Expander 7.2 Online Documentation

Introduction

EXPANDER (EXpression Analyzer and DisplayER) is a java-based tool for analysis of gene expression data. It is capable of (1) preprocessing (2) visualizing (3) clustering (4) biclustering and (5) performing downstream analysis of clusters and biclusters such as functional enrichment and promoter analysis (i.e. analysis of gene groups for enrichment of transcription factor binding sites in their promoters).

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

EXPANDER versions are available for Windows OS and for Linux/Unix OS and require the pre-installation of the Java Runtime Environment (JRE) 5.0 or later (Expander 6.05 is the first version that fully supports java 1.7). The Java Runtime Environment can be installed via: http://java.sun.com/javase/downloads/index.jsp.

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

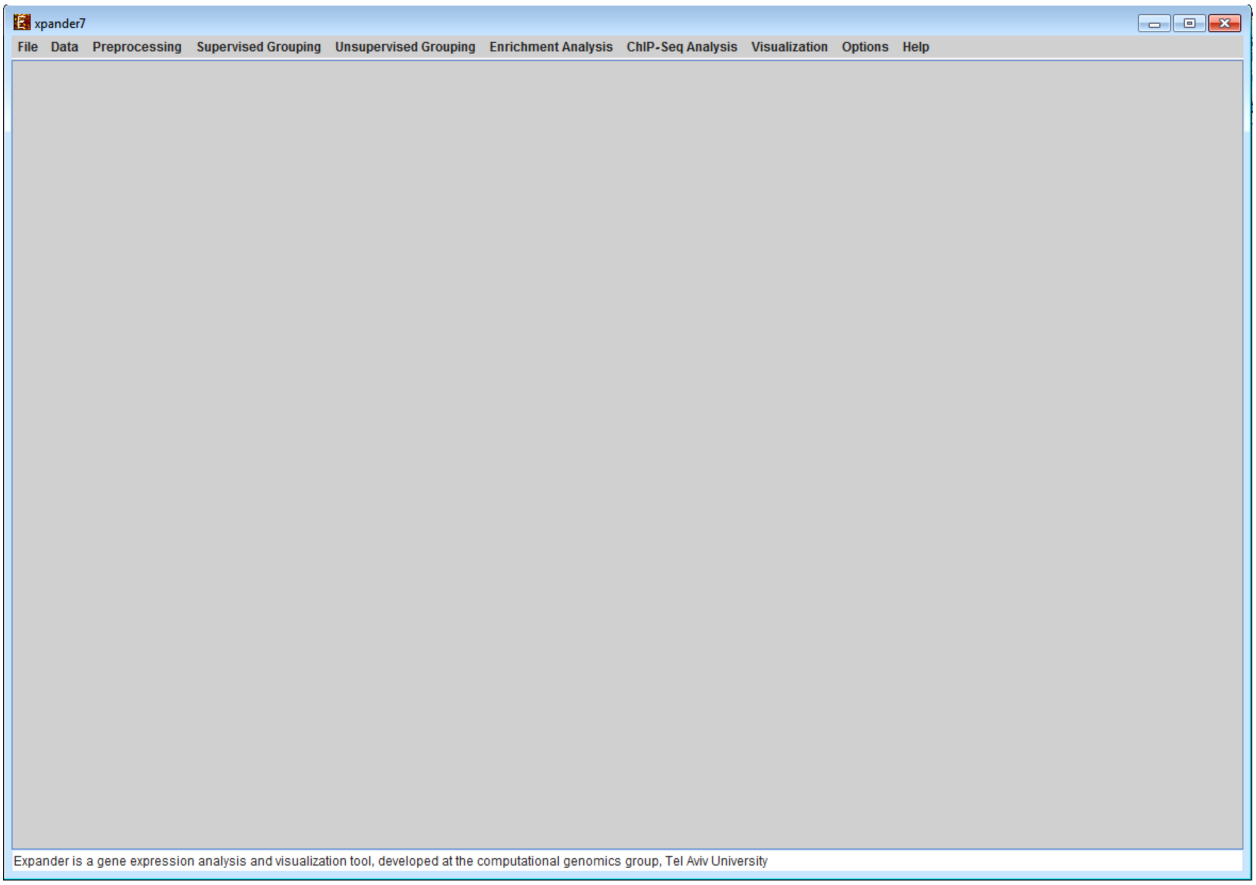
For improved network visualization please install Cytoscape 3.1.1 or higher. Cytoscape can be installed via http://www.cytoscape.org/download.php

Starting EXPANDER

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

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

Upon running the program, the main menu bar appears:



Expander is a gene expression analysis and visualization tool, developed at the computational genomics group, Tel Aviv University

Input Data

Expander operates on the following types of data:

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

Values can be either relative intensities data, expected as log 2 (R/G) values data (e.g. cDNA microarrays), RNA-Seq counts OR absolute intensities data, expected as positive expression levels (E.g. High-density oligonucleotide data). Oligonucleotide data can be loaded with/without detection calls. 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 use TopHat (http://tophat.cbcb.umd.edu/tutorial.html) to align the sequenced reads to the relevant genome, and then use Cufflinks (http://cufflinks.cbcb.umd.edu/howitworks.html) or HTSeq (http://www-huber.embl.de/users/anders/HTSeq/doc/count.html#count) to obtain gene (or transcript) expression estimates from TopHat output.

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

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

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

For details regarding the Gene ID convention that is used for each organism, refer to the Supplied Files section.

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

d) **Gene Ranking Analysis** (GSEA) – contains predefined ranked genes. In this data type, the conventional gene Ids that are used by EXPANDER in the Gene Set Enrichment Analysis are expected.

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

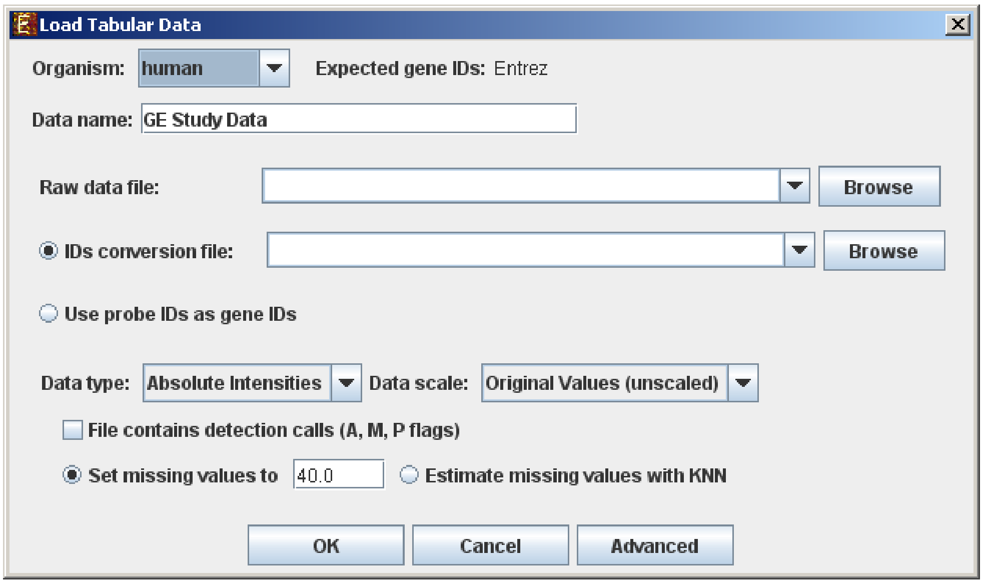
e) **ChIP-Seq data** – contains discovered peaks of Transcription Factor(TF) ChIP-Seq experiment in BED or GFF3 formats (see File Formats section).

Loading gene expression data:

Tabular Data File

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

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



Load Tabular Data

Organism: human

Expected gene IDs: Entrez

Data name: GE Study Data

Raw data file:

● IDs conversion file:

○ Use probe IDs as gene IDs

Data type: Absolute Intensities

Data scale: Original Values (unscaled)

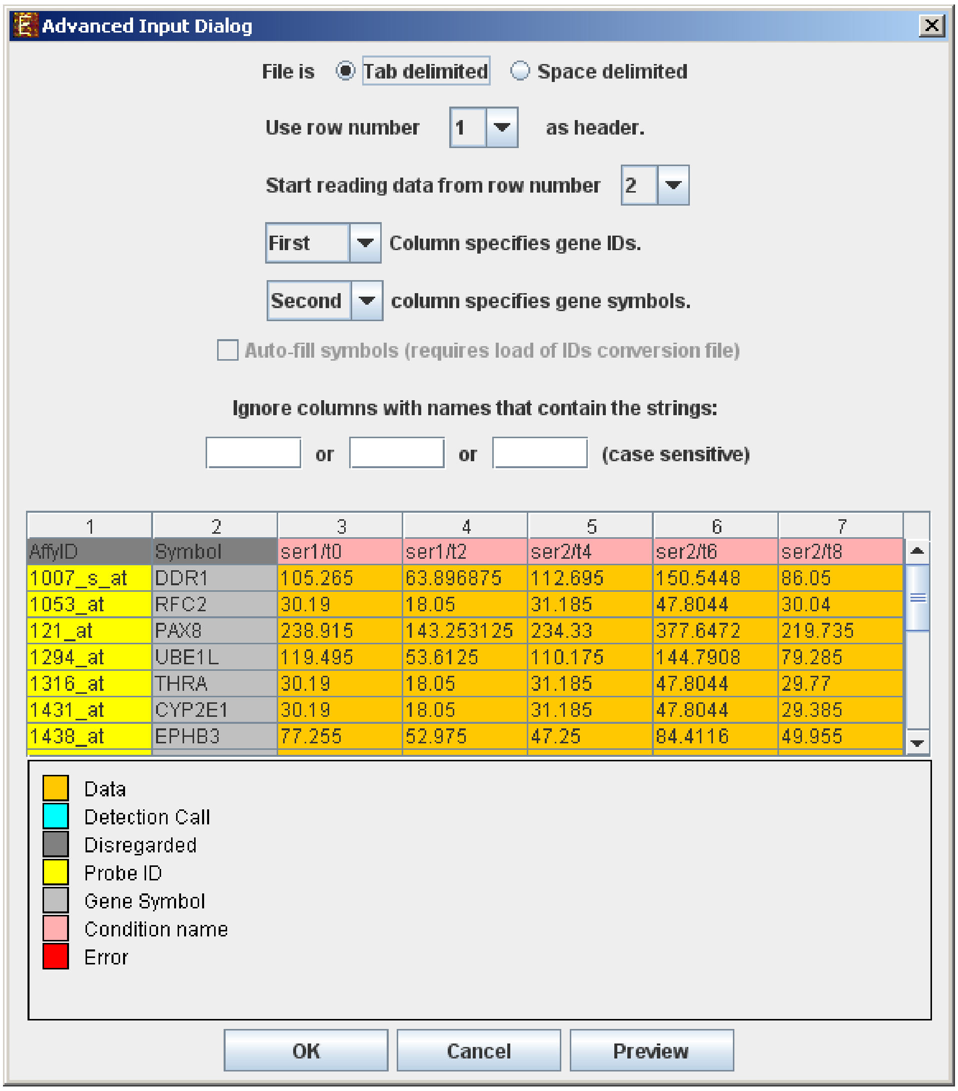
□ File contains detection calls (A, M, P flags)

● Set missing values to 40.0

○ Estimate missing values with KNN

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

Advanced Input Dialog: Upon pressing the ‘Advanced’ button after filling the ‘Raw Data File’ field, an ‘Advanced Input Dialog’ appears. This dialog box can be used in order to facilitate the data load of files that are not in the required format. The first few rows and columns of the data are displayed in a table, demonstrating the way the data is read by the program according to the current input values.



Advanced Input Dialog

File is ● Tab delimited ○ Space delimited

Use row number 1 as header.

Start reading data from row number 2

First Column specifies gene IDs.

Second column specifies gene symbols.

□ Auto-fill symbols (requires load of IDs conversion file)

Ignore columns with names that contain the strings:

[ ] or [ ] or [ ] (case sensitive)

1	2	3	4	5	6	7
AffyID	Symbol	ser1/t0	ser1/t2	ser2/t4	ser2/t6	ser2/t8
1007_s_at	DDR1	105.265	63.896875	112.695	150.5448	86.05
1053_at	RFC2	30.19	18.05	31.185	47.8044	30.04
121_at	PAX8	238.915	143.253125	234.33	377.6472	219.735
1294_at	UBE1L	119.495	53.6125	110.175	144.7908	79.285
1316_at	THRA	30.19	18.05	31.185	47.8044	29.77
1431_at	CYP2E1	30.19	18.05	31.185	47.8044	29.385
1438_at	EPHB3	77.255	52.975	47.25	84.4116	49.955

🟧 Data
🟦 Detection Call
⬛ Disregarded
🟨 Probe ID
⬜ Gene Symbol
🟫 Condition name
🟥 Error

CEL Files

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

The load of CEL requires installation of R software (see R External Application section) along with specific packages, as detailed below. An open internet connection is also required for this operation.

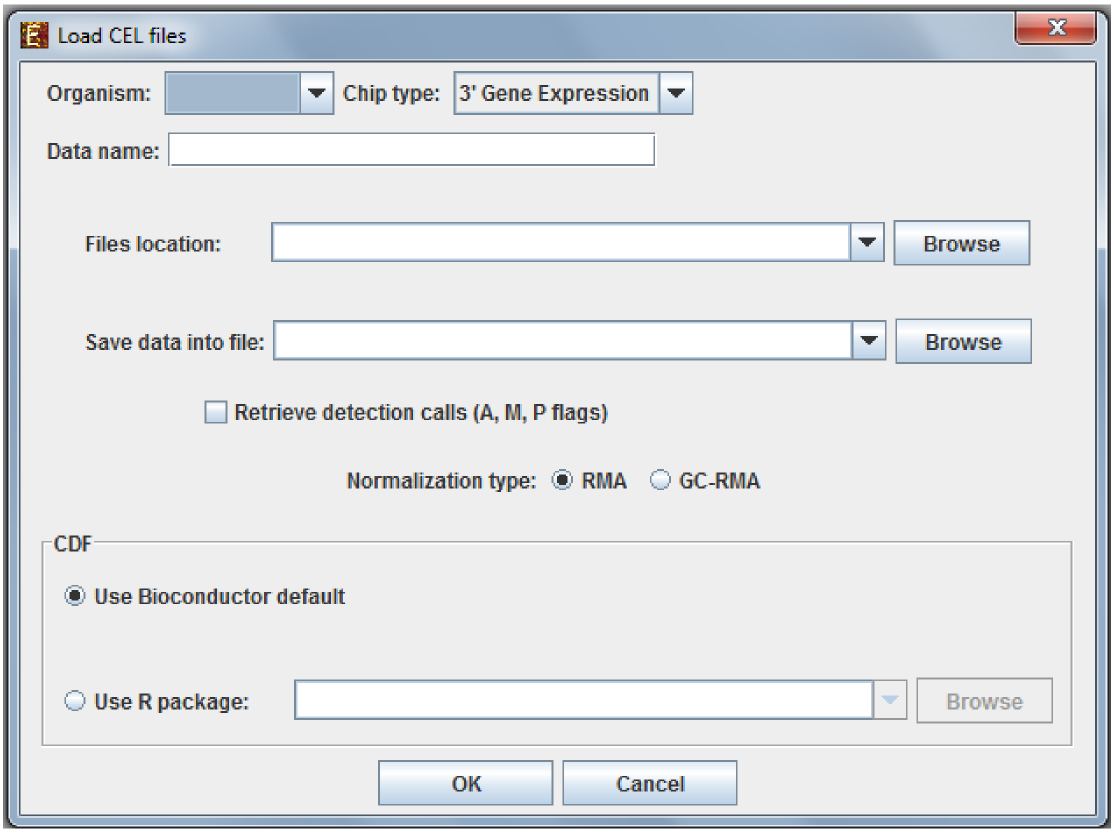
Expander supports CEL files of three chip types:

1. **3' Gene Expression** - requires Bioconductor “affy” package

2. **Whole-Transcript Gene Expression** (Gene 1.0 chips) – requires the prior installation of a cdf package for the used chip (see links below).

3. **Alternative Splicing** (Exon 1.0 chips) requires the prior installation of a cdf package for the used chip (see links below). * Please note that we estimate the overall expression for the transcript, not exon-by-exon. Therefore, this becomes 'gene data' rather than 'alternative splicing data'.

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



Load CEL files

Organism:

Chip type: 3' Gene Expression

Data name:

Files location:

Save data into file:

☐ Retrieve detection calls (A, M, P flags)

Normalization type: ● RMA ○ GC-RMA

CDF

● Use Bioconductor default

○ Use R package:

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

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

CDF environment choice: You may use the default Bioconductor CDF environment for the chips or browse to an alternative CDF package which you have already installed in R. For whole transcript and alternative splicing chips (for which there is no default Bioconductor CDF environment), you will need to supply an alternative CDF package (see links below).

Note: GC-RMA requires the probe sequence information of the chip. If you decide not to use the default Bioconductor CDF environment, and have GC-RMA as the preprocessing method, you must have the suitable probe package installed in addition to the CDF alternative package.

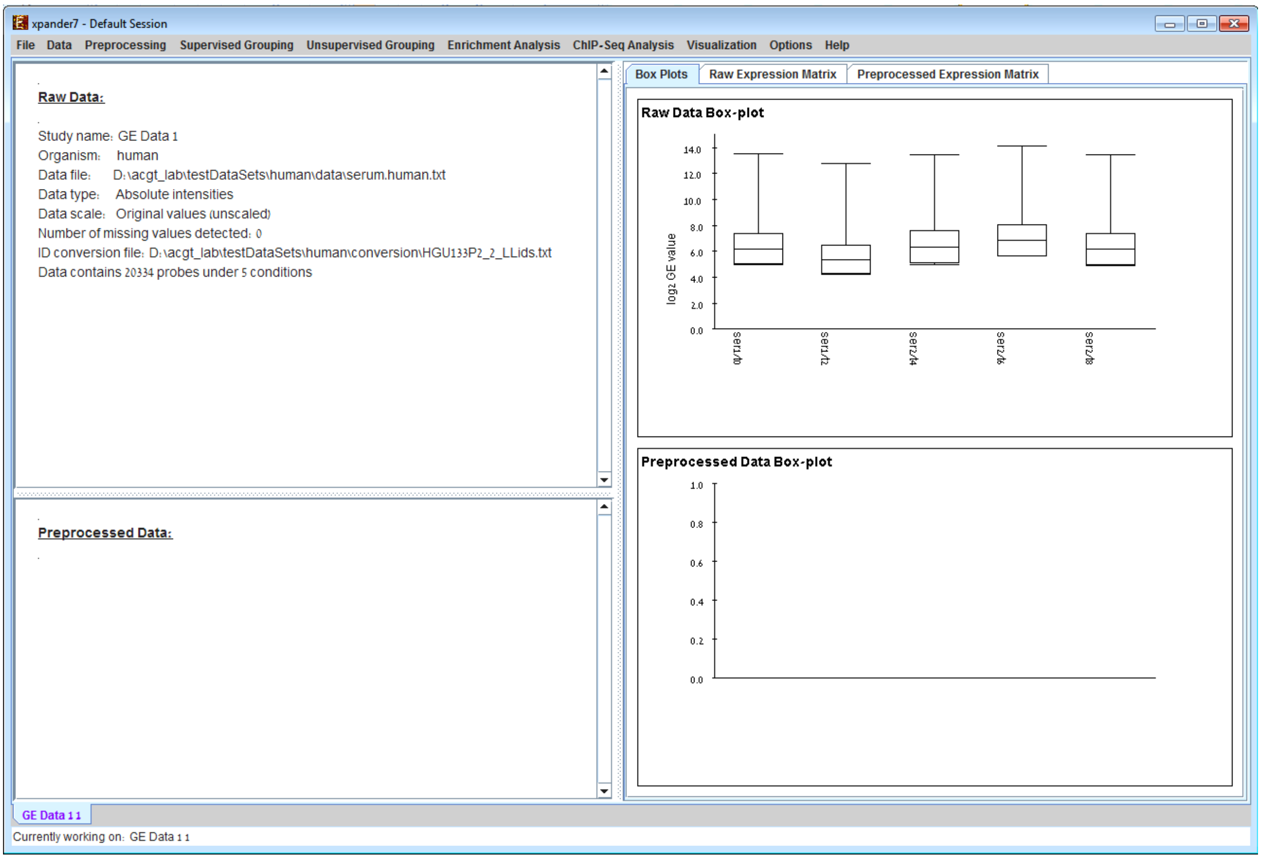
Link for downloading CDF environment packages (for 2nd option):

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

If Expander cannot find your R software, a window will appear, asking you to specify its location. Please browse to the location of your R software. In Windows, R.exe file is likely to be located in the 'bin' folder of R software. In Linux, you may type 'which R' in the command line to find R path. If you have a few versions of R installed, please make sure to point Expander to a version in which the Bioconductor “affy” package has been installed.

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

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


xpander7 - Default Session
File Data Preprocessing Supervised Grouping Unsupervised Grouping Enrichment Analysis ChIP-Seq Analysis Visualization Options Help
Box Plots
Raw Expression Matrix
Preprocessed Expression Matrix
Raw Data Box-plot
14.0
12.0
10.0
8.0
6.0
4.0
2.0
0.0
log2 GE value
ser1/t0
ser1/t2
ser2/t4
ser2/t6
ser2/t8
Preprocessed Data Box-plot
1.0
0.8
0.6
0.4
0.2
0.0
GE Data 1 1

Raw Data:

Study name: GE Data 1
Organism: human
Data file: D:\acgt_lab\testDataSets\human\data\serum.human.txt
Data type: Absolute intensities
Data scale: Original values (unscaled)
Number of missing values detected: 0
ID conversion file: D:\acgt_lab\testDataSets\human\conversion\HGU133P2_2_LLids.txt
Data contains 20334 probes under 5 conditions

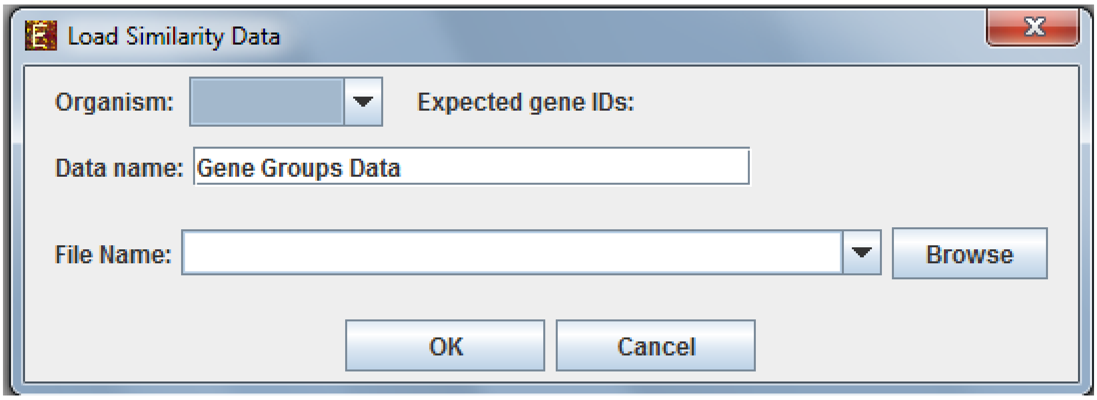
Preprocessed Data:

Currently working on: GE Data 1 1

Working on similarity data no associated expression data

To start working on similarity data (no expression data associated) select File>>New Session>> Similarity Data...

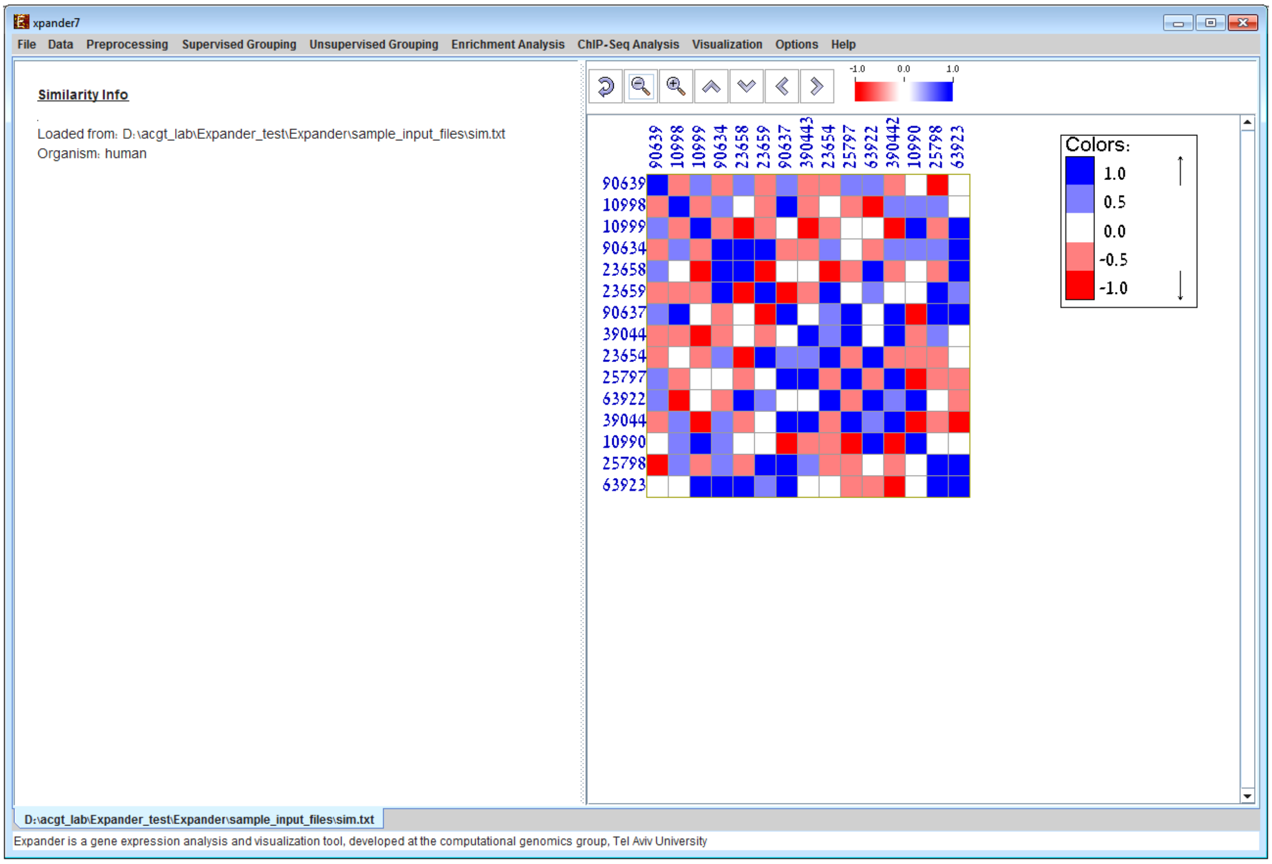
The following dialog box will appear:



Data name: Gene Groups Data

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 window



Similarity Info

Loaded from: D:\acgt_lab\Expander_test\Expander\sample_input_files\sim.txt
Organism: human

D:\acgt_lab\Expander_test\Expander\sample_input_files\sim.txt

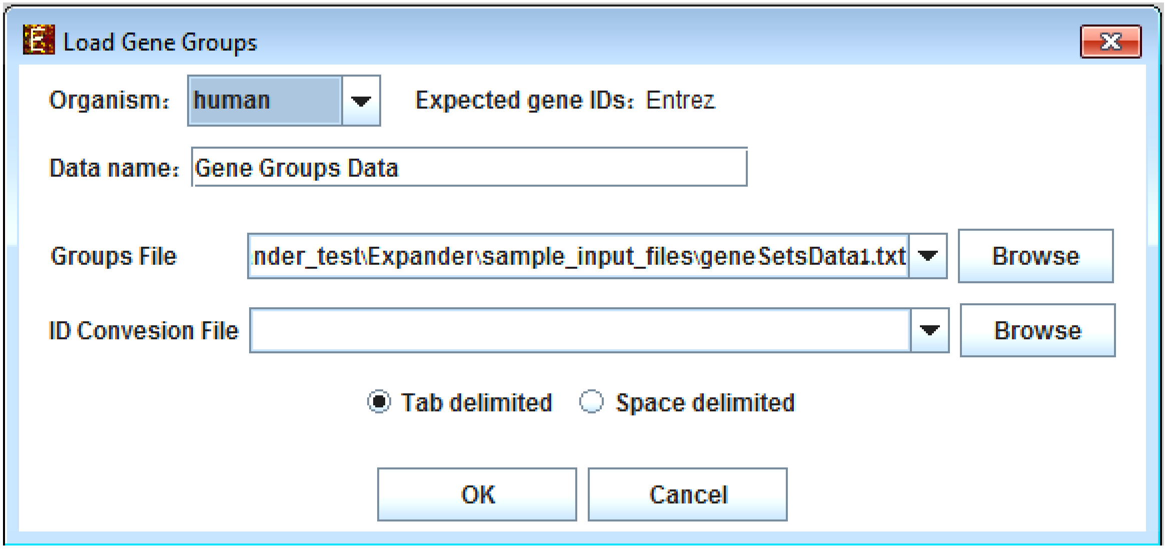
Expander is a gene expression analysis and visualization tool, developed at the computational genomics group, Tel Aviv University

Currently similarity data can only be clustered using the Hierarchical clustering procedure by selecting *Unsupervised Grouping>>Hierarchical Clustering>>Cluster...* The resulting tree can 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>>New Session*. From the submenu select *Gene Groups*.

The following dialog box will appear:



Load Gene Groups

Organism: human

Expected gene IDs: Entrez

Data name: Gene Groups Data

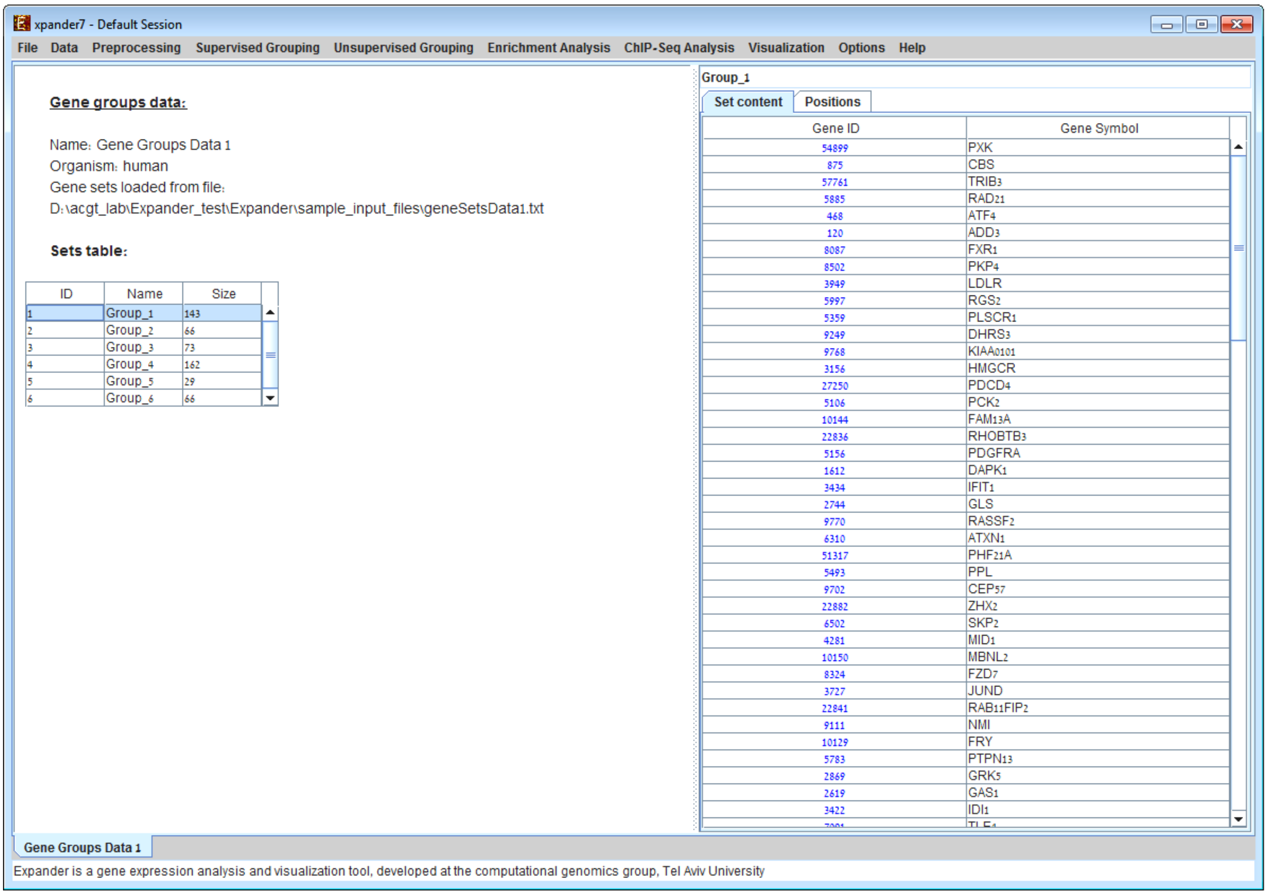
Groups File nder_test\Expander\sample_input_files\geneSetsData1.txt

ID Convesion File

● Tab delimited ○ Space delimited

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 main window (see example below). It contains information regarding the data file, and a table describing the different groups (serial number, name and size). Group names can be modified, by editing the corresponding cell in the table. Upon clicking on a row in the table, the corresponding group pane appears on the right. It contains a list of the genes in the group and a view of their chromosomal positions. If a network file has been loaded (via *Data>>Load Network*), the sub-graph, induced by the group is displayed as well.


xpander7 - Default Session
File Data Preprocessing Supervised Grouping Unsupervised Grouping Enrichment Analysis ChIP-Seq Analysis Visualization Options Help
Gene Groups Data 1

Gene groups data:

Name: Gene Groups Data 1
Organism: human
Gene sets loaded from file:
D:\acgt_lab\Expander_test\Expander\sample_input_files\geneSetsData1.txt

Sets table:

ID	Name	Size
1	Group_1	143
2	Group_2	66
3	Group_3	73
4	Group_4	162
5	Group_5	29
6	Group_6	66

Group_1

Set content | Positions

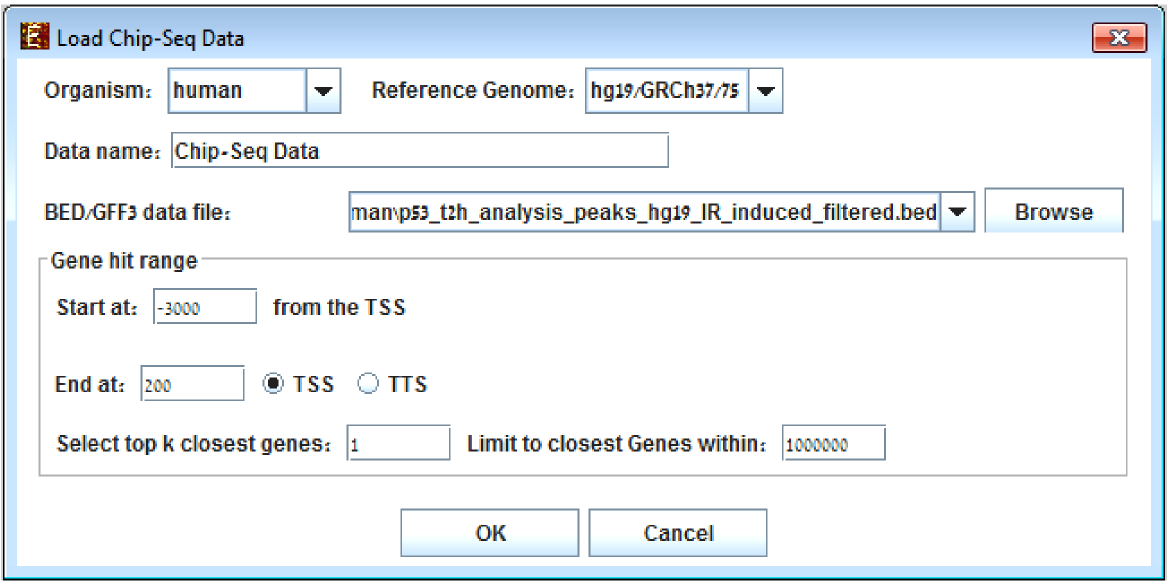
Gene ID	Gene Symbol
54899	PXK
875	CBS
57761	TRIB3
5885	RAD21
468	ATF4
120	ADD3
8087	FXR1
8502	PKP4
3949	LDLR
5997	RGS2
5359	PLSCR1
9249	DHRS3
9768	KIAA0101
3156	HMGCR
27250	PDCD4
5106	PCK2
10144	FAM13A
22836	RHOBTB3
5156	PDGFRA
1612	DAPK1
3434	IFIT1
2744	GLS
9770	RASSF2
6310	ATXN1
51317	PHF21A
5493	PPL
9702	CEP57
22882	ZHX2
6502	SKP2
4281	MID1
10150	MBNL2
8324	FZD7
3727	JUND
22841	RAB11FIP2
9111	NMI
10129	FRY
5783	PTPN13
2869	GRK5
2619	GAS1
3422	IDI1

Expander is a gene expression analysis and visualization tool, developed at the computational genomics group, Tel Aviv University

ChIP-Seq Data

To load ChIP-Seq data, select: *File >> New Session*. From the submenu select *"ChIP-Seq Data"*.

The following dialog box will appear:



Organism: human
Reference Genome: hg19/GRCh37/75
Data name: Chip-Seq Data
BED/GFF3 data file: man\p53_t2h_analysis_peaks_hg19_IR_induced_filtered.bed

Start at: -3000 from the TSS
End at: 200 ● TSS ○ TTS
Select top k closest genes: 1 Limit to closest Genes within: 1000000

Select the organism and its reference genome that correspond to the ChIP-Seq experiment.

For details regarding the data BED/GFF3 files formats see the File Formats section.

Gene hit range:

This option allows choosing the gene range to be searched for peak hit.

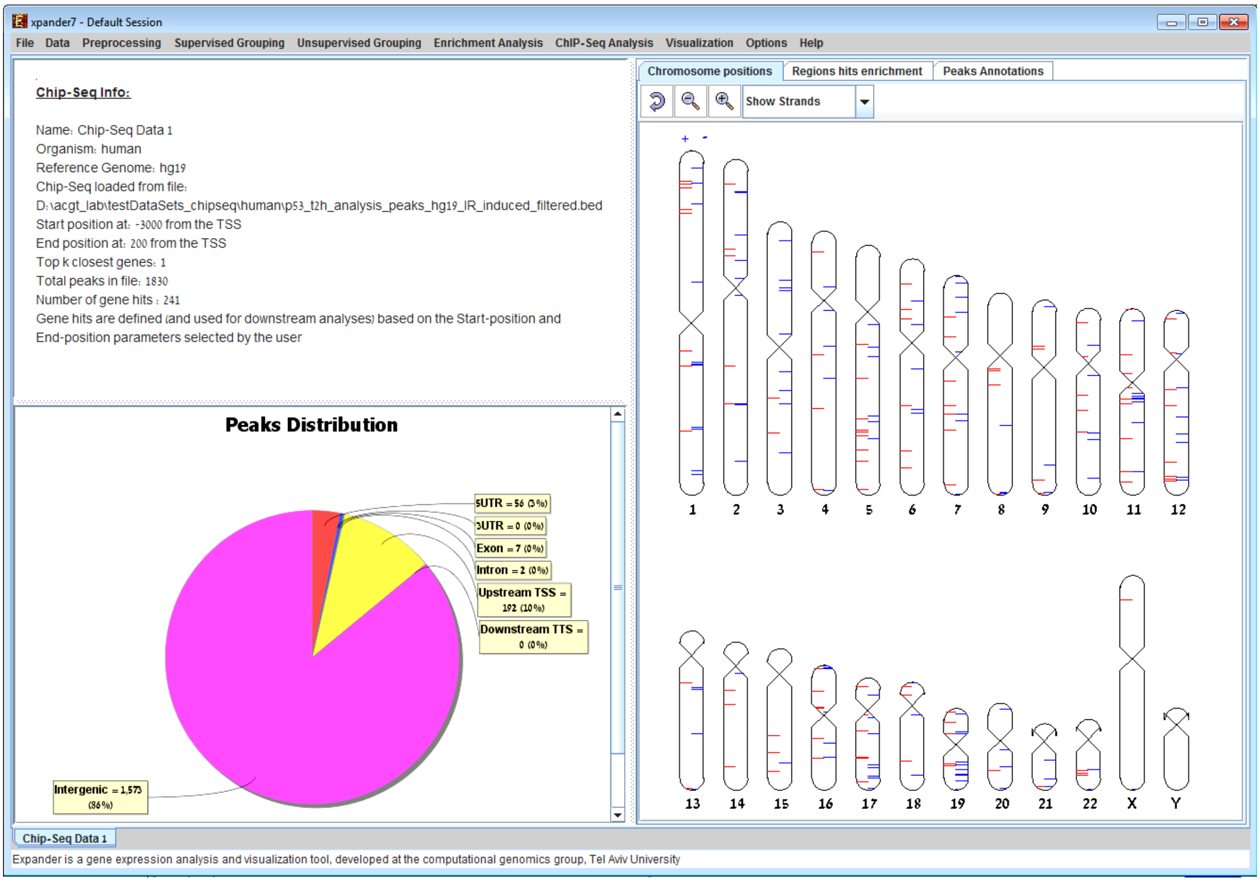
The range is selected in the following way:

Start position upstream to the Transcript Start Site (TSS) and end position downstream to the TSS or to the Transcript Termination Site (TTS).

Note that the start position is a non-positive value.

There is an option to select more than 1 closest gene to peak by changing "Selecting top k closest genes" field and to set a distance bound for k>1 closest gene to peak.

