EMBL-EBI Tutorial

Gabriella Rustici (version2, 11/08/2008)

Expression Profiler for Beginners – part 1

Expression Profiler (EP) is a web-based platform for gene expression data analysis. Individual components for data pre-processing, filtering, significant gene finding, clustering, visualization, between group analysis and other statistical tools are all available in EP, mostly implemented via integration with R [3]. The web-based design of EP supports data sharing and collaborative analysis in a secure environment. Developed tools are integrated with the microarray database ArrayExpress (AE) and form the exploratory analytical front-end to those data. Users can upload in EP their own data or data retrieved from the AE database. The users only need a web browser to use EP from their local PCs.

You will learn about:

- The basics of Expression Profiler – how to get started
- How to upload data
- How to transform the data
- How to filter data
- How to analyze the data using basic tools such as clustering and GO annotation
- How to identify differentially expressed genes using t-test

Contents:

1 The basics of Expression Profiler – how to get started
2 How to upload data
3 How to transform the data
4 How to filter the data
5 Clustering analysis in Expression Profiler
6 Gene Ontology annotation in Expression Profiler
7 Identification of differentially expressed genes using t-test analysis
8 How to obtain a data matrix for experiment E-MEXP-29

This work is licensed under the Creative Commons Attribution-Share Alike 3.0 License. To view a copy of this license, visit http://creativecommons.org/licenses/by-sa/3.0/ or send a letter to Creative Commons, 543 Howard Street, 5th Floor, San Francisco, California, 94105, USA.

1 The basics of Expression Profiler – how to get started

Go straight to the EBI's EP main page by using Tools – Microarray Analysis menu on the EBI homepage (http://www.ebi.ac.uk - Fig. 1).

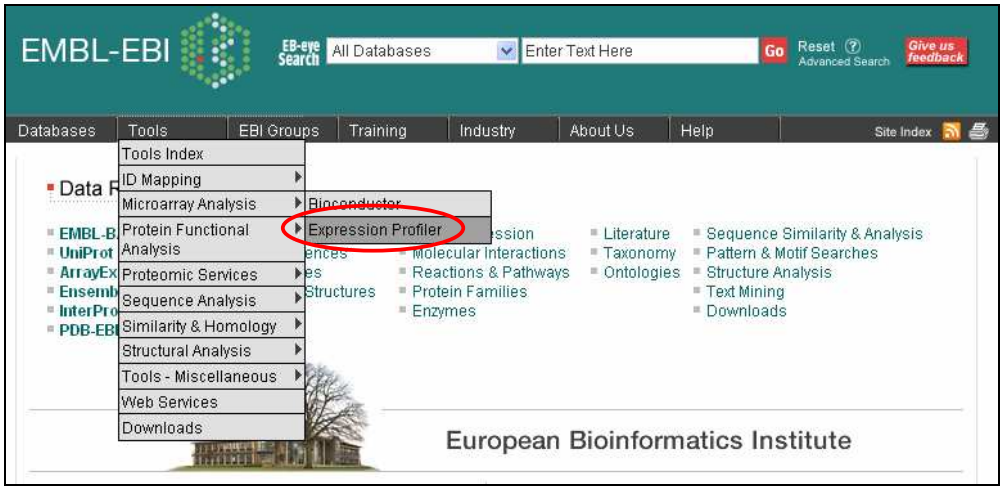


Fig. 1: Accessing Expression Profiler from the EBI homepage (http://www.ebi.ac.uk/)

This will bring you to the EP homepage (Fig. 2). If this is the first time you have used EP, you will need to fill in the new user registration page with all the details required and choose a personal user name and password. You will be able to use them each time you want to login. All the data loaded and analysis history will be saved and stored under this user login, until you decide to delete/modify it. With a 'guest login' all the data and analysis will be lost at the end of each session.

At the next login, click on the 'EP:NG Login Page' link, on the EP main page, enter your username and password and click 'login'. You will then be prompted to the data upload page.

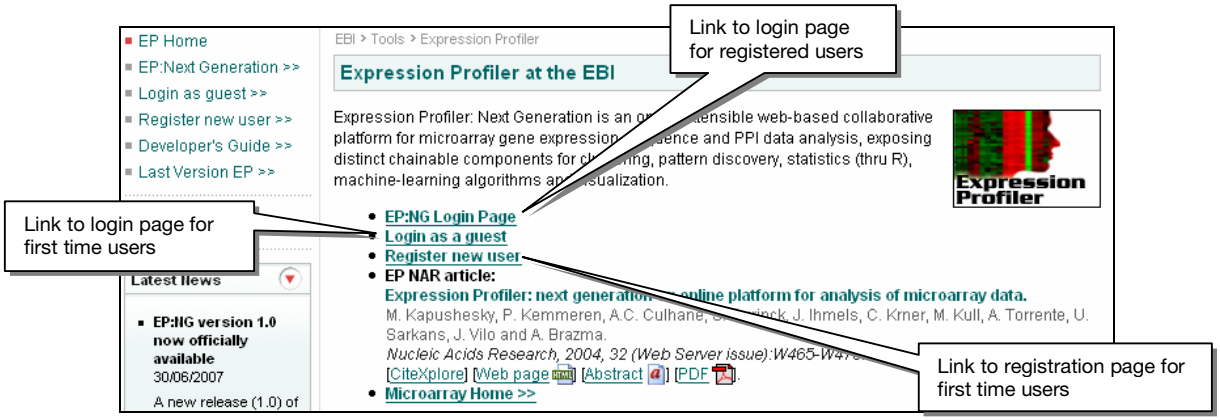


Fig. 2: Expression Profiler homepage (http://www.ebi.ac.uk/expressionprofiler/)

2 How to upload data

The Data Upload component (Fig. 3) can accept data in a number of formats including basic tab-delimited files, such as those exported by Microsoft Excel ('Tabular data' option), and **Affymetrix .CEL data files** ('Affymetrix' option). Users can also select a published dataset from the AE database through the EP interface ('ArrayExpress' option). A particular dataset can also be directly uploaded from a specific URL, for both Affymetrix and tabular data. Except for .CEL files, uploaded expression datasets must be represented as a **data matrix**, with rows and columns corresponding to genes and experimental conditions, respectively.

For this tutorial, we will use part of the normalized E-MEXP-29 dataset exported from the AE database in the 'ArrayExpress for Beginners' Tutorial. All you need is the **data matrix**, saved as .csv file. At the end of this tutorial is a quick reminder on how to obtain this file.

In this study, transcriptional profiling of stress response in fission yeast (*Schizosaccharomyces pombe*) cells was performed in order to identify genes whose expression varies following a stress stimulus. Five different types of stress were employed on wild type and mutant cells (**sty1** and **atf1** knock-out cells). The five stresses were heat, and four types of compound: heavy metal (cadmium), oxidation (hydrogen peroxide), alkylation (MMS) and osmosis (sorbitol). For each condition, cells were harvested immediately before (reference) as well as 15 and 60 min after stress treatment. A total of 67 two-channel arrays were used. Each sample was hybridized to one array (channel 1) together with a reference sample from the same culture (channel 2) [9].

