4'. The R software used for preprocessing CEL files
dealing with spaces in the path. If this is the problem, then in the expanderLog
Expander directory) there should be a message about arguments being ignored.

Loading of CEL files fail.

Answer: Make sure you have R along with the Bioconductor “affy” package inst
version which is specified in the settings “External Applications” tab (from the m
Options >> Settings >> External applications). If R location is not defined in the s
please define it (In Windows, R.exe file is likely to be located in the 'bin' folder of
Linux, you may type 'which R' in the command line to find R path). If you are usir
package as cdf source, please make sure that the package is a folder located ur
library directory and that it is the correct package for your chip. If loading of CEL

please make sure that the *Files location* which you specified in the 'Load CEL Fil
is a folder which contains CEL files and that you have write permission to that fol

When I try to run Biclustering on my data I get a failure notice.

Answer: Make sure that the 'Use option files of type' field in the SAMBA input di
empty (if it is, please re-download Expander). Also make sure that the followir
your Expander directory: ibic.opt, samba.exe.

How can I save the clustering expression patterns charts?

When the clustering results tab is open, please go to *File >> Save As Image*.

When I run Functional Analysis, Expander gets stuck.

Answer: If you are working with Expander version 4.0 or 4.0.1, please upda
version (4.0.2 and on).

When I load a session with that contains Functional Analysis results, Expander g

Answer: If you are working with Expander version 4.0 or 4.0.1 (and the sessic
with a version < 4.0), please update to a higher version (4.0.2 and on).

When I try to run Promoter Analysis no values appear in the Fingerprints file fi
dialog box.

Answer: Fingerprint files are not placed in the right directory. Fingerprint files sh
under the 'TF_fingerprints' directory that is under the Expander
name> directory. For example, the human FP file should be p
…/Expander/organisms/human/TF_fingerprints/. When downloading the organis
zip, it should be extracted into the Expander/organisms/ directory. This will au
them in the right place.

Enrichment analysis leads to a failure message box.

Answer: Errors while running group (enrichment) analysis can be caused b
problems:

a) Organism specific data (Fingerprint\annotation files) is not in the right directory.
specific data zip should be extracted into the Expander/organisms/ direct
download the relevant data by selecting from the menu: *Help >> Download Data*

b) Data contains elements that do not appear in the background set (this is only rel
background set is loaded from an external file).

Enrichment analysis produces no results (the resulting view is empty).

Answer*:* This can be caused by one of the following:

a) You are using the wrong conversion file or a conversion file that is not in the
does not map the probes to the expected type of gene Ids. The conversion f
probe ID in your data file to a gene ID that is used for enrichment analysis. A co

required when the probe Ids in your data file do not match the ones in the enric
example annotation and TF_fingerprint files that we supply).

b) You did not set the organism field in the input dialog to the organism type of y

c) You are trying to analyze only one set (e.g. the filtered data set) which you ar
background (in this case the analysis has no meaning since it is trying to detect
the cluster/bicluster in comparison to the background set).

d) You set the threshold p-value to be too strict (low).

e) Biological reason i.e., there is nothing to report regarding this specific cluster
solution or this gene sets data.

How can I save the bar charts produced by Expander, displaying the enrichment

When the results tab is open, please go to *File >> Save As Image*.

Why do certain Transcription Factors have a few accession numbers or\and a fe

Answer: The transcription factors (TFs) found enriched by Prima are presented
way:

Accession Num. in TRANSFAC DB [TF name]. For example, M00287[NF-Y]

It is possible that a TF will have a few accession numbers in TRANSFAC, w
different PWMs (position weight matrices specify the probability for observing e
at each position of the binding site, based on a set of empirically validated bind
respective TF).

It is also possible that a TF will have a few Entrez gene IDs, since a TF may be
few proteins. For example, NF-Y is a trimer, composed of 3 subunits.

Can I run Expander on Mac OS?

Expander is not designed for Mac OS. You can probably use it partially –
running its features that require the execution of exe files (CLICK, SAMBA
and PRIMA). The exe files are only suitable for Windows and Linux / Unix

When I try to save a session, Expander fails and returns an XStream error mess

If you are using java version 1.7, please switch to version 6. We currently have
this problem, that occurs with java1.7 and XStream, which is an external packa
using. We will do our best to resolve it in the coming future. In order to
Expander.bat files to use a java 6 version do the following:

1) Make sure that jave 6 (or 5) is installed on your PC by exploring the “Program Fi
(or “Program Files/Java”) directory. In it there should be a subdirectory by the na
(otherwise please install java6 from http://java.sun.com/javase/downloads/index.
2) In the Expander directory right click on one of the Expander.bat files (the one
and select “Edit”.

3) In the file type the path of the java6 exe file instead of the word java. E.g. C:\Program Files (x86)\Java\jre6 then put the text: "C:\Pr (x86)\Java\jre6\bin\java.exe" (including the quotes (“”)) instead of the word java.
4) Remove the text "–client" from the file
5) Save and close the file

This section will be updated as we get user feedbacks and problems.

Please refer all questions/comments to Expander@cs.tau.ac.il.