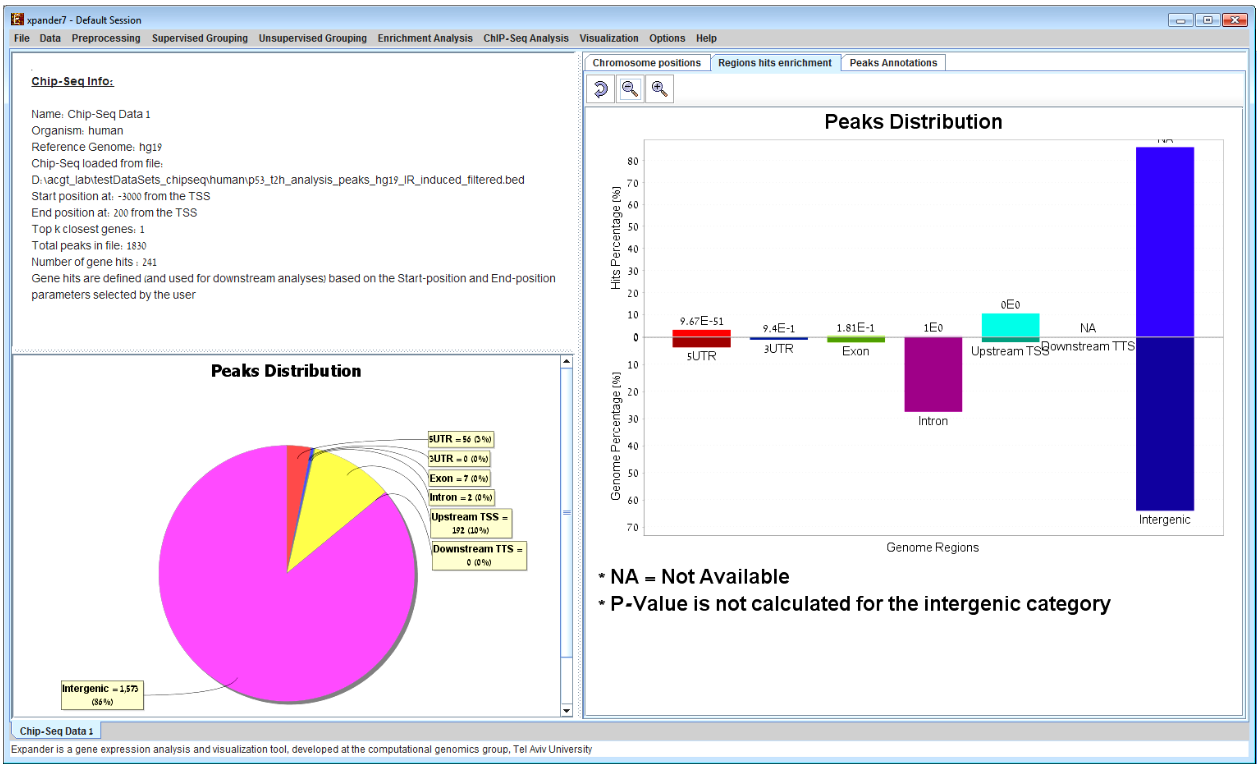
After loading a ChIP-Seq data, a ‘ChIP-Seq Study Data’ display tab is added to the main window (see example below). It contains information regarding the raw data file, a pie chart, a chromosome visualization of the mapped genes positions, a region hits enrichment bar chart, and a peaks annotations table. The pie chart contains the peaks distribution of the first closest gene hits. Each peak was mapped to the closest gene with regard to the TSS and mapped to one of the following regions: Upstream of the TSS, 5UTR, Exon, Intron, 3UTR, Downstream of the TTS or Intergenic (i.e., regions between genes). Chromosome visualization displays mapped genes to peaks with option to show or not to show the gene's strands. Region hits enrichments displays the enrichment test using binomial p-value was performed in order to evaluate the randomness of peaks falling in a specific region (excluding "Intergenic" region), for example, it can be seen that under 5UTR bar, peaks fall by random in this region with p-value 3.87E-133. This option is currently available only for human and mouse datasets.



Chip-Seq Info:

Name: Chip-Seq Data 1
Organism: human
Reference Genome: hg19
Chip-Seq loaded from file:
D:\acgt_lab\testDataSets_chipseq\human\p53_t2h_analysis_peaks_hg19_IR_induced_filtered.bed
Start position at: -3000 from the TSS
End position at: 200 from the TSS
Top k closest genes: 1
Total peaks in file: 1830
Number of gene hits : 241
Gene hits are defined (and used for downstream analyses) based on the Start-position and End-position parameters selected by the user

Expander is a gene expression analysis and visualization tool, developed at the computational genomics group, Tel Aviv University



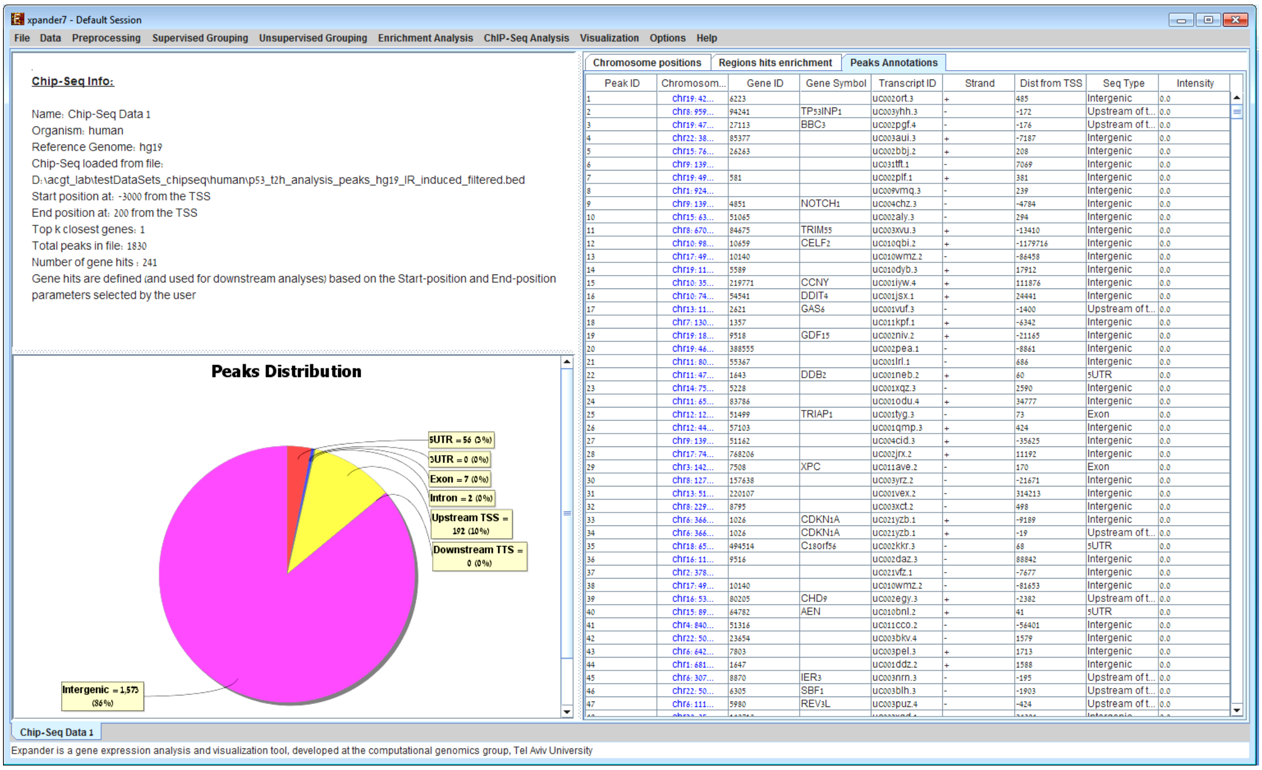
Chip-Seq Info:

Name: Chip-Seq Data 1
Organism: human
Reference Genome: hg19
Chip-Seq loaded from file:
D:\acgt_lab\testDataSets_chipseq\human\p53_t2h_analysis_peaks_hg19_IR_induced_filtered.bed
Start position at: -3000 from the TSS
End position at: 200 from the TSS
Top k closest genes: 1
Total peaks in file: 1830
Number of gene hits : 241
Gene hits are defined (and used for downstream analyses) based on the Start-position and End-position parameters selected by the user

Expander is a gene expression analysis and visualization tool, developed at the computational genomics group, Tel Aviv University

The peaks annotation table (see below) displays details for each peak – Peak ID (row number in ChIP-Seq file), Chromosome Position of the peak with a link to UCSC genome browser, Gene ID of the mapped gene to peak (blank if the peak was not mapped to a gene), Gene symbol, UCSC Transcript ID, Strand, Distance from TSS (negative/positive for upstream/downstream), Sequence type for the peak's mapped region and Intensity" or "Q-value" depending on the uploaded file format (BED or GFF3).

Pie chart, chromosome visualization and Region hits enrichment bar chart can be increased or decreased using the mouse scroll wheel.



Chip-Seq Info:

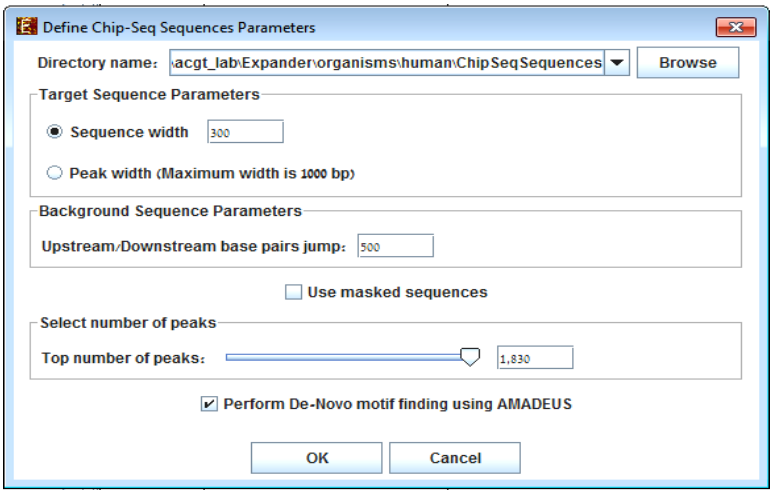
Name: Chip-Seq Data 1
Organism: human
Reference Genome: hg19
Chip-Seq loaded from file:
D:\acgt_lab\testDataSets_chipseq\human\p53_t2h_analysis_peaks_hg19_IR_induced_filtered.bed
Start position at: -3000 from the TSS
End position at: 200 from the TSS
Top k closest genes: 1
Total peaks in file: 1830
Number of gene hits : 241
Gene hits are defined (and used for downstream analyses) based on the Start-position and End-position parameters selected by the user

Peak ID	Chromosom...	Gene ID	Gene Symbol	Transcript ID	Strand	Dist from TSS	Seq Type	Intensity
1	chr19: 42...	6223		uc002ort.3	+	485	Intergenic	0.0
2	chr8: 959...	94241	TP53INP1	uc003yhh.3	-	-172	Upstream of t...	0.0
3	chr19: 47...	27113	BBC3	uc002pgf.4	-	-176	Upstream of t...	0.0
4	chr22: 38...	85377		uc003aui.3	+	-7187	Intergenic	0.0
5	chr15: 76...	26263		uc002bbj.2	+	208	Intergenic	0.0
6	chr9: 139...			uc031tft.1	-	7069	Intergenic	0.0
7	chr19: 49...	581		uc002plf.1	+	381	Intergenic	0.0
8	chr1: 924...			uc009vmq.3	-	239	Intergenic	0.0
9	chr9: 139...	4851	NOTCH1	uc004chz.3	-	-4784	Intergenic	0.0
10	chr15: 63...	51065		uc002aly.3	-	294	Intergenic	0.0
11	chr8: 670...	84675	TRIM55	uc003xvu.3	+	-13410	Intergenic	0.0
12	chr10: 98...	10659	CELF2	uc010qbi.2	+	-1179716	Intergenic	0.0
13	chr17: 49...	10140		uc010wmz.2	-	-86458	Intergenic	0.0
14	chr19: 11...	5589		uc010dyb.3	+	17912	Intergenic	0.0
15	chr10: 35...	219771	CCNY	uc001iyw.4	+	111876	Intergenic	0.0
16	chr10: 74...	54541	DDIT4	uc001jsx.1	+	24441	Intergenic	0.0
17	chr13: 11...	2621	GAS6	uc001vuf.3	-	-1400	Upstream of t...	0.0
18	chr7: 130...	1357		uc011kpf.1	+	-6342	Intergenic	0.0
19	chr19: 18...	9518	GDF15	uc002niv.2	+	-21165	Intergenic	0.0
20	chr19: 46...	388555		uc002pea.1	-	-8861	Intergenic	0.0
21	chr11: 80...	55367		uc001lrl.1	-	686	Intergenic	0.0
22	chr11: 47...	1643	DDB2	uc001neb.2	+	60	5UTR	0.0
23	chr14: 75...	5228		uc001xqz.3	-	2590	Intergenic	0.0
24	chr11: 65...	83786		uc001odu.4	+	34777	Intergenic	0.0
25	chr12: 12...	51499	TRIAP1	uc001tyg.3	-	73	Exon	0.0
26	chr12: 44...	57103		uc001qmp.3	+	424	Intergenic	0.0
27	chr9: 139...	51162		uc004cid.3	+	-35625	Intergenic	0.0
28	chr17: 74...	768206		uc002jrx.2	+	11192	Intergenic	0.0
29	chr3: 142...	7508	XPC	uc011ave.2	-	170	Exon	0.0
30	chr8: 127...	157638		uc003yrz.2	-	-21671	Intergenic	0.0
31	chr13: 51...	220107		uc001vex.2	-	314213	Intergenic	0.0
32	chr8: 229...	8795		uc003xct.2	-	498	Intergenic	0.0
33	chr6: 366...	1026	CDKN1A	uc021yzb.1	+	-9189	Intergenic	0.0
34	chr6: 366...	1026	CDKN1A	uc021yzb.1	+	-19	Upstream of t...	0.0
35	chr18: 65...	494514	C18orf56	uc002kkr.3	-	68	5UTR	0.0
36	chr16: 11...	9516		uc002daz.3	-	88842	Intergenic	0.0
37	chr2: 378...			uc021vfz.1	-	-7677	Intergenic	0.0
38	chr17: 49...	10140		uc010wmz.2	-	-81653	Intergenic	0.0
39	chr16: 53...	80205	CHD9	uc002egy.3	+	-2382	Upstream of t...	0.0
40	chr15: 89...	64782	AEN	uc010bnl.2	+	41	5UTR	0.0
41	chr4: 840...	51316		uc011cco.2	-	-56401	Intergenic	0.0
42	chr22: 50...	23654		uc003bkv.4	-	1579	Intergenic	0.0
43	chr6: 642...	7803		uc003pel.3	+	1713	Intergenic	0.0
44	chr1: 681...	1647		uc001ddz.2	+	1588	Intergenic	0.0
45	chr6: 307...	8870	IER3	uc003nrn.3	-	-195	Upstream of t...	0.0
46	chr22: 50...	6305	SBF1	uc003blh.3	-	-1903	Upstream of t...	0.0
47	chr6: 111...	5980	REV3L	uc003puz.4	-	-424	Upstream of t...	0.0

Expander is a gene expression analysis and visualization tool, developed at the computational genomics group, Tel Aviv University

Fetch ChIP-Seq peaks sequences

To fetch FASTA sequences for the peaks select ChIP-Seq Analysis >> Motif analysis >> Fetch ChIP-Seq Sequences. The following dialog box will appear:



Directory name: \acgt_lab\Expander\organisms\human\ChipSeqSequences

Target Sequence Parameters

● Sequence width 300

○ Peak width (Maximum width is 1000 bp)

Background Sequence Parameters

Upstream/Downstream base pairs jump: 500

□ Use masked sequences

Select number of peaks

Top number of peaks: 1,830

☑ Perform De-Novo motif finding using AMADEUS

Select the directory path where the created files will be inserted (Default folder is

<path to>/Expander/organisms/<Selected organism>/ChipSeqSequences/).

Sequence parameters:

Please see image demonstration below.

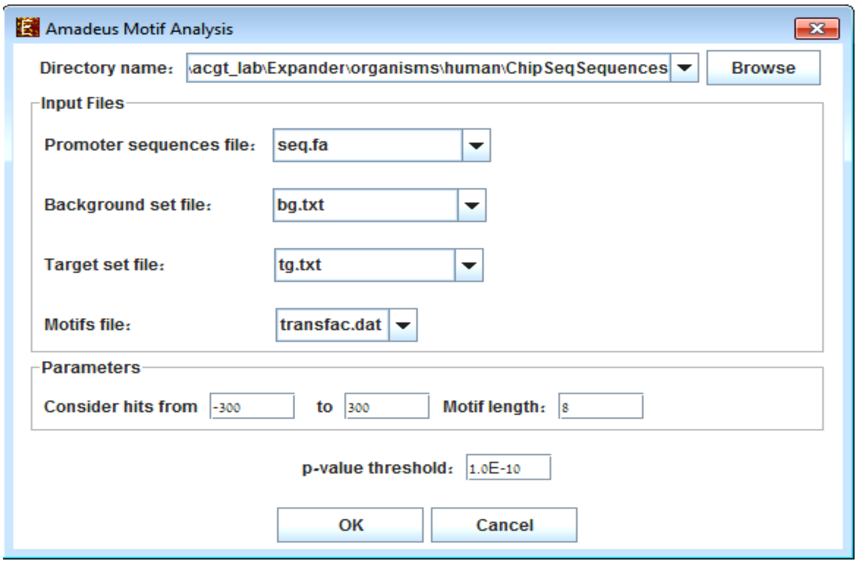
- Sequence Width - by selected width (Default is 300 bps) or by peak width.
- Upstream/Downstream base pairs jump (Default is 500 bps) – For each peak, two background sequences of length 300 bps are created 500 bps upstream and downstream from the middle position of the peak.
- Use masked sequences – Fetch sequences from repetitive masked FASTA genome.

Select number of peaks:

- Top number of peaks – top peaks are selected by their score (Intensity or q-value). If the no score was added then the peaks are selected according to their row line position in the file.
- Peak minimum score – available only in BED format. Peaks are selected if their score is above the selected minimum score.
- Maximum q-value – available only in GFF3 format. Peaks are selected if their q-value is below the selected maximum q-value.

Option to perform AMADEUS De-Novo motif finding:

If selected then after creating the sequences the following AMADEUS dialog box will appear:



Amadeus Motif Analysis

Directory name: \acgt_lab\Expander\organisms\human\ChipSeqSequences

Input Files

Promoter sequences file: seq.fa

Background set file: bg.txt

Target set file: tg.txt

Motifs file: transfac.dat

Parameters

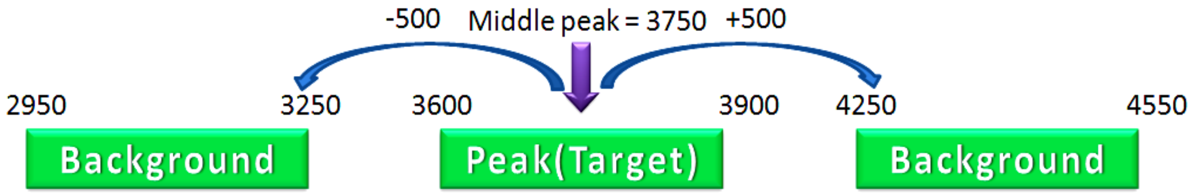
Consider hits from -300 to 300 Motif length: 8

p-value threshold: 1.0E-10

The created 3 files will be filled in the relevant 'Input Files' fields. Under 'Parameters' a search region upstream/downstream with respect to the middle peak and a motif length should be defined. After clicking 'OK', An Amadeus visualization will be created. For further explanation on the visualization please refer to AMADEUS section.

After clicking OK, 3 files will be created in the selected folder:

- sequences.fa – contains the FASTA sequences of the selected peaks and their two background sequences.
- target.txt – contains identifiers of the selected peaks.
- background.txt – contains identifiers of the selected peaks and their background identifiers.



Notes

Fetching non-masked sequences process should take on average ~1-2 seconds for about ~5000 sequences of length 300 bps.

Fetching masked sequences might take longer time – on average ~10 seconds for ~5000 sequences of length 300 bps.

Preprocessing GE Data

The following preprocessing operations can be performed using EXPANDER:

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

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

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

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

Normalization can be performed using the following schemes:

a) **Quantile normalization** (*Preprocessing >> Normalization >> Quantile*), in which the whole data is used.

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

For more details regarding the normalization schemes see the References section.

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

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

Probe filtration can be performed using the following schemes:

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

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

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

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

e) **Detection calls** (*Preprocessing >> Filter Probes >> Detection calls*): in this method probes/genes are filtered according to the number of expression signals for which the detection 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 Subset*): the filtered set is loaded from an external txt file (for details regarding the format please see the File Formats section).

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

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

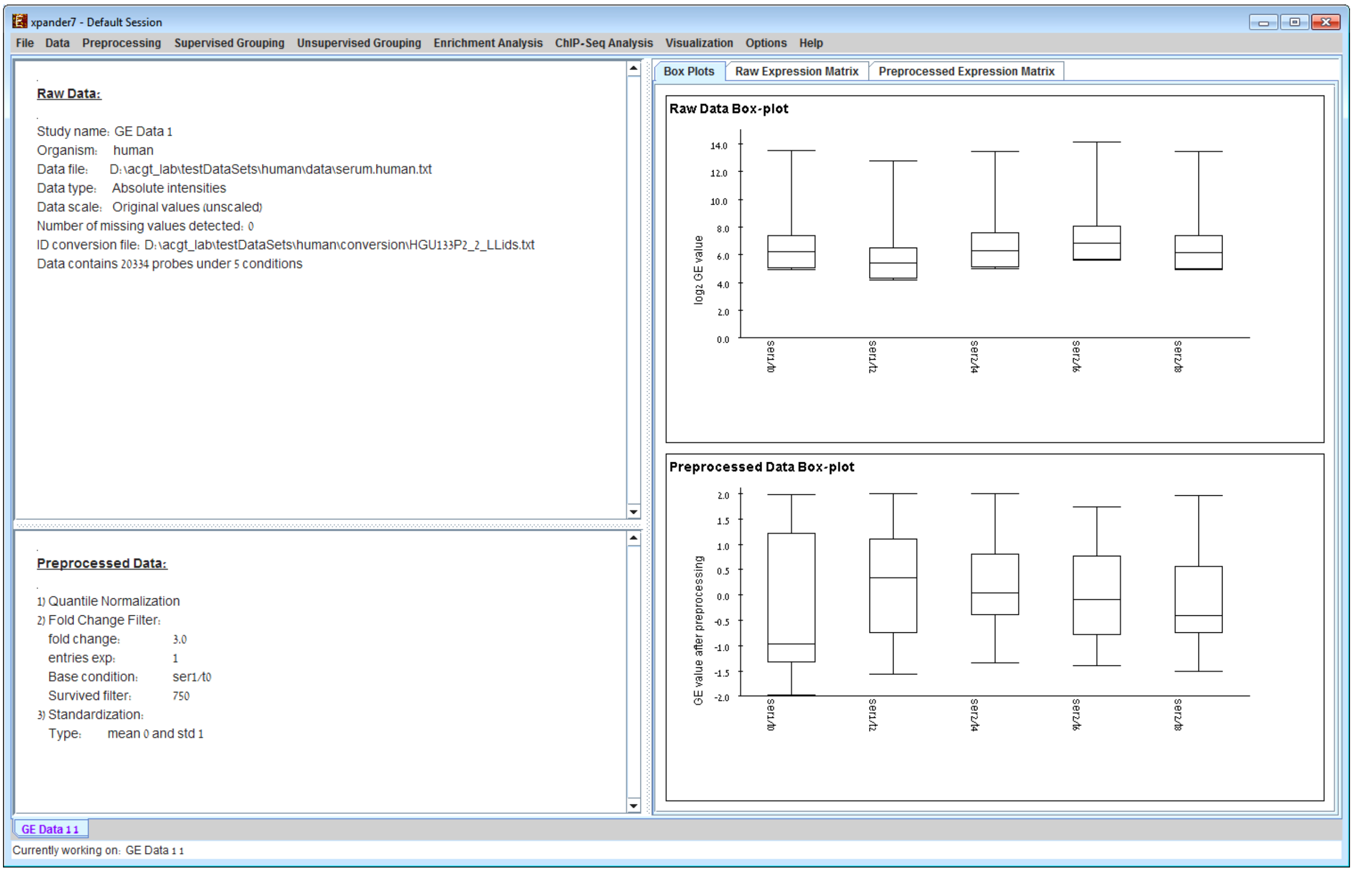
9) **Standardization**: When expression values between different genes are very different, but general expression patterns are similar (high Pearson Correlation values), we would expect to see this similarity when looking on a pattern display. Since the absolute values of expression are different, a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

Standardization can be performed using the following schemes:

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

b) **Fixed norm** (*Preprocessing >> Standardization >> Fixed Norm*) - normalizes each expression pattern to have a fixed norm i.e. expression levels are divided by the norm of that expression vector (the root of sum of squares of that vector). This method is appropriate when different mean values or variances are expected for different patterns (e.g. when working on conditions and expecting larger variance in later phases of a response.

After performing a preprocessing operation, the information regarding the operation is added to the ‘Preprocessed Data’ section in the ‘Session Data’ tab. In addition, the ‘Preprocessed Data box plot’ and ‘Preprocessed Expression Matrix’ are automatically updated according to the new values in the data.


xpander7 - Default Session
File Data Preprocessing Supervised Grouping Unsupervised Grouping Enrichment Analysis ChIP-Seq Analysis Visualization Options Help
Box Plots
Raw Expression Matrix
Preprocessed Expression Matrix
Raw Data Box-plot
log2 GE value
14.0
12.0
10.0
8.0
6.0
4.0
2.0
0.0
ser1/t0
ser1/t2
ser2/t4
ser2/t6
ser2/t8
Preprocessed Data Box-plot
GE value after preprocessing
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
ser1/t0
ser1/t2
ser2/t4
ser2/t6
ser2/t8
GE Data 1 1

Raw Data:

Study name: GE Data 1

Organism: human

Data file: D:\acgt_lab\testDataSets\human\data\serum.human.txt

Data type: Absolute intensities

Data scale: Original values (unscaled)

Number of missing values detected: 0

ID conversion file: D:\acgt_lab\testDataSets\human\conversion\HGU133P2_2_LLids.txt

Data contains 20334 probes under 5 conditions

Preprocessed Data:

1) Quantile Normalization

2) Fold Change Filter:

fold change: 3.0

entries exp: 1

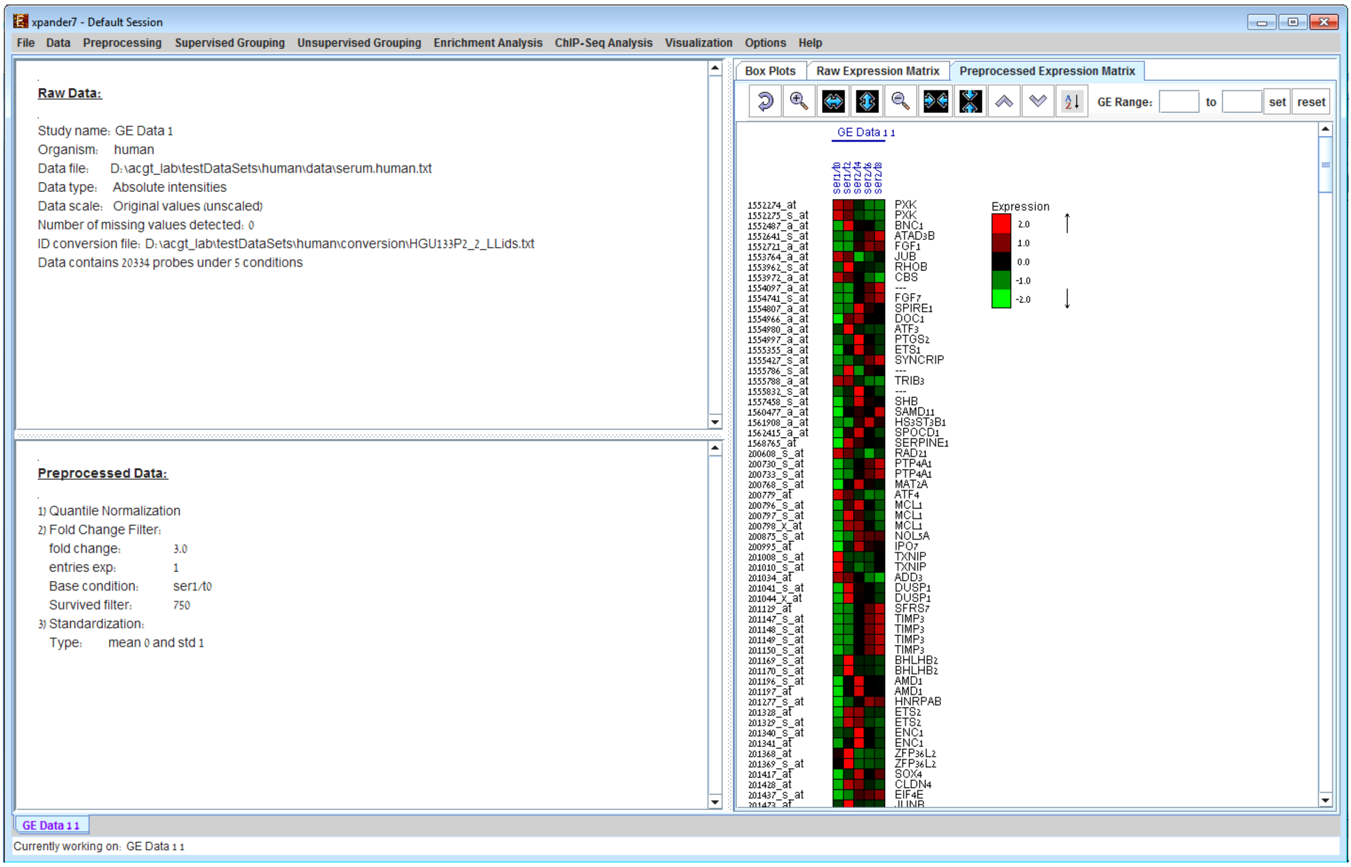
Base condition: ser1/t0

Survived filter: 750

3) Standardization:

Type: mean 0 and std 1

Currently working on: GE Data 1 1



Raw Data:

Study name: GE Data 1

Organism: human

Data file: D:\acgt_lab\testDataSets\human\data\serum.human.txt

Data type: Absolute intensities

Data scale: Original values (unscaled)

Number of missing values detected: 0

ID conversion file: D:\acgt_lab\testDataSets\human\conversion\HGU133P2_2_LLids.txt

Data contains 20334 probes under 5 conditions

Preprocessed Data:

1) Quantile Normalization

2) Fold Change Filter:

fold change: 3.0

entries exp: 1

Base condition: ser1/t0

Survived filter: 750

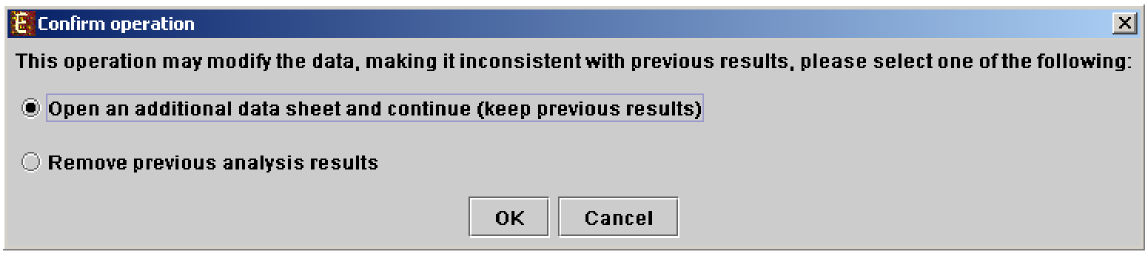
3) Standardization:

Type: mean 0 and std 1

Currently working on: GE Data 1 1

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

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



Confirm operation

This operation may modify the data, making it inconsistent with previous results, please select one of the following:

● Open an additional data sheet and continue (keep previous results)

○ Remove previous analysis results

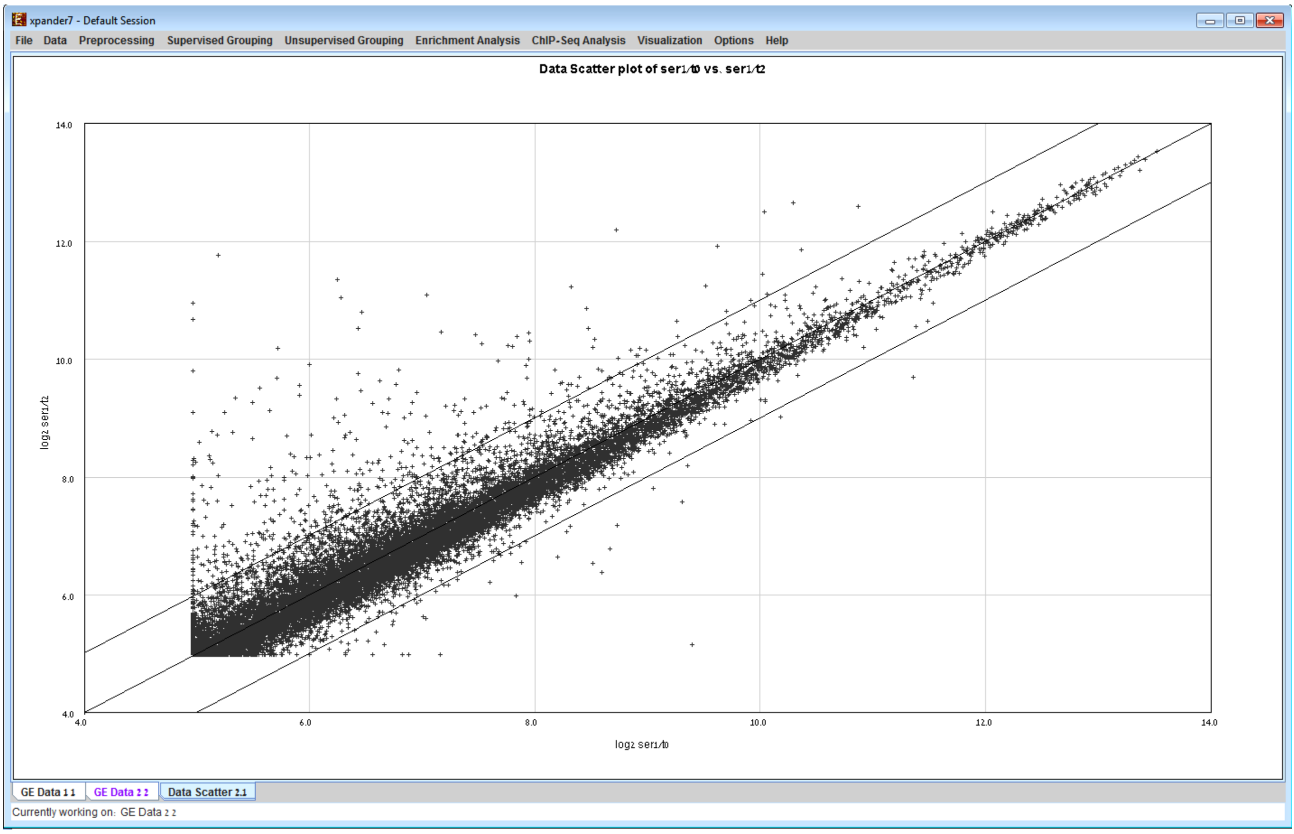
Upon choosing to open an additional data sheet, a new data set view tab called 'Data Sheet 2' is added to the main frame. The title of this tab is highlighted (colored in purple), indicating that it is now the active data sheet (i.e. all further operations refer to this data sheet). The active data sheet is automatically changed according to the selected (front) visualization tab.

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

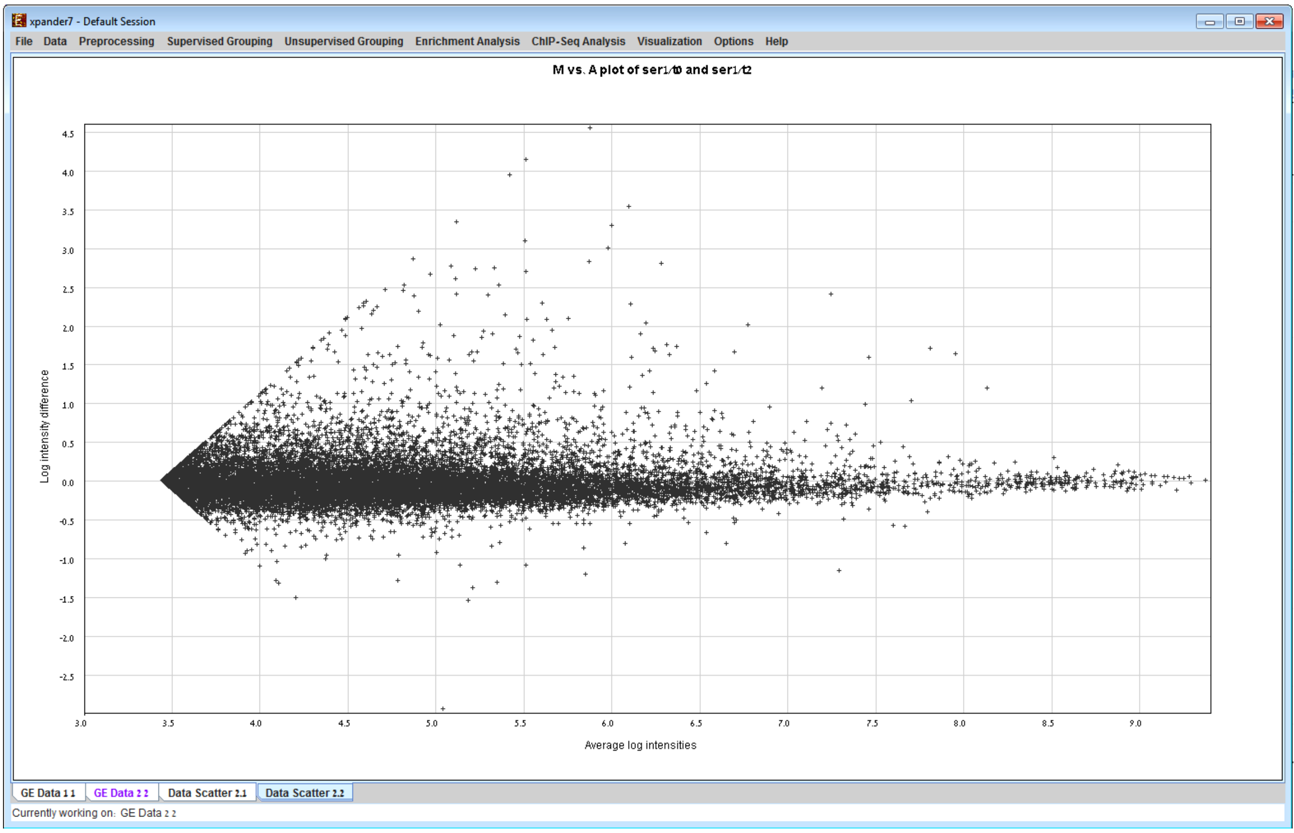
Gene Expression Data Plots

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

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



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



Differential Expression Analysis

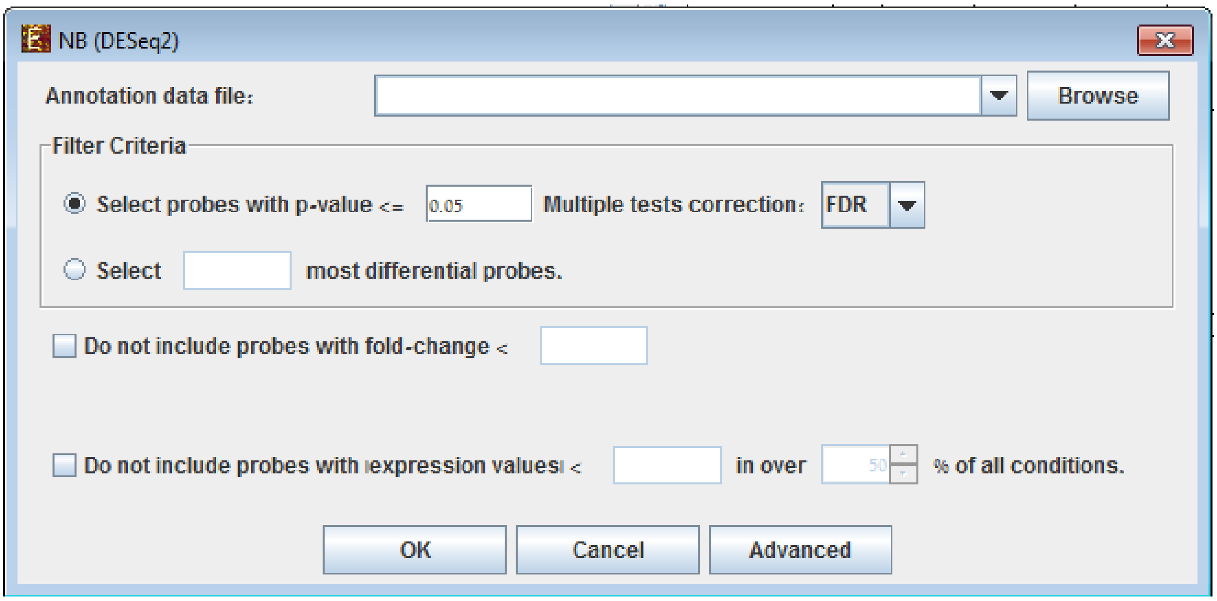
The goal in this analysis is to detect groups of genes that demonstrate differential expression between two/more condition groups.

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

b) **Wilcoxon\Mann-Whitney two sample rank sum test -** (*Supervised Grouping >> Differential Expression >> Ranksum test*): When using this method, genes can be assigned into one of two groups (up-regulated and down-regulated), depending on the definitions of Ranksum test parameters. This method is nearly as efficient as the t-test on normal distributions of expression values but has a greater efficiency than the t-test on non-normal distributions of expression values.

Negative binomial (DESeq2) – Negative Binomial distribution test for RNA-seq count data (*Supervised Grouping >> Differential Expression >> NB (DESeq2)*): this method is used to demonstrate differential expression between 2 condition subsets for RNA-seq count data where for each probe i and condition j in the expression matrix the value is a non-negative integer. The probes are then assigned into two groups (up-regulated and down-regulated). DESeq2 requires annotation data table file which includes the attributes/categories and their corresponding labels for each condition in the RNA-seq data. For details regarding the annotation data file format see the File Formats section.