To simplify things, in this tutorial, we are only using a subset of this dataset, consisting of 8 conditions: 4 wild type untreated (2 biological replica, 15 and 60 min) and 4 wild type treated with 0.5 mM hydrogen peroxide (2 biological replica, 15 and 60 min). The **data matrix** we are about to upload contains *n* rows (where *n* = total number of genes) and 8 columns, each corresponding to one experimental condition. An entry in the matrix is a ratio of expression levels of one gene under one condition. Each ratio is calculated by dividing the **probe intensity** value in channel 1 (sample) by the **probe intensity** value in channel 2 (reference). The data was normalized by the authors and it is taken as such [9].

Fill in the 'Tabular data' upload page as shown in Fig. 3 and click 'Execute'.

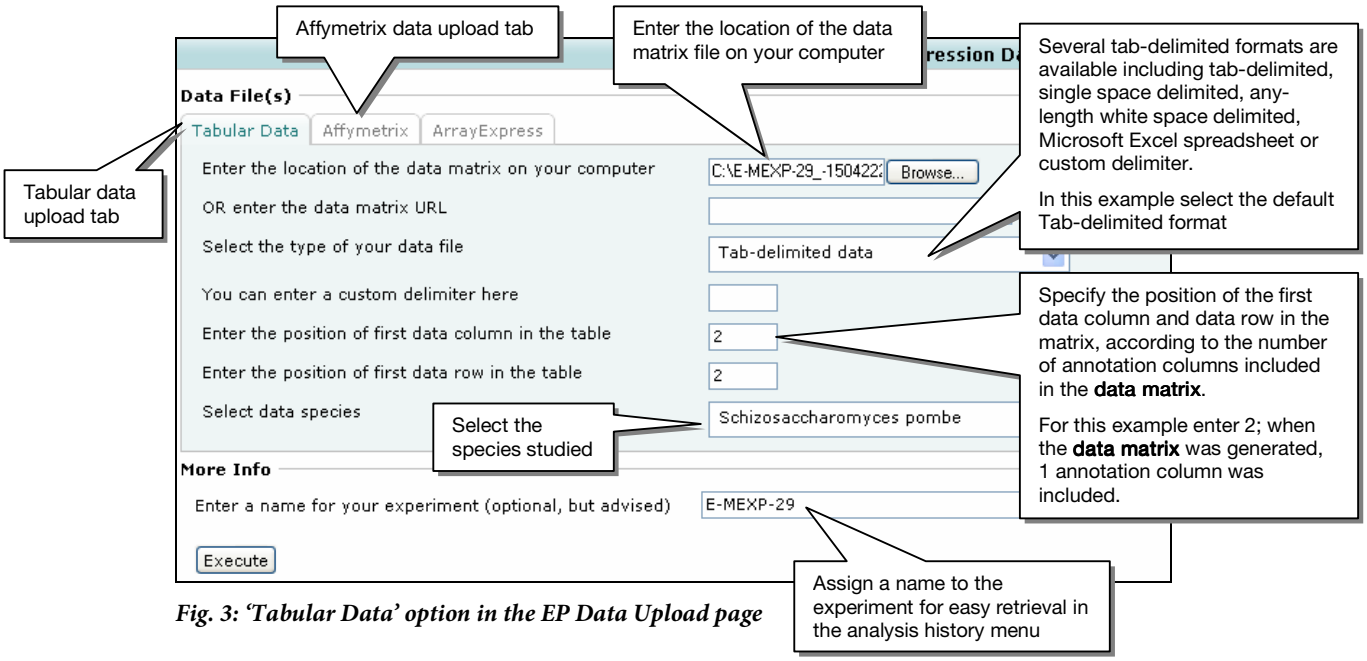


Fig. 3: 'Tabular Data' option in the EP Data Upload page

After a successful microarray data import, the EP Data Selection view is displayed (Fig. 4).

This view has three sections:

- 'Current dataset', where the user's folder structure, current dataset selection and ongoing analysis history are displayed. EP stores all parameters, results and graphics files for every performed analysis step (Fig. 4, yellow box). These can be retrieved at any stage in the analysis by clicking the 'View action output' icon next to the respective analysis step; this is the last icon on the right-hand side. Additional icons allow the user to view the entire dataset as well as row and column headers.
- 'Descriptive statistics', where data visualization graphics are provided such as a plot of **perfect match (PM) probe intensities** (log-scale) for **Affymetrix arrays** or distribution **density histograms**, for **one- and two-channel experiments** (ratios and log-ratios) (Fig. 4, green box).
- A bottom menu, which changes according to which EP analysis component the user selects from the main menu (Fig. 4.1). After the data import, the 'Subselection' menu is shown by default (Fig. 4, orange box).

The loaded dataset is now selected in the 'Current dataset' window, the normalised ratio distribution is shown in the 'Descriptive statistics' plot and the data is now available for further **pre-processing** and analysis.