Copyrights Information

References

R: R Development Core Team (2008). R: A language and environment computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900 http://www.R-project.org

affy package for CEL files preprocessing: Gautier, L., Cope, L., Bolstad, B. M R. A. 2004. affy---analysis of Affymetrix GeneChip data at the probe level. Bioin (Feb. 2004), 307-315

gcrma R package: Jean(ZHIJIN) Wu and Rafael Irizarry with contribution MacDonald Jeff Gentry (). gcrma: Background Adjustment Using Sequence package version 2.14.1.

Quantile normalization: Bolstad, B. M. Irizarry, R. A. Astrand, M. and Sp Comparison of Normalization Methods for High Density Oligonucleotide Array I Variance and Bias. Bioinformatics 19(2):185-193, 2003

Non-linear baseline normalization: Schadt, E., C. Li, B. Eliss, and W. H. V extraction and normalization algorithms for high-density oligonucleotide gene ex data. J. Cell. Biochem. 84(S37),120–125, 2002

SAM (Significance Analysis of Microarray):

V. Tusher., R. Tibshirani., and G. Chu. Significance analysis of microarrays appli ionizing radiation response. PNAS, 98: 5116-5121, 2001

R. Tibshirani, G. Chu, T. Hastie and Balasubramanian Narasimhan (). samr: SAI Analysis of Microarrays. R package version 1.26. http://www-stat.stanford.edu/~t

K-Means clustering algorithm: Tavazoie, S., Hughes, J. D., Campbell, M. J., C Church, G. M. Systematic determination of genetic network architecture. Nat Ge 285, 1999

SOM clustering algorithm: Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., K Dmitrovsky, E., Lander, E. S., and Golub, T. R. Interpreting patterns of gene e self-organizing maps: methods and application to hematopoietic differentiation. Sci U S A, 96, 2907-2912, 1999

CLICK clustering algorithm: Sharan, R. and Shamir, R. CLICK: a clustering applications to gene expression analysis. Proc Int Conf Intell Syst Mol Biol 8, 307

ISA Biclustering algorithm: Bergmann, S., Ihmels, J., Barkai, N.
Iterative signature algorithm for the analysis of large-scale gene expression data Stat Nonlin Soft Matter Phys 2003 Mar; 67(3 Pt 1)

SAMBA biclustering algorithm: Tanay, A. Sharan, R. and Shamir, R. Discover significant biclusters in gene expression data. Bioinformatics, 18(1), 136-144, 20

Matisse network grouping: Ulitsky, I. and Shamir, R. MATISSE: Identificatio modules using network topology and high-throughput data. BMC Systems Biolo (2007)

Degas network grouping:

Ulitsky, I., Karp , R.M. and Shamir, R.
Detecting Disease-Specific Dysregulated Pathways Via Analysis of Clinical Expr *Proceedings of RECOMB 2008*, pp. 347--359, LNBI 4955, Springer, Berlin, (200

Context Scores for miRNA enrichment analysis: Grimson, A., Kai-How Farh,

Johnston, W., Garrett-Engele, F., P Lim, L., P Bartel, D. MicroRNA Targeting Sp
Mammals: Determinants beyond Seed Pairing. Molecular Cell, 27:91-105 (2007)
PRIMA algorithm: Elkon, R., Linhart, C. Sharan, R. Samir, R. and Shiloh, Y. Ge
Silico Identification of Transcriptional Regulators Controlling the Cell Cycle in
Genome Research, Vol. 13(5), pp. 773-780, 2003.

Spike software and DB: R. Elkon, R. Vesterman, N. Amit, I. Ulitsky, I. Zohar
Mass, N. Orlev, G. Sternberg, R. Blekhman, J. Assa, Y. Shiloh and R.Shar
database, visualization and analysis tool of cellular signaling pathways. BMC
*2008, **9:**11*

Spike home page: http://www.cs.tau.ac.il/~spike/

Agglomerative algorithm for hierarchical clustering: Eisen, M. B., Spellma
Cluster analysis and display of genome-wide expression patterns. Proc Natl A
95(25), 14863-8, 1998

**TF binding site profiles that were used to generate the supplied yeast
files:** Harbison, C.T., D.B. Gordon, T.I. Lee, N.J. Rinaldi, K.D. Macisaac, T.W.
Hannett, J.B. Tagne, D.B. Reynolds, J. Yoo, E.G. Jennings, J. Zeitlinger, D.K
Kellis, P.A. Rolfe, K.T. Takusagawa, E.S. Lander, D.K. Gifford, E. Fraenkel, ar
Transcriptional regulatory code of a eukaryotic genome. Nature 431: 99-104, 20(

expressionData1.txt sample input file: Spellman, P. T., Sherlock, G., et al. C
identification of cell cycle-regulated genes of the yeast Saccharomyces
microarray hybridization. Mol Biol Cell 9(12), 3273-97, 1998

Gene Set Enrichment Analysis: Aravind Subramanian, Pablo Tamayo, Vamsi
Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Sc
Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichme
knowledge-based approach for interpreting genome-wide expression profiles. PN
no. 43, 15545–15550, 2005