When selecting *Supervised Grouping >> Differential Expression >> NB (DESeq2),* the following dialog box will appear:



NB (DESeq2)

Annotation data file:

Filter Criteria

● Select probes with p-value <= 0.05 Multiple tests correction: FDR

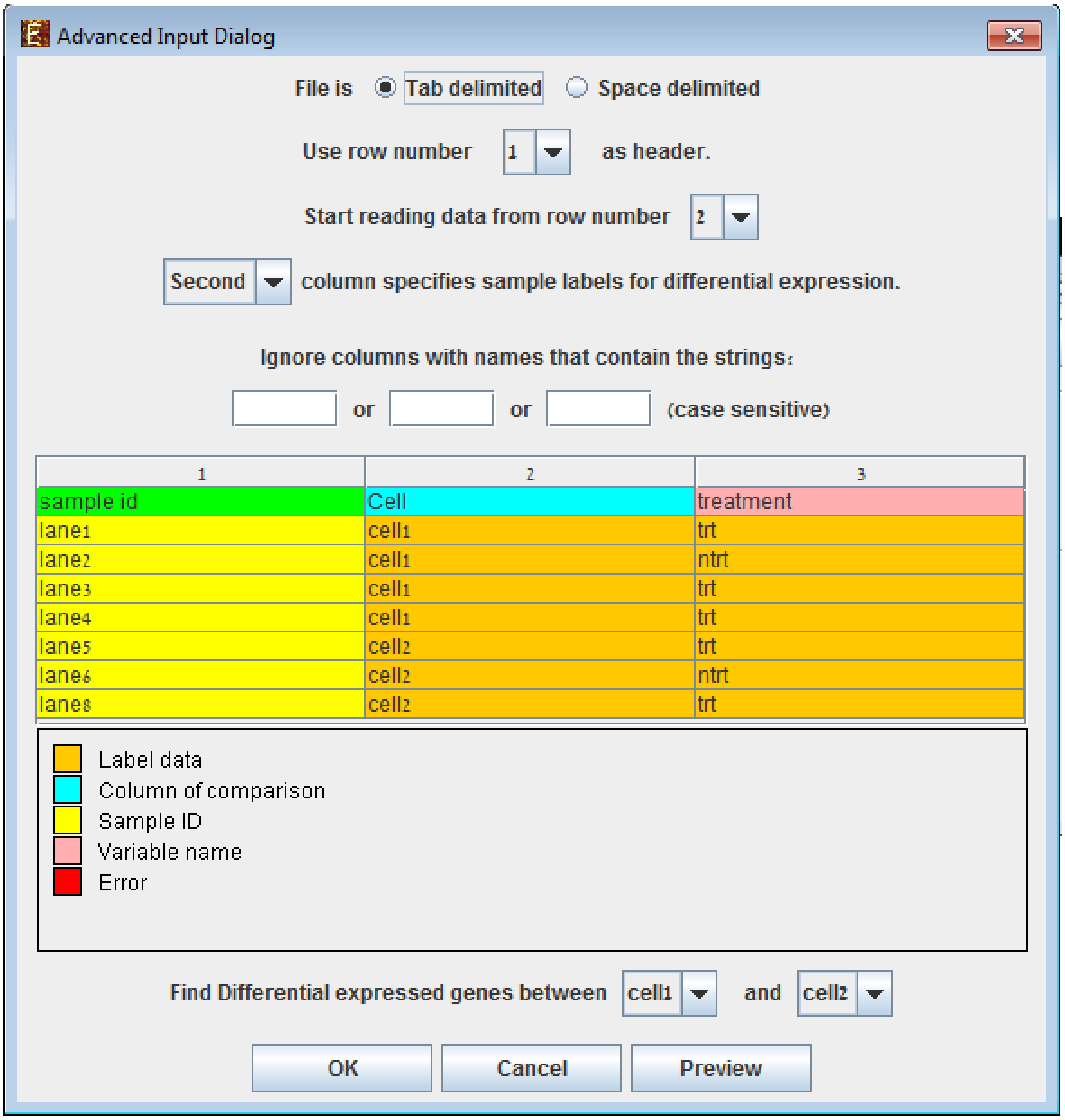
○ Select most differential probes.

□ Do not include probes with fold-change <

□ Do not include probes with expression values < in over 50 % of all conditions.

Annotation data file should be given.

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



Advanced Input Dialog

File is ● Tab delimited ○ Space delimited

Use row number 1 as header.

Start reading data from row number 2

Second column specifies sample labels for differential expression.

Ignore columns with names that contain the strings:

or or (case sensitive)

1	2	3
sample id	Cell	treatment
lane1	cell1	trt
lane2	cell1	ntrt
lane3	cell1	trt
lane4	cell1	trt
lane5	cell2	trt
lane6	cell2	ntrt
lane8	cell2	trt

🟧 Label data
🟦 Column of comparison
🟨 Sample ID
■ Variable name
🟥 Error

Find Differential expressed genes between cell1 and cell2

For further information regarding DESeq2 please refer to References. Before using DESeq2, please make sure you have **R software along with the “DESeq2" package** installed (see R External Application section).

c) **Negative binomial (edgeR) –** Negative Binomial distribution test for RNA-seq count data (Supervised Grouping >> Differential Expression >> NB (edgeR)): this method is used to demonstrate differential expression between 2 condition subsets for RNA-seq count data where for each probe i and condition j in the expression matrix the value is a non-negative integer. As part of the test 3 different dispersion options are given: Tagwise – for a large amount of samples (>6), where a different dispersion is calculated for each probe, Common – for small amount of samples (<6), where the same dispersion is given for each probe, Poisson – a special case of NB where dispersion = 0 for all probes. The probes are then assigned into two groups (up-regulated and down-regulated). For further information regarding edgeR please refer to References. Before using edgeR, please make sure you have **R software along with the “edgeR” and "limma" packages** installed (see R External Application section).

d) **SAM -** Significance Analysis of Microarray (*Supervised Grouping >> Differential Expression >> SAM*): this method detects probes that demonstrate differential expression between conditions subsets. You may choose 2 or more subsets (multi-class tests are supported). The probes are then assigned into two groups (up-regulated and down-regulated) if 2 condition groups are tested or into one group of differentially expressed otherwise. SAM uses permutations to get an ’empirical’ estimate for the FDR of the reported differential genes (for details see the References section). Before using SAM, please make sure you have **R software along with the “samr” package** installed (see R External Application section).

After performing differential expression grouping analysis, a solution visualization tab is added to the main window. It contains the following views:

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

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

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

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

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

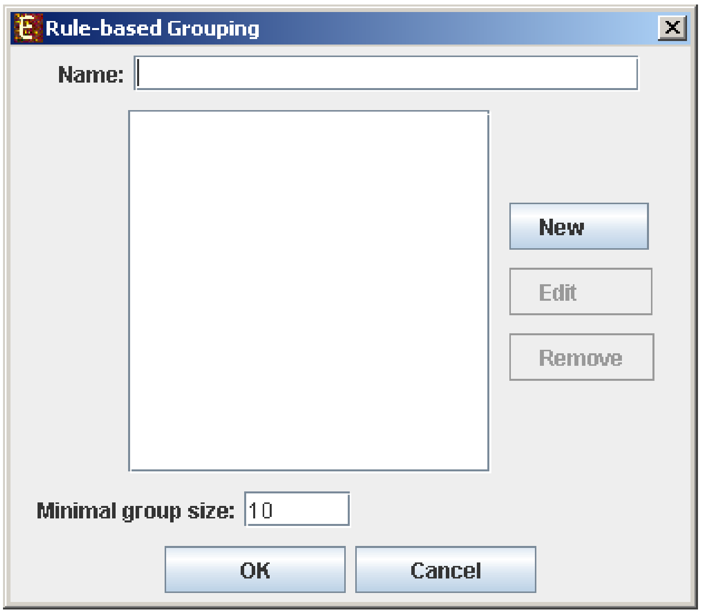
Upon selecting a group, the corresponding pane is displayed on the right. It contains a list of probes, p-values/q-values, fold-change, probe patterns, expression matrix (heat map) and the chromosomal locations of the genes. Similarity matrices for probes within the cluster as well as for conditions are also displayed in this tab, if the relevant options in the display settings are selected (see the Settings section). If a network file has been loaded (via Data>>Load Network), the sub-graph, induced by the cluster is also displayed in the group pane.

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

A differential expression solution can be saved using the File >> Export to text..., and reloaded using the Grouping Supervised Grouping >> Differential Expression >> Load Solution.

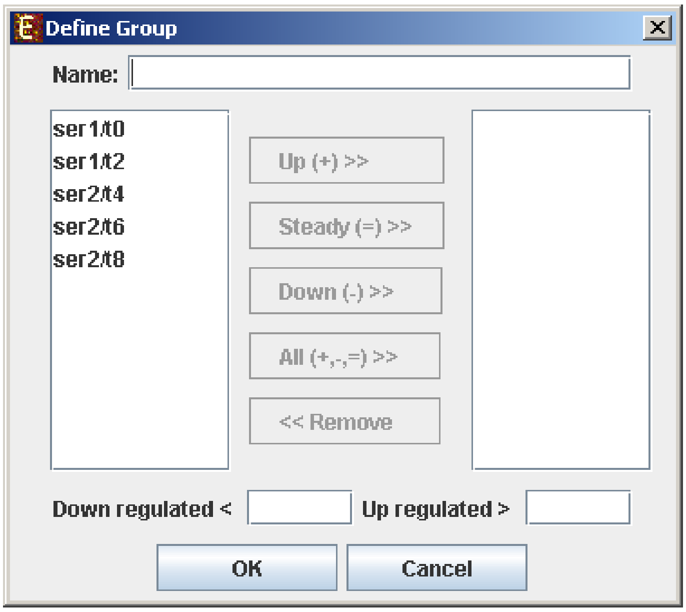
Defining a group according to a rule

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



Minimal group size: 10

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



ser1/t0
ser1/t2
ser2/t4
ser2/t6
ser2/t8

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

Defining a group according to similarity to a selected probe

This can be done by selecting *Supervised Grouping >> Group by Pattern Similarity.* An input dialog box allows setting the similarity measure (Pearson correlation, Spearman correlation or Euclidean distance) and reference probe ID as well as the expected group size. The visualization for this 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 genes that are assigned to the same cluster should have similar expression patterns, while genes assigned to different clusters should have non-similar expression patterns. Usually there is no one solution that is the ‘true’ mathematical solution for this problem, but a 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 clusters).

After operating one of the clustering algorithms a clustering results view appears. The view contains information about the solution and its quality including the method and parameters that were used to obtain it, number of clusters, number of singletons (probes that were not assigned to any cluster), overall homogeneity and separation, as well as the size and homogeneity of each cluster. This summary can be used to compare different solutions.

In order to apply a clustering algorithm to the data, select the required algorithm from the *Unsupervised Grouping >> Clustering* menu (options are: **KMeans**, **CLICK**, **SOM**). You can also use the agglomerative hierarchical clustering algorithm by extracting a partition from an existing hierarchical tree, by selecting *Unsupervised Grouping >> Hierarchical Clustering>> Generate Groups* (For details about building such a tree, please go to Hierarchical Clustering).

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

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

Fill the required input data in the algorithm input dialog box and press the 'Ok' button. 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 required 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 the References section.

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

Information regarding the clustering algorithm, number of clusters, number of un-clustered elements (singletons), and numerical measures of the clustering quality, including:

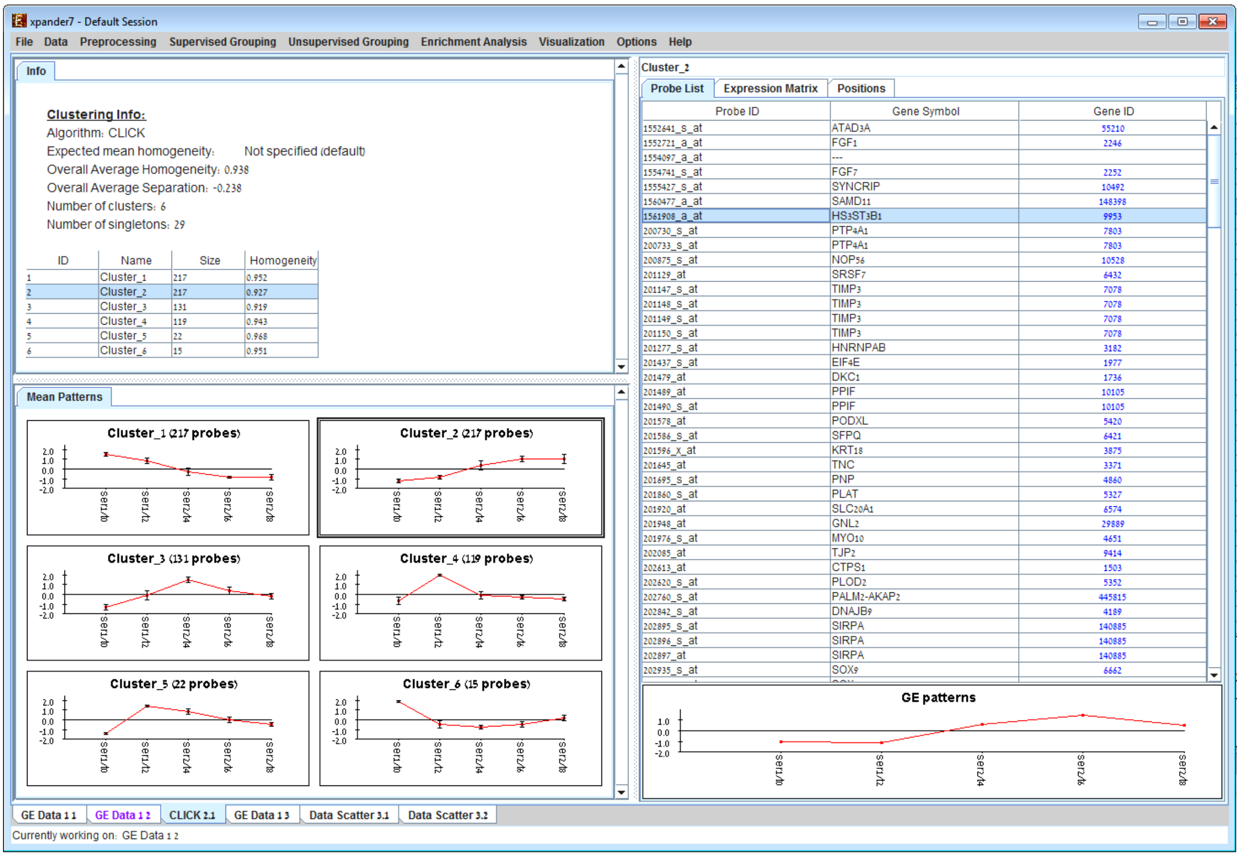
d) Overall average homogeneity - calculated as the average value of similarity between each element and the center of the cluster to which it has been assigned, weighted according to the size of the cluster.

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

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

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

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

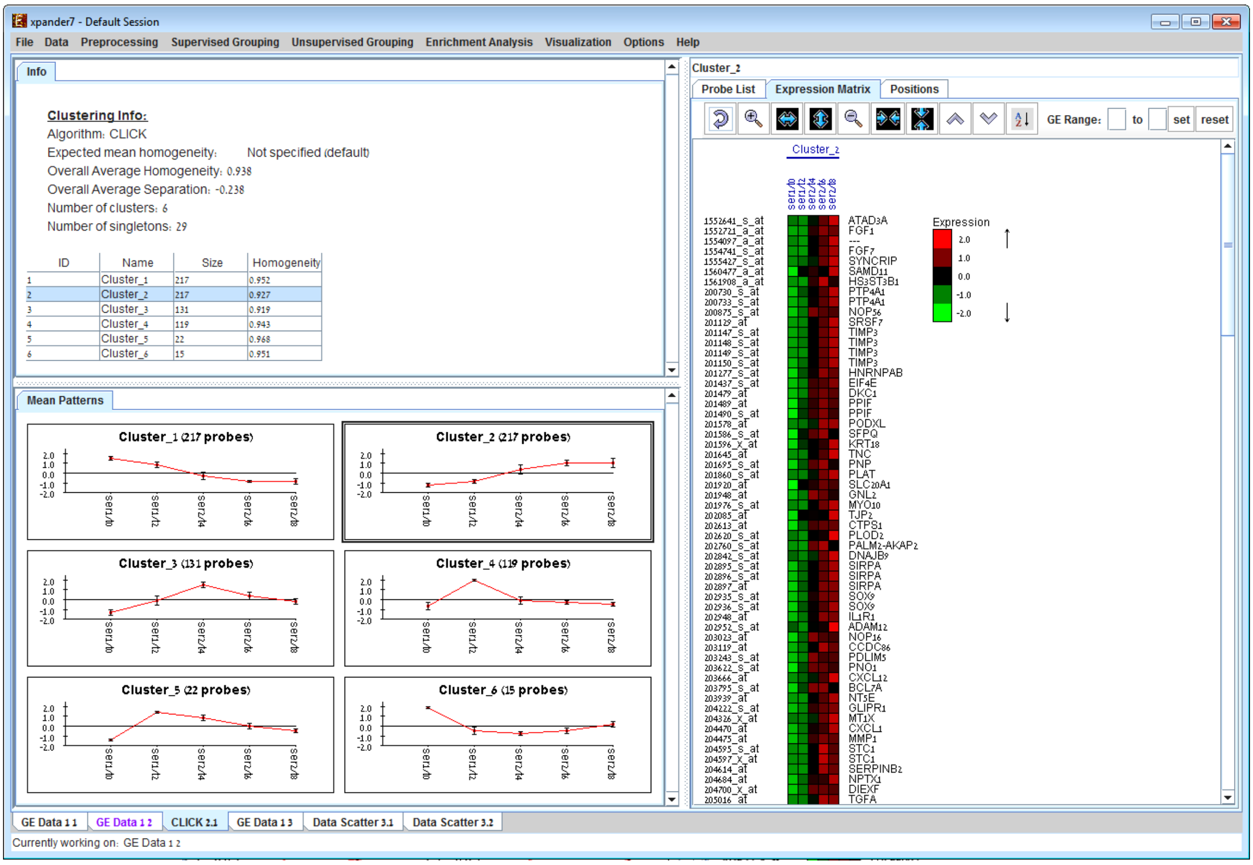


Clustering Info:
Algorithm: CLICK
Expected mean homogeneity: Not specified (default)
Overall Average Homogeneity: 0.938
Overall Average Separation: -0.238
Number of clusters: 6
Number of singletons: 29

ID	Name	Size	Homogeneity
1	Cluster_1	217	0.952
2	Cluster_2	217	0.927
3	Cluster_3	131	0.919
4	Cluster_4	119	0.943
5	Cluster_5	22	0.968
6	Cluster_6	15	0.951

Probe ID	Gene Symbol	Gene ID
1552641_s_at	ATAD3A	55210
1552721_a_at	FGF1	2246
1554097_a_at	---	
1554741_s_at	FGF7	2252
1555427_s_at	SYNCRIP	10492
1560477_a_at	SAMD11	148398
1561908_a_at	HS3ST3B1	9953
200730_s_at	PTP4A1	7803
200733_s_at	PTP4A1	7803
200875_s_at	NOP56	10528
201129_at	SRSF7	6432
201147_s_at	TIMP3	7078
201148_s_at	TIMP3	7078
201149_s_at	TIMP3	7078
201150_s_at	TIMP3	7078
201277_s_at	HNRNPAB	3182
201437_s_at	EIF4E	1977
201479_at	DKC1	1736
201489_at	PPIF	10105
201490_s_at	PPIF	10105
201578_at	PODXL	5420
201586_s_at	SFPQ	6421
201596_x_at	KRT18	3875
201645_at	TNC	3371
201695_s_at	PNP	4860
201860_s_at	PLAT	5327
201920_at	SLC20A1	6574
201948_at	GNL2	29889
201976_s_at	MYO10	4651
202085_at	TJP2	9414
202613_at	CTPS1	1503
202620_s_at	PLOD2	5352
202760_s_at	PALM2-AKAP2	445815
202842_s_at	DNAJB9	4189
202895_s_at	SIRPA	140885
202896_s_at	SIRPA	140885
202897_at	SIRPA	140885
202935_s_at	SOX9	6662

Currently working on: GE Data 1.2



Clustering Info:
Algorithm: CLICK
Expected mean homogeneity: Not specified (default)
Overall Average Homogeneity: 0.938
Overall Average Separation: -0.238
Number of clusters: 6
Number of singletons: 29

ID	Name	Size	Homogeneity
1	Cluster_1	217	0.952
2	Cluster_2	217	0.927
3	Cluster_3	131	0.919
4	Cluster_4	119	0.943
5	Cluster_5	22	0.968
6	Cluster_6	15	0.951

Currently working on: GE Data 1.2

After performing enrichment analysis (for details see the Enrichment Analysis Tools), if enrichment has been detected in the selected cluster, the corresponding histogram and analysis information are added to the single cluster view.

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

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

Hierarchical Clustering and Visualization

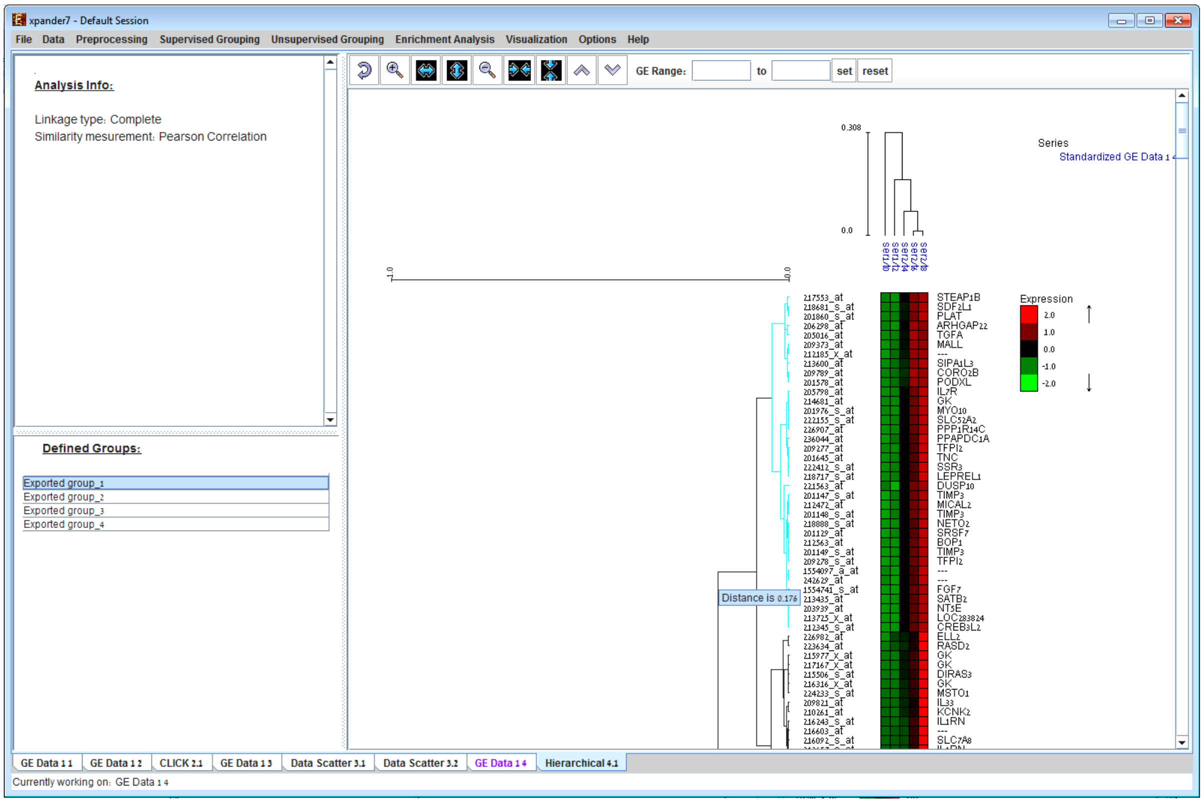
This tool uses the agglomerative algorithm to calculate a dendrogram tree for all expression patterns (probe patterns) and/or profiles (condition profiles). The type of linkage (manner in which the distance between a new node and the rest of the nodes is calculated) used in the algorithm can be set via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

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

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

If condition attributes file has been loaded for the analyzed dataset, a matrix representation of these attributes will be displayed below the expression matrix (heatmap).

For details regarding the condition attributes file format, refer to the File Formats section.



Analysis Info:

Linkage type: Complete
Similarity mesurement: Pearson Correlation

Defined Groups:

Exported group_1
Exported group_2
Exported group_3
Exported group_4

Currently working on: GE Data 1 4

Upon clicking on the vertical tree, a corresponding sub tree is highlighted (selected) and can be defined as a group by right clicking on the same location and selecting the “Export group” option from the right click menu. The sub tree is then added as a group of the bottom left panel of the display.

Upon selecting one of the groups that have been previously defined and added to the list on 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 the corresponding group in the bottom left panel and selecting remove group.

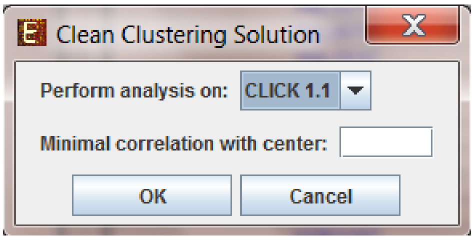
Manually selected groups can then be defined as a grouping solution by selecting *Unsupervised Grouping>>Hierarchical Clustering>>Generated Groups>> From Selected subtrees.*

A hierarchical clustering dendrogram can be exported to a Newick format text file by selecting *File>>Export to text...* when the relevant solution tab is selected.

Clustering solution cleaning

This feature allows removing elements (i.e. probes) from a clustering solution in order to obtain higher levels of homogeneity within each cluster. It can be applied on an existing clustering solution, and results in the generation of a new “cleaner” version of the solution.

To perform cluster cleaning select *Unsupervised Grouping>>Clustering>>Clean Clusters.* The following dialog box will appear:



Perform analysis on: CLICK 1.1

Minimal correlation with center:

In the dialog box, select the clustering solution on which the feature should be operated, specify the minimal required correlation between each probe and the cluster center and press OK.

After cluster cleaning is performed, a new clustering solution visualization tab is added to the main window. The new tab will be named by the original solution with the extension “ – cleaned by <c>” where c = the correlation threshold selected by the user.

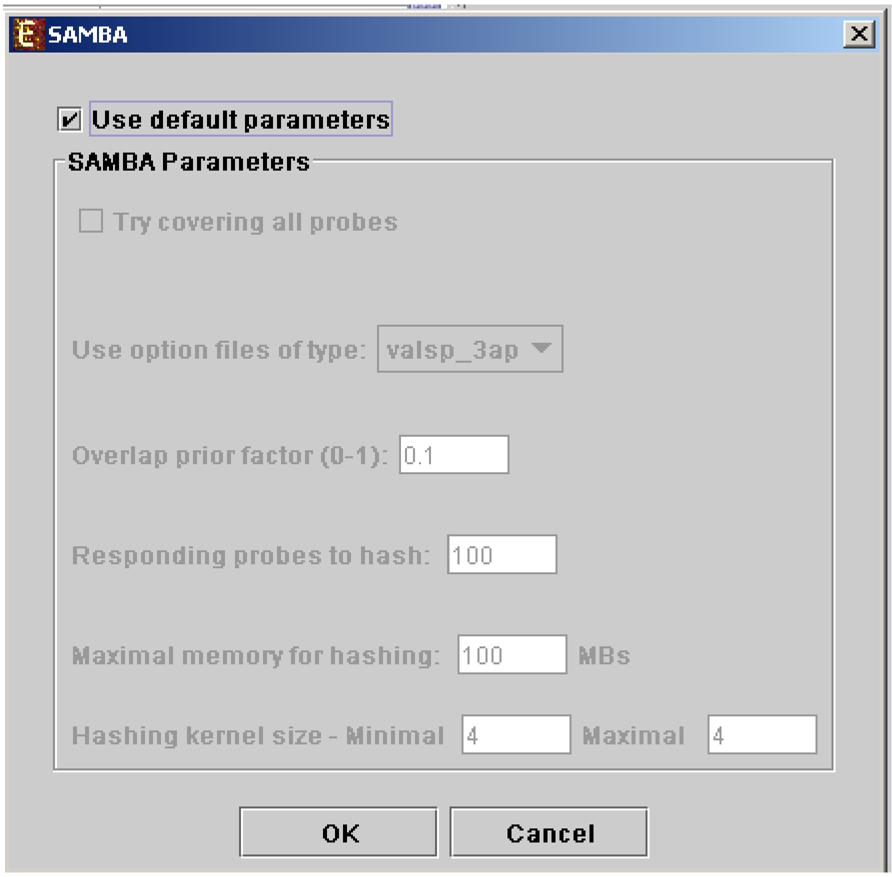
Biclustering GE Data

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

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

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

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



SAMBA

☑ Use default parameters

SAMBA Parameters

□ Try covering all probes

Use option files of type: valsp_3ap

Overlap prior factor (0-1): 0.1

Responding probes to hash: 100

Maximal memory for hashing: 100 MBs

Hashing kernel size - Minimal 4 Maximal 4

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

Field	Description
Use default parameters	When checked, biclustering parameters (described below) are set automatically (this option is recommended unless the user is familiar with the parameters).
Option files type	The user can select one out of 6 options. The following table describes the advantages and disadvantages of each option: Option name | fast performance | less memory required: | Flexible | Robust- can handle normalization problems and non gene-expression

<table>
<tr><td></td><td>
<table>
<tr><td></td><td></td><td></td><td></td><td>data</td></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</td></tr>
</table>

	number will significantly increase the algorithm run time (may also produce better results).
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.

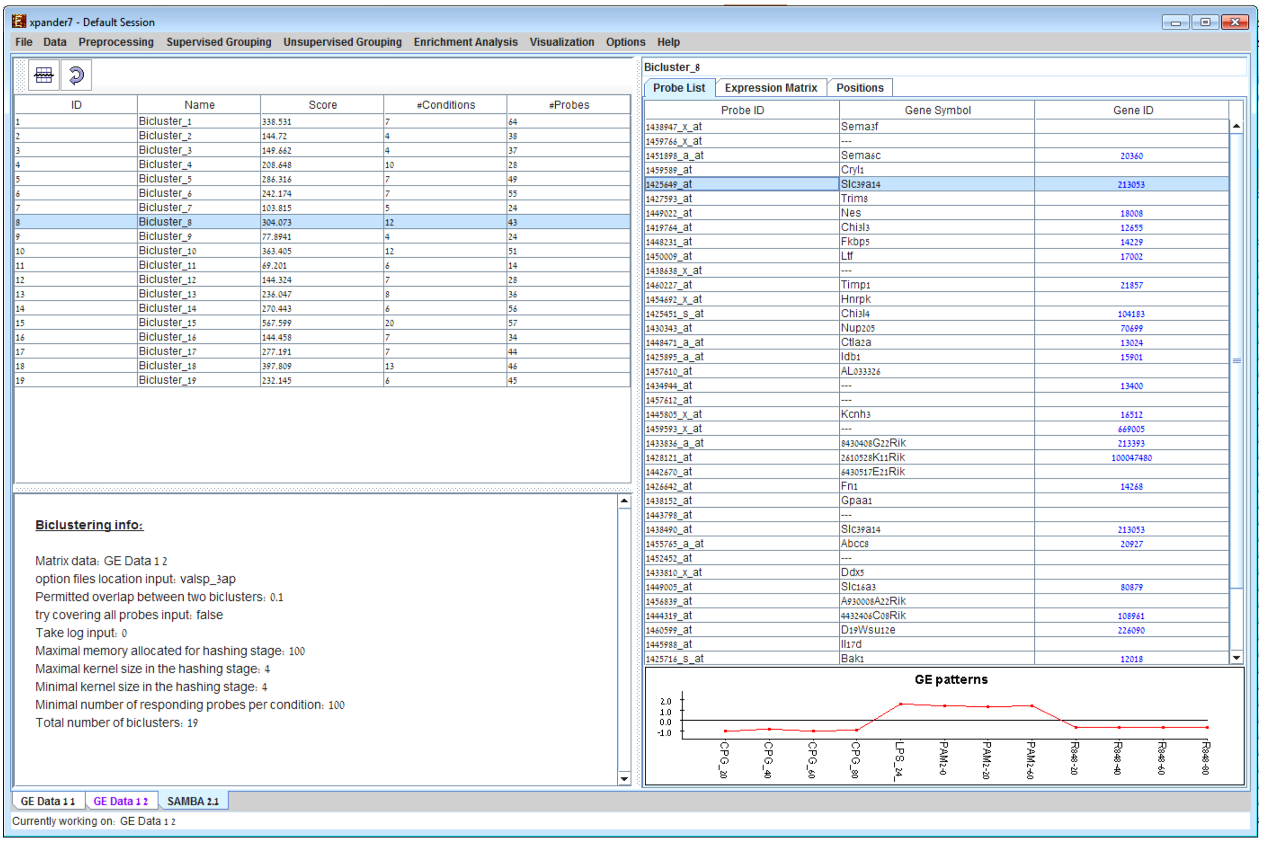
Upon clicking ‘OK’ in the dialog box, the SAMBA algorithm is operated on the dataset.

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

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

g) Biclusters table – contains the following information for each bicluster: serial number, name, score, number of probes genes and number of conditions. The name of a bicluster can be changed by editing the corresponding cell in the table. The score is given by the SAMBA algorithm and is size-dependent, thus, it is not recommended to use it to compare the quality of two biclusters of different sizes. The table can be filtered to display a subset of the biclusters by clicking on the ‘Filter’ (▦) button in the toolbar. Filtering can be performed according to: Score, number of probes and number of conditions.

Upon selecting a bicluster (from the biclusters table), the corresponding pane is displayed on the right. It contains a list of probes, probe patterns, expression matrix (heat map) and the chromosomal locations of the genes. Similarity matrices for probes within the cluster as well as for conditions are also displayed in this tab, if the relevant options in the display settings are selected (see the Settings section). If a network file has been loaded (via *Data>>Load Network*), the sub-graph, induced by the cluster is also displayed in the cluster pane.



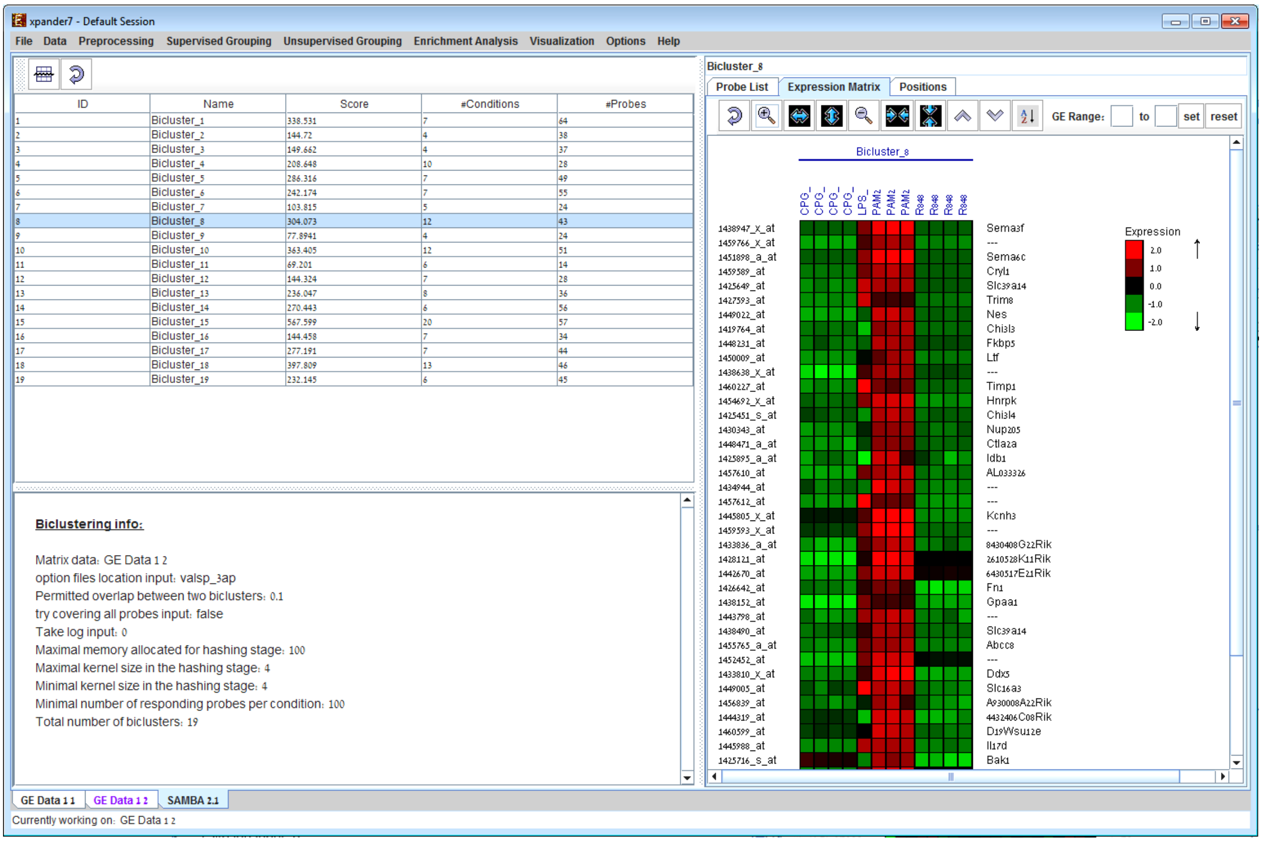
ID	Name	Score	#Conditions	#Probes
1	Bicluster_1	338.531	7	64
2	Bicluster_2	144.72	4	38
3	Bicluster_3	149.662	4	37
4	Bicluster_4	208.648	10	28
5	Bicluster_5	286.316	7	49
6	Bicluster_6	242.174	7	55
7	Bicluster_7	103.815	5	24
8	Bicluster_8	304.073	12	43
9	Bicluster_9	77.8941	4	24
10	Bicluster_10	363.405	12	51
11	Bicluster_11	69.201	6	14
12	Bicluster_12	144.324	7	28
13	Bicluster_13	236.047	8	36
14	Bicluster_14	270.443	6	56
15	Bicluster_15	567.599	20	57
16	Bicluster_16	144.458	7	34
17	Bicluster_17	277.191	7	44
18	Bicluster_18	397.809	13	46
19	Bicluster_19	232.145	6	45

Biclustering info:

Matrix data: GE Data 1 2
option files location input: valsp_3ap
Permitted overlap between two biclusters: 0.1
try covering all probes input: false
Take log input: 0
Maximal memory allocated for hashing stage: 100
Maximal kernel size in the hashing stage: 4
Minimal kernel size in the hashing stage: 4
Minimal number of responding probes per condition: 100
Total number of biclusters: 19

Probe ID	Gene Symbol	Gene ID
1438947_x_at	Sema3f	
1459766_x_at	---	
1451898_a_at	Sema6c	20360
1459589_at	Cryl1	
1425649_at	Slc39a14	213053
1427593_at	Trim8	
1449022_at	Nes	18008
1419764_at	Chi3l3	12655
1448231_at	Fkbp5	14229
1450009_at	Ltf	17002
1438638_x_at	---	
1460227_at	Timp1	21857
1454692_x_at	Hnrpk	
1425451_s_at	Chi3l4	104183
1430343_at	Nup205	70699
1448471_a_at	Ctla2a	13024
1425895_a_at	Idb1	15901
1457610_at	AL033326	
1434944_at	---	13400
1457612_at	---	
1445805_x_at	Kcnh3	16512
1459593_x_at	---	669005
1433836_a_at	8430408G22Rik	213393
1428121_at	2610528K11Rik	100047480
1442670_at	6430517E21Rik	
1426642_at	Fn1	14268
1438152_at	Gpaa1	
1443798_at	---	
1438490_at	Slc39a14	213053
1455765_a_at	Abcc8	20927
1452452_at	---	
1433810_x_at	Ddx5	
1449005_at	Slc16a3	80879
1456839_at	A930008A22Rik	
1444319_at	4432406C08Rik	108961
1460599_at	D19Wsu12e	226090
1445988_at	Il17d	
1425716_s_at	Bak1	12018

Currently working on: GE Data 1 2

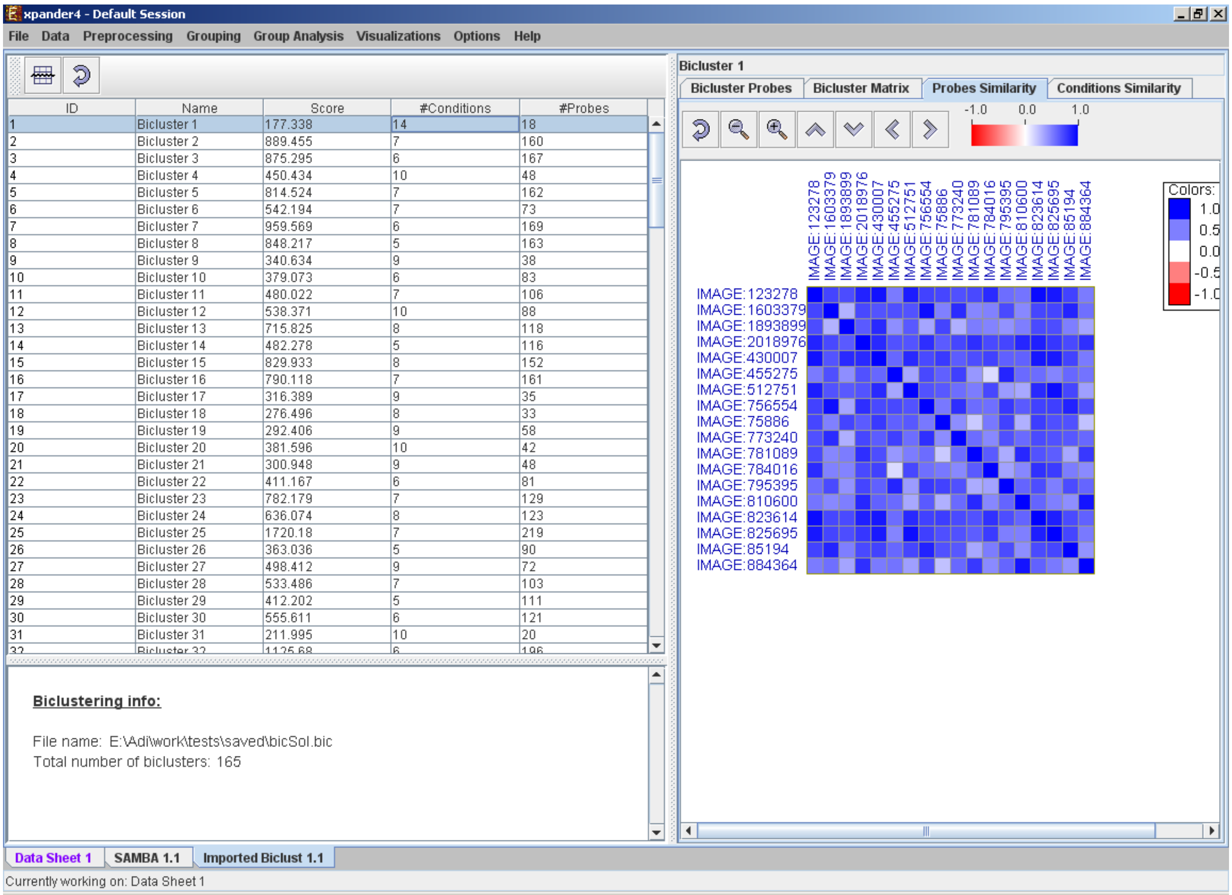


ID	Name	Score	#Conditions	#Probes
1	Bicluster_1	338.531	7	64
2	Bicluster_2	144.72	4	38
3	Bicluster_3	149.662	4	37
4	Bicluster_4	208.648	10	28
5	Bicluster_5	286.316	7	49
6	Bicluster_6	242.174	7	55
7	Bicluster_7	103.815	5	24
8	Bicluster_8	304.073	12	43
9	Bicluster_9	77.8941	4	24
10	Bicluster_10	363.405	12	51
11	Bicluster_11	69.201	6	14
12	Bicluster_12	144.324	7	28
13	Bicluster_13	236.047	8	36
14	Bicluster_14	270.443	6	56
15	Bicluster_15	567.599	20	57
16	Bicluster_16	144.458	7	34
17	Bicluster_17	277.191	7	44
18	Bicluster_18	397.809	13	46
19	Bicluster_19	232.145	6	45

Biclustering info:

Matrix data: GE Data 1 2
option files location input: valsp_3ap
Permitted overlap between two biclusters: 0.1
try covering all probes input: false
Take log input: 0
Maximal memory allocated for hashing stage: 100
Maximal kernel size in the hashing stage: 4
Minimal kernel size in the hashing stage: 4
Minimal number of responding probes per condition: 100
Total number of biclusters: 19

Currently working on: GE Data 1 2



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
32	Bicluster 32	1125.68	6	196

Biclustering info:

File name: E:\Adi\work\tests\saved\bicSol.bic
Total number of biclusters: 165

Currently working on: Data Sheet 1

After performing enrichment analysis (for details see the Enrichment Analysis Tools section), if enrichment has been detected in the selected bicluster, the corresponding histogram and analysis information are added to the single bicluster view, and a column is added to the expression matrix display for each enrichment class, stating for each probe, whether it belongs to that class.