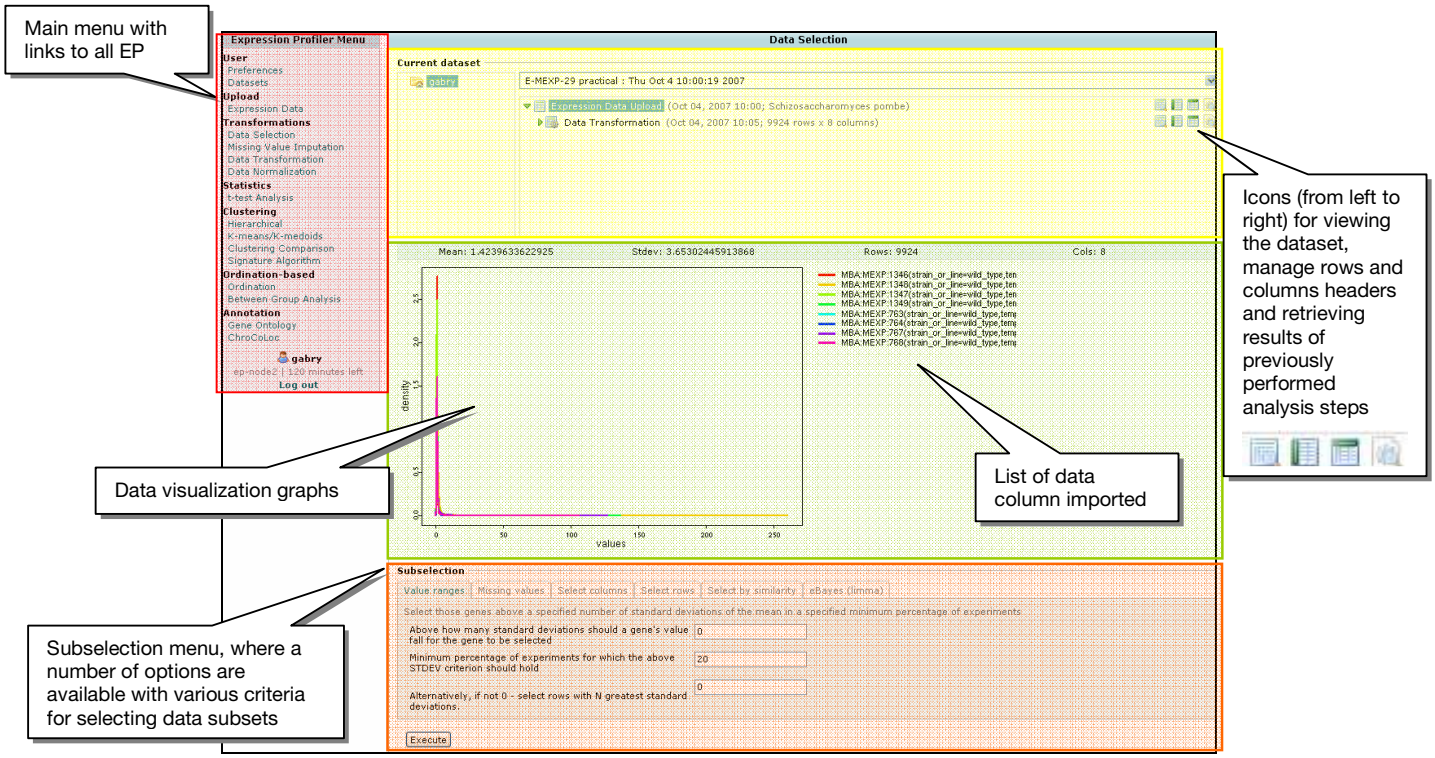


Fig. 4: EP Data selection view after uploading expression data from experiment E-MEXP-29. This window is divided in 3 mains sections: current dataset (yellow box), descriptive statistics (green box) and subselection menu (orange box). The EP main menu, on the top left hand side, is highlighted in red.

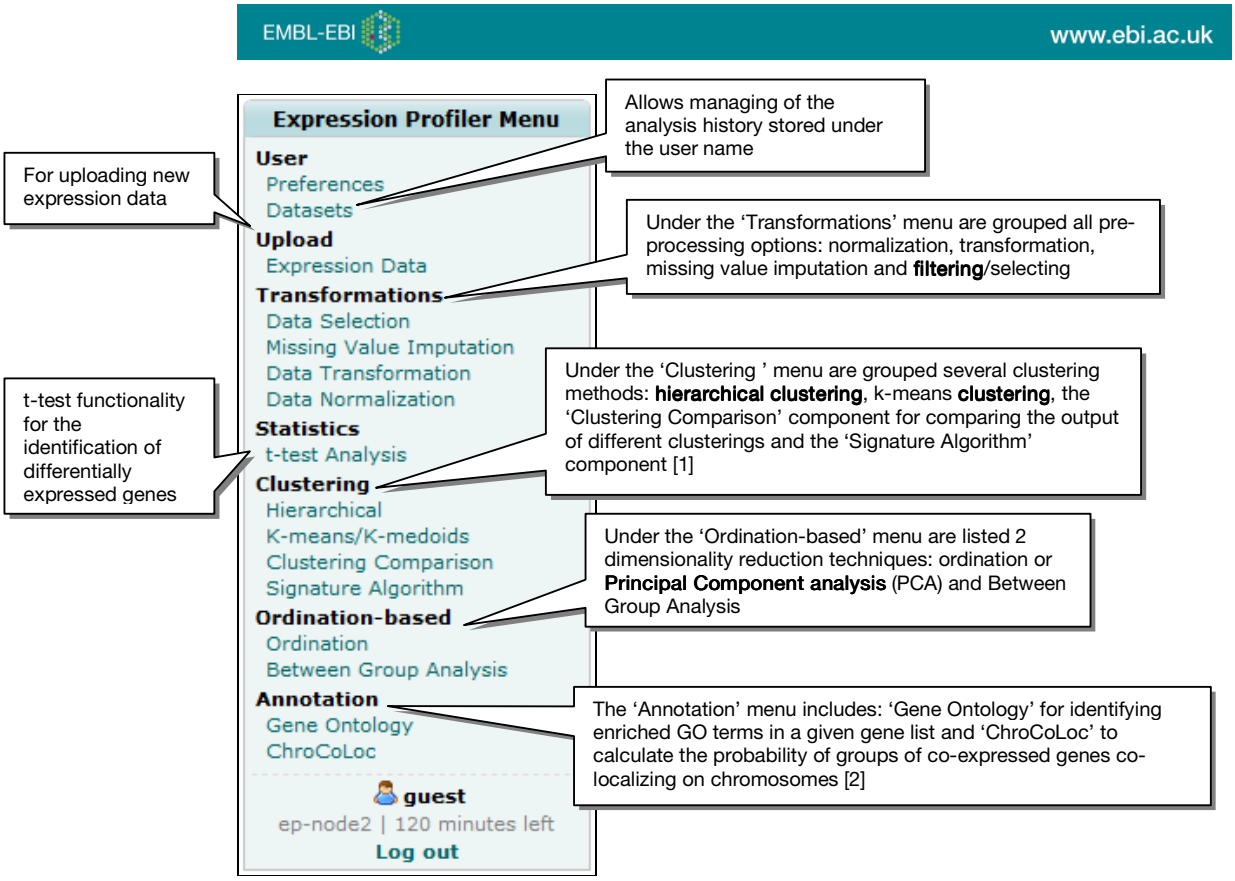


Fig. 4.1: EP main menu with links to all EP data analysis components.

3 How to transform the data

The E-MEXP-29 dataset is already normalised so we do not need to run any **normalisation** algorithm. Observe that the normalised ratio intensity distribution is centred on 1 (mean = 1.42, Fig. 4.2). In this dataset, the expression value for each gene and time point was calculated, relative to a reference sample; as result, the median ratio over the course of the experiment is centred on 1.

If you take a closer look to the ratio intensity distribution (Fig. 4.2), you will see how all repressed genes (those genes that had higher intensity vales in the reference) have ratios compressed between 0 and 1 while all induced genes (those genes that had higher intensity vales in the sample) have ratios between 1 and >250. Clearly, repressed and induced genes are not equally represented and the compression of ratios between 0 and 1 causes problems with mathematical techniques for analyzing and comparing gene expression patterns.

This problem can be easily solved by log-transforming the ratio data. The key attribute of log-transformed expression data is that equally sized gene induction and repression receive equal treatment, visually and mathematically. For this reason, microarray data is normally transformed from ratios to log-ratios.

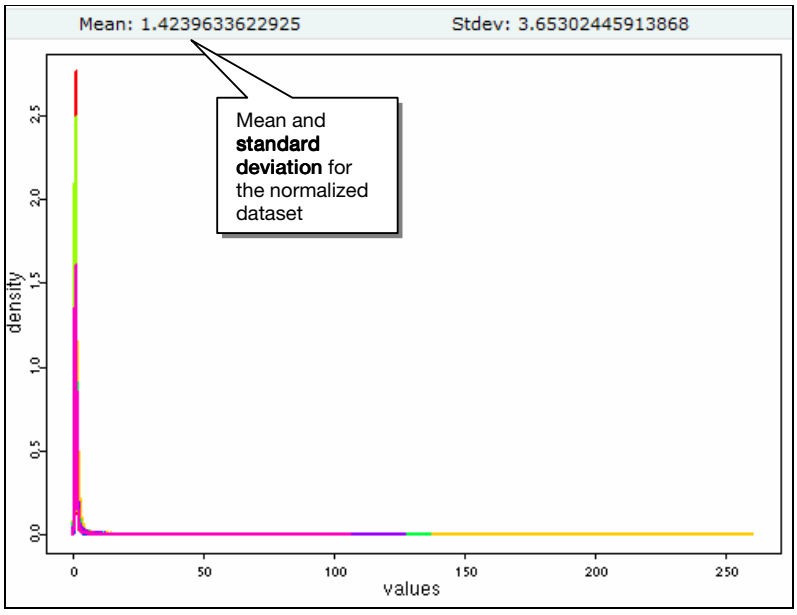


Fig. 4.2: Ratio intensity distribution plot for E-MEXP-29.

In the EP main menu, click on 'Data transformation', under the 'Transformations' menu (Fig. 4.1). Several types of transformations are available (Fig. 5). They allow converting the data from one format to the other, as required by the user.

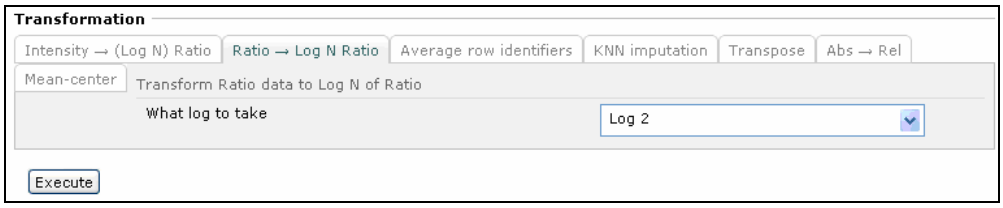


Fig. 5: Transformation menu in EP – 'Ratio to Log N ratio' is selected

The description of all transformation available, from left to right, is as follows:

Transformation type	**Description**
Intensity to (Log N) Ratio	For taking a set of two-channel arrays, dividing every channel 1 column by the respective channel 2 column, and then, optionally taking a logarithm of the ratio
Ratio to Log N Ratio	For log-transforming the selected dataset
Average row identifiers	For replacing multiple rows with the same identifier with a single row, containing the column-wise averages
K-Nearest Neighbour (KNN) Imputation	For filling in the missing values in the **data matrix** [10]
Transpose	For switching the rows and columns of the matrix
Absolute to Relative	For converting from absolute expression values to relative ones, either relative to a specified column of the dataset, or relative to

	the gene's mean.
Mean-center	For rescaling the rows and/or columns of the matrix to zero-mean. It can be used for running ordination-based methods (e.g. PCA)

In this case, we will apply a log2 transformation to the normalised E-MEXP-29 dataset by clicking on the 'Ratio to Log N Ratio' tab and selecting 'log 2' in the 'What log to take' drop down menu (Fig. 5). Click 'Execute'.

The result of data transformation will be displayed in a new window as a 'Dataset **heatmap**'. Explore the changes in the expression value **distribution plot** after transformation by going back (by clicking on the 'Data selection' in the main menu) to the previous window (Fig. 6). Observe that the log-ratios distribution is now centred on 0 (mean = 0.083). Now both repressed and induced genes are equally represented allowing us to perform further analysis.

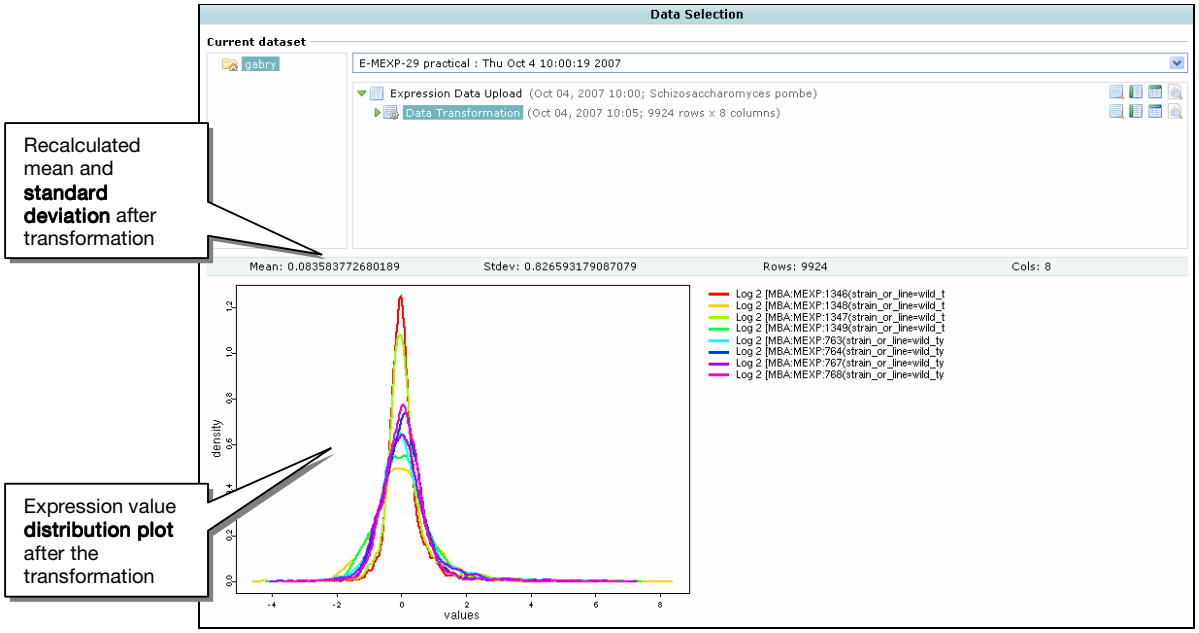


Fig. 6: Log-ratio intensity distribution plot for E-MEXP-29.

4 How to filter the data

In the EP main menu, click on 'Data selection', under the 'Transformations' menu (Fig. 4.1).

The 'Data selection' components provide several basic mechanisms to select, at any stage of the analysis, genes and conditions that might be of particular interest. The description of all selection options available is as follows:

Selection type	**Description**
Select rows and Select columns	For sub-selecting a slice of the gene expression matrix by row or column names (partial word matching can be used for this filter)
Missing values	For **filtering** out rows of the matrix with more than a specified

	percentage of the values marked as NA (not available).
Value ranges	For selecting genes above a specified number of **standard deviation**s of the mean in a minimum percentage of experiments. Alternatively, it can be used to sub-select the top N genes with greatest **standard deviation**s; an input box is provided to specify the value N.
Select by similarity	Provides the functionality to supply a list of genes and, for each of those, select a specified number of most similarly expressed ones in the same dataset, merging the results in one list
eBayes (limma)	Provides a simple interface to the eBayes function from the limma Bioconductor package [11]. It allows searching for **differentially expressed** genes in predefined sample groups, which are determined by discriminating factors.

Statistical analysis of microarray data can be significantly affected by the presence of missing values, especially when working with two-colour arrays, as in this example. Therefore, it is important to estimate missing values as accurately as possible before performing any analysis, One option is to use one of the 'Missing Value Imputation' methods available under the 'Transformations' menu [10, 12] (Fig. 4.1). Alternatively, we can simply filter the data allowing only for a small percentage of missing values in the dataset.

Click on 'Data selection', under the 'Transformations' menu (Fig. 4.1) and select the 'Missing values' tab. Choose 10 as maximum percentage of missing values allowed and click 'Execute'.

The result of data selection will be displayed in a new window as a 'Dataset **heatmap**'. Go back to the previous window (by clicking on the 'Data selection' in the main menu) and observe how many genes are eliminated after **filtering**. 8908 genes out of 9924 made the cut (Fig. 7).