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

Network Based Grouping of GE Data

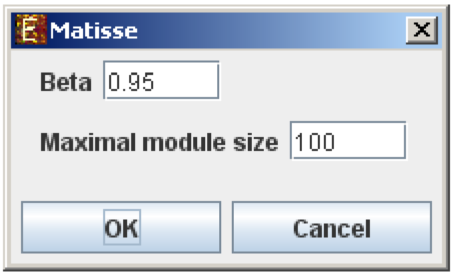
The goal here is to detect groups of genes that demonstrate similar expression patterns and are also highly connected in a given interactions network.

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

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

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

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



Beta 0.95

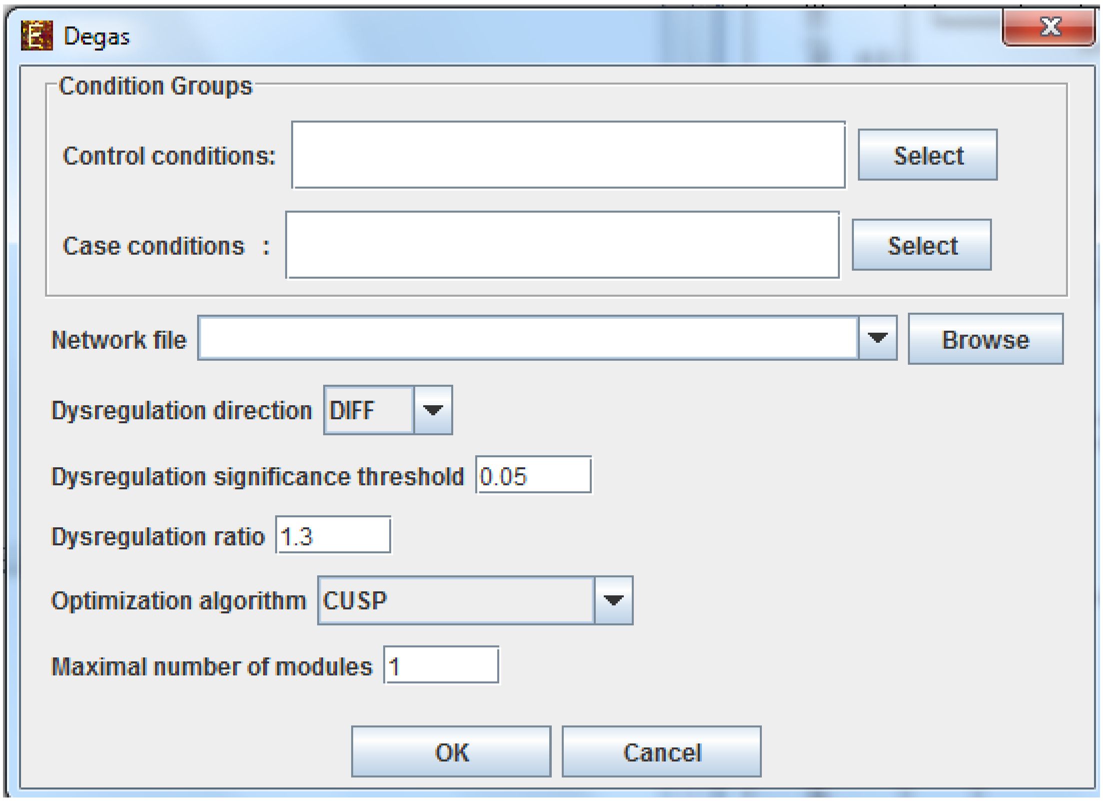
Maximal module size 100

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 dataset.

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



Control conditions:

Case conditions :

Network file

Dysregulation direction DIFF

Dysregulation significance threshold 0.05

Dysregulation ratio 1.3

Optimization algorithm CUSP

Maximal number of modules 1

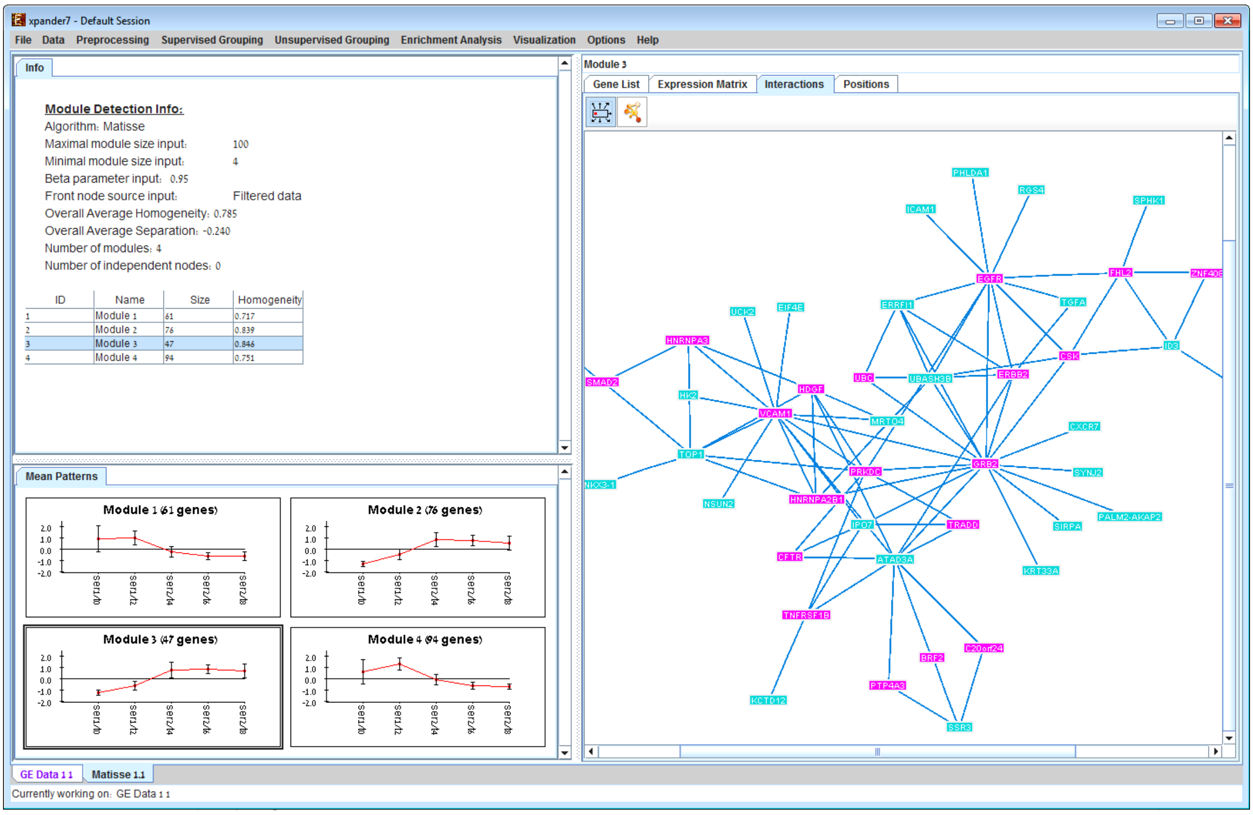
It enables the configuration of some of the parameters for the algorithm:

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 dataset.

After running network-based clustering, the solution is displayed in a new tab, which is added to the main window. The view is similar to the clustering results display. However it contains an additional interactions view tab for each module showing the sub-network that is formed by the module. If the Cytoscape network analysis software is installed and is running, it can be used for more advanced visualizations and analysis of this sub-network by clicking on the Cytoscape tool-button () placed at the top of the interactions view tab. In the display, back nodes (genes that appear in the network, but not in the GE data) are marked in pink.



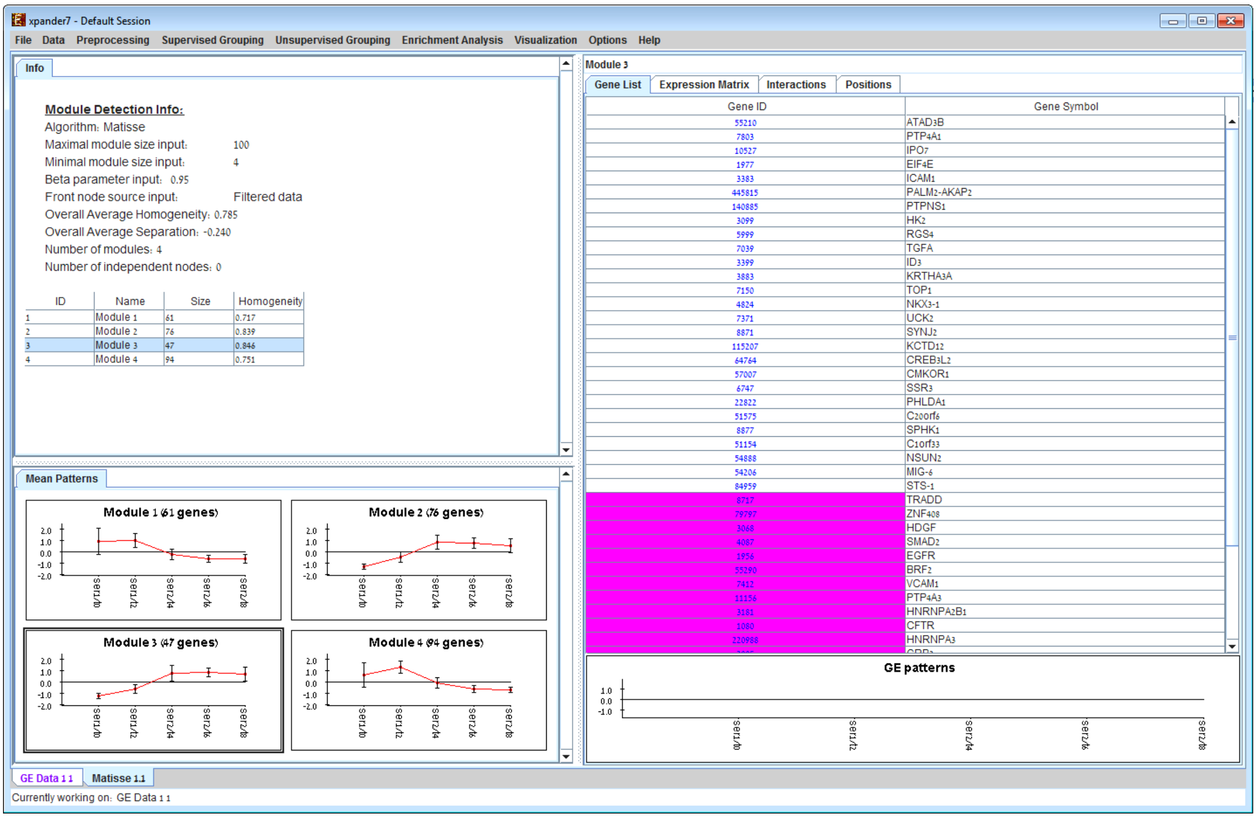
xpander7 - Default Session

Module Detection Info:

Algorithm: Matisse
Maximal module size input: 100
Minimal module size input: 4
Beta parameter input: 0.95
Front node source input: Filtered data
Overall Average Homogeneity: 0.785
Overall Average Separation: -0.240
Number of modules: 4
Number of independent nodes: 0

ID	Name	Size	Homogeneity
1	Module 1	61	0.717
2	Module 2	76	0.839
3	Module 3	47	0.846
4	Module 4	94	0.751

Currently working on: GE Data 1 1



Module Detection Info:

Algorithm: Matisse

Maximal module size input: 100

Minimal module size input: 4

Beta parameter input: 0.95

Front node source input: Filtered data

Overall Average Homogeneity: 0.785

Overall Average Separation: -0.240

Number of modules: 4

Number of independent nodes: 0

ID	Name	Size	Homogeneity
1	Module 1	61	0.717
2	Module 2	76	0.839
3	Module 3	47	0.846
4	Module 4	94	0.751

Gene ID	Gene Symbol
55210	ATAD3B
7803	PTP4A1
10527	IPO7
1977	EIF4E
3383	ICAM1
445815	PALM2-AKAP2
140885	PTPNS1
3099	HK2
5999	RGS4
7039	TGFA
3399	ID3
3883	KRTHA3A
7150	TOP1
4824	NKX3-1
7371	UCK2
8871	SYNJ2
115207	KCTD12
64764	CREB3L2
57007	CMKOR1
6747	SSR3
22822	PHLDA1
51575	C20orf6
8877	SPHK1
51154	C1orf33
54888	NSUN2
54206	MIG-6
84959	STS-1
8717	TRADD
79797	ZNF408
3068	HDGF
4087	SMAD2
1956	EGFR
55290	BRF2
7412	VCAM1
11156	PTP4A3
3181	HNRNPA2B1
1080	CFTR
220988	HNRNPA3

Currently working on: GE Data 1 1

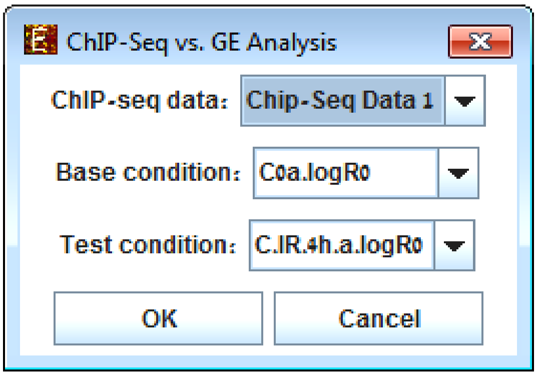
After performing enrichment analysis (for details see the Enrichment Analysis Tools section), if enrichment has been detected in the selected module, the corresponding histogram and analysis information are added to the single module view, and a column is added to the expression matrix display for each enrichment class, stating for each probe, whether it belongs to that class.

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

Integrative analysis of ChIP-Seq and Gene Expression Data

EXPANDER has a "ChIP-Seq Analysis" menu item which provides the following tools for the joint analysis after loading both ChIP-seq and gene expression data:

1) Examining gene expression distribution within ChIP-Seq-associated genes: this is done via ChIP-Seq Analysis >> Integration with expression data >> ChIP-Seq vs. GE analysis. Upon selecting this option, the following dialog box will appear:

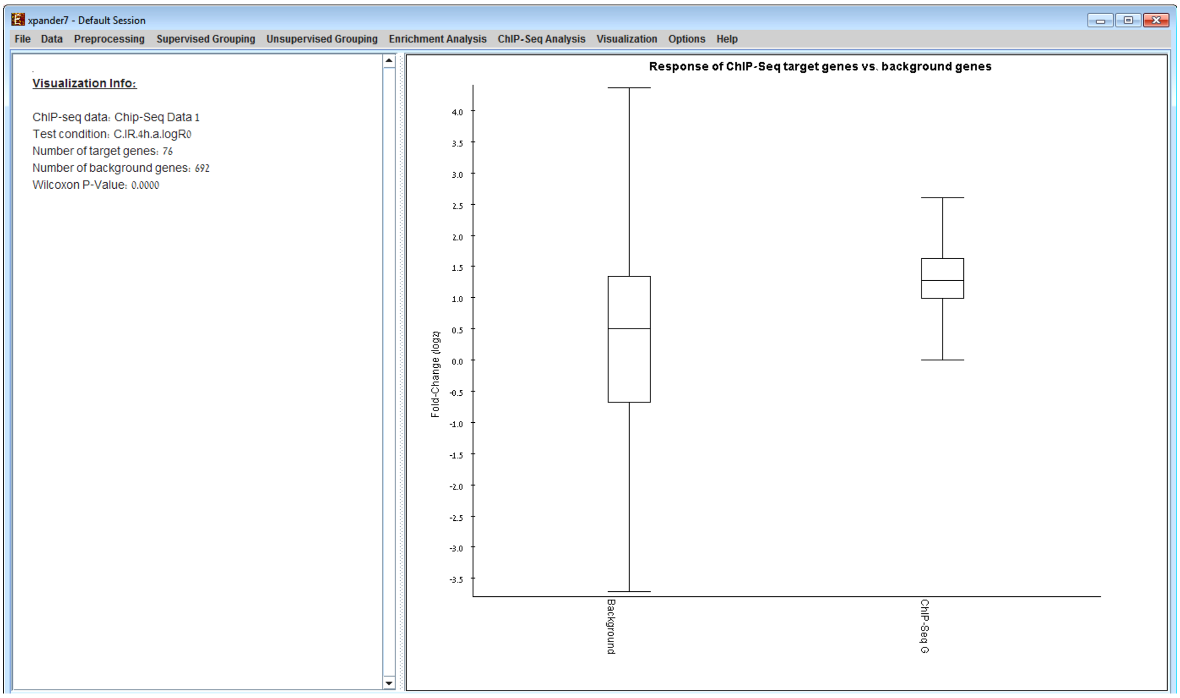


ChIP-seq data: Chip-Seq Data 1

Base condition: C0a.logR0

Test condition: C.IR.4h.a.logR0

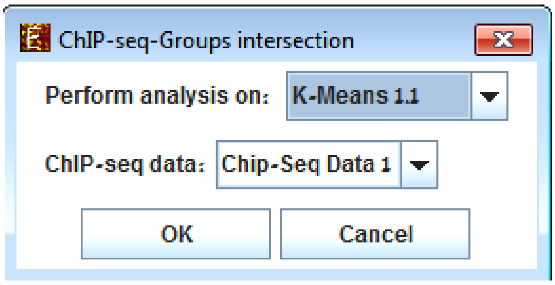
In the dialog box select the relevant ChIP-seq data set name, the base-condition and the test-condition, and click OK. A box plot showing the distribution of test-base GE ratios will be displayed for the set of chip-seq genes (right) and the rest of the genes in the gene-expression data set (left).


xpander7 - Default Session
File Data Preprocessing Supervised Grouping Unsupervised Grouping Enrichment Analysis ChIP-Seq Analysis Visualization Options Help
Response of ChIP-Seq target genes vs. background genes
Fold-Change (log2)
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
-3.5
Background
ChIP-Seq G

Visualization Info:

ChIP-seq data: Chip-Seq Data 1
Test condition: C.IR.4h.a.logR0
Number of target genes: 76
Number of background genes: 692
Wilcoxon P-Value: 0.0000

2) Extracting a grouping solution from ChIP-seq-gene-groups intersection: this is done via ChIP-Seq Analysis >> Integration with expression data >> *ChIP-Seq Intersection.* Upon selecting this option the following dialog box is displayed:



Perform analysis on: K-Means 1.1

ChIP-seq data: Chip-Seq Data 1

In the dialog box select the relevant grouping solution and ChIP-seq data set name, and click OK. A grouping solution visualization tab will be added to the main window. It will contain a gene-group for each non-empty intersection between a gene-group in the original solution and the chIP-Seq genes. Visualization is similar to clustering visualization.

3) Gene Set Enrichment Analysis: this is done via ChIP-Seq Analysis >> Integration with expression data >> GSEA. ChIP-Seq data can be selected as "Grouping solution". For further information please refer to: Gene Set Enrichment Analysis (GSEA) .

4) ChIP-Seq Enrichment: this is done via ChIP-Seq Analysis >> Integration with expression data >> ChIP-Seq Enrichment. For further information please refer to: ChIP-Seq Enrichment Analysis .

Group Enrichments Analysis Tools

- Functional Analysis
- Promoter Analysis – PRIMA, AMADEUS
- Location Enrichment Analysis
- miRNA Targets Enrichment Analysis
- ChIP-Seq Enrichment Analysis
- Pathway Enrichment Analysis
- General Enrichment Analysis
- Network Based Enrichment Analysis

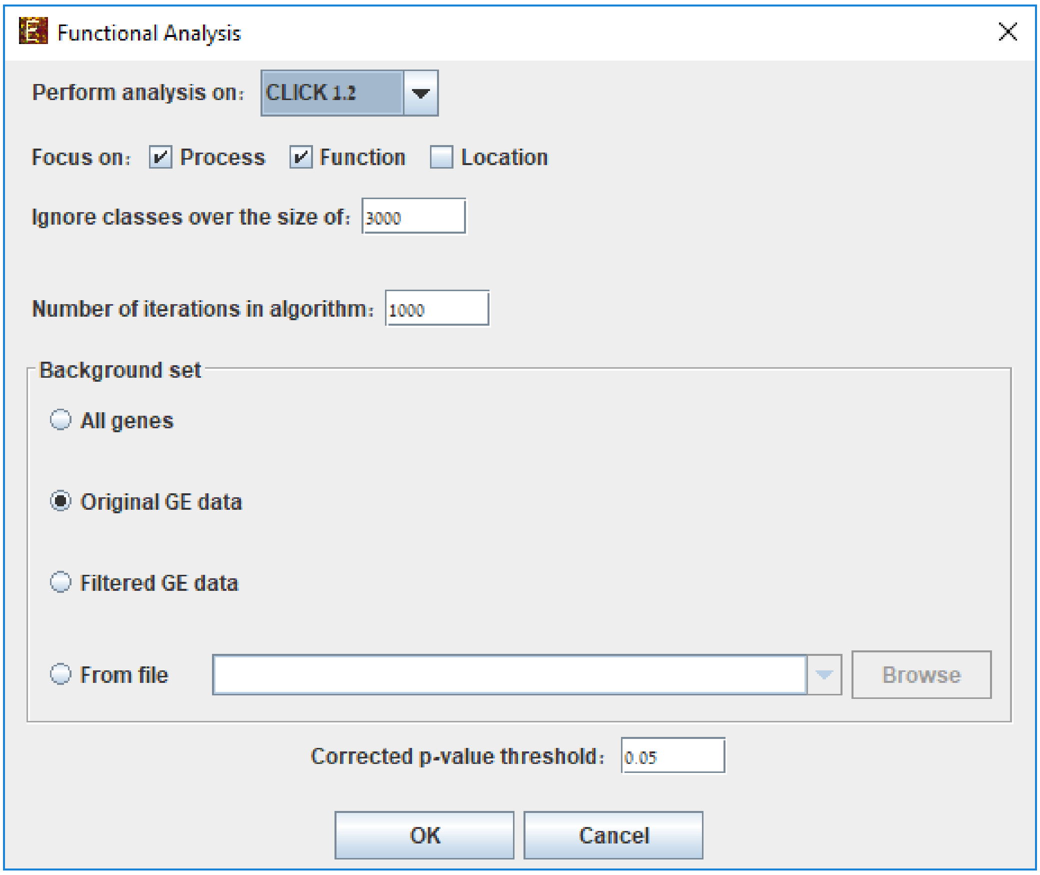
- Gene Set Enrichment Analysis (GSEA)

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

Functional Analysis

This tool performs basic statistical analysis on the distribution of functions of genes within each cluster. The functions of the genes are determined according to annotation files (GO), which can be downloaded from the EXPANDER download page (see the Supplied Files section). To perform this analysis, Expander utilizes the TANGO software, which performs hyper-geometric enrichment tests and corrects for multiple testing by bootstrapping and estimating the empirical p-value distribution for the evaluated sets.

Before operating functional analysis the annotation files for the relevant organism should be downloaded from the download page (more details at introduction of Group Analysis Tools). To perform the analysis, select *Enrichment Analysis >> Functional Analysis >> TANGO*. The following dialog box will appear:



Functional Analysis

Perform analysis on: CLICK 1.2

Focus on: ☑ Process ☑ Function □ Location

Ignore classes over the size of: 3000

Number of iterations in algorithm: 1000

Background set

○ All genes

● Original GE data

○ Filtered GE data

○ From file

Corrected p-value threshold: 0.05

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

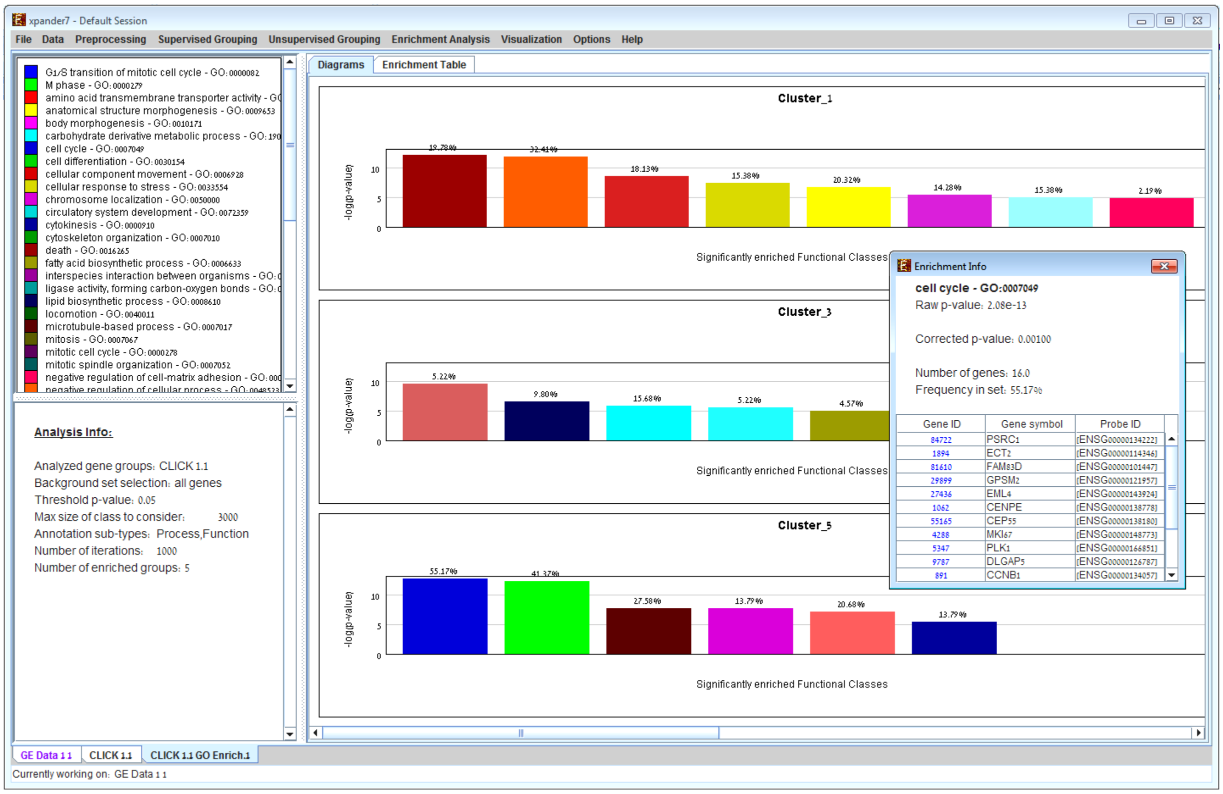
Field	Description
Perform analysis on	The grouping solution on which the analysis will be performed.
Include back nodes	Include genes that are part of the module' but not included in the GE data (Relevant only if the analysis is performed on modules, detected by network based algorithm)
Focus on	Can 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 tab is added to the main window. It contains general information regarding the analysis, a sort-able table holding all detected enrichments (set ID, functional class, p-value, etc.) and a multi-histogram panel along with a color index (mapping each color to a corresponding functional class). The multi-histogram panel contains one histogram for each probe/gene set/group in which enrichment has been detected. Each histogram contains a column for each significant (more frequent than would be expected by random) functional class. The definition of significant depends on the user’s selection of threshold p-value i.e., a functional class is considered significantly enriched in a 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 (i.e. height = -log(raw p-value)). The frequency in set (frequency of genes of a functional class within the examined set, in %) is written on top of the column. Upon clicking on a column, a dialog box is displayed containing the class name, raw p-value, corrected p-value, and a list of the genes in the cluster/bi-cluster that belong to the class. Upon clicking on one of the gene Ids in the table, a relevant web page with information regarding this gene is displayed. The display tool tip shows the cluster number, size and homogeneity.



■ G1/S transition of mitotic cell cycle - GO:0000082
■ M phase - GO:0000279
■ amino acid transmembrane transporter activity - GO
■ anatomical structure morphogenesis - GO:0009653
■ body morphogenesis - GO:0010171
■ carbohydrate derivative metabolic process - GO:190
■ cell cycle - GO:0007049
■ cell differentiation - GO:0030154
■ cellular component movement - GO:0006928
■ cellular response to stress - GO:0033554
■ chromosome localization - GO:0050000
■ circulatory system development - GO:0072359
■ cytokinesis - GO:0000910
■ cytoskeleton organization - GO:0007010
■ death - GO:0016265
■ fatty acid biosynthetic process - GO:0006633
■ interspecies interaction between organisms - GO:0
■ ligase activity, forming carbon-oxygen bonds - GO:0
■ lipid biosynthetic process - GO:0008610
■ locomotion - GO:0040011
■ microtubule-based process - GO:0007017
■ mitosis - GO:0007067
■ mitotic cell cycle - GO:0000278
■ mitotic spindle organization - GO:0007052
■ negative regulation of cell-matrix adhesion - GO:000
■ negative regulation of cellular process - GO:0048523

Analysis Info:

Analyzed gene groups: CLICK 1.1
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 groups: 5

Enrichment Info

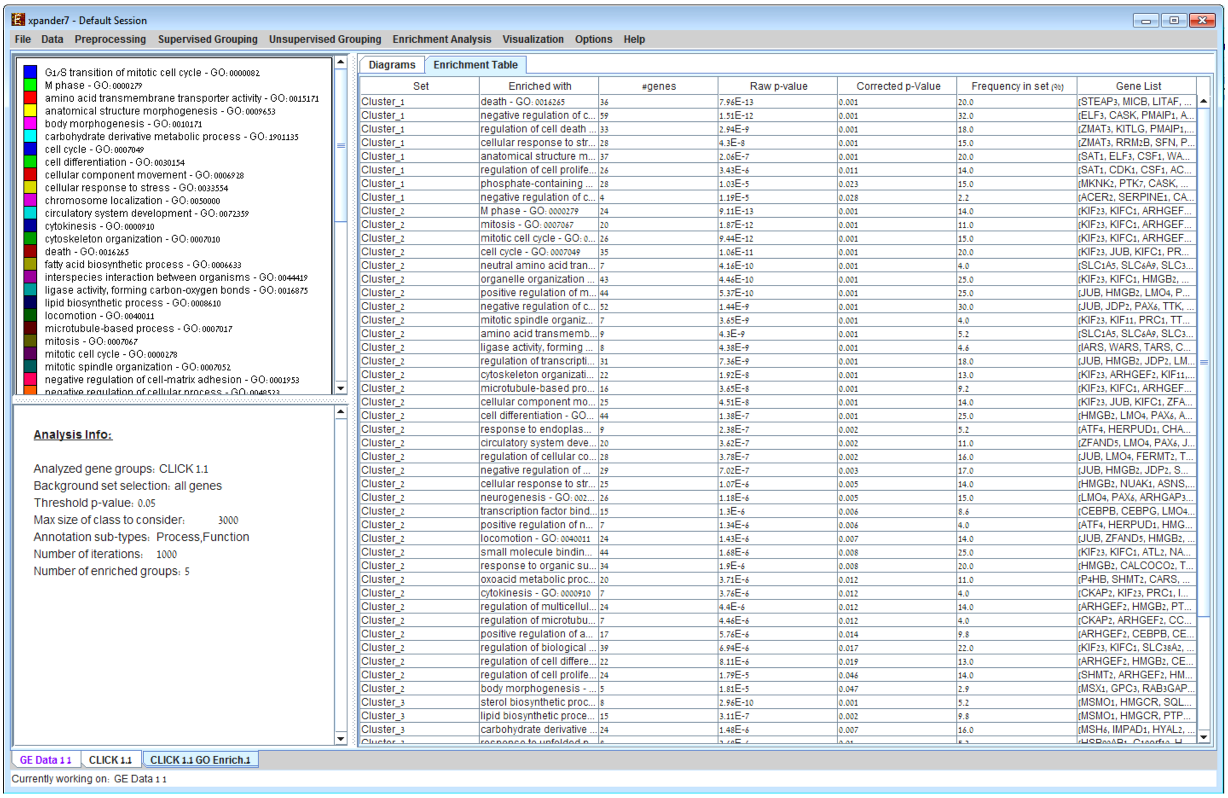
cell cycle - GO:0007049
Raw p-value: 2.08e-13

Corrected p-value: 0.00100

Number of genes: 16.0
Frequency in set: 55.17%

Gene ID	Gene symbol	Probe ID
84722	PSRC1	[ENSG00000134222]
1894	ECT2	[ENSG00000114346]
81610	FAM83D	[ENSG00000101447]
29899	GPSM2	[ENSG00000121957]
27436	EML4	[ENSG00000143924]
1062	CENPE	[ENSG00000138778]
55165	CEP55	[ENSG00000138180]
4288	MKI67	[ENSG00000148773]
5347	PLK1	[ENSG00000166851]
9787	DLGAP5	[ENSG00000126787]
891	CCNB1	[ENSG00000134057]

Currently working on: GE Data 1.1



- 🟦 G1/S transition of mitotic cell cycle - GO:0000082
- 🟩 M phase - GO:0000279
- 🟥 amino acid transmembrane transporter activity - GO:0015171
- 🟨 anatomical structure morphogenesis - GO:0009653
- 🟪 body morphogenesis - GO:0010171
- 🟦 carbohydrate derivative metabolic process - GO:1901135
- 🟦 cell cycle - GO:0007049
- 🟩 cell differentiation - GO:0030154
- 🟥 cellular component movement - GO:0006928
- 🟨 cellular response to stress - GO:0033554
- 🟪 chromosome localization - GO:0050000
- 🟦 circulatory system development - GO:0072359
- 🟦 cytokinesis - GO:0000910
- 🟩 cytoskeleton organization - GO:0007010
- 🟥 death - GO:0016265
- 🟨 fatty acid biosynthetic process - GO:0006633
- 🟪 interspecies interaction between organisms - GO:0044419
- 🟦 ligase activity, forming carbon-oxygen bonds - GO:0016875
- 🟦 lipid biosynthetic process - GO:0008610
- 🟩 locomotion - GO:0040011
- 🟥 microtubule-based process - GO:0007017
- 🟨 mitosis - GO:0007067
- 🟪 mitotic cell cycle - GO:0000278
- 🟦 mitotic spindle organization - GO:0007052
- 🟥 negative regulation of cell-matrix adhesion - GO:0001953
- 🟧 negative regulation of cellular process - GO:0048523

Analysis Info:

Analyzed gene groups: CLICK 1.1
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 groups: 5

Set	Enriched with	#genes	Raw p-value	Corrected p-Value	Frequency in set (%)	Gene List
Cluster_1	death - GO:0016265	36	7.96E-13	0.001	20.0	[STEAP3, MICB, LITAF, ...
Cluster_1	negative regulation of c...	59	1.51E-12	0.001	32.0	[ELF3, CASK, PMAIP1, A...
Cluster_1	regulation of cell death ...	33	2.94E-9	0.001	18.0	[ZMAT3, KITLG, PMAIP1,...
Cluster_1	cellular response to str...	28	4.3E-8	0.001	15.0	[ZMAT3, RRM2B, SFN, P...
Cluster_1	anatomical structure m...	37	2.06E-7	0.001	20.0	[SAT1, ELF3, CSF1, WA...
Cluster_1	regulation of cell prolife...	26	3.43E-6	0.011	14.0	[SAT1, CDK1, CSF1, AC...
Cluster_1	phosphate-containing ...	28	1.03E-5	0.023	15.0	[MKNK2, PTK7, CASK, ...
Cluster_1	negative regulation of c...	4	1.19E-5	0.028	2.2	[ACER2, SERPINE1, CA...
Cluster_2	M phase - GO:0000279	24	9.11E-13	0.001	14.0	[KIF23, KIFC1, ARHGEF...
Cluster_2	mitosis - GO:0007067	20	1.87E-12	0.001	11.0	[KIF23, KIFC1, ARHGEF...
Cluster_2	mitotic cell cycle - GO:0...	26	9.44E-12	0.001	15.0	[KIF23, KIFC1, ARHGEF...
Cluster_2	cell cycle - GO:0007049	35	1.06E-11	0.001	20.0	[KIF23, JUB, KIFC1, PR...
Cluster_2	neutral amino acid tran...	7	4.16E-10	0.001	4.0	[SLC1A5, SLC6A9, SLC3...
Cluster_2	organelle organization ...	43	4.46E-10	0.001	25.0	[KIF23, KIFC1, HMGB2, ...
Cluster_2	positive regulation of m...	44	5.37E-10	0.001	25.0	[JUB, HMGB2, LMO4, P...
Cluster_2	negative regulation of c...	52	1.44E-9	0.001	30.0	[JUB, JDP2, PAX6, TTK, ...
Cluster_2	mitotic spindle organiz...	7	3.65E-9	0.001	4.0	[KIF23, KIF11, PRC1, TT...
Cluster_2	amino acid transmemb...	9	4.3E-9	0.001	5.2	[SLC1A5, SLC6A9, SLC3...
Cluster_2	ligase activity, forming ...	8	4.38E-9	0.001	4.6	[IARS, WARS, TARS, C...
Cluster_2	regulation of transcripti...	31	7.36E-9	0.001	18.0	[JUB, HMGB2, JDP2, LM...
Cluster_2	cytoskeleton organizati...	22	1.92E-8	0.001	13.0	[KIF23, ARHGEF2, KIF11,...
Cluster_2	microtubule-based pro...	16	3.65E-8	0.001	9.2	[KIF23, KIFC1, ARHGEF...
Cluster_2	cellular component mo...	25	4.51E-8	0.001	14.0	[KIF23, JUB, KIFC1, ZFA...
Cluster_2	cell differentiation - GO...	44	1.38E-7	0.001	25.0	[HMGB2, LMO4, PAX6, A...
Cluster_2	response to endoplas...	9	2.38E-7	0.002	5.2	[ATF4, HERPUD1, CHA...
Cluster_2	circulatory system deve...	20	3.62E-7	0.002	11.0	[ZFAND5, LMO4, PAX6, J...
Cluster_2	regulation of cellular co...	28	3.78E-7	0.002	16.0	[JUB, LMO4, FERMT2, T...
Cluster_2	negative regulation of ...	29	7.02E-7	0.003	17.0	[JUB, HMGB2, JDP2, S...
Cluster_2	cellular response to str...	25	1.07E-6	0.005	14.0	[HMGB2, NUAK1, ASNS,...
Cluster_2	neurogenesis - GO:002...	26	1.18E-6	0.005	15.0	[LMO4, PAX6, ARHGAP3...
Cluster_2	transcription factor bind...	15	1.3E-6	0.006	8.6	[CEBPB, CEBPG, LMO4...
Cluster_2	positive regulation of n...	7	1.34E-6	0.006	4.0	[ATF4, HERPUD1, HMG...
Cluster_2	locomotion - GO:0040011	24	1.43E-6	0.007	14.0	[JUB, ZFAND5, HMGB2, ...
Cluster_2	small molecule bindin...	44	1.68E-6	0.008	25.0	[KIF23, KIFC1, ATL2, NA...
Cluster_2	response to organic su...	34	1.9E-6	0.008	20.0	[HMGB2, CALCOCO2, T...
Cluster_2	oxoacid metabolic proc...	20	3.71E-6	0.012	11.0	[P4HB, SHMT2, CARS, ...
Cluster_2	cytokinesis - GO:0000910	7	3.76E-6	0.012	4.0	[CKAP2, KIF23, PRC1, I...
Cluster_2	regulation of multicellul...	24	4.4E-6	0.012	14.0	[ARHGEF2, HMGB2, PT...
Cluster_2	regulation of microtubu...	7	4.46E-6	0.012	4.0	[CKAP2, ARHGEF2, CC...
Cluster_2	positive regulation of a...	17	5.76E-6	0.014	9.8	[ARHGEF2, CEBPB, CE...
Cluster_2	regulation of biological ...	39	6.94E-6	0.017	22.0	[KIF23, KIFC1, SLC38A2, ...
Cluster_2	regulation of cell differe...	22	8.11E-6	0.019	13.0	[ARHGEF2, HMGB2, CE...
Cluster_2	regulation of cell prolife...	24	1.79E-5	0.046	14.0	[SHMT2, ARHGEF2, HM...
Cluster_2	body morphogenesis - ...	5	1.81E-5	0.047	2.9	[MSX1, GPC3, RAB3GAP...
Cluster_3	sterol biosynthetic proc...	8	2.96E-10	0.001	5.2	[MSMO1, HMGCR, SQL...
Cluster_3	lipid biosynthetic proce...	15	3.11E-7	0.002	9.8	[MSMO1, HMGCR, PTP...
Cluster_3	carbohydrate derivative ...	24	1.48E-6	0.007	16.0	[MSH6, IMPAD1, HYAL2, ...