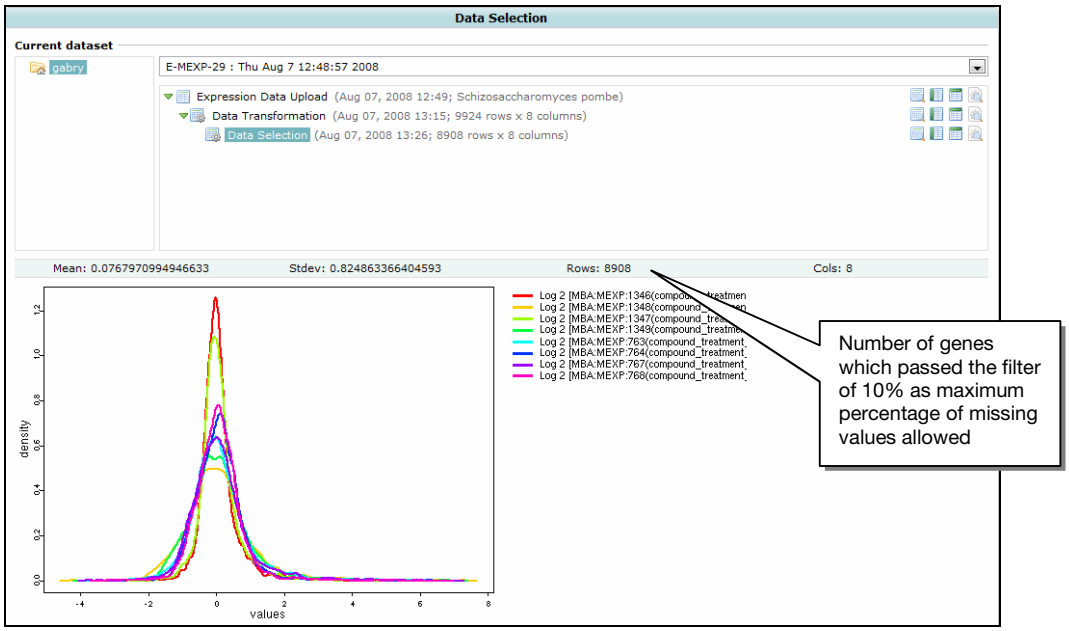


Fig. 7: Log-ratio intensity distribution plot for E-MEXP-29 after filtering for a maximum of 10% missing values in the dataset

We will now use the 'Value ranges' option to select for the top 200 genes with N greatest **standard deviation**s. These are likely to be the genes with the most interesting expression patterns.

Select the 'Value ranges' tab in the Subselection menu, type 200 in the bottom text box and click 'Execute' (Fig. 8).

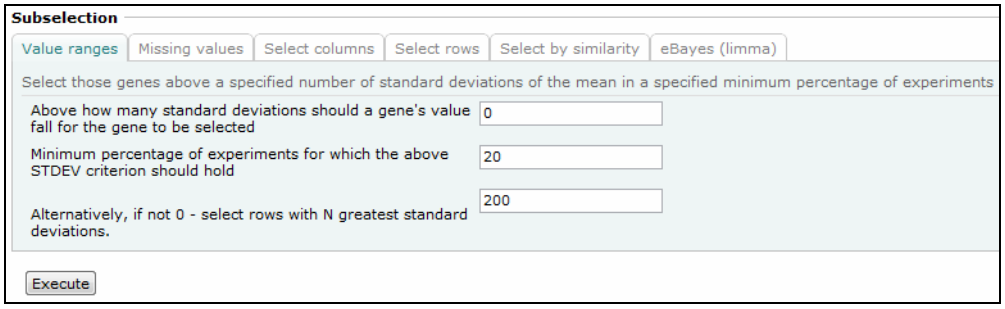


Fig. 8: 'Data selection' menu in EP – Value ranges option is selected. In this case, we want to retrieve the top 200 genes with the N greatest standard deviations. Alternatively, we could select for genes above a specified number of standard deviations of the mean in a minimum percentage of conditions

The result of the 'Value ranges' search will be displayed in a new window as a 'Dataset **heatmap**'. Observe the genes **distribution plot** by going back to the EP data selection view (Fig. 9).

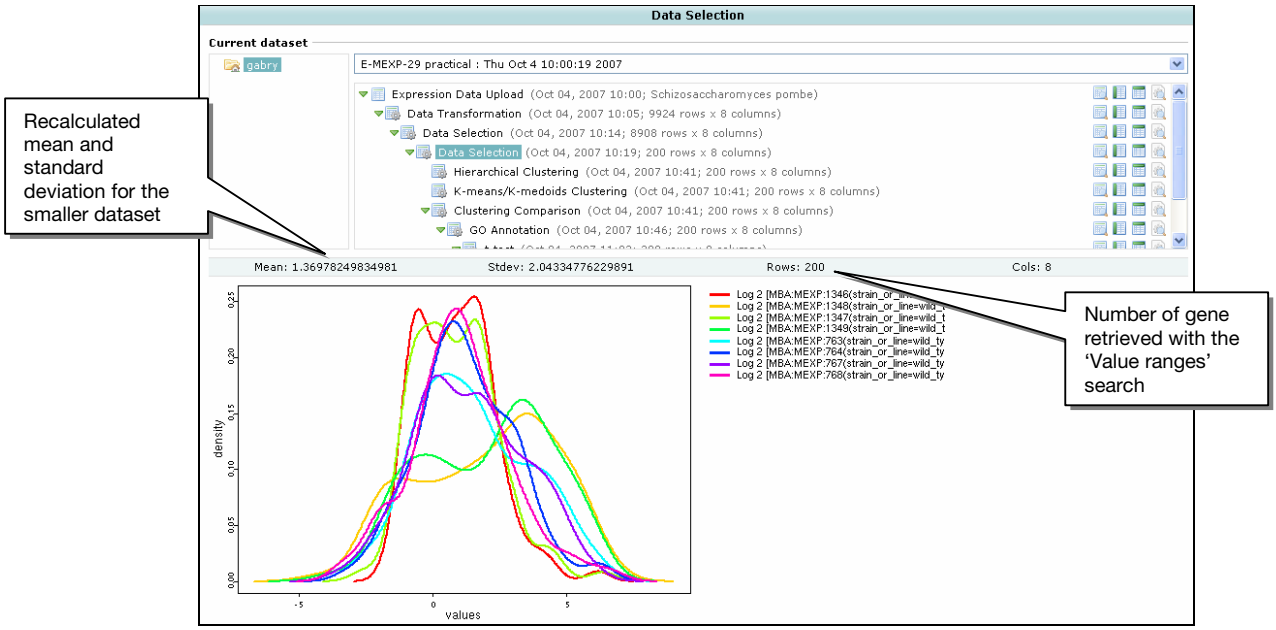


Fig. 9: Log-ratio intensity distribution plot for E-MEXP-29 after using the 'Value ranges' option to select the top 200 genes with the N greatest standard deviations

We can now use **clustering** to visualize patterns of gene expression among the 200 selected genes.

5 Clustering analysis in Expression Profiler

Clustering is an extremely popular analytical approach for identifying and visualizing patterns of gene expression in microarray datasets.

EP provides fast implementations of two **clustering** methods: **hierarchical clustering** and flat **partitioning**, as well as a novel approach for comparing the results of such **clustering** algorithms in the 'Clustering Comparison' component. The 'Signature Algorithm' component is an alternative approach to **clustering**-like analysis, based on the method by Ihmels et al. [1].

All **clustering** methods aim at grouping objects, such as genes, together, according to some measure of similarity, so that objects within one group or cluster are more similar to each other than to objects in other groups. **Clustering** analysis involves one essential elementary concept: the definition of similarity between objects, also known as **distance measure**. EP implements a wide variety of **distance measure**s for **clustering** analysis (all **distance measure**s can be found in the 'Distance measure' drop down menu – Fig. 10). The **Euclidean distance** and the **Correlation-based distance** represent the 2 most commonly applied measures of similarity. It is recommended combining different **clustering** methods and **distance measure**s to find the optimal combination for each dataset.

Go to the 'Clustering' component and click on 'Hierarchical' (Fig. 4.1).

Hierarchical clustering is an agglomerative approach in which single expression profiles are joined to form groups, which are further joined until the process has been completed, forming a single hierarchical tree. See Fig. 10 for all **Hierarchical clustering** options available.

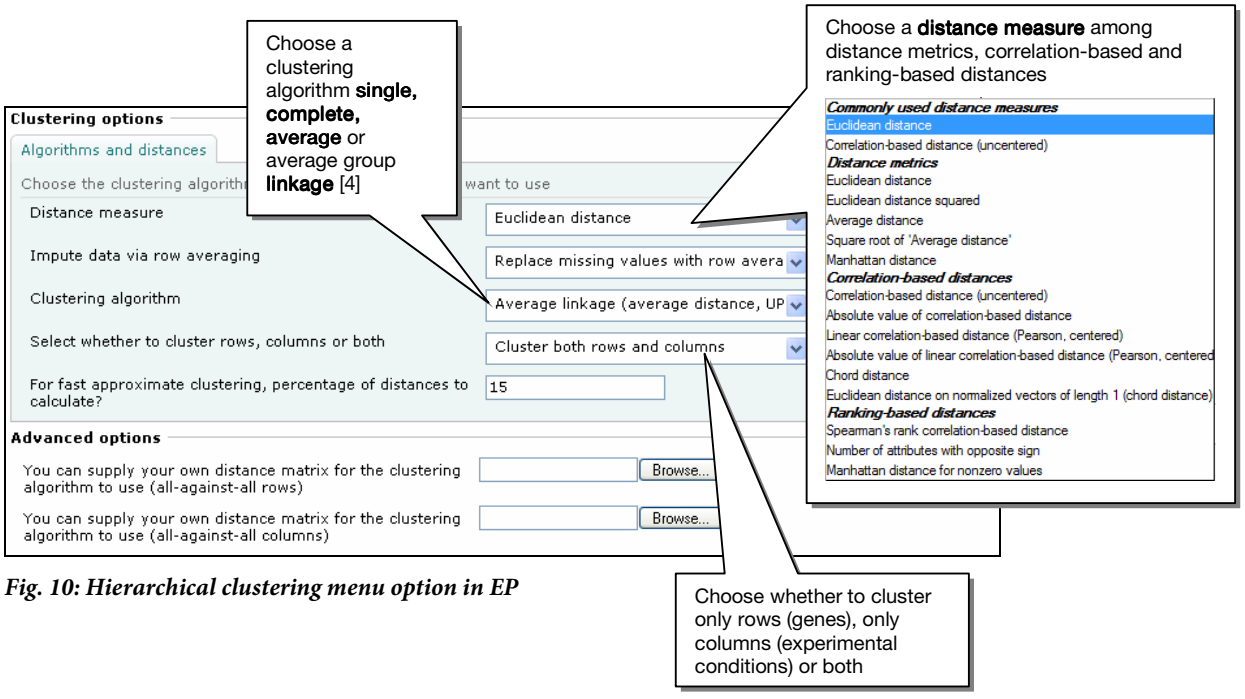


Fig. 10: Hierarchical clustering menu option in EP

The '**Hierarchical clustering**' output provides a visual display of the generated hierarchy in the form of a **dendrogram** or tree, attached to a **heatmap** representation of the clustered matrix (Fig. 13, left side).

Now click on '**K-means-K-medoids**' option. This component provides 2 flat **partitioning** methods, similar in their design. Both K-groups approaches are based on the idea that, for a

specified number K, K initial objects are chosen as cluster centers, the remaining objects in the dataset are iteratively reshuffled around these centers and new centers are chosen to maximize the similarity within each cluster, at the same time maximizing the dissimilarity between clusters. The **clustering** options available for 'K-means/K-medoids' are shown in Fig 11.

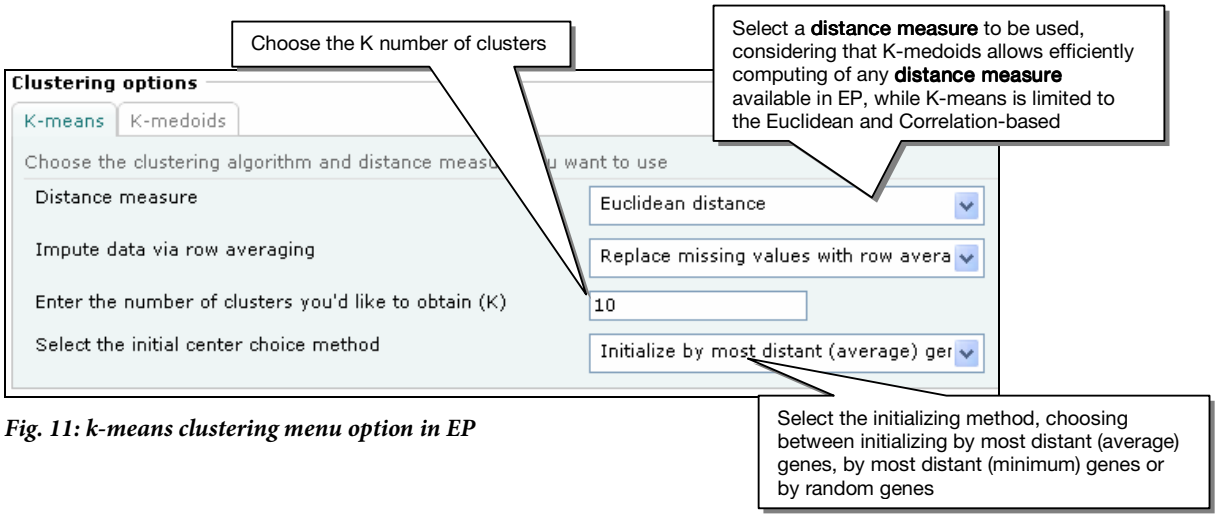


Fig. 11: k-means clustering menu option in EP

In the 'K-means/K-medoids' output, each cluster is visualized by a **heatmap** and a multi-gene lineplot (Fig. 13, right side). In addition, the list of genes present in each cluster is provided (Fig. 14).

A commonly encountered problem with **hierarchical clustering** is that it is difficult to identify branches within the hierarchy that form tight clusters. Similarly, in the case of flat **partitioning**, the determination of the K number of desired clusters is often arbitrary and unguided.