Currently working on: GE Data 1 1

Annotation files for each organism are updated on a regular basis (for more information, refer to the Supplied Files section).

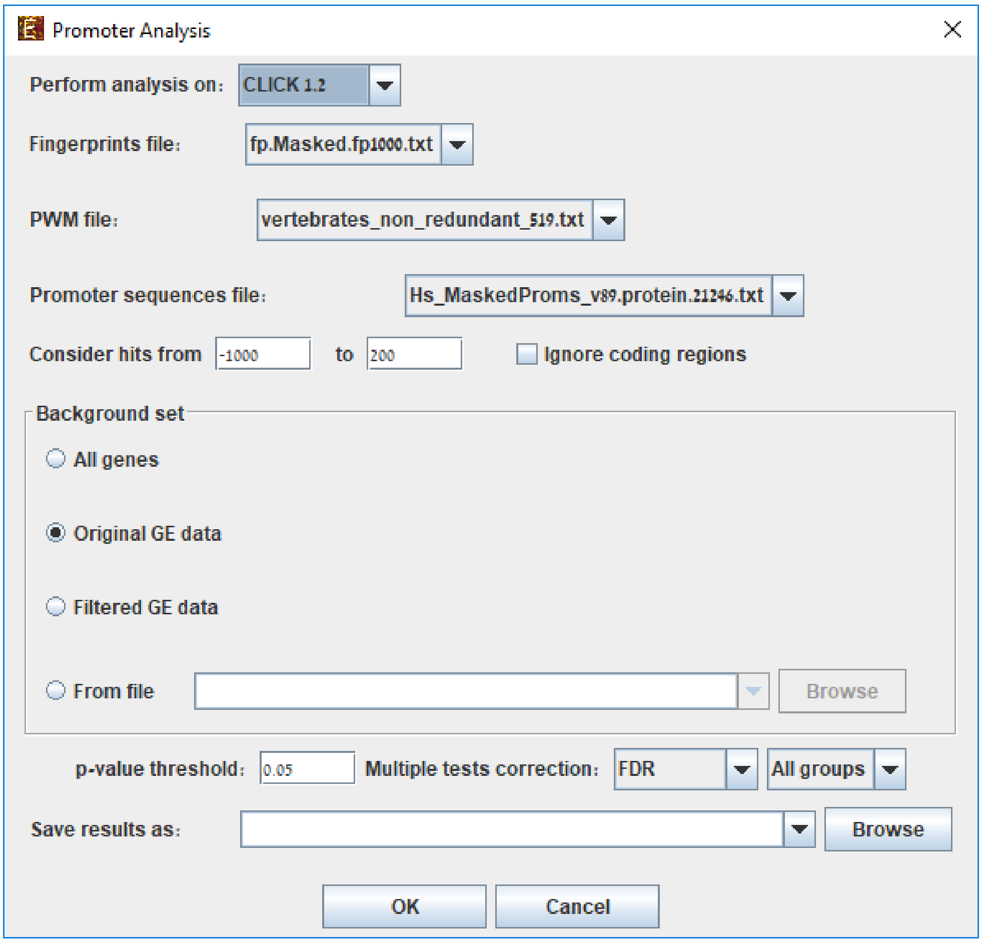
The results of this analysis can be exported to a text file by selecting *File>>Export to text* when the corresponding view is the selected tab OR by using the *File>>Save All* option, which will export all solutions within a session to text and image files.

Promoter Analysis

PRIMA

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

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



Perform analysis on: CLICK 1.2

Fingerprints file: fp.Masked.fp1000.txt

PWM file: vertebrates_non_redundant_519.txt

Promoter sequences file: Hs_MaskedProms_v89.protein.21246.txt

Consider hits from -1000 to 200 □ Ignore coding regions

Background set

○ All genes

● Original GE data

○ Filtered GE data

○ From file

p-value threshold: 0.05 Multiple tests correction: FDR All groups

Save results as:

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

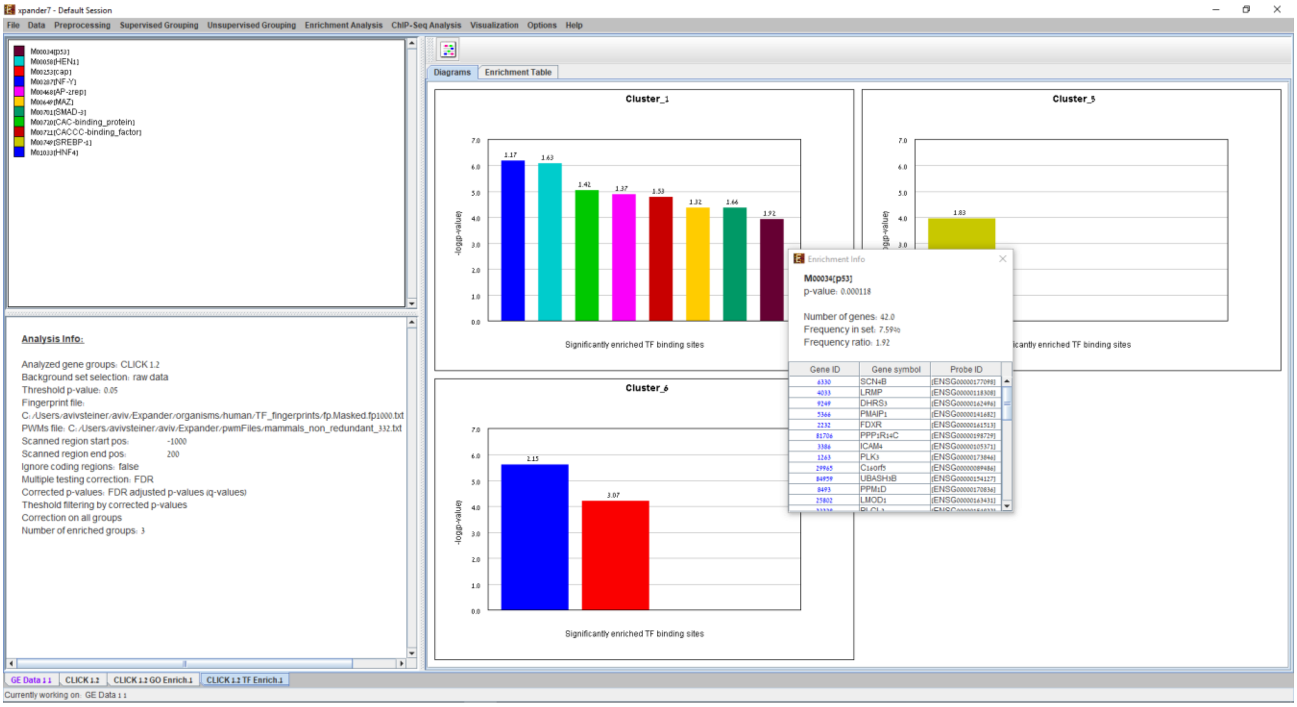
Field	Description
Perform analysis on	The grouping solution on which the analysis will be performed.
Include back nodes	Include genes that are part of the module' but not included in the GE data (Relevant only if the analysis is performed on modules, detected by

	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 FDR, Bonferroni or None (when set to Bonferroni/FDR 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 tab is added to the main window. It contains general information regarding the analysis, a sort-able table holding all detected enrichments (set ID, TF binding site, p-value, etc.) and a multi-histogram panel along with a color index (mapping each color to a corresponding TF binding site). The multi-histogram panel contains one histogram for each probe/gene set/group in which enrichment has been detected. Each histogram contains a column for each significant (more frequent than would be expected by random) TF binding site. The definition of significant depends on the user's selection of threshold p-value. i.e., a TF binding site is considered significantly enriched in a cluster/bicluster if its corrected p-value is lower than the preset threshold p-value.

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

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



Analysis Info:

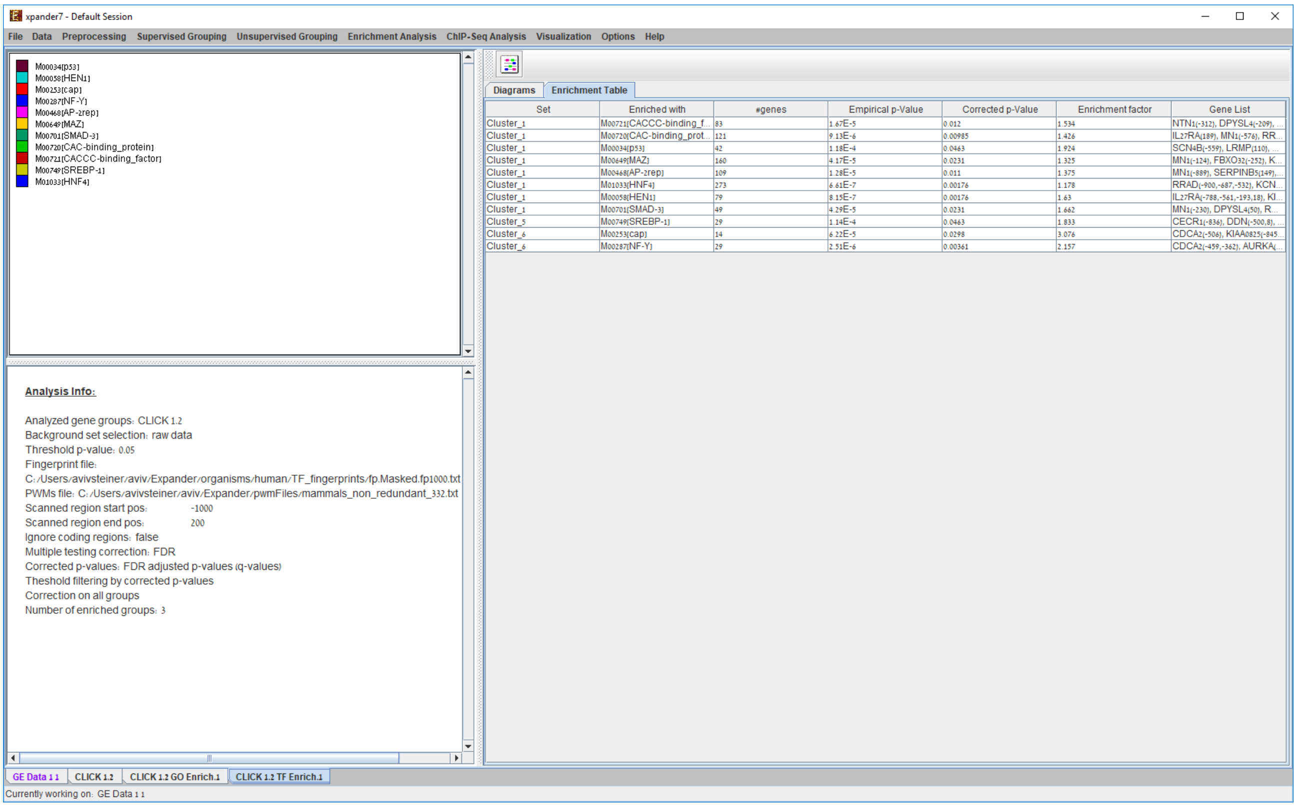
Analyzed gene groups: CLICK 1.2
Background set selection: raw data
Threshold p-value: 0.05
Fingerprint file:
C: /Users/avivsteiner/aviv/Expander/organisms/human/TF_fingerprints/fp.Masked.fp1000.txt
PWMs file: C: /Users/avivsteiner/aviv/Expander/pwmFiles/mammals_non_redundant_332.txt
Scanned region start pos: -1000
Scanned region end pos: 200
Ignore coding regions: false
Multiple testing correction: FDR
Corrected p-values: FDR adjusted p-values (q-values)
Theshold filtering by corrected p-values
Correction on all groups
Number of enriched groups: 3

M00034[p53]
p-value: 0.000118

Number of genes: 42.0
Frequency in set: 7.59%
Frequency ratio: 1.92

Gene ID	Gene symbol	Probe ID
6330	SCN4B	[ENSG00000177098]
4033	LRMP	[ENSG00000118308]
9249	DHRS3	[ENSG00000162496]
5366	PMAIP1	[ENSG00000141682]
2232	FDXR	[ENSG00000161513]
81706	PPP1R14C	[ENSG00000198729]
3386	ICAM4	[ENSG00000105371]
1263	PLK3	[ENSG00000173846]
29965	C16orf5	[ENSG00000089486]
84959	UBASH3B	[ENSG00000154127]
8493	PPM1D	[ENSG00000170836]
25802	LMOD1	[ENSG00000163431]

Currently working on: GE Data 1 1



🟫 M00034[p53]
🟦 M00058[HEN1]
🟥 M00253[cap]
🟦 M00287[NF-Y]
🟪 M00468[AP-2rep]
🟨 M00649[MAZ]
🟩 M00701[SMAD-3]
🟩 M00720[CAC-binding_protein]
🟥 M00721[CACCC-binding_factor]
🟨 M00749[SREBP-1]
🟦 M01033[HNF4]

Set	Enriched with	#genes	Empirical p-Value	Corrected p-Value	Enrichment factor	Gene List
Cluster_1	M00721[CACCC-binding_f...	83	1.67E-5	0.012	1.534	NTN1(-312), DPYSL4(-209), ...
Cluster_1	M00720[CAC-binding_prot...	121	9.13E-6	0.00985	1.426	IL27RA(189), MN1(-576), RR...
Cluster_1	M00034[p53]	42	1.18E-4	0.0463	1.924	SCN4B(-559), LRMP(110), ...
Cluster_1	M00649[MAZ]	160	4.17E-5	0.0231	1.325	MN1(-124), FBXO32(-252), K...
Cluster_1	M00468[AP-2rep]	109	1.28E-5	0.011	1.375	MN1(-889), SERPINB5(149),...
Cluster_1	M01033[HNF4]	273	6.61E-7	0.00176	1.178	RRAD(-900,-687,-532), KCN...
Cluster_1	M00058[HEN1]	79	8.15E-7	0.00176	1.63	IL27RA(-788,-561,-193,18), KI...
Cluster_1	M00701[SMAD-3]	49	4.29E-5	0.0231	1.662	MN1(-230), DPYSL4(50), R...
Cluster_5	M00749[SREBP-1]	29	1.14E-4	0.0463	1.833	CECR1(-836), DDN(-500,8), ...
Cluster_6	M00253[cap]	14	6.22E-5	0.0298	3.076	CDCA2(-506), KIAA0825(-845...
Cluster_6	M00287[NF-Y]	29	2.51E-6	0.00361	2.157	CDCA2(-459,-362), AURKA(...

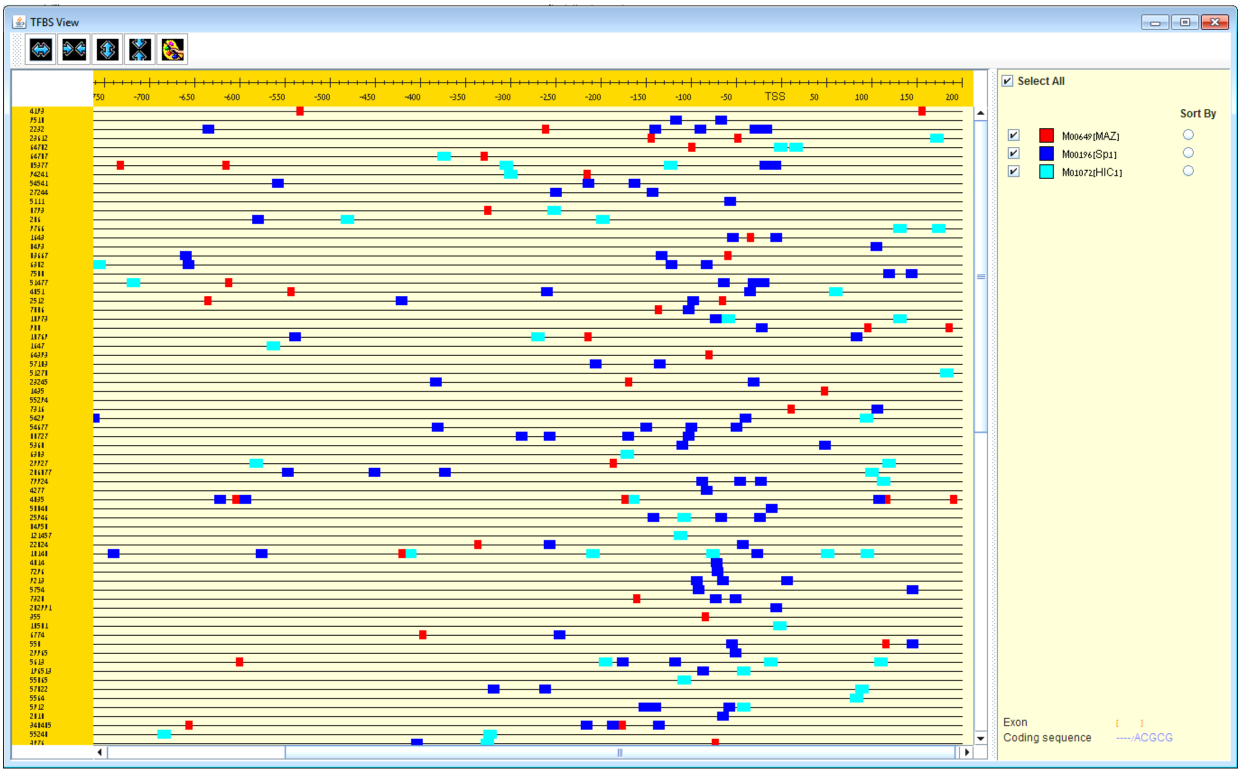
Analysis Info:

Analyzed gene groups: CLICK 1.2
Background set selection: raw data
Threshold p-value: 0.05
Fingerprint file:
C:/Users/avivsteiner/aviv/Expander/organisms/human/TF_fingerprints/fp.Masked.fp1000.txt
PWMs file: C:/Users/avivsteiner/aviv/Expander/pwmFiles/mammals_non_redundant_332.txt
Scanned region start pos: -1000
Scanned region end pos: 200
Ignore coding regions: false
Multiple testing correction: FDR
Corrected p-values: FDR adjusted p-values (q-values)
Theshold filtering by corrected p-values
Correction on all groups
Number of enriched groups: 3

Currently working on: GE Data 1 1

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

Promoter and TF fingerprint files for each organism are updated on a regular basis (for more information, refer to the Supplied Files section).

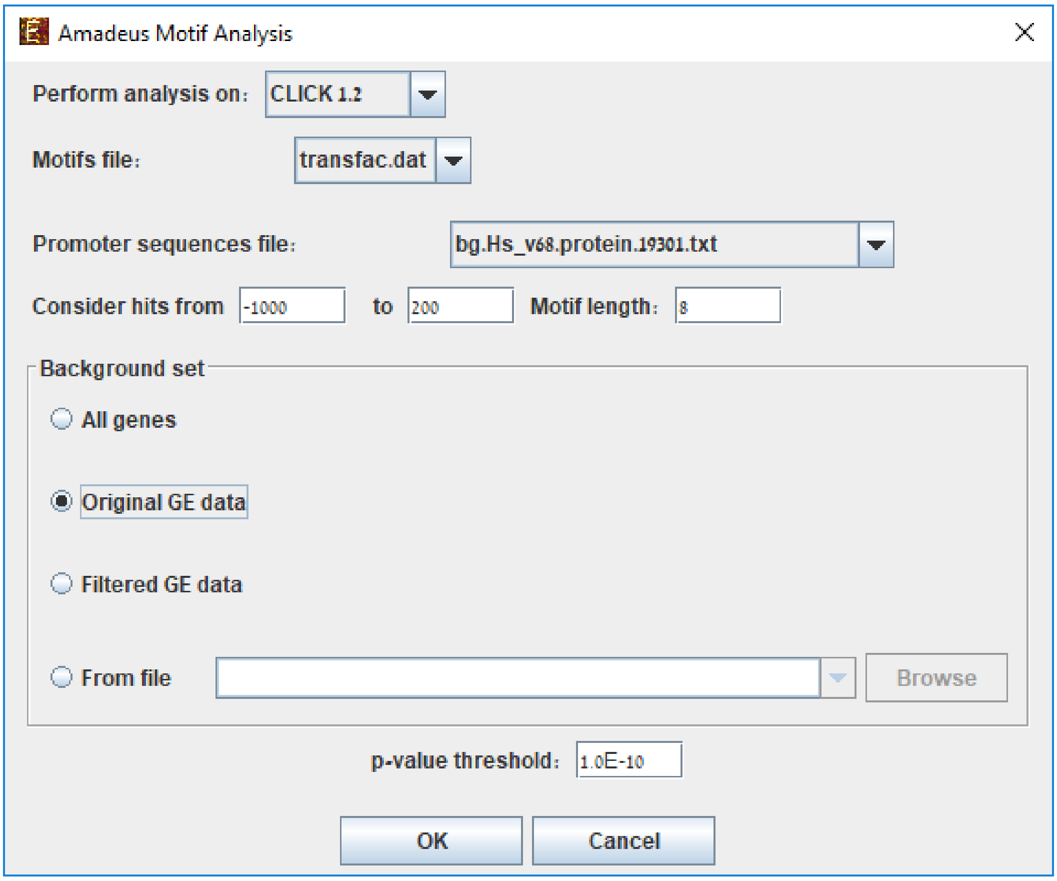


AMADEUS

Another option for performing promoter analysis, is finding enriched motifs using AMADEUS. Amadeus is a tool for de novo motif discovery. It seeks for motifs that are enriched in the promoters of a target set of genes compared to the background set. Such analysis can be applied to other sets of sequences (e.g., ChIP-Seq peaks, enhancers, etc.).

In order to perform motif enrichments analysis, select Enrichment Analysis >> Promoter Analysis >> AMADEUS.

The following dialog box will appear:



Amadeus Motif Analysis

Perform analysis on: CLICK 1.2

Motifs file: transfac.dat

Promoter sequences file: bg.Hs_v68.protein.19301.txt

Consider hits from -1000 to 200 Motif length: 8

Background set

○ All genes

● Original GE data

○ Filtered GE data

○ From file

p-value threshold: 1.0E-10

The different parameters that can be set via this dialog box are:

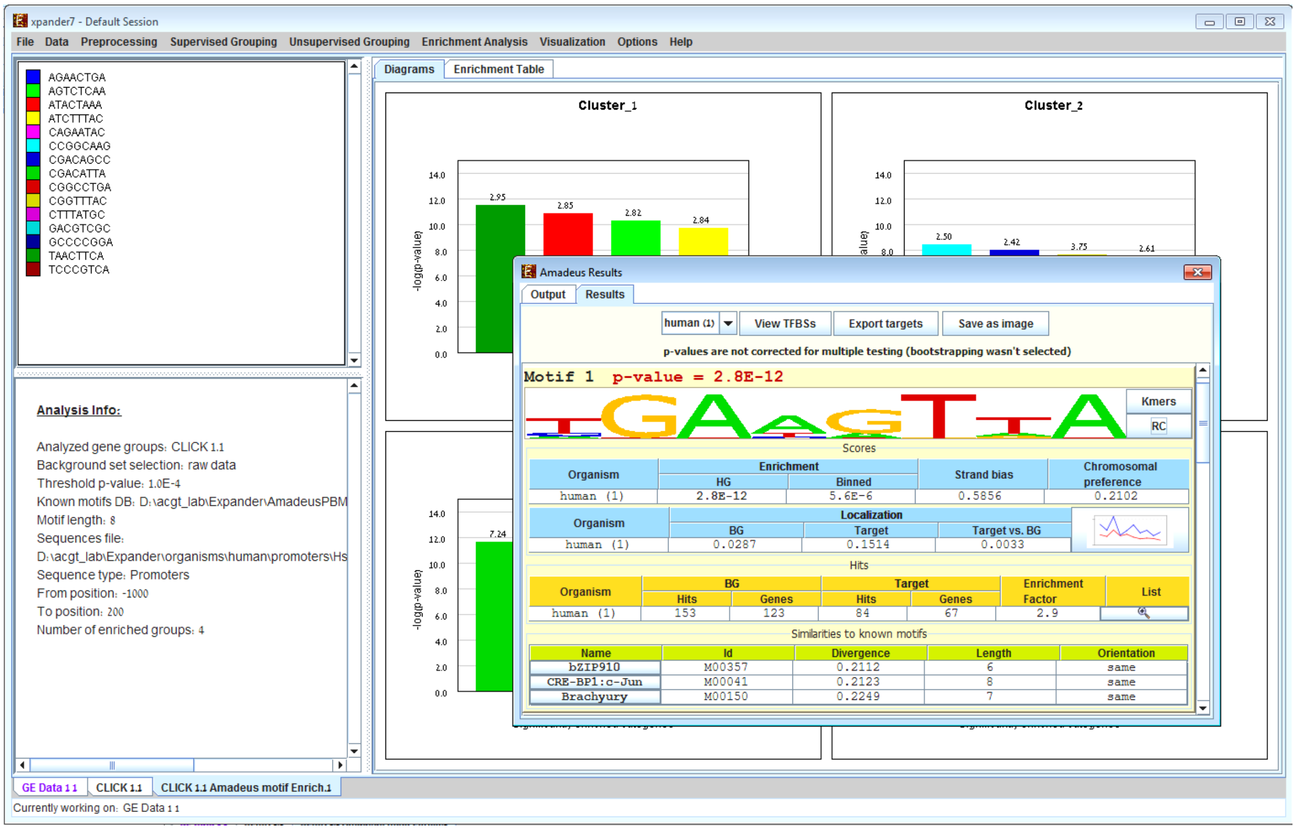
Field	Description
Perform analysis on	The grouping solution on which the analysis will be performed.
Motifs file	A matrix table of known motifs in TRANSFAC format. The default is transfac.dat, a public release of TRANSFAC from *2005*.
Promoter sequences file	Contains the promoter sequences in fasta format. Automatically set according to the selection of the organism. Can be 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).
P-value threshold	A motif will be considered significantly enriched in a tested set if its corrected p-value is lower than this threshold.

After AMDEUS analysis is performed, an Amadeus motif solution visualization tab is added to the main window. It contains general information regarding the analysis, a sortable table holding all detected enrichments (set ID, Motif binding site, p-value, etc.) and a multi-histogram panel along with a color index (mapping each color to a corresponding motif). The multi-histogram panel contains one histogram for each probe/gene set/group in which enrichment has been detected. Each histogram contains a column for each significantly enriched motif. The significance depends on the user's selection of p-value threshold. i.e., a motif is considered significantly enriched in a set if its corrected p-value is lower than the preset p-value threshold.

The height of a column is proportional to the significance of this enrichment (i.e. height = -log(p-value)), and the frequency ratio (frequency in the target set divided by frequency in the background set) is written on top of the column.

Upon clicking on a column, a dialog box is displayed containing the full AMADEUS graphical output relevant to the clicked set: This output shows rich information for every significant motif detected for that set. For further explanation regarding AMDEUS output please refer to AMADEUS/ALLEGRO site under section 8.



Analysis Info:

Analyzed gene groups: CLICK 1.1
Background set selection: raw data
Threshold p-value: 1.0E-4
Known motifs DB: D:\acgt_lab\Expander\AmadeusPBM
Motif length: 8
Sequences file:
D:\acgt_lab\Expander\organisms\human\promoters\Hs
Sequence type: Promoters
From position: -1000
To position: 200
Number of enriched groups: 4

p-values are not corrected for multiple testing (bootstrapping wasn't selected)

Motif 1 p-value = 2.8E-12

Scores

Organism	Enrichment		Strand bias	Chromosomal preference
	HG	Binned		
human (1)	2.8E-12	5.6E-6	0.5856	0.2102

Organism	Localization		
	BG	Target	Target vs. BG
human (1)	0.0287	0.1514	0.0033

Hits

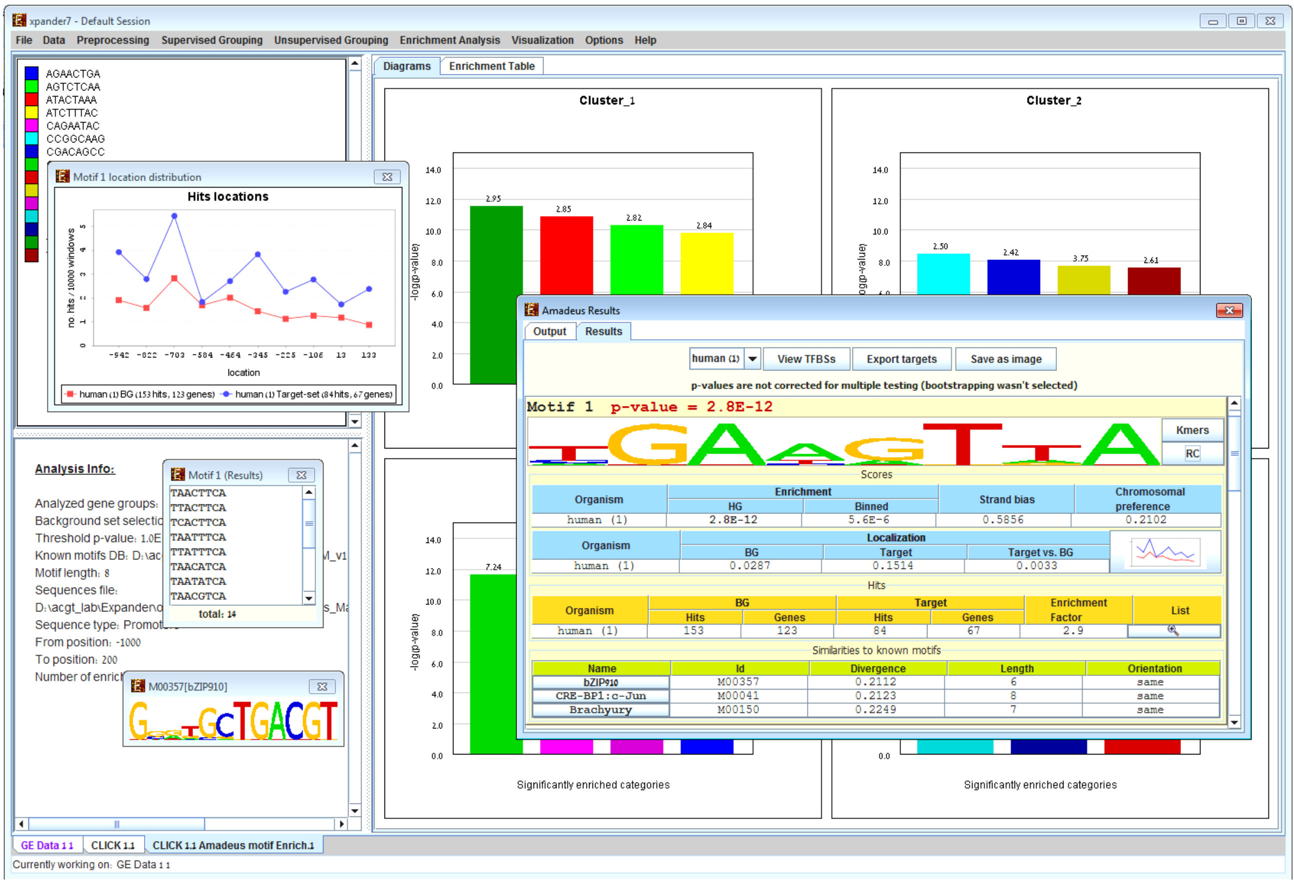
Organism	BG		Target		Enrichment Factor	List
	Hits	Genes	Hits	Genes		
human (1)	153	123	84	67	2.9	🔍

Similarities to known motifs

Name	Id	Divergence	Length	Orientation
bZIP910	M00357	0.2112	6	same
CRE-BP1:c-Jun	M00041	0.2123	8	same
Brachyury	M00150	0.2249	7	same

Currently working on: GE Data 1 1

Clicking on some fields in the graphical output window opens additional windows with more information, e.g. frequency of motifs along the promoter regions in the target and background set, a list of motif k-mer, and the motif logo:



xpander7 - Default Session
File Data Preprocessing Supervised Grouping Unsupervised Grouping Enrichment Analysis Visualization Options Help
AGAACTGA
AGTCTCAA
ATACTAAA
ATCTTTAC
CAGAATAC
CCGGCAAG
CGACAGCC
Motif 1 location distribution
Hits locations
no hits / 10000 windows
0
1
2
3
4
5
-942 -822 -703 -584 -464 -345 -225 -106 13 133
location
human (1) BG (153 hits, 123 genes)
human (1) Target-set (84 hits, 67 genes)
Diagrams
Enrichment Table
Cluster_1
Cluster_2
-log(p-value)
14.0
12.0
10.0
8.0
6.0
4.0
2.0
0.0
2.95
2.85
2.82
2.84
2.50
2.42
3.75
2.61
7.24
Significantly enriched categories
Significantly enriched categories
GE Data 1 1
CLICK 1.1
CLICK 1.1 Amadeus motif Enrich.1
Currently working on: GE Data 1 1

Analysis Info:

Analyzed gene groups:
Background set selectio
Threshold p-value: 1.0E
Known motifs DB: D:\ac M_v1
Motif length: 8
Sequences file:
D:\acgt_lab\Expander\o s_Ma
Sequence type: Promot
From position: -1000
To position: 200
Number of enric

Motif 1 (Results)

TAACTTCA
TTACTTCA
TCACTTCA
TAATTTCA
TTATTTCA
TAACATCA
TAATATCA
TAACGTCA

total: 14

M00357[bZIP910]

Amadeus Results

Output | Results

human (1) | View TFBSs | Export targets | Save as image

p-values are not corrected for multiple testing (bootstrapping wasn't selected)

Motif 1 p-value = 2.8E-12

Kmers | RC

Scores

Organism	Enrichment		Strand bias	Chromosomal preference
	HG	Binned		
human (1)	2.8E-12	5.6E-6	0.5856	0.2102

Organism	Localization		
	BG	Target	Target vs. BG
human (1)	0.0287	0.1514	0.0033

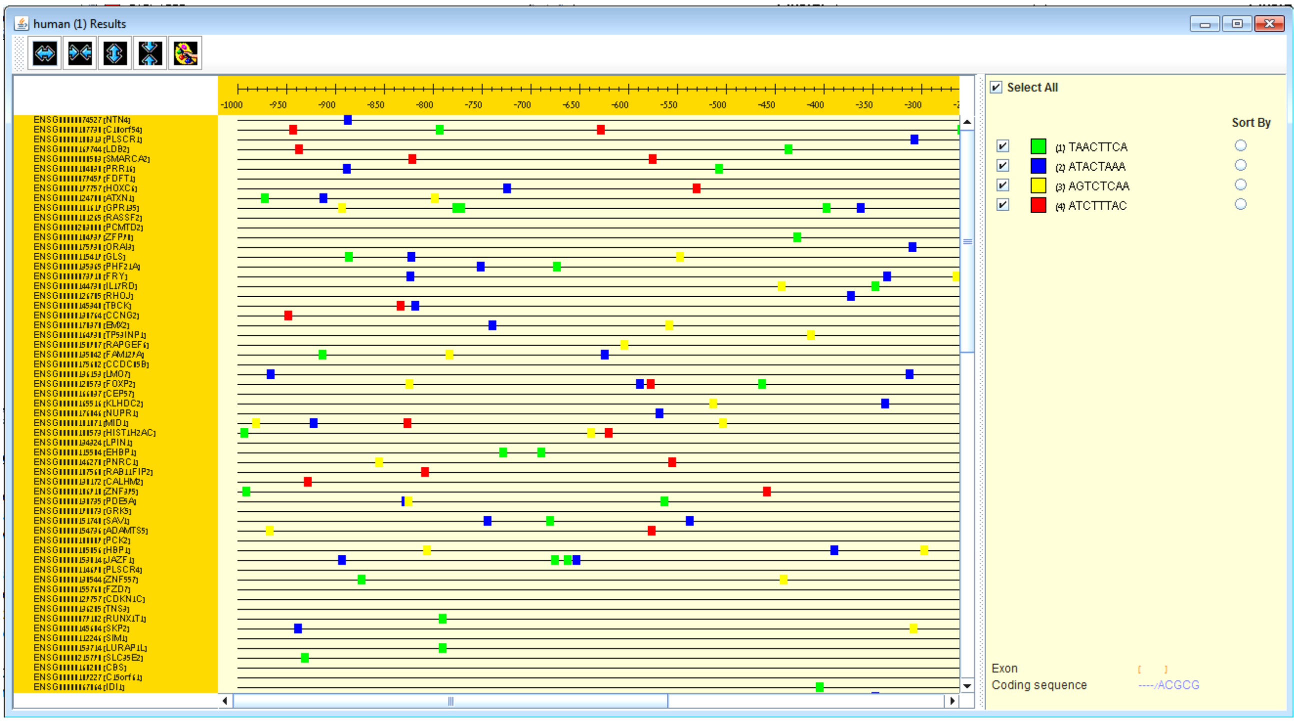
Hits

Organism	BG		Target		Enrichment Factor	List
	Hits	Genes	Hits	Genes		
human (1)	153	123	84	67	2.9	

Similarities to known motifs

Name	Id	Divergence	Length	Orientation
bZIP910	M00357	0.2112	6	same
CRE-BP1:c-Jun	M00041	0.2123	8	same
Brachyury	M00150	0.2249	7	same

Upon clicking on "View TFBSs" button inside AMADEUS visualization the following dialog box is displayed marking the binding sites of the enriched motifs on the promoters:



☑ Select All

☑	🟩	(1) TAACTTCA
☑	🟦	(2) ATACTAAA
☑	🟨	(3) AGTCTCAA
☑	🟥	(4) ATCTTTAC

ENSG00000074527 [NTN4]
ENSG00000182919 [C11orf54]
ENSG00000188313 [PLSCR1]
ENSG00000169744 [LDB2]
ENSG00000080503 [SMARCA2]
ENSG00000184838 [PRR16]
ENSG00000079459 [FDFT1]
ENSG00000197757 [HOXC6]
ENSG00000124788 [ATXN1]
ENSG00000181619 [GPR135]
ENSG00000101265 [RASSF2]
ENSG00000203880 [PCMTD2]
ENSG00000186660 [ZFP91]
ENSG00000175938 [ORAI3]
ENSG00000115419 [GLS]
ENSG00000135365 [PHF21A]
ENSG00000073910 [FRY]
ENSG00000144730 [IL17RD]
ENSG00000126785 [RHOJ]
ENSG00000145348 [TBCK]
ENSG00000138764 [CCNG2]
ENSG00000170370 [EMX2]
ENSG00000164938 [TP53INP1]
ENSG00000158987 [RAPGEF6]
ENSG00000134590 [FAM127A]
ENSG00000175602 [CCDC85B]
ENSG00000136153 [LMO7]
ENSG00000128573 [FOXP2]
ENSG00000166037 [CEP57]
ENSG00000165516 [KLHDC2]
ENSG00000176046 [NUPR1]
ENSG00000101871 [MID1]
ENSG00000180573 [HIST1H2AC]
ENSG00000134324 [LPIN1]
ENSG00000115504 [EHBP1]
ENSG00000146278 [PNRC1]
ENSG00000107560 [RAB11FIP2]
ENSG00000138172 [CALHM2]
ENSG00000186918 [ZNF395]
ENSG00000138735 [PDE5A]
ENSG00000198873 [GRK5]
ENSG00000151748 [SAV1]
ENSG00000154736 [ADAMTS5]
ENSG00000100889 [PCK2]
ENSG00000105856 [HBP1]
ENSG00000153814 [JAZF1]
ENSG00000114698 [PLSCR4]
ENSG00000130544 [ZNF557]
ENSG00000155760 [FZD7]
ENSG00000129757 [CDKN1C]
ENSG00000136205 [TNS3]
ENSG00000079102 [RUNX1T1]
ENSG00000145604 [SKP2]
ENSG00000112246 [SIM1]
ENSG00000153714 [LURAP1L]
ENSG00000215790 [SLC35E2]
ENSG00000160200 [CBS]
ENSG00000189227 [C15orf61]
ENSG00000067064 [IDI1]

Promoter and TF fingerprint files for each organism are updated on a regular basis (for more information, refer to the Supplied Files section).

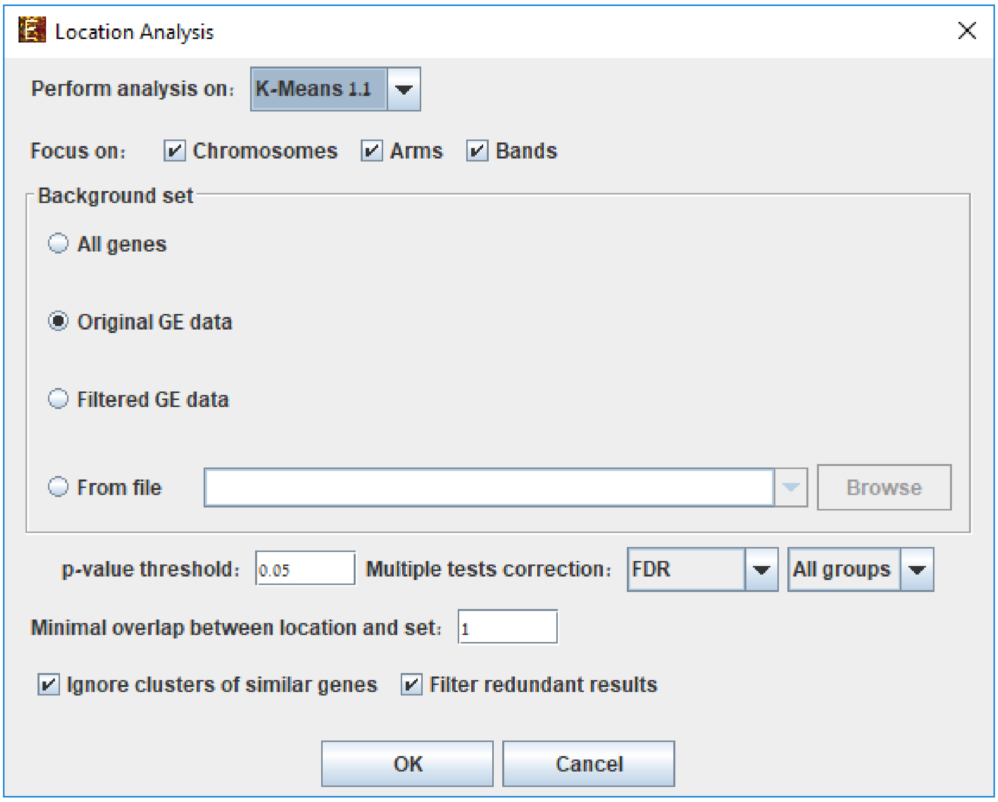
The results of this analysis can be exported to text by selecting File>>Export to text when the corresponding view is the selected tab.