The 'Clustering Comparison' component provides an algorithm and a visual depiction of a mapping between a **dendrogram** and a set of flat clusters [5] or between a pair of flat **partitioning** clusters. The **clustering** comparison component not only provides an informative insight into the structure of the tree by highlighting the branches that best correspond to one or more flat clusters from the **partitioning**, but also can be useful when comparing the **hierarchical clustering** to a predefined functionally meaningful grouping of the genes.

We will now run the 'Clustering Comparison' algorithm, utilizing the list of 200 genes generated at the end of session 4 and compare how the **hierarchical clustering** and the k-means **clustering** perform on the same 200 gene list.

First, select this list in the analysis history menu. Then click on '**Clustering** Comparison', under the **Clustering** menu (Fig. 4.1), and fill the 'Clustering comparison' parameter session as shown in Fig. 12. Once done click 'Execute'.

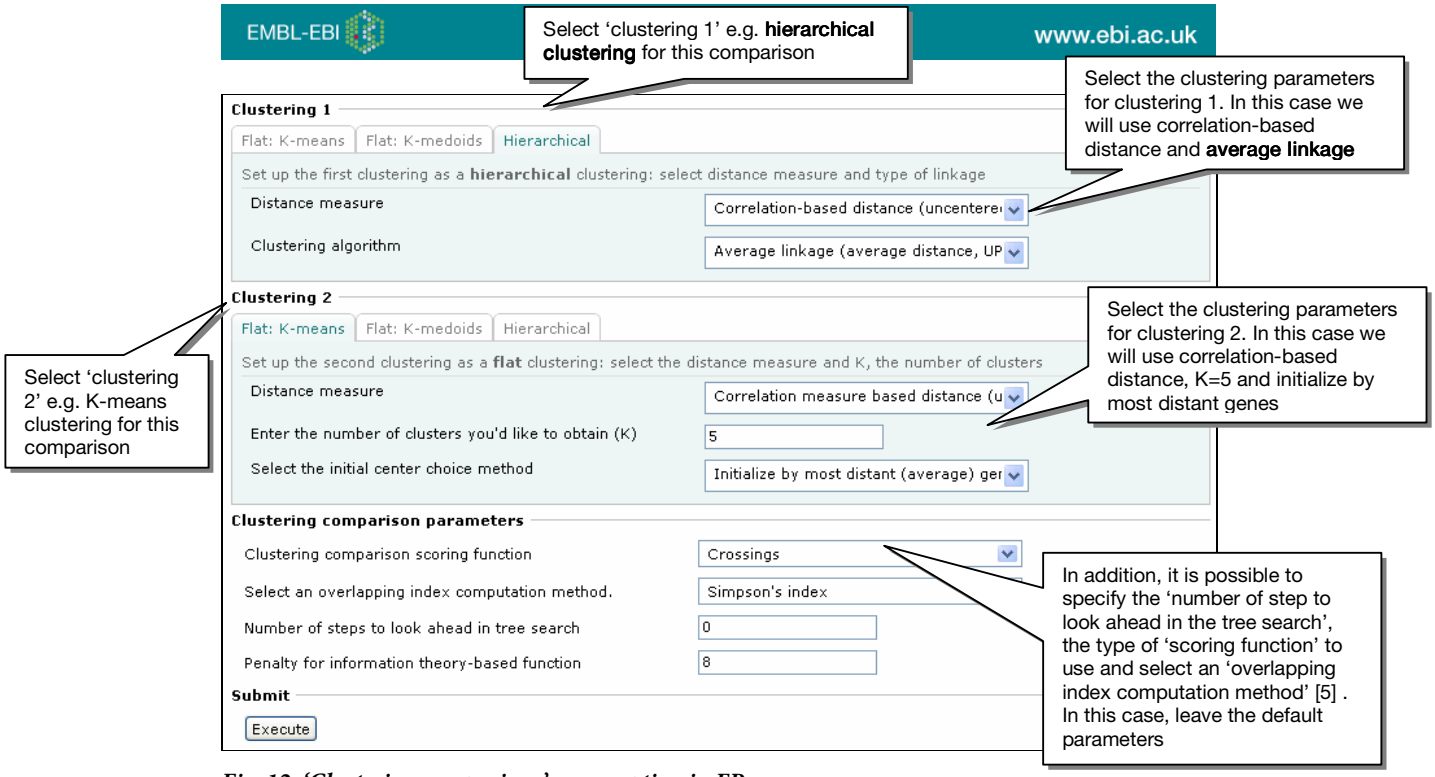


Fig. 12: 'Clustering comparison' menu option in EP

The new window shows a graphical visualization of the **clustering** comparison (Fig. 13).

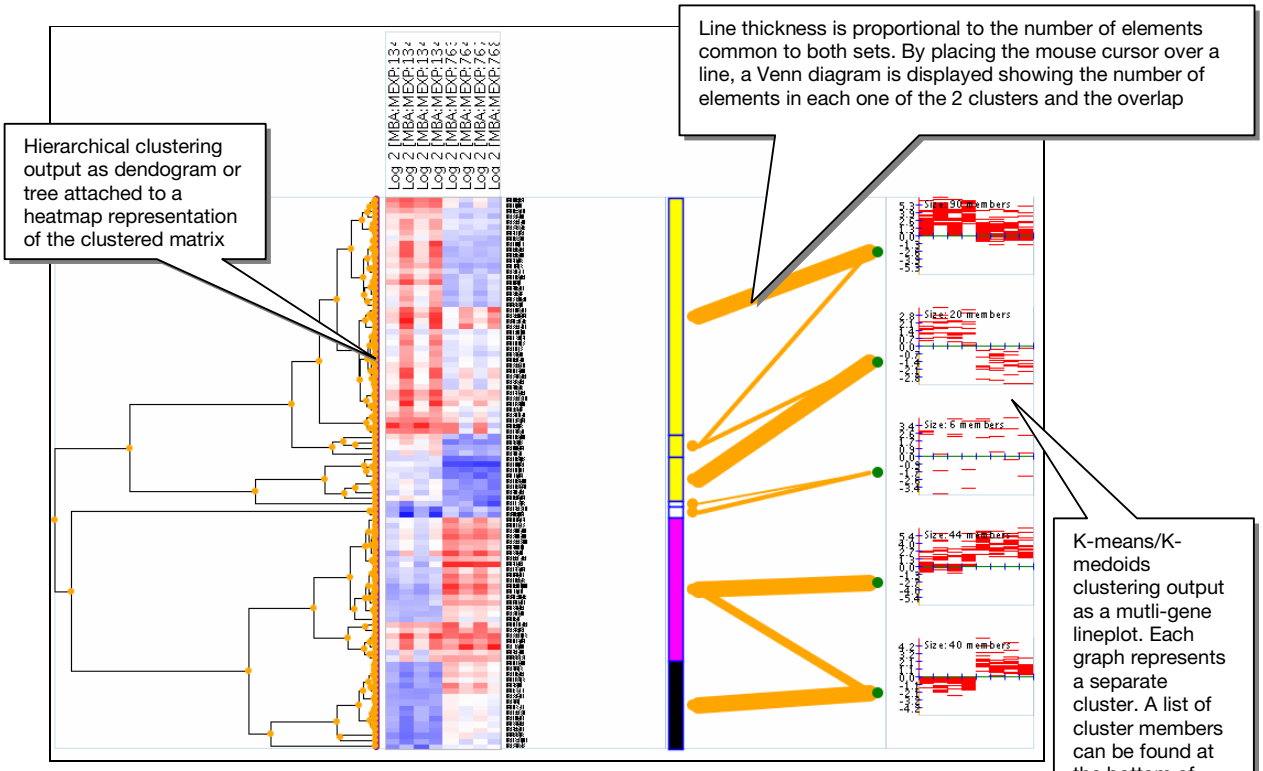


Fig. 13: Clustering comparison graphical output. The comparison was run between hierarchical clustering (correlation-based distance, average linkage) and k-means clustering (correlation-based distance, k=5.) Line thickness is proportional to the number of elements common to both sets.