Location Enrichment Analysis

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

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

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



Location Analysis

Perform analysis on: K-Means 1.1

Focus on: ☑ Chromosomes ☑ Arms ☑ Bands

Background set

○ All genes

● Original GE data

○ Filtered GE data

○ From file

p-value threshold: 0.05 Multiple tests correction: FDR All groups

Minimal overlap between location and set: 1

☑ Ignore clusters of similar genes ☑ Filter redundant results

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

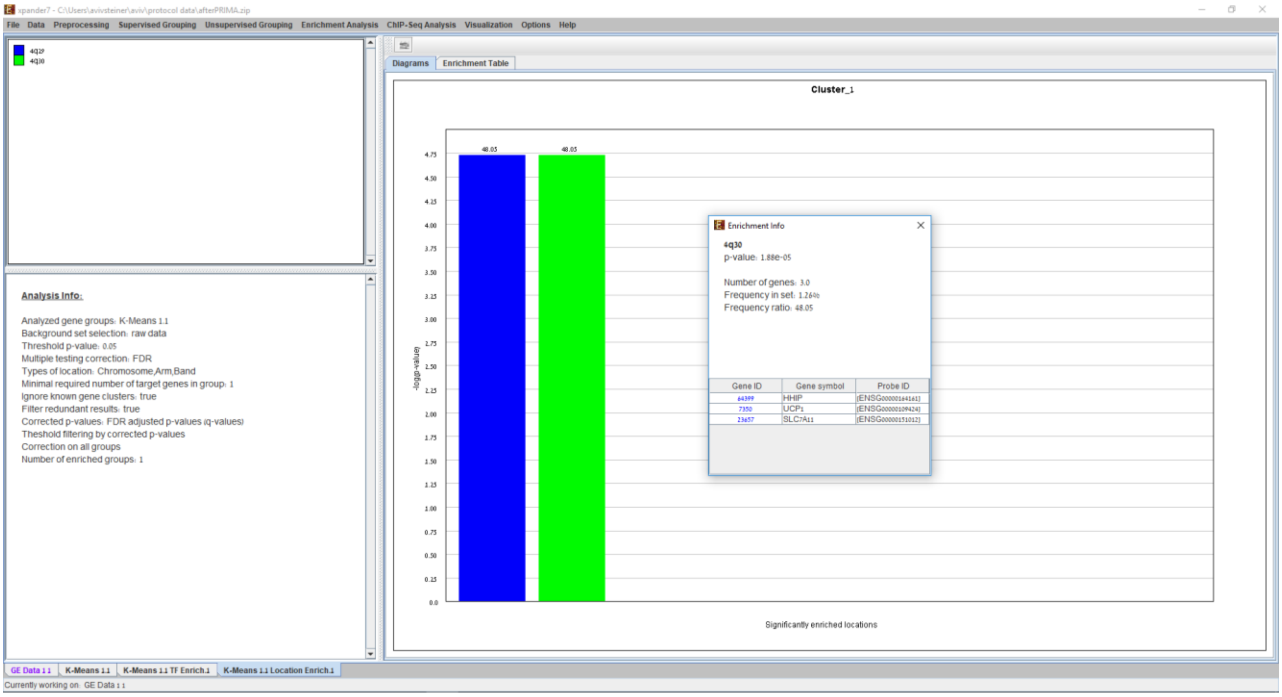
Field	Description
Perform analysis on	The grouping solution on which the analysis will be performed.
Include back nodes	Include genes that are part of the module' but not included in the GE data (Relevant only if the analysis is performed on modules, detected by network based algorithm)
Focus on (Chromosomes, 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 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 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	If selected, genes from known homology

similar genes*	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 is added to the main window. It contains general information regarding the analysis, a sort-able table holding all detected enrichments (set ID, enrichment category, p-value, etc.) and a multi-histogram panel along with a color index (mapping each color to a corresponding location). The multi-histogram panel contains one histogram for each probe/gene group in which enrichment has been detected. Each histogram contains a column for each significant (more frequent than would be expected by random) location. The definition of significant depends on the user's selection of threshold p-value i.e., a category is considered significantly enriched in a 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 (i.e. height = -log(raw p-value)), and the frequency ratio (frequency in set divided by frequency in background) is written on top of the column. Upon clicking on a column, a dialog box is displayed containing the location, corrected p-value, and a list of the genes in the group that are mapped to this location. Upon clicking on one of the gene Ids in the table, a relevant web page with information regarding this gene is displayed.



Analysis Info:

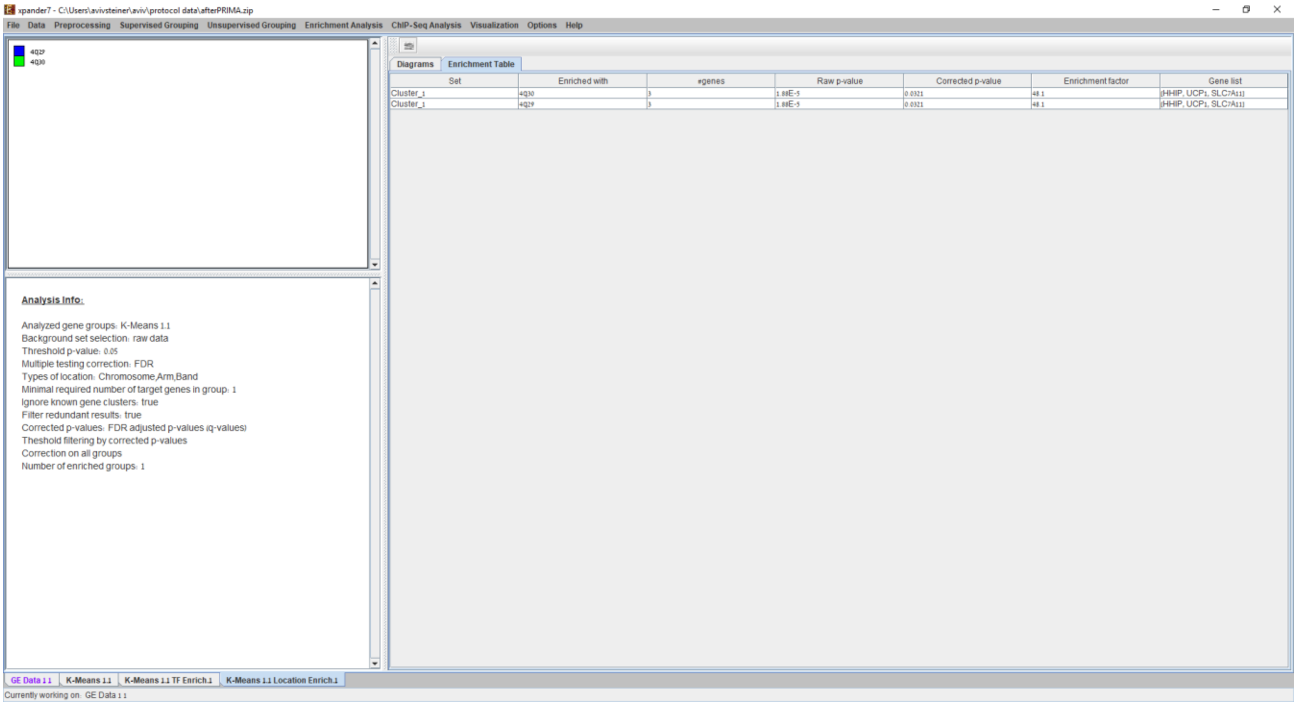
Analyzed gene groups: K-Means 1.1
Background set selection: raw data
Threshold p-value: 0.05
Multiple testing correction: FDR
Types of location: Chromosome,Arm,Band
Minimal required number of target genes in group: 1
Ignore known gene clusters: true
Filter redundant results: true
Corrected p-values: FDR adjusted p-values (q-values)
Theshold filtering by corrected p-values
Correction on all groups
Number of enriched groups: 1

4q30
p-value: 1.88e-05

Number of genes: 3.0
Frequency in set: 1.26%
Frequency ratio: 48.05

Gene ID	Gene symbol	Probe ID
64399	HHIP	[ENSG00000164161]
7350	UCP1	[ENSG00000109424]
23657	SLC7A11	[ENSG00000151012]

Currently working on: GE Data 1.1



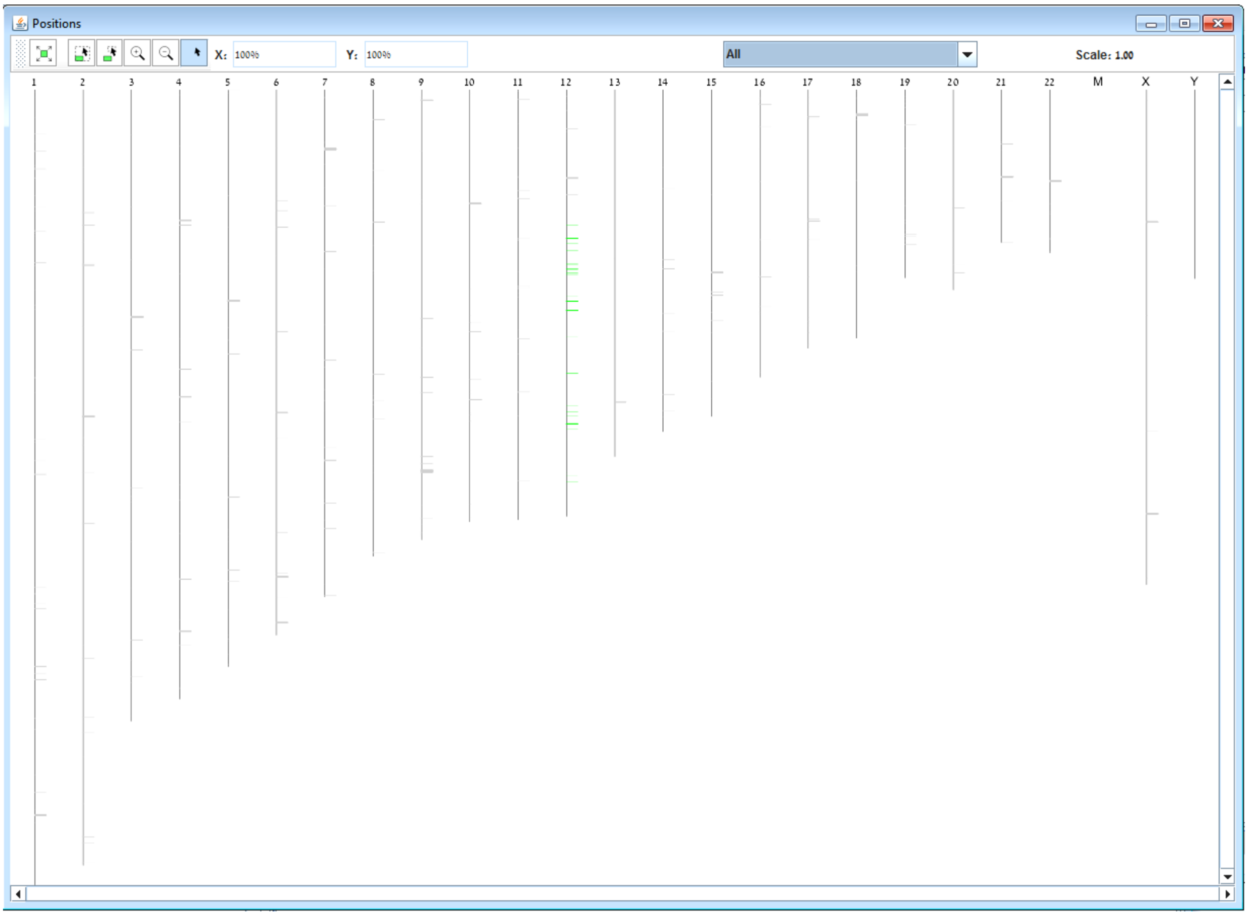
Set	Enriched with	#genes	Raw p-value	Corrected p-value	Enrichment factor	Gene list
Cluster_1	4Q30	3	1.88E-5	0.0321	48.1	[HHIP, UCP1, SLC7A11]
Cluster_1	4Q29	3	1.88E-5	0.0321	48.1	[HHIP, UCP1, SLC7A11]

Analysis Info:

Analyzed gene groups: K-Means 1.1
Background set selection: raw data
Threshold p-value: 0.05
Multiple testing correction: FDR
Types of location: Chromosome,Arm,Band
Minimal required number of target genes in group: 1
Ignore known gene clusters: true
Filter redundant results: true
Corrected p-values: FDR adjusted p-values (q-values)
Theshold filtering by corrected p-values
Correction on all groups
Number of enriched groups: 1

Currently working on: GE Data 1.1

After performing location enrichment analysis, the locations can be viewed by selecting *Enrichment Analysis >> Location Analysis >> View Locations* OR by pressing the toolbar button (▤). After selecting the gene group (cluster/bi-cluster etc.) to be viewed, a separate frame is displayed, containing an image of all chromosomes on which the positions of the genes in the group are marked. If the gene is located on an area that was identified to be enriched in that group, its position is marked in the same color to this area the enrichment results histogram.



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

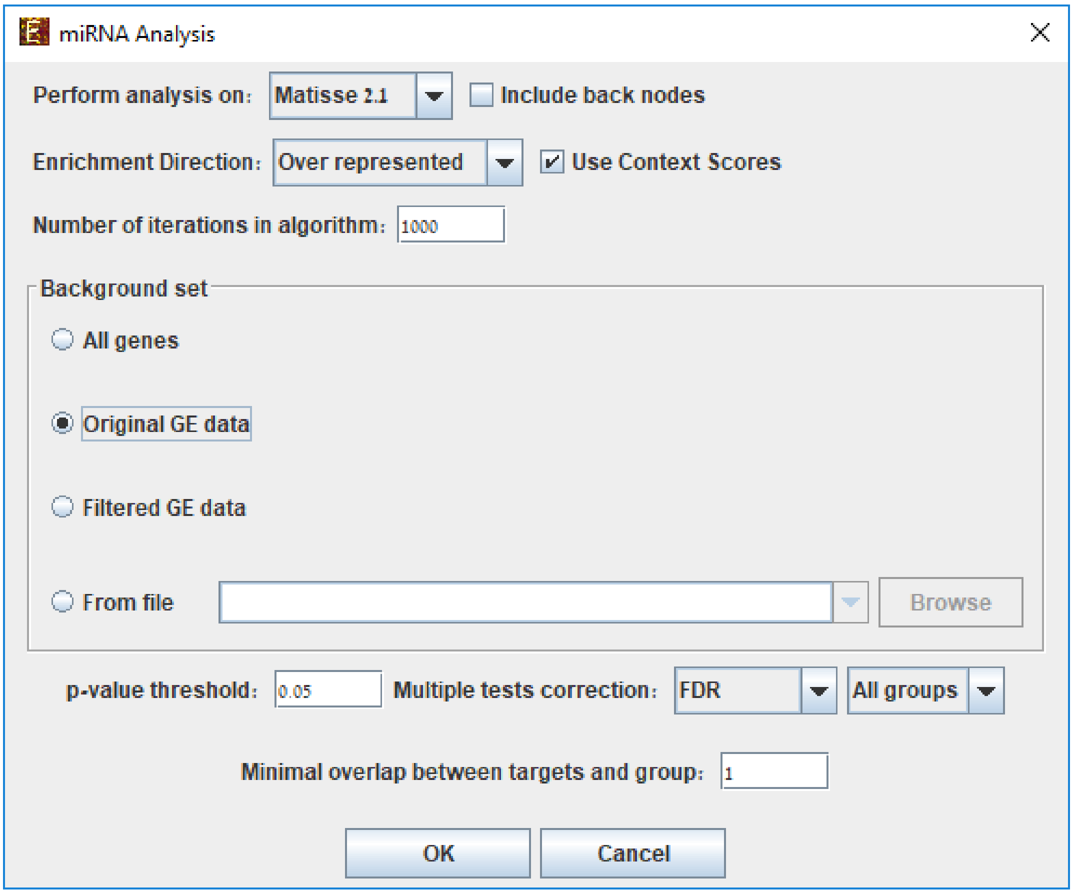
miRNA Targets Enrichment Analysis

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

Before operating miRNA enrichment analysis, the location data for the relevant organism should be downloaded from the download page (more details at introduction of Enrichment

Analysis Tools). In this analysis, hyper-geometric enrichment tests are performed, and the results can be (if requested) corrected for multiple testing using the FDR/Bonferroni correction.

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



miRNA Analysis

Perform analysis on: Matisse 2.1 □ Include back nodes

Enrichment Direction: Over represented ☑ Use Context Scores

Number of iterations in algorithm: 1000

Background set

○ All genes

● Original GE data

○ Filtered GE data

○ From file

p-value threshold: 0.05 Multiple tests correction: FDR All groups

Minimal overlap between targets and group: 1

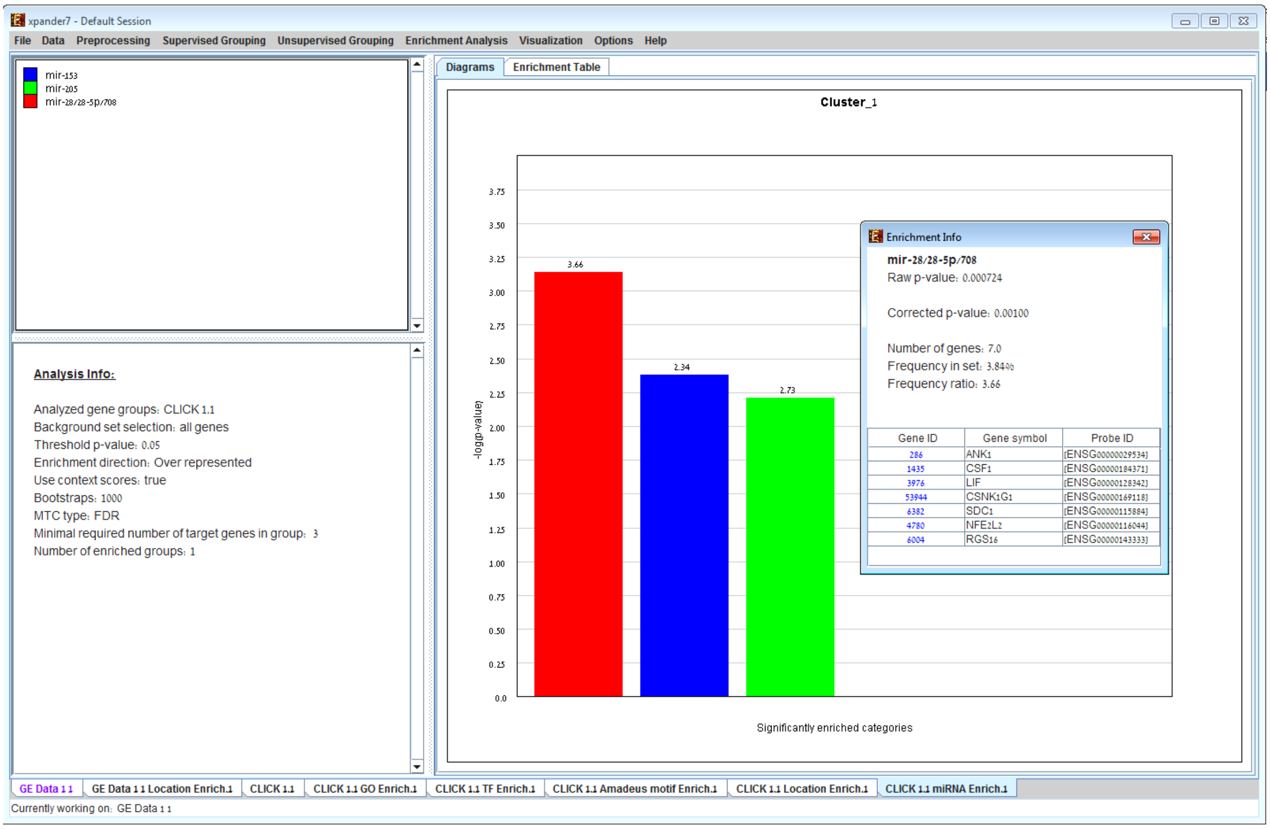
The following table specifies the different parameters that can be set via this dialog box:

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	Allows to choose between searching for over-represented targets and searching for under-

Direction	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 is added to the main window. It contains general information regarding the analysis, a sort-able table holding all detected enrichments (group name, enriched miRNA target, p-value, etc.) and a multi-histogram panel along with a color index (mapping each color to a corresponding miRNA). The multi-histogram panel contains one histogram for each probe/gene group in which enrichment has been detected. Each histogram contains a column for each significant (more frequent than would be expected by random) miRNA target. The definition of significant depends on the user's selection of threshold p-value i.e., an mRNA target is considered significantly enriched in a group of genes if its corrected p-value is lower than the selected threshold p-value.

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



Analysis Info:

Analyzed gene groups: CLICK 1.1
Background set selection: all genes
Threshold p-value: 0.05
Enrichment direction: Over represented
Use context scores: true
Bootstraps: 1000
MTC type: FDR
Minimal required number of target genes in group: 3
Number of enriched groups: 1

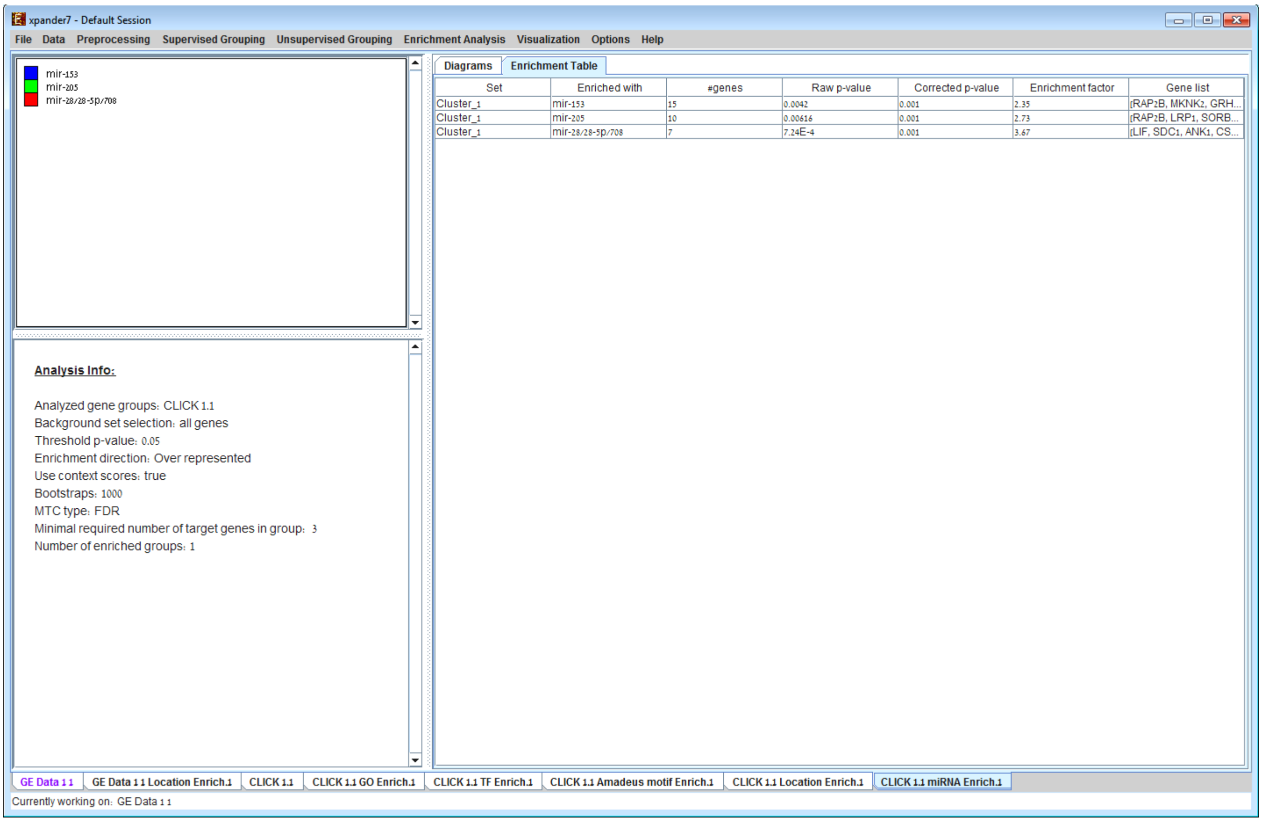
mir-28/28-5p/708
Raw p-value: 0.000724

Corrected p-value: 0.00100

Number of genes: 7.0
Frequency in set: 3.84%
Frequency ratio: 3.66

Gene ID	Gene symbol	Probe ID
286	ANK1	[ENSG00000029534]
1435	CSF1	[ENSG00000184371]
3976	LIF	[ENSG00000128342]
53944	CSNK1G1	[ENSG00000169118]
6382	SDC1	[ENSG00000115884]
4780	NFE2L2	[ENSG00000116044]
6004	RGS16	[ENSG00000143333]

Currently working on: GE Data 1 1



xpander7 - Default Session
File Data Preprocessing Supervised Grouping Unsupervised Grouping Enrichment Analysis Visualization Options Help
🟦 mir-153
🟩 mir-205
🟥 mir-28/28-5p/708
Diagrams
Enrichment Table
GE Data 1 1
GE Data 1 1 Location Enrich.1
CLICK 1.1
CLICK 1.1 GO Enrich.1
CLICK 1.1 TF Enrich.1
CLICK 1.1 Amadeus motif Enrich.1
CLICK 1.1 Location Enrich.1
CLICK 1.1 miRNA Enrich.1

Set	Enriched with	#genes	Raw p-value	Corrected p-value	Enrichment factor	Gene list
Cluster_1	mir-153	15	0.0042	0.001	2.35	[RAP2B, MKNK2, GRH...
Cluster_1	mir-205	10	0.00616	0.001	2.73	[RAP2B, LRP1, SORB...
Cluster_1	mir-28/28-5p/708	7	7.24E-4	0.001	3.67	[LIF, SDC1, ANK1, CS...

Analysis Info:

Analyzed gene groups: CLICK 1.1
Background set selection: all genes
Threshold p-value: 0.05
Enrichment direction: Over represented
Use context scores: true
Bootstraps: 1000
MTC type: FDR
Minimal required number of target genes in group: 3
Number of enriched groups: 1

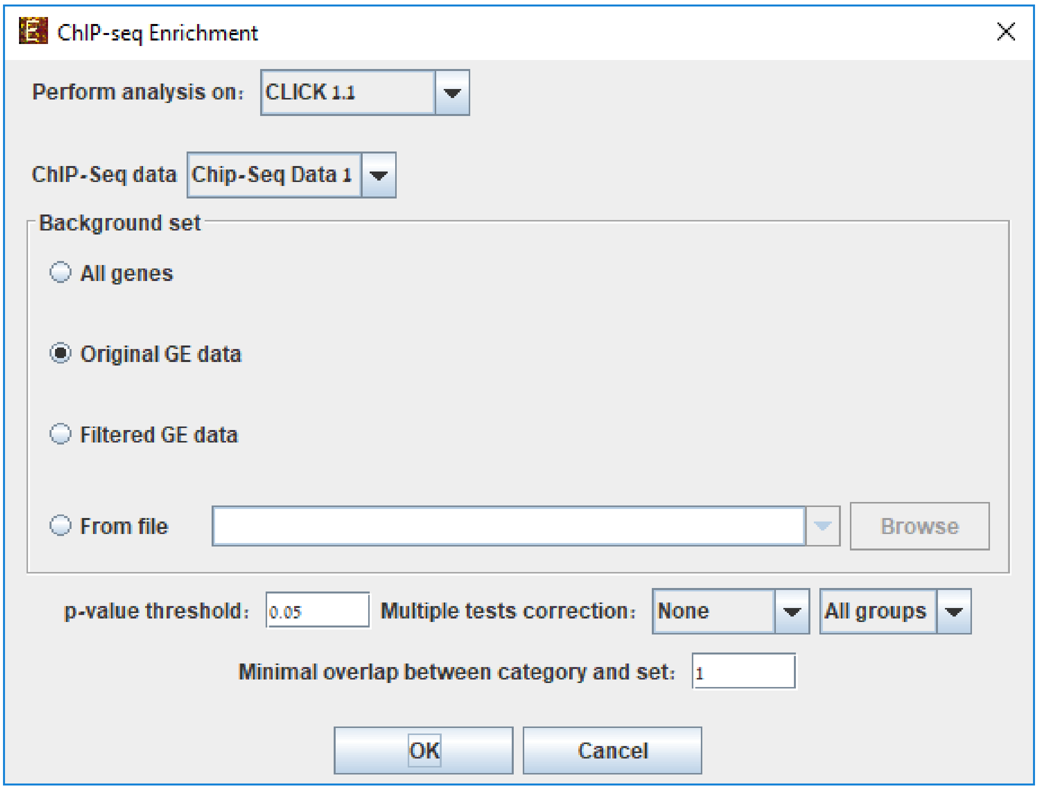
Currently working on: GE Data 1 1

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

ChIP-Seq Enrichment Analysis

This tool performs a statistical analysis to test for significant representation of the genes closest to ChIP-Seq data peaks within each group.

To perform the analysis, select *Enrichment Analysis >> ChIP-Seq Enrichment*. The following dialog box will appear:



ChIP-seq Enrichment

Perform analysis on: CLICK 1.1

ChIP-Seq data Chip-Seq Data 1

Background set

○ All genes

● Original GE data

○ Filtered GE data

○ From file

p-value threshold: 0.05 Multiple tests correction: None All groups

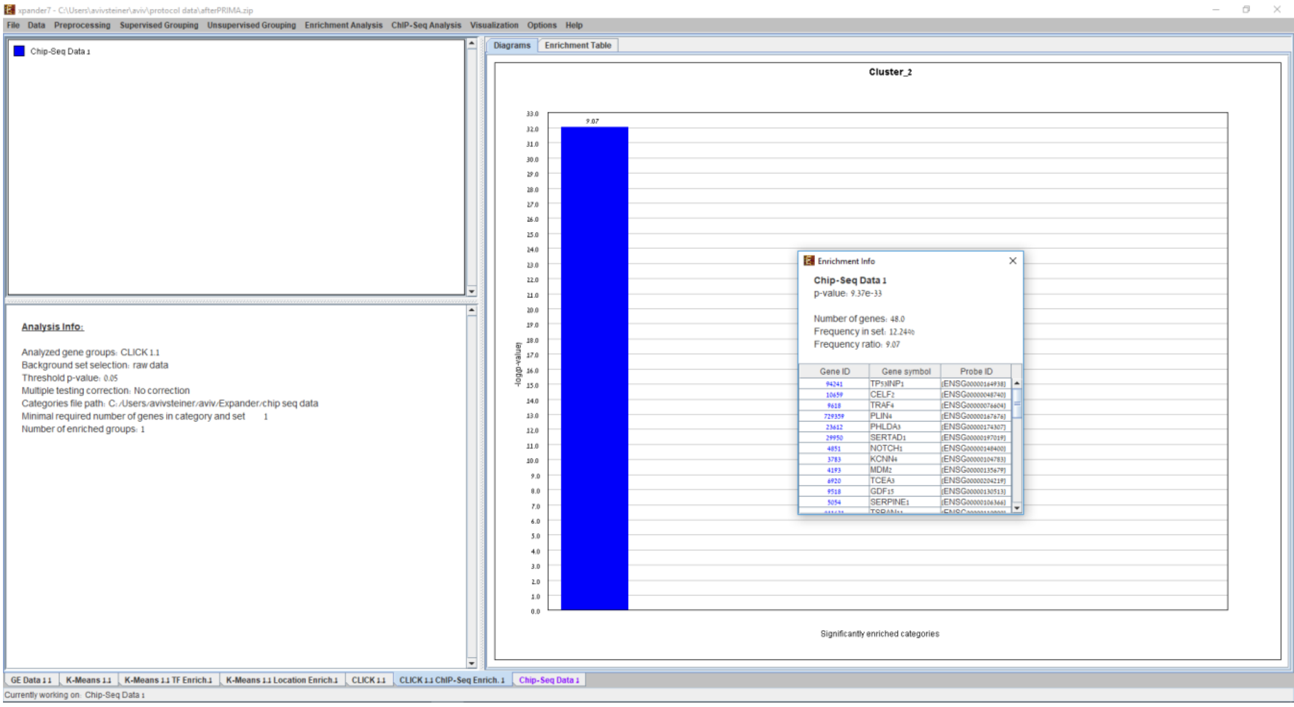
Minimal overlap between category and set: 1

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

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 protein coding 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	ChIP-Seq representation 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 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 category and set	The minimal number of genes from a cluster/bi-cluster expected to be part of the set of ChIP-Seq closest genes to peaks

After the analysis is performed, an enrichment solution visualization tab is added to the main window. It contains general information about the analysis, a sorted table holding all detected enrichments (group name, name of enriched ChIP-Seq data, p-value, etc.) and a multi-histogram panel which contains one histogram for each probe/gene group in which enrichment has been detected. The definition of significant depends on the user's selection of threshold p-value i.e., a ChIP-Seq data-set is considered significantly enriched in a group of genes if its corrected p-value is lower than the selected threshold p-value.
The height of the column is proportional to the significance of this enrichment (i.e. height = -log (raw p-value)), and the frequency ratio (frequency in set divided by frequency in background) is written on top of the column. Upon clicking on a column, a dialog box is displayed containing the enrichment information, corrected p-value and a list of the genes in the group that are included in the corresponding ChIP-Seq data. Upon clicking on one of the gene Ids in the table, a relevant web page with information regarding this gene is displayed.



Analysis Info:

Analyzed gene groups: CLICK 1.1
Background set selection: raw data
Threshold p-value: 0.05
Multiple testing correction: No correction
Categories file path: C:/Users/avivsteiner/aviv/Expander/chip seq data
Minimal required number of genes in category and set 1
Number of enriched groups: 1

Enrichment Info

Chip-Seq Data 1
p-value: 9.37e-33

Number of genes: 48.0
Frequency in set: 12.24%
Frequency ratio: 9.07

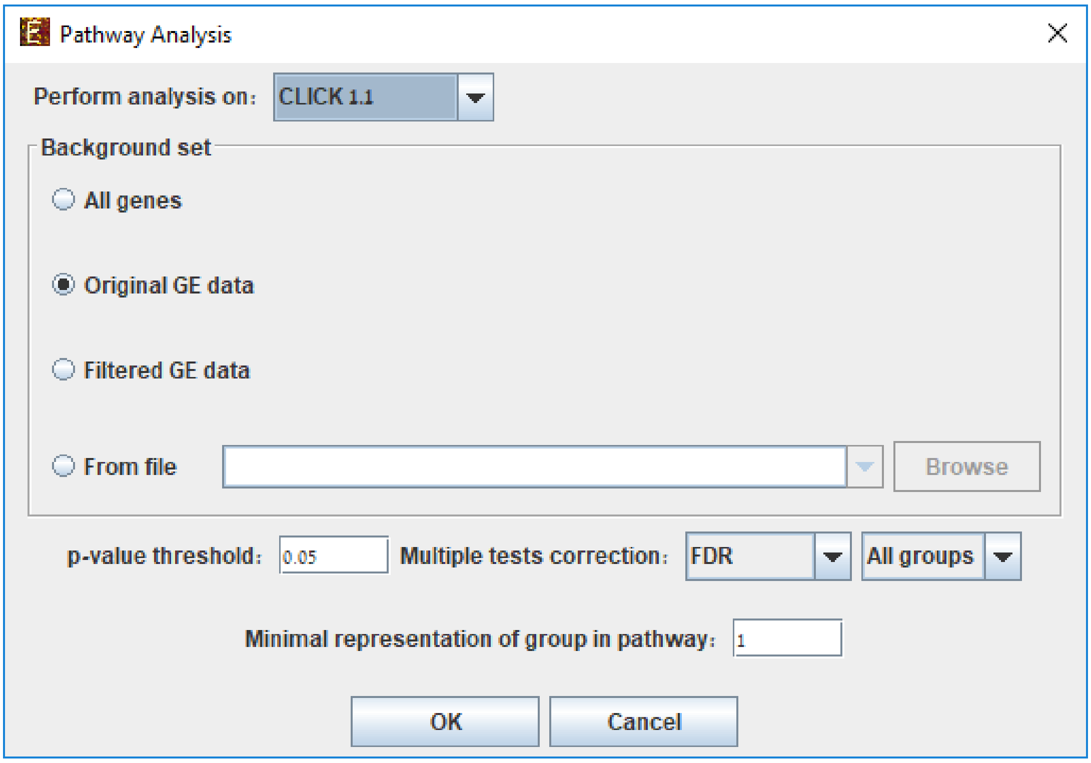
Gene ID	Gene symbol	Probe ID
94241	TP53INP1	[ENSG0000164938]
10659	CELF2	[ENSG0000048740]
9618	TRAF4	[ENSG0000076604]
729359	PLIN4	[ENSG0000167676]
23612	PHLDA3	[ENSG0000174307]
29950	SERTAD1	[ENSG0000197019]
4851	NOTCH1	[ENSG0000148400]
3783	KCNN4	[ENSG0000104783]
4193	MDM2	[ENSG0000135679]
6920	TCEA3	[ENSG0000204219]
9518	GDF15	[ENSG0000130513]
5054	SERPINE1	[ENSG0000106366]

Currently working on: Chip-Seq Data 1

Pathway Enrichment Analysis

This tool performs a statistical analysis on the representation of KEGG and WikiPathways pathway maps within each group. The KEGG and WikiPathways information is supplied in organism-specific data files, which can be downloaded from the EXPANDER download-page (see the Supplied Files section In this analysis, hyper-geometric enrichment tests are performed, and the results can be (if requested) corrected for multiple testing using the FDR/Bonferroni correction.

To perform the analysis, select *Enrichment Analysis >> Pathway Analysis >> KEGG or Enrichment Analysis >> Pathway Analysis >> WikiPathways* . The following dialog box will appear:



Pathway Analysis

Perform analysis on: CLICK 1.1

Background set

○ All genes

● Original GE data

○ Filtered GE data

○ From file

p-value threshold: 0.05

Multiple tests correction: FDR

All groups

Minimal representation of group in pathway: 1

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

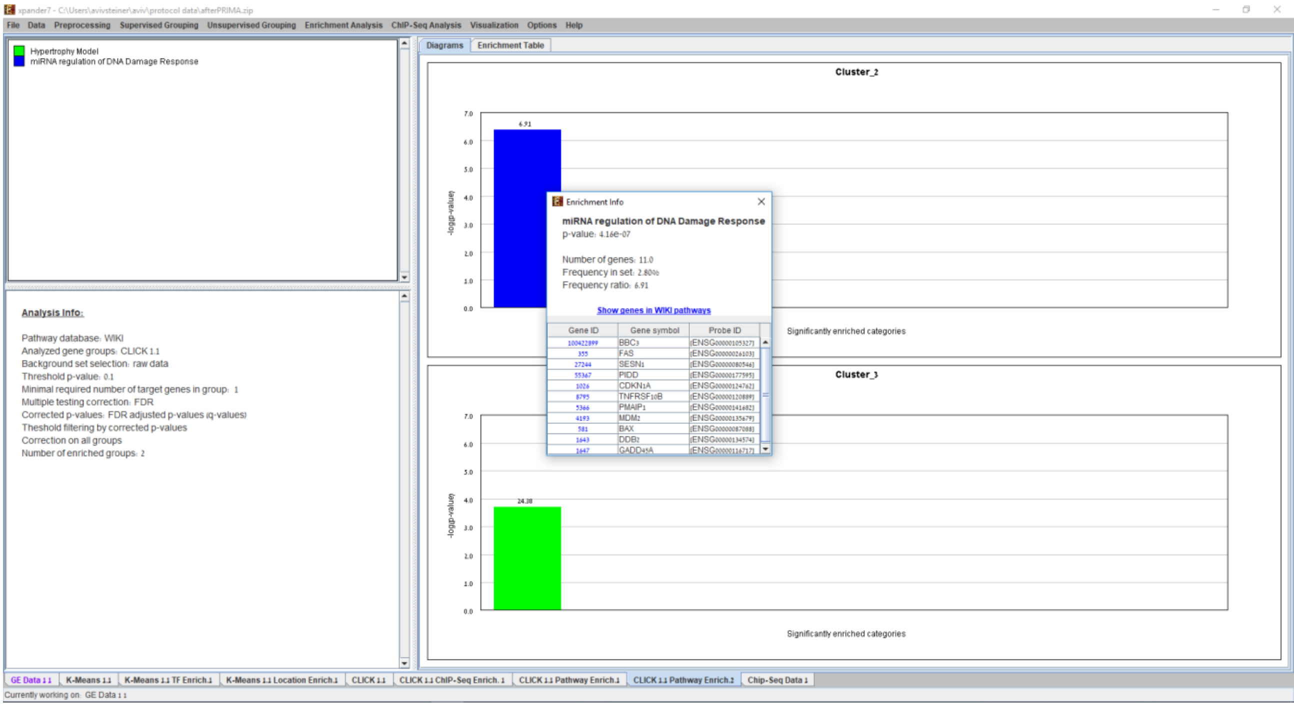
Field	Description
Perform analysis on	The grouping solution on which the analysis will be performed.
Background set	Determines the set of genes that will be used as background in the analysis. Options are: all genes (of the relevant organism), original input

	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 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 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 added to the main window. It contains general information about the analysis, a sorted table holding all detected pathways (group name, enriched pathway target, p-value, etc.) and multi-histogram panel along with a color index (mapping each color to a corresponding pathway). The multi-histogram panel contains one histogram for each probe/gene group in which enrichment has been detected. Each histogram contains a column for each significant (more frequent than would be expected by random) pathway target. The definition of significant depends on the user's selection of threshold p-value i.e., a pathway target is considered significantly enriched in a group of genes if its corrected p-value is lower than the selected threshold p-value.

The height of the column is proportional to the significance of this enrichment (i.e. height = -log(raw p-value)), and the frequency ratio (frequency in set divided by frequency in background) is written on top of the column. Upon clicking on a column, a dialog box is displayed containing the pathway name, corrected p-value, link to the relevant pathway map web page, and a list of the genes in the group that are included in the corresponding pathway.

Upon clicking on one of the gene Ids in the table, a relevant web page with information regarding this gene is displayed.


xpander7 - C:\Users\avivsteiner\aviv\protocol data\afterPRIMA.zip
File Data Preprocessing Supervised Grouping Unsupervised Grouping Enrichment Analysis ChIP-Seq Analysis Visualization Options Help
Diagrams
Enrichment Table
Cluster_2
Cluster_3
-log(p-value)
Significantly enriched categories
6.91
24.38
GE Data 1.1
K-Means 1.1
K-Means 1.1 TF Enrich.1
K-Means 1.1 Location Enrich.1
CLICK 1.1
CLICK 1.1 ChIP-Seq Enrich. 1
CLICK 1.1 Pathway Enrich.1
CLICK 1.1 Pathway Enrich.2
Chip-Seq Data 1
Currently working on: GE Data 1.1

🟩 Hypertrophy Model
🟦 miRNA regulation of DNA Damage Response

Analysis Info:

Pathway database: WIKI
Analyzed gene groups: CLICK 1.1
Background set selection: raw data
Threshold p-value: 0.1
Minimal required number of target genes in group: 1
Multiple testing correction: FDR
Corrected p-values: FDR adjusted p-values (q-values)
Theshold filtering by corrected p-values
Correction on all groups
Number of enriched groups: 2

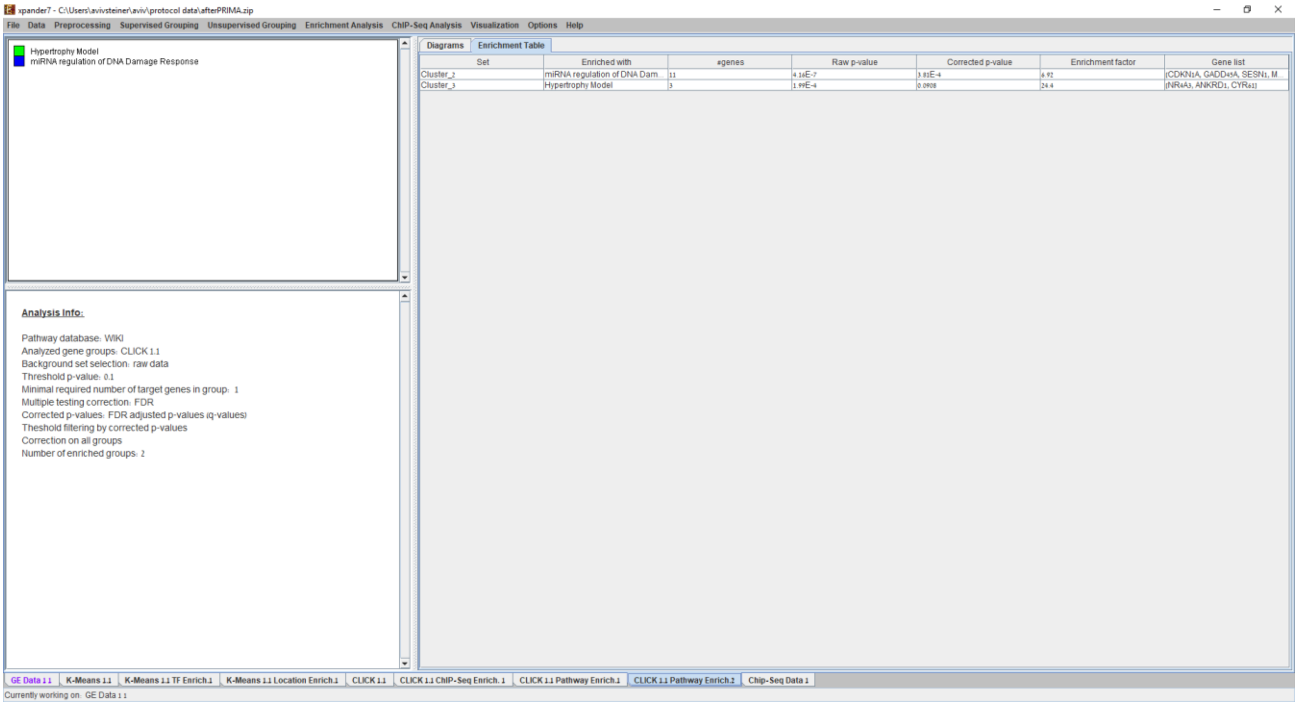
Enrichment Info

miRNA regulation of DNA Damage Response
p-value: 4.16e-07

Number of genes: 11.0
Frequency in set: 2.80%
Frequency ratio: 6.91

Show genes in WIKI pathways

Gene ID	Gene symbol	Probe ID
100422899	BBC3	[ENSG00000105327]
355	FAS	[ENSG00000026103]
27244	SESN1	[ENSG00000080546]
55367	PIDD	[ENSG00000177595]
1026	CDKN1A	[ENSG00000124762]
8795	TNFRSF10B	[ENSG00000120889]
5366	PMAIP1	[ENSG00000141682]
4193	MDM2	[ENSG00000135679]
581	BAX	[ENSG00000087088]
1643	DDB2	[ENSG00000134574]
1647	GADD45A	[ENSG00000116717]


xpander7 - C:\Users\avivsteiner\aviv\protocol data\afterPRIMA.zip
File Data Preprocessing Supervised Grouping Unsupervised Grouping Enrichment Analysis ChIP-Seq Analysis Visualization Options Help
Diagrams
Enrichment Table
GE Data 1.1
K-Means 1.1
K-Means 1.1 TF Enrich.1
K-Means 1.1 Location Enrich.1
CLICK 1.1
CLICK 1.1 ChIP-Seq Enrich. 1
CLICK 1.1 Pathway Enrich.1
CLICK 1.1 Pathway Enrich.2
Chip-Seq Data 1
Currently working on: GE Data 1.1

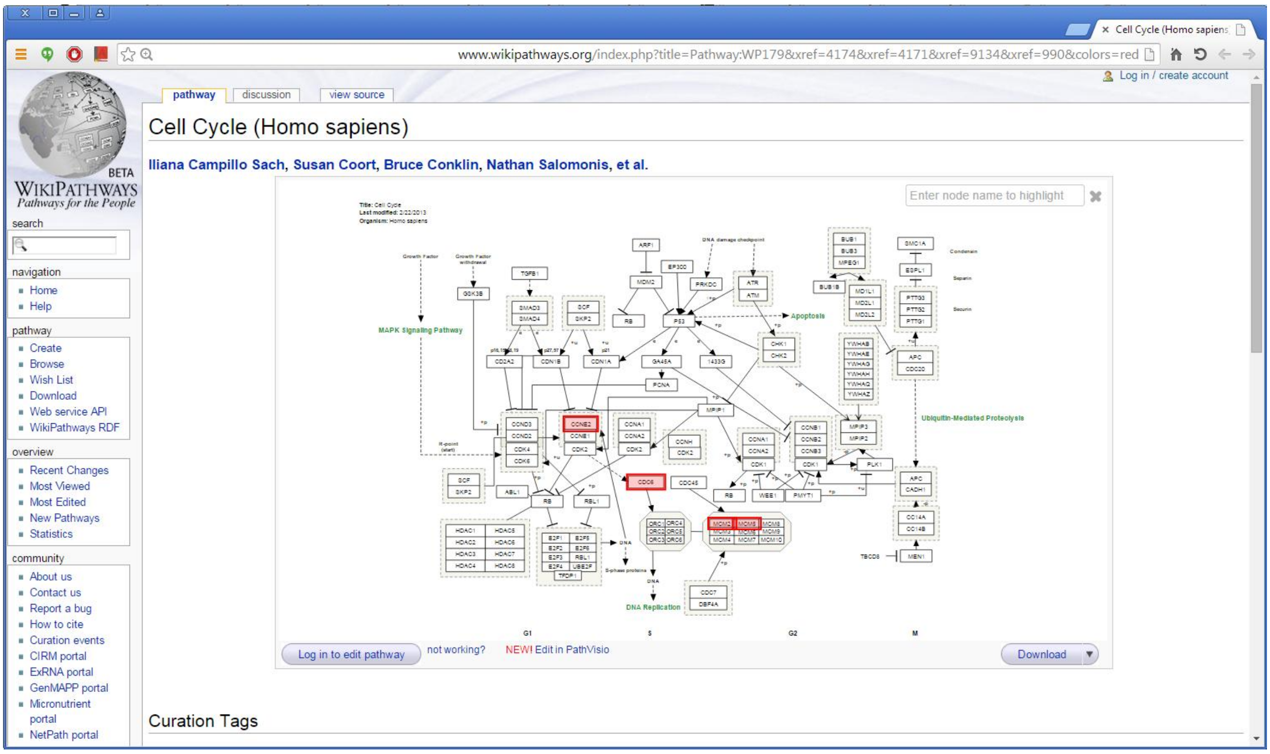
🟩 Hypertrophy Model
🟦 miRNA regulation of DNA Damage Response

Analysis Info:

Pathway database: WIKI
Analyzed gene groups: CLICK 1.1
Background set selection: raw data
Threshold p-value: 0.1
Minimal required number of target genes in group: 1
Multiple testing correction: FDR
Corrected p-values: FDR adjusted p-values (q-values)
Theshold filtering by corrected p-values
Correction on all groups
Number of enriched groups: 2

Set	Enriched with	#genes	Raw p-value	Corrected p-value	Enrichment factor	Gene list
Cluster_2	miRNA regulation of DNA Dam...	11	4.16E-7	3.81E-4	6.92	[CDKN1A, GADD45A, SESN1, M...
Cluster_3	Hypertrophy Model	3	1.99E-4	0.0908	24.4	[NR4A3, ANKRD1, CYR61]

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

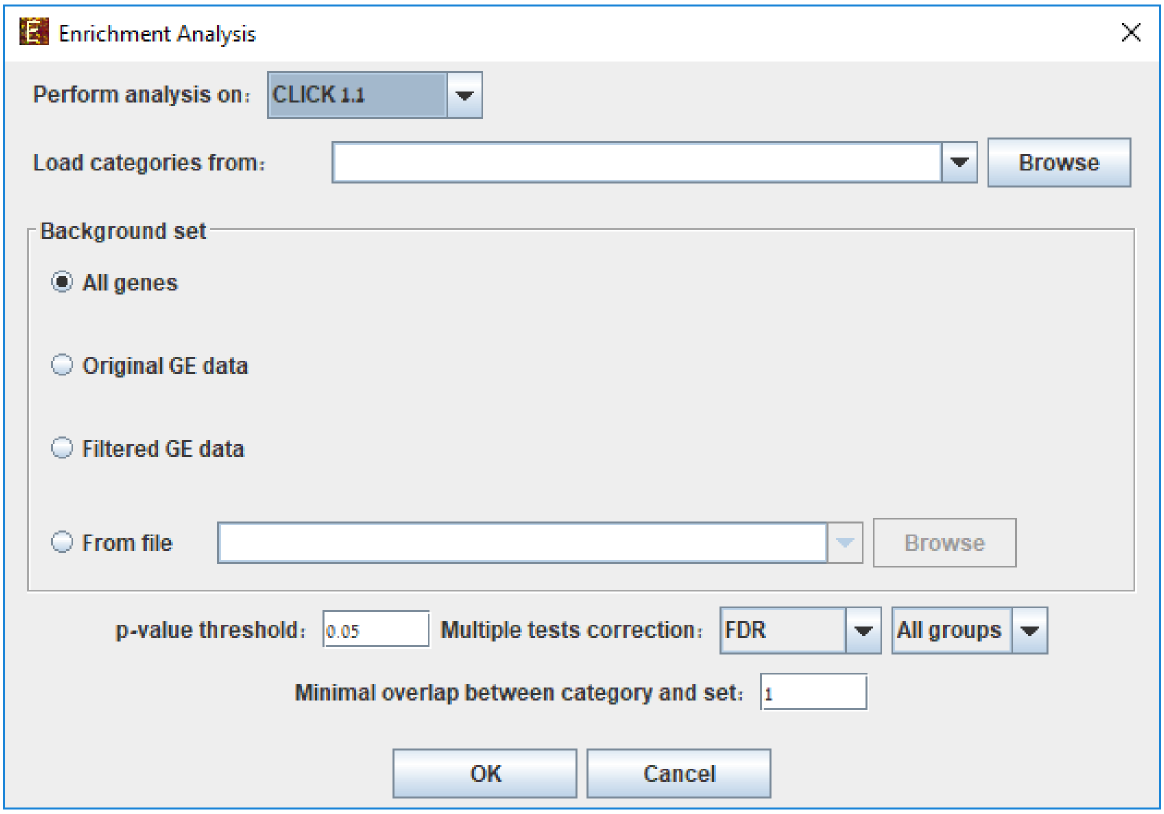


The results of this analysis can be exported to a text file by selecting *File>>Export to text* when the corresponding view is the selected tab. OR by using the *File>>Save All* option, which will export all solutions within a session to text and image files.

Custom Enrichment Analysis

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

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



Perform analysis on: CLICK 1.1

Load categories from:

Background set

● All genes

○ Original GE data

○ Filtered GE data

○ From file

p-value threshold: 0.05 Multiple tests correction: FDR All groups

Minimal overlap between category and set: 1

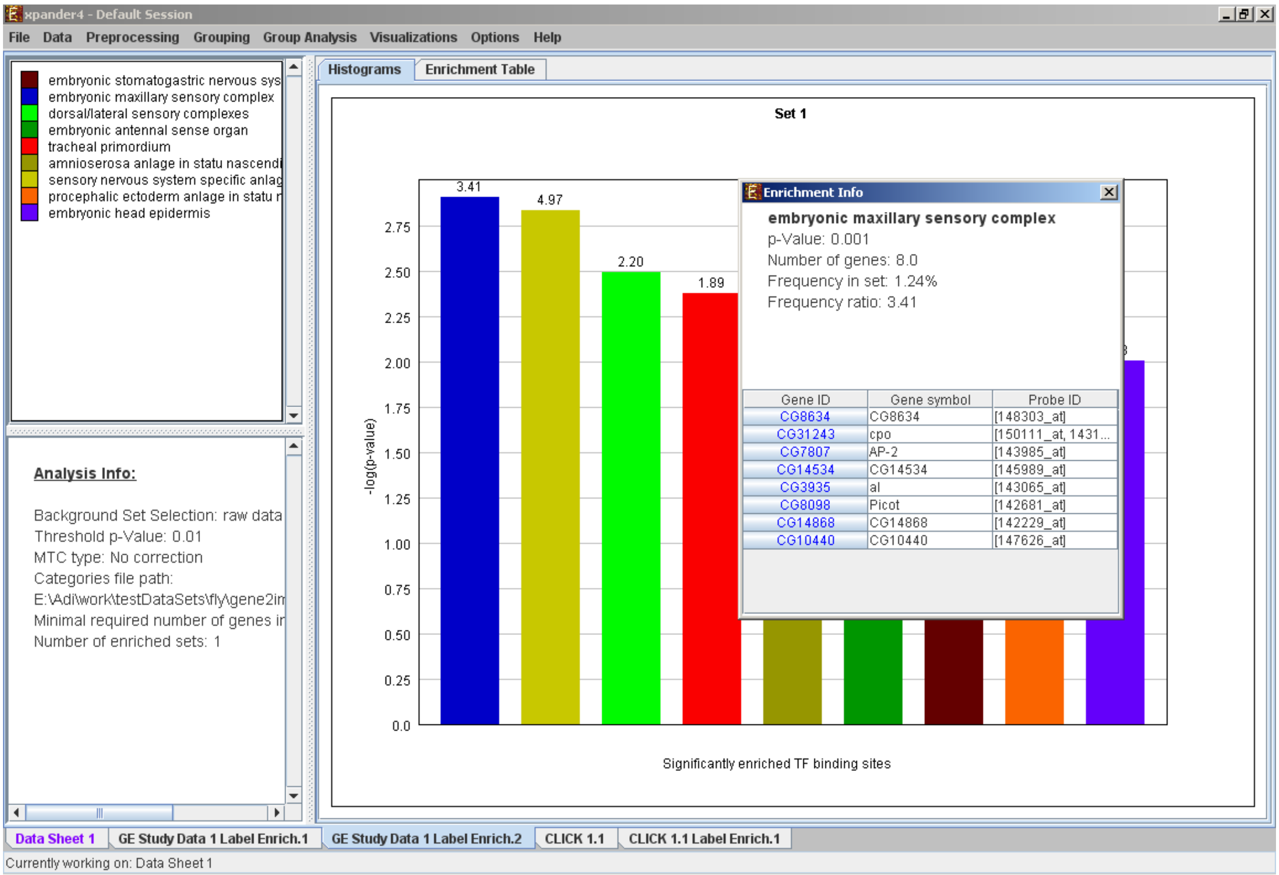
The following table specifies the different parameters that can be set via this dialog box:

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 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 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 enrichment to be accepted.

After the analysis is performed an enrichment analysis solution visualization tab is added to the main window. It contains general information regarding the analysis, a sort-able table holding all detected enrichments (set ID, enrichment category, p-value, etc.) and a multi-histogram panel along with a color index (mapping each color to a corresponding category). The multi-histogram panel contains one histogram for each probe/gene set/group in which enrichment has been detected. Each histogram contains a column for each significant (more frequent than would be expected by random) category. The definition of significant depends on the user's selection of threshold p-value i.e., a category is considered significantly enriched in a 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 (i.e. height = -log(raw p-value)), and the frequency ratio (frequency in set divided by frequency in background) is written on top of the column. Upon clicking on a column, a dialog box is displayed containing the class name, corrected p-value, and a list of the genes in the cluster/bi-cluster that belong to the category. Upon clicking on one of the gene Ids in the table, a relevant web page with information regarding this gene is displayed. The display tool tip shows the cluster number, size and homogeneity.



xpander4 - Default Session
File
Data
Preprocessing
Grouping
Group Analysis
Visualizations
Options
Help
🟫 embryonic stomatogastric nervous sys
🟦 embryonic maxillary sensory complex
🟩 dorsal/lateral sensory complexes
🟩 embryonic antennal sense organ
🟥 tracheal primordium
🟨 amnioserosa anlage in statu nascendi
🟨 sensory nervous system specific anlag
🟧 procephalic ectoderm anlage in statu n
🟪 embryonic head epidermis
Histograms
Enrichment Table
Set 1
3.41
4.97
2.20
1.89
3
2.75
2.50
2.25
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.0
-log(p-value)
Significantly enriched TF binding sites
Enrichment Info
Data Sheet 1
GE Study Data 1 Label Enrich.1
GE Study Data 1 Label Enrich.2
CLICK 1.1
CLICK 1.1 Label Enrich.1

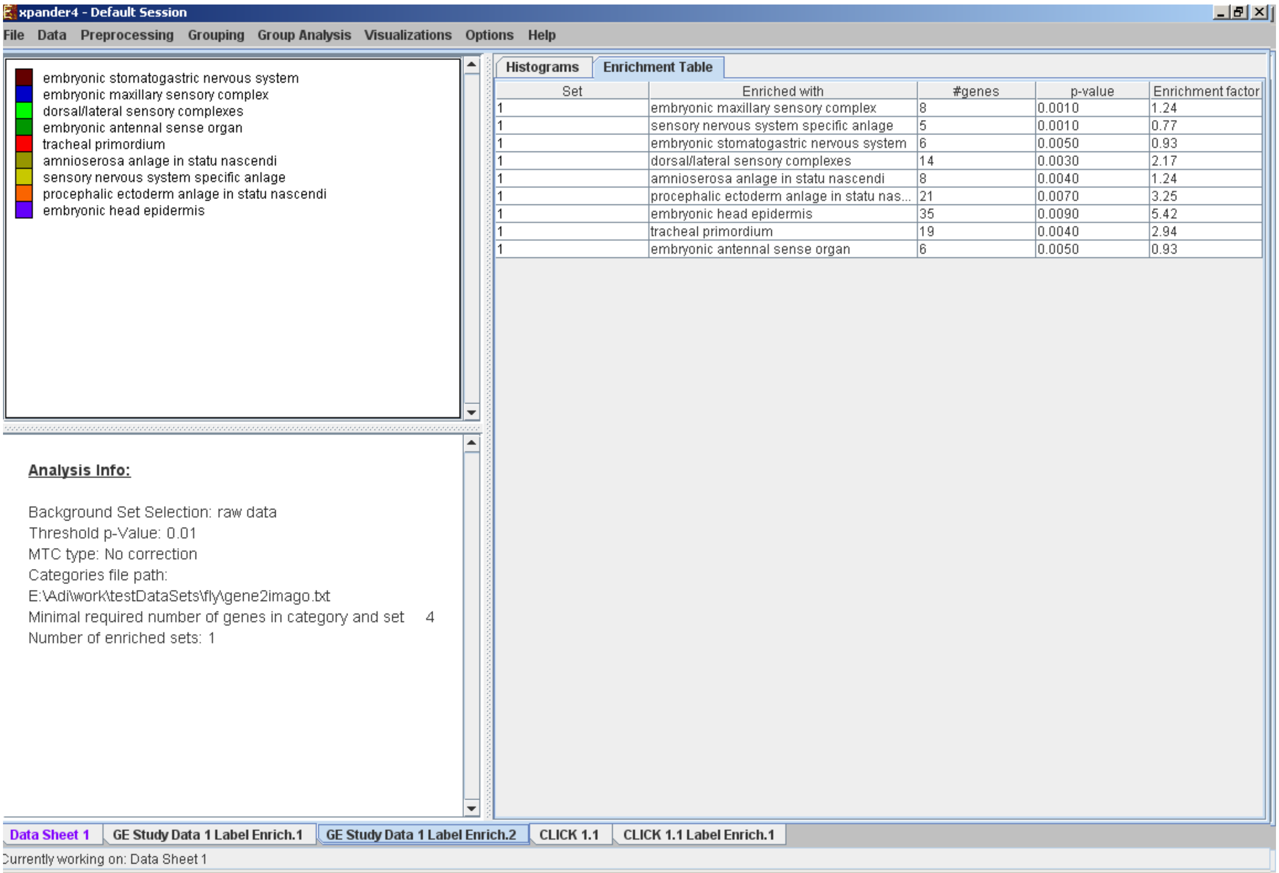
Analysis Info:

Background Set Selection: raw data
Threshold p-Value: 0.01
MTC type: No correction
Categories file path:
E:\Adi\work\testDataSets\fly\gene2in
Minimal required number of genes in
Number of enriched sets: 1

embryonic maxillary sensory complex
p-Value: 0.001
Number of genes: 8.0
Frequency in set: 1.24%
Frequency ratio: 3.41

Gene ID	Gene symbol	Probe ID
CG8634	CG8634	[148303_at]
CG31243	cpo	[150111_at, 1431...
CG7807	AP-2	[143985_at]
CG14534	CG14534	[145989_at]
CG3935	al	[143065_at]
CG8098	Picot	[142681_at]
CG14868	CG14868	[142229_at]
CG10440	CG10440	[147626_at]

Currently working on: Data Sheet 1



🟥	embryonic stomatogastric nervous system
🟦	embryonic maxillary sensory complex
🟩	dorsal/lateral sensory complexes
🟩	embryonic antennal sense organ
🟥	tracheal primordium
🟨	amnioserosa anlage in statu nascendi
🟨	sensory nervous system specific anlage
🟧	procephalic ectoderm anlage in statu nascendi
🟪	embryonic head epidermis

Set	Enriched with	#genes	p-value	Enrichment factor
1	embryonic maxillary sensory complex	8	0.0010	1.24
1	sensory nervous system specific anlage	5	0.0010	0.77
1	embryonic stomatogastric nervous system	6	0.0050	0.93
1	dorsal/lateral sensory complexes	14	0.0030	2.17
1	amnioserosa anlage in statu nascendi	8	0.0040	1.24
1	procephalic ectoderm anlage in statu nas...	21	0.0070	3.25
1	embryonic head epidermis	35	0.0090	5.42
1	tracheal primordium	19	0.0040	2.94
1	embryonic antennal sense organ	6	0.0050	0.93

Analysis Info:

Background Set Selection: raw data
Threshold p-Value: 0.01
MTC type: No correction
Categories file path:
E:\Adi\work\testDataSets\fly\gene2imago.txt
Minimal required number of genes in category and set 4
Number of enriched sets: 1

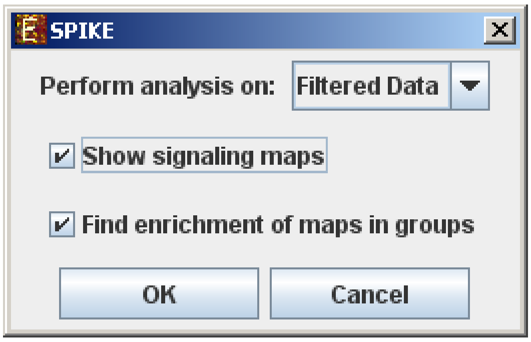
Currently working on: Data Sheet 1

The results of this analysis can be exported to a text file by selecting *File>>Export to text* when the corresponding view is the selected tab. OR by using the *File>>Save All* option, which will 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 that are induced by the analyzed gene groups. It also enables the user to search for statistical enrichment of these groups in highly curated signaling maps. To perform this task, Expander interfaces with the SPIKE software and database. For further information regarding the SPIKE software see the References section.

To perform the analysis on one/more of the gene groups defined in Expander (i.e. clusters, bi-clusters, modules, loaded gene sets or filtered data), select *Enrichment Analysis >> Network >> SPIKE>>Gene Groups*. The following dialog box will appear:



Perform analysis on: Filtered Data

☑ Show signaling maps

☑ Find enrichment of maps in groups

The following table specifies the different parameters that can be set via this dialog 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 operated for the first time, the launch takes a few minutes, since it has to build a local database. From 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 an existing enrichment solution in Expander. I.e. a group of genes that has a common annotation that was found to be enriched by one of the enrichment analysis operations. In order to operate SPIKE on such a *group, select: Enrichment Analysis>>Network>> SPIKE>>Enrichment Derived Sets*.

Gene Set Enrichment Analysis (GSEA)

GSEA (Subramanian et al 2005) considers experiments with genome-wide expression profiles from samples belonging to two classes. Genes are ranked based on the correlation between their expression and the differential expression between classes distinction or pre-ranked by the user.

Given an *a priori* defined set of genes *S*, the goal of GSEA is to determine whether the members of *S* are randomly distributed throughout the ranked list of genes (L) or primarily found at the top or bottom. It is expected that sets related to the phenotypic distinction will tend 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 the degree to which a set *S* is overrepresented at the extremes (top or bottom) of the entire ranked list *L*. The score is calculated by walking down the list *L*, increasing a running-sum statistic when we encounter a gene in *S* and decreasing it when we encounter genes not in *S.* The magnitude of the increment depends on the correlation of the gene with the phenotype. The enrichment score is the maximum deviation from zero encountered in the random walk. It corresponds to a weighted Kolmogorov–Smirnov-like statistic.

Step 2: Estimation of Significance Level of ES. An estimation of the statistical significance (nominal *P*-value) of the *ES* is done by using an empirical phenotype-based permutation test procedure that preserves the complex correlation structure of the gene expression data. Specifically, the phenotype labels are permuted again and the *ES* of the gene set for the permuted data is re-computed, which generates a null distribution for the *ES.* If the user provided a pre-ranked list of genes then a random shuffling of the ranked list is done instead. The empirical, nominal *P* value of the observed *ES* is then calculated relative to this null distribution. Importantly, the permutation of class labels preserves gene-gene correlations and, thus, provides a more biologically reasonable assessment of significance than would be obtained by permuting genes.

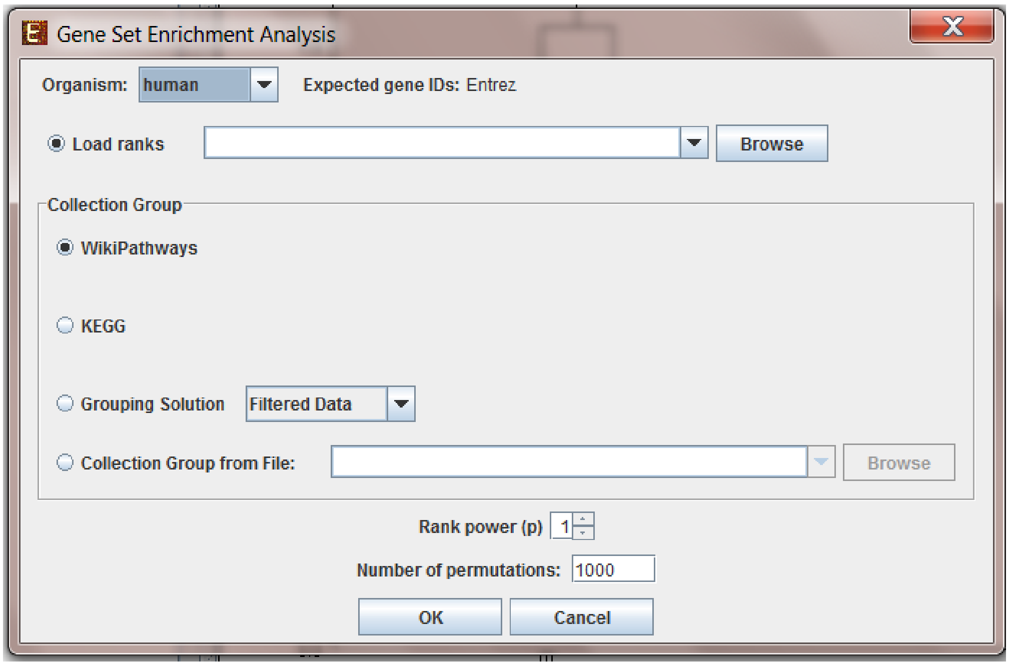
There are 2 ways to perform GSEA:

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->Gene Ranking Analysis (GSEA).*

The following dialog box will appear:


Gene Set Enrichment Analysis

Organism: human Expected gene IDs: Entrez

● Load ranks

Collection Group

● WikiPathways

○ KEGG

○ Grouping Solution: Filtered Data

○ Collection Group from File:

Rank power (p): 1

Number of permutations: 1000

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

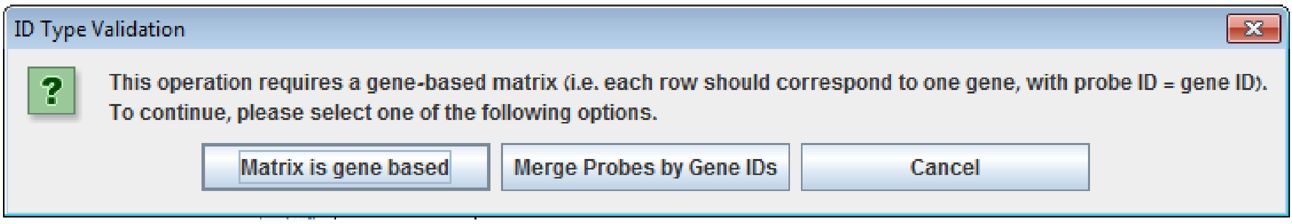
Field	Description
Load ranks	User pre-ranked list of genes file composed of two columns – first with genes and second with values
Collection Group	Can be chosen between: WikiPathways, KEGG, a gene-groups solution which was generated in the current session or an external file with gene sets*
Rank Power (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 the results.

GSEA on a gene expression data

To perform the analysis on the gene expression, select *Enrichment Analysis* >> run GSEA…

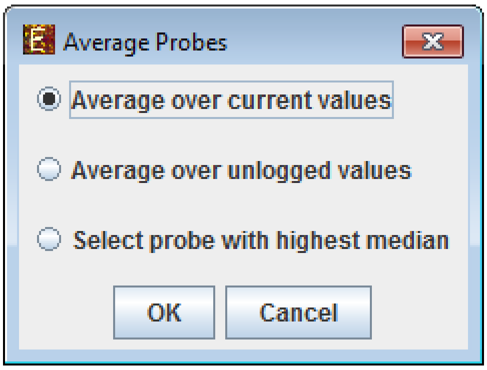
The following dialog box will appear:


ID Type Validation

This operation requires a gene-based matrix (i.e. each row should correspond to one gene, with probe ID = gene ID). To continue, please select one of the following options.

The user can choose between "Matrix is gene based" (i.e each row should correspond to one 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 "Average Probes" will appear:

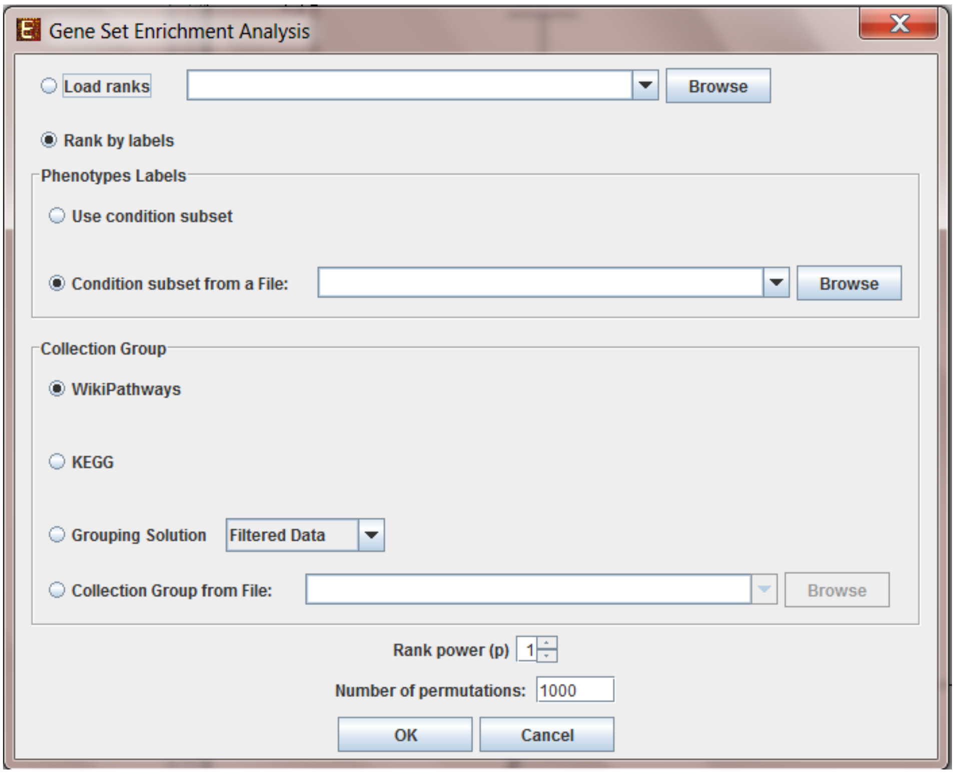

Average Probes

● Average over current values

○ Average over unlogged values

○ Select probe with highest median

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



○ Load ranks

● Rank by labels

Phenotypes Labels

○ Use condition subset

● Condition subset from a File:

Collection Group

● WikiPathways

○ KEGG

○ Grouping Solution Filtered Data

○ Collection Group from File:

Rank power (p) 1

Number of permutations: 1000

The following table specifies the different parameters that can be set via this dialog 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, a gene-groups solution which was generated in the current session or an external 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 be loaded via "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

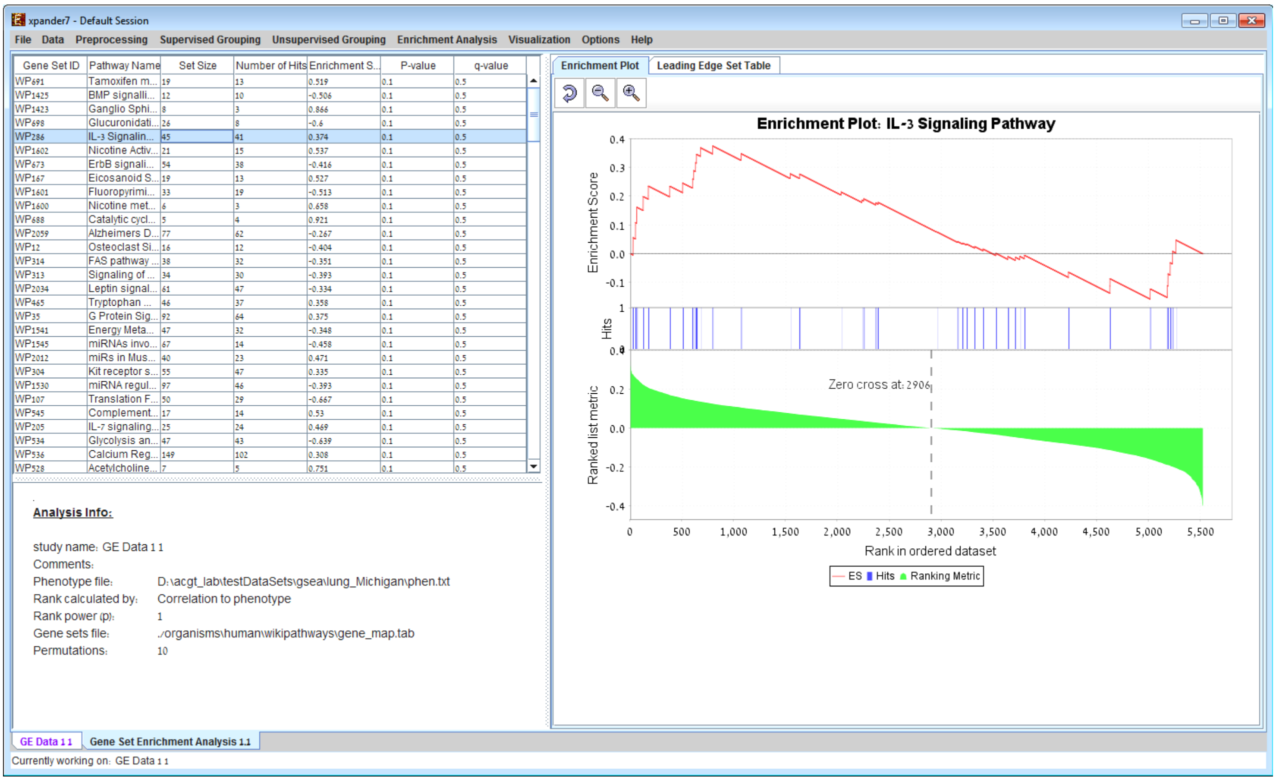
C6.gmt – Oncogenic signatures

C7.gmt – Immunologic signatures

For further information, please refer to: http://www.broadinstitute.org/gsea/msigdb/index.jsp

Results

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

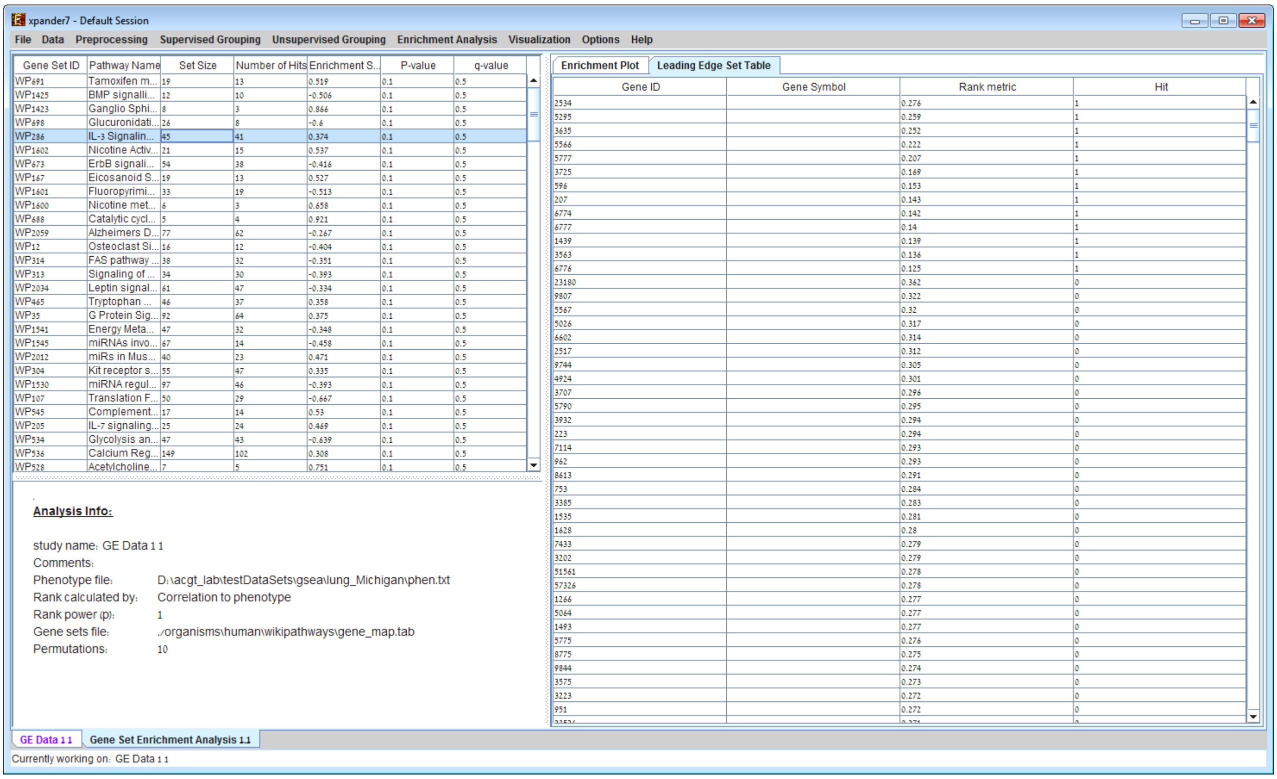


Gene Set ID	Pathway Name	Set Size	Number of Hits	Enrichment S...	P-value	q-value
WP691	Tamoxifen m...	19	13	0.519	0.1	0.5
WP1425	BMP signalli...	12	10	-0.506	0.1	0.5
WP1423	Ganglio Sphi...	8	3	0.866	0.1	0.5
WP698	Glucuronidati...	26	8	-0.6	0.1	0.5
WP286	IL-3 Signalin...	45	41	0.374	0.1	0.5
WP1602	Nicotine Activ...	21	15	0.537	0.1	0.5
WP673	ErbB signali...	54	38	-0.416	0.1	0.5
WP167	Eicosanoid S...	19	13	0.527	0.1	0.5
WP1601	Fluoropyrimi...	33	19	-0.513	0.1	0.5
WP1600	Nicotine met...	6	3	0.658	0.1	0.5
WP688	Catalytic cycl...	5	4	0.921	0.1	0.5
WP2059	Alzheimers D...	77	62	-0.267	0.1	0.5
WP12	Osteoclast Si...	16	12	-0.404	0.1	0.5
WP314	FAS pathway ...	38	32	-0.351	0.1	0.5
WP313	Signaling of ...	34	30	-0.393	0.1	0.5
WP2034	Leptin signal...	61	47	-0.334	0.1	0.5
WP465	Tryptophan ...	46	37	0.358	0.1	0.5
WP35	G Protein Sig...	92	64	0.375	0.1	0.5
WP1541	Energy Meta...	47	32	-0.348	0.1	0.5
WP1545	miRNAs invo...	67	14	-0.458	0.1	0.5
WP2012	miRs in Mus...	40	23	0.471	0.1	0.5
WP304	Kit receptor s...	55	47	0.335	0.1	0.5
WP1530	miRNA regul...	97	46	-0.393	0.1	0.5
WP107	Translation F...	50	29	-0.667	0.1	0.5
WP545	Complement...	17	14	0.53	0.1	0.5
WP205	IL-7 signaling...	25	24	0.469	0.1	0.5
WP534	Glycolysis an...	47	43	-0.639	0.1	0.5
WP536	Calcium Reg...	149	102	0.308	0.1	0.5
WP528	Acetylcholine...	7	5	0.751	0.1	0.5

Analysis Info:

study name: GE Data 1 1
Comments:
Phenotype file: D:\acgt_lab\testDataSets\gsea\lung_Michigan\phen.txt
Rank calculated by: Correlation to phenotype
Rank power (p): 1
Gene sets file: ./organisms\human\wikipathways\gene_map.tab
Permutations: 10

Currently working on: GE Data 1 1


xpander7 - Default Session
File Data Preprocessing Supervised Grouping Unsupervised Grouping Enrichment Analysis Visualization Options Help
Enrichment Plot
Leading Edge Set Table
GE Data 1 1
Gene Set Enrichment Analysis 1.1

Gene Set ID	Pathway Name	Set Size	Number of Hits	Enrichment S...	P-value	q-value
WP691	Tamoxifen m...	19	13	0.519	0.1	0.5
WP1425	BMP signalli...	12	10	-0.506	0.1	0.5
WP1423	Ganglio Sphi...	8	3	0.866	0.1	0.5
WP698	Glucuronidati...	26	8	-0.6	0.1	0.5
WP286	IL-3 Signalin...	45	41	0.374	0.1	0.5
WP1602	Nicotine Activ...	21	15	0.537	0.1	0.5
WP673	ErbB signali...	54	38	-0.416	0.1	0.5
WP167	Eicosanoid S...	19	13	0.527	0.1	0.5
WP1601	Fluoropyrimi...	33	19	-0.513	0.1	0.5
WP1600	Nicotine met...	6	3	0.658	0.1	0.5
WP688	Catalytic cycl...	5	4	0.921	0.1	0.5
WP2059	Alzheimers D...	77	62	-0.267	0.1	0.5
WP12	Osteoclast Si...	16	12	-0.404	0.1	0.5
WP314	FAS pathway ...	38	32	-0.351	0.1	0.5
WP313	Signaling of ...	34	30	-0.393	0.1	0.5
WP2034	Leptin signal...	61	47	-0.334	0.1	0.5
WP465	Tryptophan ...	46	37	0.358	0.1	0.5
WP35	G Protein Sig...	92	64	0.375	0.1	0.5
WP1541	Energy Meta...	47	32	-0.348	0.1	0.5
WP1545	miRNAs invo...	67	14	-0.458	0.1	0.5
WP2012	miRs in Mus...	40	23	0.471	0.1	0.5
WP304	Kit receptor s...	55	47	0.335	0.1	0.5
WP1530	miRNA regul...	97	46	-0.393	0.1	0.5
WP107	Translation F...	50	29	-0.667	0.1	0.5
WP545	Complement...	17	14	0.53	0.1	0.5
WP205	IL-7 signaling...	25	24	0.469	0.1	0.5
WP534	Glycolysis an...	47	43	-0.639	0.1	0.5
WP536	Calcium Reg...	149	102	0.308	0.1	0.5
WP528	Acetylcholine...	7	5	0.751	0.1	0.5

Analysis Info:

study name: GE Data 1 1
Comments:
Phenotype file: D:\acgt_lab\testDataSets\gsea\lung_Michigan\phen.txt
Rank calculated by: Correlation to phenotype
Rank power (p): 1
Gene sets file: ./organisms\human\wikipathways\gene_map.tab
Permutations: 10

Gene ID	Gene Symbol	Rank metric	Hit
2534		0.276	1
5295		0.259	1
3635		0.252	1
5566		0.222	1
5777		0.207	1
3725		0.169	1
596		0.153	1
207		0.143	1
6774		0.142	1
6777		0.14	1
1439		0.139	1
3563		0.136	1
6776		0.125	1
23180		0.362	0
9807		0.322	0
5567		0.32	0
5026		0.317	0
6602		0.314	0
2517		0.312	0
9744		0.305	0
4924		0.301	0
3707		0.296	0
5790		0.295	0
3932		0.294	0
223		0.294	0
7114		0.293	0
962		0.293	0
8613		0.291	0
753		0.284	0
3385		0.283	0
1535		0.281	0
1628		0.28	0
7433		0.279	0
3202		0.279	0
51561		0.278	0
57326		0.278	0
1266		0.277	0
5064		0.277	0
1493		0.277	0
5775		0.276	0
8775		0.275	0
9844		0.274	0
3575		0.273	0
3223		0.272	0
951		0.272	0

Currently working on: GE Data 1 1

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

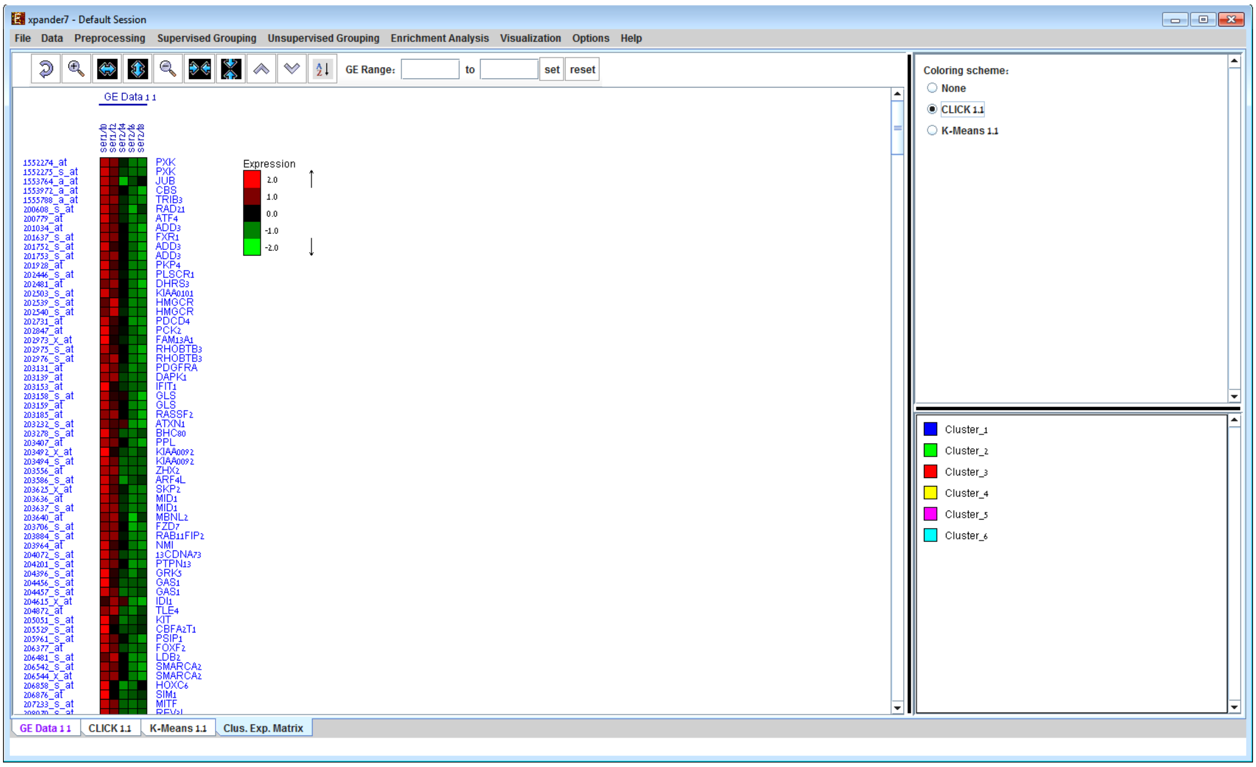
The results in the tables can be saved as image file by selecting *File>>Save As Image* when the corresponding view is the selected tab OR by using the *File>>Save All* option, which will export all solutions within a session to text and image files.

Matrix Visualizations

An expression matrix (Heat-map) visualization is integrated in many of EXPANDER's displays. This visualization is similar to the red-green matrix representation of Eisen et al (1998). All it does is to render the gene-expression data on the screen in color, where green indicates under expression, and red indicates over expression. Color rendering can be configured by the user in one of the following manners: (a) by setting the range (top and bottom values) of rendered values (default values are set according to the data scale, e.g. 40-1000 for non-standardized absolute intensities data) or (b) by setting the percent of values, which are to be disregarded as extreme values from each edge (by default set to 5%). The manner of color scale configuration (i.e. (a) vs. (b)) can be set via the 'Data Matrix View' tab in the 'Display Settings' dialog box, available from *Options* >> *Settings*. The red/green coloring scheme can be changed to blue/yellow (using *Options* >> *Settings* >> *Display* >> 'Data Matrix View' tab).

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

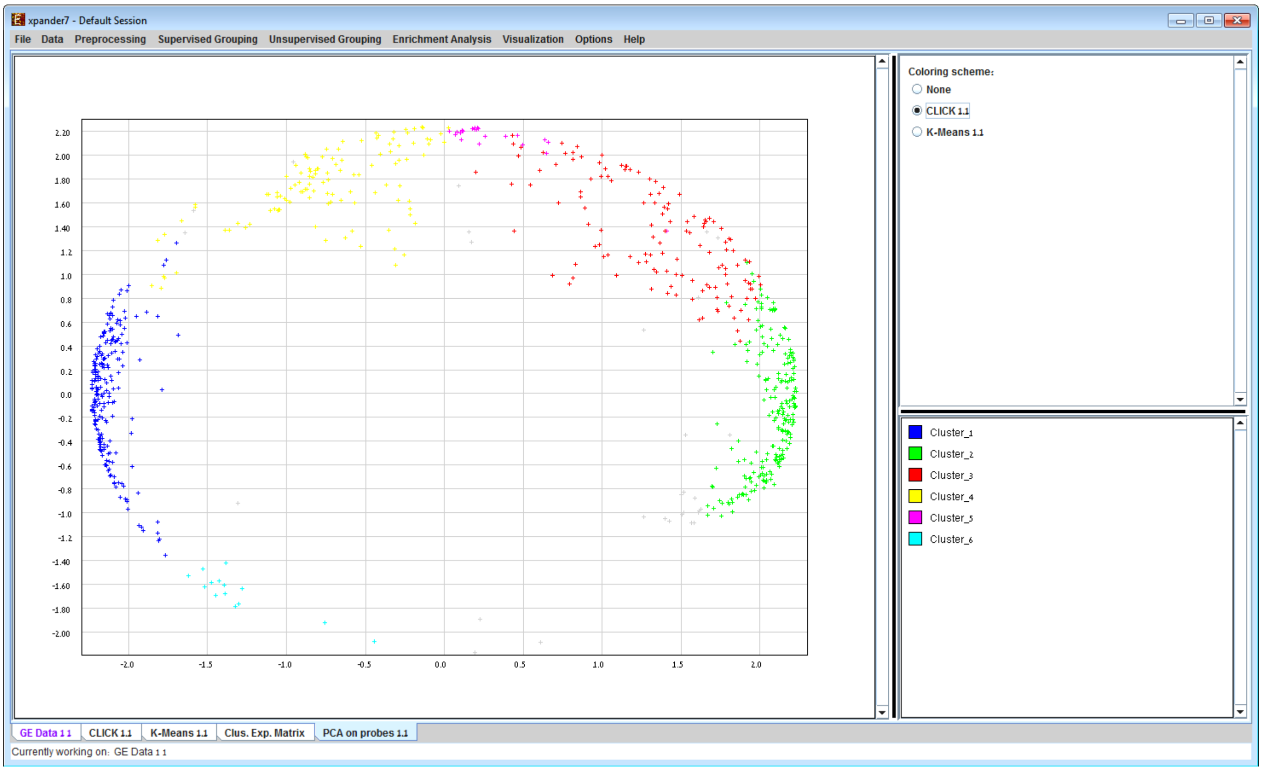
Upon selecting *Visualization >> Clustered Expression Matrix*, a clustered expression matrix visualization tab is added to the main window. The probes are ordered in their original order. If a clustering solution has been previously created, its' name appears next to a radio button in the top right panel. Upon pressing this button, the order of the probes in the display changes and probe IDs are colored according to the clusters. The color index at the bottom right panel, 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 2 dimensional space, so that each expression vector is represented by a dot on an XY scatter chart. The transformation is based on the PCA (Principal Component Analysis) algorithm. To operate the tool, select *Visualization* >> *PCA*.

If a clustering solution has been previously created, its' name appears next to a radio button in the top right panel. Upon pressing this button, the color of each dot in the display changes according to the cluster assignment of the corresponding probe. The color index at the bottom right panel, maps each color to the index of the corresponding cluster.



Coloring scheme:

○ None

● CLICK 1.1

○ K-Means 1.1

🟦 Cluster_1

🟩 Cluster_2

🟥 Cluster_3

🟨 Cluster_4

🟪 Cluster_5

🟦 Cluster_6

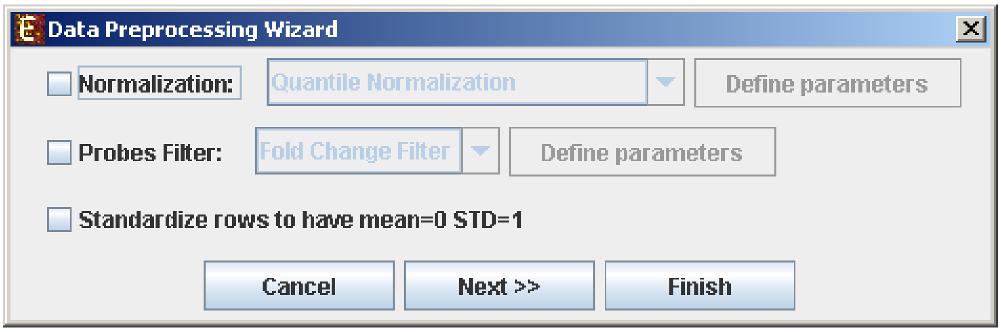
Currently working on: GE Data 1.1

Analysis Wizard

Expander allows performing an automatic analysis on a loaded dataset by using the 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 to define the required preprocessing operations:

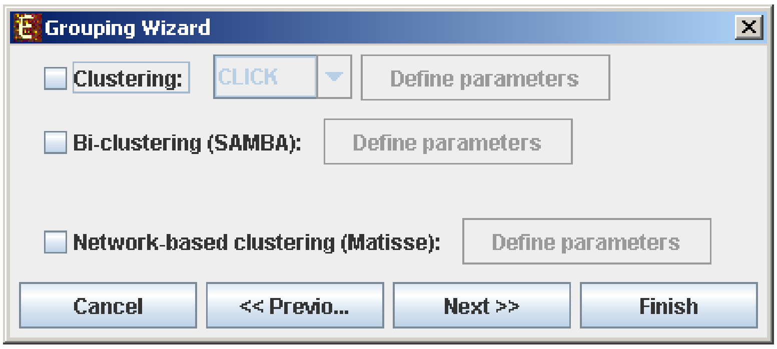


□ Normalization: Quantile Normalization

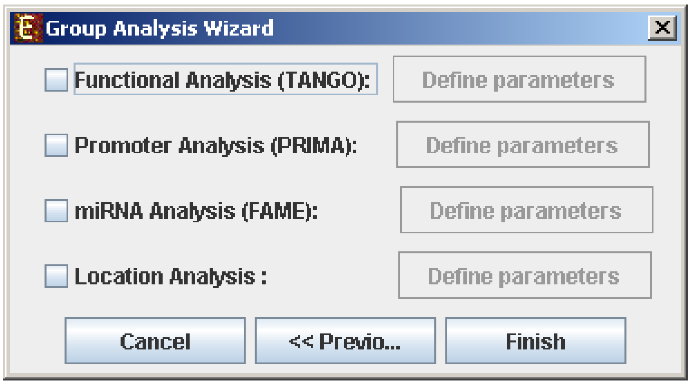
□ Probes Filter: Fold Change Filter

□ Standardize rows to have mean=0 STD=1

For some of the stages, parameters can be defined by pressing the corresponding "Define parameters" button. Upon pressing the "Next>>" button, the following dialog box will appear, allowing to define the required grouping operations:



Upon pressing the "Next>>" button, the following dialog box will appear, allowing to define the required enrichment analysis operations:

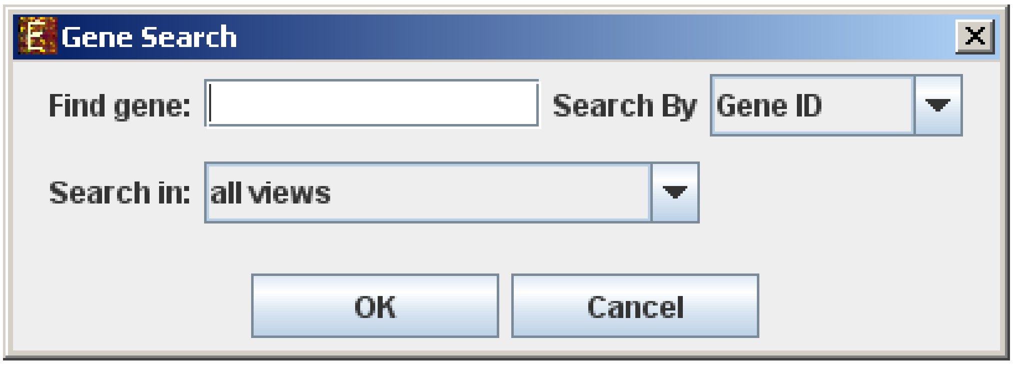


Upon pressing the "Finish" button (in any one of the dialog boxes) the entire set of operations defined by the user is performed by Expander, and the corresponding visualizations are generated.

Additional Options

Searching for a gene/probe in the display

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



Find gene:

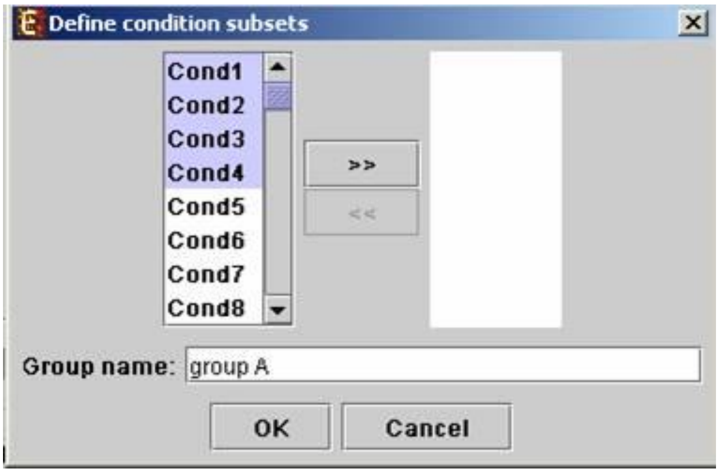
Search By Gene ID

Search in: all views

Please type the ID of the gene (can also be symbol/probe ID depending on the selection in the "Search By" combo-box) in the corresponding text box. Note that you must type the entire name or ID, not part of it. After pressing the "OK" button, a window will appear containing text describing the number of items detected in each of the searched views (number of "hits" in each view). In addition, the corresponding elements will be highlighted in all searched displays.

Defining condition subsets

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



Group name: group A

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

Saving and loading sessions

A set of analysis operations performed on one data set can be saved by selecting *File >> Save Session*. It can later be reloaded by selecting *File >> Load Session*. Loading a previously saved session will bring up all analysis output and visualizations that had been generated in that session, and the user will be able to continue working where he had previously stopped.

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 the relevant view is selected OR by right clicking on the tab title of the relevant view and selecting *Close* from the popup menu.

Docking a view into a separate frame

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

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

Accessing the EXPANDER download page

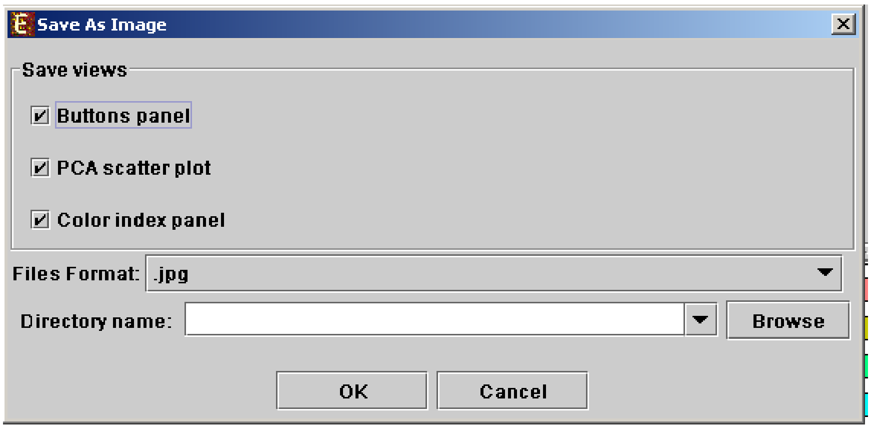
The Expander download page can be accessed directly by selecting *Help >> Open Download Page*, while the machine is connected to the Internet.

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-script). This can be done by selecting *File >> Save As Image*. Upon selecting this option, a dialog box, similar to the following is displayed. In the dialog box the saved images (sections of the view), image files format, and destination directory name are input.



Save As Image

Save views

☑ Buttons panel

☑ PCA scatter plot

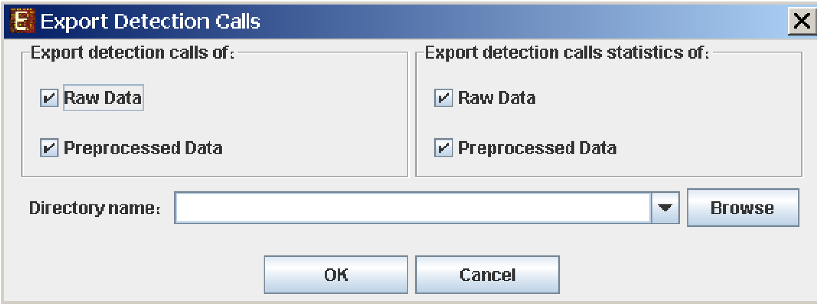
☑ Color index panel

Files Format: .jpg

Directory name:

Exporting detection calls information

The detection calls info of the raw and preprocessed data can be exported into text files, by selecting *Data* >> *Export Detection Calls*. Upon selecting this option, the following dialog box is displayed. You may export the detection calls and also the statistics of detection calls (percent of P, M and A calls per condition), for raw data and for preprocessed data.



Export Detection Calls

Export detection calls of:

☑ Raw Data

☑ Preprocessed Data

Export detection calls statistics of:

☑ Raw Data

☑ Preprocessed Data

Directory name:

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 a string like ‘geneSymbol’ and a tab delimiter, followed by the names of all conditions separated by tab delimiters. The symbol column is optional – if the file does not contain a symbol column, please specify it in the Advanced Input Dialog box (see Input Data section).

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

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

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

If the data is not in the above format, it may be possible to load it using the 'Advanced' dialog box, which appears upon pressing the 'Advanced' button in the Expression Data load dialog 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 a string like 'geneSymbol' and a tab delimiter, followed by the names of all conditions and detection signals columns alternately, separated by tab delimiters. Each title of condition is followed by a title of its detection column.

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

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

*For example see files ‘expressionWithDetection.txt’ in the Expander/sample_input_files/ directory.

If the data is not in the above format, it may be possible to load it using the ‘Advanced’ dialog box, which appears upon pressing the ‘Advanced’ button in the Expression Data load dialog 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 name/number of its set (separated by tabs/spaces). The gene IDs are expected to be of the same convention used in the GO annotation and TF fingerprint files. For details regarding the Gene ID convention that is used for each organism, refer to the Supplied files section.

*For example see file ‘geneSetsData1.txt’ under the Expander/sample_input_files/ directory (see Sample input files for more details).

Similarity file format:

1) Suffix: no limitations

2) Separating token: tab delimiter.

3) Format:

1st line: First field is empty followed by tab delimiter, followed by probe/gene/identifier ids separated with tab delimiters.

Next lines: First field in line i should be the identifier of column i, followed by fields with values between (-1) and (1) describing the similarity between the identifiers. A similarity value of 0 between two identifiers represents no similarity. A similarity value of 1 between two identifiers represents a complete similarity and a similarity value of (-1) between two identifiers represents the opposite similarity.

*For example see file ‘sim.txt’ under the Expander/sample_input_files/ directory.

Gene Rank file format:

1) Suffix: no limitations

2) Format: Each line contains a gene ID and a rank number (separated by tab/space) where the highest gene is ranked 1. The gene IDs are expected to be of the same convention used in

the Gene Set Enrichment Analysis. For details regarding the Gene ID convention that is used for each organism, refer to the Supplied files section.

ChIP-Seq file format:

1) Suffix: BED or GFF3

2) Format: Please refer to the following links explaining the formats:

- BED
- GFF3 – note that "Score" field is Q-value and can range between 0-1 or be in –log(10) values.

Note that the files should not contain any headers and should contain only the peaks data.

Probes Filter file format:

Each line contains a single identifier. Identifiers can be probe Ids, gene Ids OR gene symbols (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 separator and the corresponding gene ID (e.g. Entrez/Locus-Link ids for mouse and human genes and ORF codes for yeast). The second field can be left blank, indicating no conversion for that probe ID.

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

condition attributes:

1) Suffix: no limitations

2) Separating token: tab delimiter.

3) Format: First line contains the headers where the first field can be left blank or given any name (e.g., "Attribute") and the next fields are the condition names as in the loaded expression dataset. The next lines contain in the first field the attribute identifier/name and the other fields are the labels of the attribute (e.g., "Treatment<tab>ntrt<tab>trt<tab>trt<tab>ntrt …") for each condition.

*For example see file ‘conditionAttribute.txt’ under the Expander/sample_input_files/ directory

Clustering files format:

1) Suffix: no limitations.

2) Format: Each line contains the probeID, a tab separator and name/number of its cluster. The number 0 is reserved for probes that are left unclustered. The file does not have to contain all probes in the data. If a probe does not appear in the file, it is automatically set as unclustered.

*For example see file ‘expressionData1Clustering.sol’ (a clustering solution for the data file’ expressionData1.txt’) under the Expander/sample_input_files/ directory (see Sample Input 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 tab delimiter (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' for probe) and 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 be of the same convention used in the annotation and TF fingerprint files for the organism you are working on (please refer to the Supplied Files section).

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

1) Suffix: no limitation.

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

DESeq2 annotation data file format:

1) Suffix: no limitation.

2) Separating token: tab delimiter.

3) Format:

1st header line: First field can be given any name (e.g., "Condition ID") or be left empty, followed by the attributes/categories names.

Next lines: First field is a condition ID or can be left empty, followed by the labels for each attribute/category. Note that DESeq2 assumes that line i corresponds to column i in the expression data. Make sure that the number of lines (not including the header line) matches to the number of conditions in the expression data.

*For example see file ‘deseqAnnotation.txt' under the Expander/sample_input_files/ directory

Sample Input Files

Several sample files are provided under Expander/sample_input_files/. These files include:

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

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

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

expressionWithDetection.txt – A gene expression data file with detection calls, that was generated in the Affymetrix technology. This dataset was generated in an experiment that was conducted in our laboratory on human cells. Affymetrix GeneChip HGU133 Plus 2.0 arrays were used for this experiment.

expressionData1Clustering.sol – A clustering solution that was generated by Expander for 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 annotation files, TF fingerprint files, promoter sequences, miRNA target scan files, chromosomal position files, and biological pathway files, taken from the KEGG database*. These files should be extracted 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**

Human	B	545M	GO@EBI (May 2017)	Ensembl database release 89	Target Scan website version 5	UCSC genome browser on August 2012
Baker's yeast	B	72.2M	NCBI (May 2017)	Ensembl database release 89	-	-
S. pombe	B	72.3M	NCBI (May 2017)	Ensembl database release 36	-	-
Listeria monocytosenes EGD-e	B	1.74M	Blast2GO -February 2009	Not available	-	-
Mouse	B	419M	NCBI (May 2017)	Ensembl database release 89	Target Scan website version 5	UCSC genome browser on August 2012
Rat	B	262M	NCBI (May 2017)	Ensembl database release 89	-	UCSC genome browser on Mar 2012
Fly	B	235M	NCBI (May 2017)	Ensembl database release 89	Target Scan website version 5	UCSC genome browser on Mar 2012

C-elegans	278Mb	NCBI (May 2017)	Ensembl database release 89	Target Scan website version 5	UCSC genome browser on Mar 2012
Arabidopsis	345MB	NCBI (May 2017)	Ensembl database release 36	-	-
Zebra Fish	333Mb	NCBI (May 2017)	Ensembl database release 89	-	UCSC genome browser on August 2012
Chicken	90.4Mb	NCBI (May 2017)	Ensembl database release 89	-	-
Tomato	347Mb	SGN database on June 2017	Ensembl database release 36	-	UCSC genome browser on August 2012
A.Fumigat	79.3M	GO@EBI	Kevin	-	-

us	b	(May 2017)	Verstrepen's lab (via private communication with Nir Osherov)		
E. coli	32MB	GO@EBI (June 2017)	UCSC microbes	-	-
Rice	360MB	Go database (July 2012)	Ensembl database release 36	-	-
Leishmania	2MB	Zilberstein D. lab Technion - Israel Nov 2011	-	-	-

* Users of this product may not download large quantities of KEGG Data.

Gene ID conversion files:

Gene ID conversion files for many of the Affymetrix chips can be downloaded from the Expander download page. The files map each Affymetrix Id into the corresponding gene Id. Conversion files are generated and added to the download page according to user requests. If you can't find the file you need here, please look it up in the download page, and contact us if 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.sif	IntAct database (http://www.ebi.ac.uk/intact/)
Human	ppi.IntAct.sif	IntAct database (2017)
Mouse	Expander.mmu.IntAct.sif	IntAct database (http://www.ebi.ac.uk/intact/)
Mouse	ppi.IntAct.sif	IntAct database (2017)
Rat	Expander.rno.IntAct.sif	IntAct database (http://www.ebi.ac.uk/intact/)
Rat	ppi.IntAct.sif	IntAct database (2017)
Zebrafish	ppi.IntAct.sif	IntAct database (2017)
C.elegans	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)
C.elegans	ppi.IntAct.sif	IntAct database (2017)
Fly	ppi.IntAct.sif	IntAct database (2017)
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)
Yeast	ppi.IntAct.sif	IntAct database (2017)
S.pombe	ppi.IntAct.sif	IntAct database (2017)
Arabidopsis	Expander.ath.TAIR.sif	TAIR database (http://www.arabidopsis.org/)
Arabidopsis	ppi.IntAct.sif	IntAct database (2017)
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)
E. coli	ppi.IntAct.sif	IntAct database (2017)
Rice	ppi.IntAct.sif	IntAct database (2017)

Settings

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

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

Clustering Results View – Contains check boxes that configure the following 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 configures whether 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 matrix values.
- The red/green coloring scheme can be changed to blue/yellow.

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

R External Application

The CEL file preprocessing and the newly added SAM filter utilities require the pre-installation of one of the recent versions of R, a free software environment for statistical computing and 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, asking you to specify your R software location. Please browse to the location of your R software. In Windows, R.exe file is likely to be located in the 'bin' folder of R software. In Linux, you may type 'which R' in the command line to find R path. If you have a few versions of R installed, please make sure to point Expander to a version in which the necessary packages have been installed.

You may also specify R location from the menu: *Options* >> *Settings* >> *External applications*.

To use R utilities, please make sure there are no white spaces in the path of Expander directory (or the CEL files directory, if loading CEL files). For example, if the name of the Expander folder is 'Expander 5', change it to 'Expander_5'. If Expander is under "Program Files" it should be moved to another location, because of the space between "Program" and "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 directory. If you are loading CEL files, check also that you have 'write' permission to the *Files location* which you specified in the 'Load CEL Files' dialog box.

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

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

If you disapprove automatically installation of R packages then please refer to "Manually installation of R packages" section.

Manually installation of R packages

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

1. Run R.
2. In the R frame\window type the text: **source("http://bioconductor.org/biocLite.R")**
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/biocLite.R")**

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 (to install only "samr", follow instructions number 1, 6 and 7).

To install the ‘edgeR’ package:

15. In the R frame\window type the text: **biocLite("edgeR")**

16. Press ‘Enter’.

To install the ‘DESeq2’ package:

17. In the R frame\window type the text: **biocLite("DESeq2")**

18. Press ‘Enter’.

FAQ

Linux/Unix problems

Click clustering, Samba bi-clustering, Tango and Prima Promoter analysis algorithms fail when 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)

When I run Functional Analysis, Expander gets stuck.

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

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

When I try to run the promoter\functional analysis, I get a failure message box.

Promoter\Functional analysis produces no results (the resulting view is empty).

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

Why do certain Transcription Factors have a few accession numbers or\and a few gene IDs?

Saving sessions

When I try to save a session Expander fails and returns an XStream error message.

Others

Can I run Expander on Mac OS?

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

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

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' window is displayed), please check if there is a folder with a space in its name somewhere in the path of Expander (or the CEL files) directory. For example, if the name of the Expander folder is 'Expander 4', change it to 'Expander_4'. The R software used for preprocessing CEL files has a problem dealing with spaces in the path. If this is the problem, then in the expanderLog.txt file (in your 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 installed in the R version which is specified in the settings “External Applications” tab (from the menu select *Options* >> *Settings* >> *External applications*). If R location is not defined in the settings, please define it (In Windows, R.exe file is likely to be located in the 'bin' folder of R software. In Linux, you may type 'which R' in the command line to find R path). If you are using an R package as cdf source, please make sure that the package is a folder located under your R library directory and that it is the correct package for your chip. If loading of CEL files still fails, please make sure that the *Files location* which you specified in the 'Load CEL Files' dialog box, is a folder which contains CEL files and that you have write permission to that folder.

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 dialog box is not empty (if it is, please re-download Expander). Also make sure that the following files exist in 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 update to a higher version (4.0.2 and on).

When I load a session with that contains Functional Analysis results, Expander gets stuck.

Answer: If you are working with Expander version 4.0 or 4.0.1 (and the session was created 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 field of the input dialog box.

Answer: Fingerprint files are not placed in the right directory. Fingerprint files should be placed under the 'TF_fingerprints' directory that is under the Expander/organism/<org name> directory. For example, the human FP file should be placed under: …/Expander/organisms/human/TF_fingerprints/. When downloading the organism specific data zip, it should be extracted into the Expander/organisms/ directory. This will automatically put them in the right place.

Enrichment analysis leads to a failure message box.

Answer: Errors while running group (enrichment) analysis can be caused by the following problems:

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

b) Data contains elements that do not appear in the background set (this is only relevant when the 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 right format or does not map the probes to the expected type of gene Ids. The conversion file maps each probe ID in your data file to a gene ID that is used for enrichment analysis. A conversion file is required when the probe Ids in your data file do not match the ones in the enrichment files (for 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 your data.

c) You are trying to analyze only one set (e.g. the filtered data set) which you are using also as background (in this case the analysis has no meaning since it is trying to detect enrichments in 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 clustering/biclustering solution or this gene sets data.

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

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 few gene IDs?

Answer: The transcription factors (TFs) found enriched by Prima are presented in the following 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, which represent different PWMs (position weight matrices specify the probability for observing each nucleotide at each position of the binding site, based on a set of empirically validated binding sites of the respective TF).

It is also possible that a TF will have a few Entrez gene IDs, since a TF may be composed of a 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 – without running its features that require the execution of exe files (CLICK, SAMBA, TANGO 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 message

If you are using java version 1.7, please switch to version 6. We currently have no solution for this problem, that occurs with java1.7 and XStream, which is an external package that we are using. We will do our best to resolve it in the coming future. In order to configure the 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 Files (x86)/Java” (or “Program Files/Java”) directory. In it there should be a subdirectory by the name jre6 or jre5 (otherwise please install java6 from http://java.sun.com/javase/downloads/index.jsp.)
2) In the Expander directory right click on one of the Expander.bat files (the one you are using) and select “Edit”.
3) In the file type the path of the java6 exe file instead of the word java. E.g. if your path is C:\Program Files (x86)\Java\jre6 then put the text: “C:\Program Files (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

Copyrights © Tel-Aviv University, Israel (2003).

This product uses the FreeHEP Java Library, which is distributed under the LGPL license. FreeHEP copyright holders: CERN, Geneva, Switzerland SLAC, Stanford, California, U.S.A. University of California Santa Cruz, U.S.A.

This product uses the XStream Java Library, which is distributed under the BSD license (see BSD.txt). Copyright holders: (c) Joe Walnes 2003-2005

This product uses the Caryoscope java component, which is distributed under the MIT license (see MIT_Caryoscope.txt). Copyright holders: Copyright 2003-2004, Ihab A.B. Awad; Copyright 2006, Anjalee Sujanani; Stanford University.

A portion of the user interface code is due to Sun Microsystems.Inc. Copyright 1994-2004 Sun Microsystems, Inc. All Rights Reserved. The following license rules apply to that portion:

‘Neither the name of Sun Microsystems, Inc. or the names of contributors may be used to endorse or promote products derived from this software without specific prior written permission.

This software is provided ‘AS IS,’ without a warranty of any kind. ALL EXPRESS OR IMPLIED CONDITIONS, REPRESENTATIONS AND WARRANTIES, INCLUDING ANY IMPLIED WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NON-INFRINGEMENT, ARE HEREBY EXCLUDED. SUN MICROSYSTEMS, INC. (‘SUN’) AND ITS LICENSORS SHALL NOT BE LIABLE FOR ANY DAMAGES SUFFERED BY LICENSEE AS A RESULT OF USING, MODIFYING OR DISTRIBUTING THIS SOFTWARE OR ITS DERIVATIVES. IN NO EVENT WILL SUN OR ITS LICENSORS BE LIABLE FOR ANY LOST REVENUE, PROFIT OR DATA, OR FOR DIRECT, INDIRECT, SPECIAL, CONSEQUENTIAL, INCIDENTAL OR PUNITIVE DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF THE USE OF OR INABILITY TO USE THIS SOFTWARE, EVEN IF SUN HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

You acknowledge that this software is not designed, licensed or intended for use in the design, construction, operation or maintenance of any nuclear facility.’

References

R: R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

affy package for CEL files preprocessing: Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. 2004. affy---analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20, 3 (Feb. 2004), 307-315

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

limma package: Smyth, GK (2005). Limma: linear models for microarray data. In: 'Bioinformatics and Computational Biology Solutions using R and Bioconductor'. R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds), Springer, New York, pages 397-420.

edgeR package: Robinson MD, McCarthy DJ and Smyth GK (2010). edgeR: a Bioconductor

package for differential expression analysis of digital gene expression data. Bioinformatics 26,139-140.

DESeq2 package: Love MI, Huber W and Anders S (2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” *Genome Biology*, **15**, pp. 550. http://doi.org/10.1186/s13059-014-0550-8.

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

Non-linear baseline normalization: Schadt, E., C. Li, B. Eliss, and W. H. Wong. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array 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 applied to the ionizing radiation response. PNAS, 98: 5116-5121, 2001

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

K-Means clustering algorithm: Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M. Systematic determination of genetic network architecture. Nat Genet, 22: 281-285, 1999

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

CLICK clustering algorithm: Sharan, R. and Shamir, R. CLICK: a clustering algorithm with applications to gene expression analysis. Proc Int Conf Intell Syst Mol Biol 8, 307-16, 2000

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

SAMBA biclustering algorithm: Tanay, A. Sharan, R. and Shamir, R. Discovering statistically significant biclusters in gene expression data. Bioinformatics, 18(1), 136-144, 2002

Matisse network grouping: Ulitsky, I. and Shamir, R. MATISSE: Identification of functional modules using network topology and high-throughput data. BMC Systems Biology, vol 1, No. 8 (2007)

Degas network grouping:

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

Context Scores for miRNA enrichment analysis: Grimson, A., Kai-How Farh, F. K Johnston, W., Garrett-Engele, F., P Lim, L., P Bartel, D. MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing. Molecular Cell, 27:91-105 (2007)
PRIMA algorithm: Elkon, R., Linhart, C. Sharan, R. Samir, R. and Shiloh, Y. Genome-Wide In Silico Identification of Transcriptional Regulators Controlling the Cell Cycle in Human Cells. Genome Research, Vol. 13(5), pp. 773-780, 2003.

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

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

Agglomerative algorithm for hierarchical clustering: Eisen, M. B., Spellman, P. T. et al. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95(25), 14863-8, 1998

TF binding site profiles that were used to generate the supplied yeast TF fingerprint files: Harbison, C.T., D.B. Gordon, T.I. Lee, N.J. Rinaldi, K.D. Macisaac, T.W. Danford, N.M. Hannett, J.B. Tagne, D.B. Reynolds, J. Yoo, E.G. Jennings, J. Zeitlinger, D.K. Pokholok, M. Kellis, P.A. Rolfe, K.T. Takusagawa, E.S. Lander, D.K. Gifford, E. Fraenkel, and R.A. Young. Transcriptional regulatory code of a eukaryotic genome. Nature 431: 99-104, 2004

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

Gene Set Enrichment Analysis: Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS vol. 102, no. 43, 15545–15550, 2005