On the left panel are shown the **dendrogram** and the **heatmap** generated by the **hierarchical clustering** algorithm. On the right panel are shown the 5 clusters (as multi gene lineplots) generated by the K-means algorithm. The **clustering** comparison correspondence is displayed in the central part of the graph. It is depicted as lines of varying thickness, mapping sub-branches of the tree to flat **clustering** superclusters. The thicker is the line, the bigger is the overlap between a sub-branch of the tree and a k-cluster.

The comparison algorithm can be re-run changing the **clustering** parameters (e.g. optimizing the number of K clusters) and observing changes in the output. Then, individual clusters can be selected for further analysis such as **Gene Ontology** term enrichment.

Go to the analysis history drop down menu and click on the 'View action output' icon corresponding to the 'K-means/K-medoids **clustering**' (as shown in Fig. 14).

This will allow you to retrieve the output of the k-means **clustering**, including information regarding the members of each cluster.

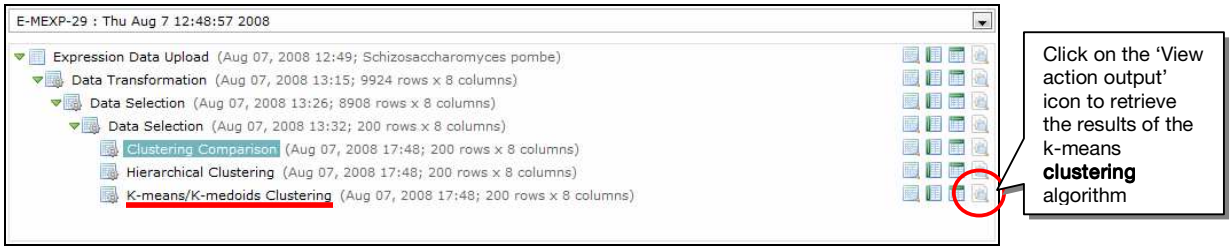


Fig. 14: How to retrieve the output of k-means clustering from the analysis history menu

Scroll down to the bottom of the new page and click on the header corresponding to the cluster of interest (e.g. cluster 4). This will open up a text box with a list of all gene members of the selected cluster (Fig. 15). Highlight all gene identifiers in the cluster and right click on the selection to copy the gene list to the clipboard.

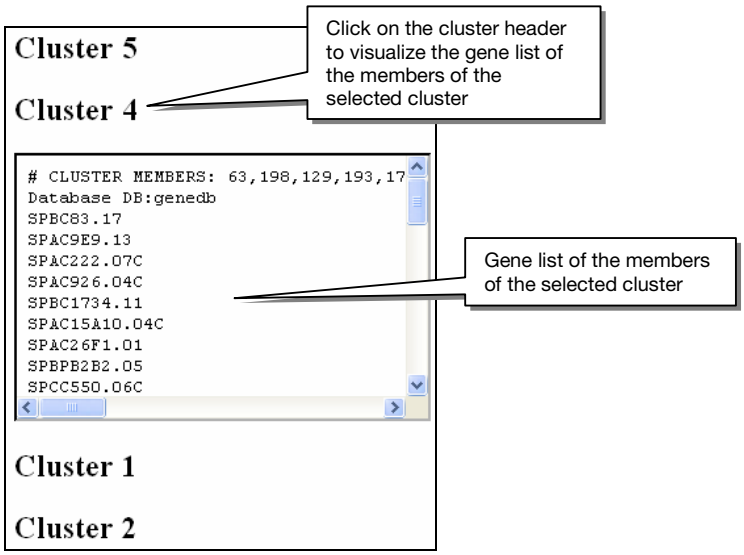


Fig. 15: List of all gene included in cluster 4, following K-means clustering

6 Gene Ontology annotation in Expression Profiler

Gene Ontology (GO) is a controlled vocabulary used to describe the biology of a gene product in any organism [13]. There are 3 independent sets of vocabularies, or ontologies, which describe the *molecular function* of a gene product, the *biological process* in which the gene product participates and the *cellular component* where the gene product can be found.

Once a subset of genes of interest has been identified, through one or several of the approaches described so far, the user can look for GO terms enriched in the given gene list (e.g. a particular cluster obtained with flat **partitioning**). Identification of over-represented GO terms among a given list of genes can help understanding the functional relevance of these genes in the biological process being studied.

We will now look for GO terms enriched in cluster 4. Go to the 'Annotation' component and click on the 'Gene Ontology' link (Fig. 4.1). Paste the list of cluster 4 members in the Gene IDs text box and fill the remaining fields as shown in Fig. 16. Then click 'Execute'.

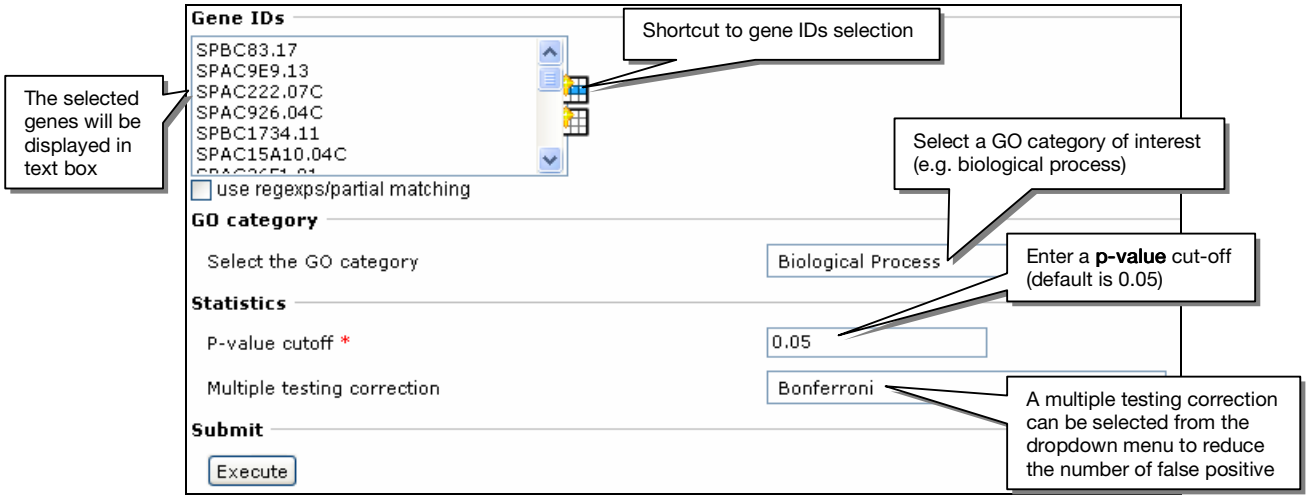
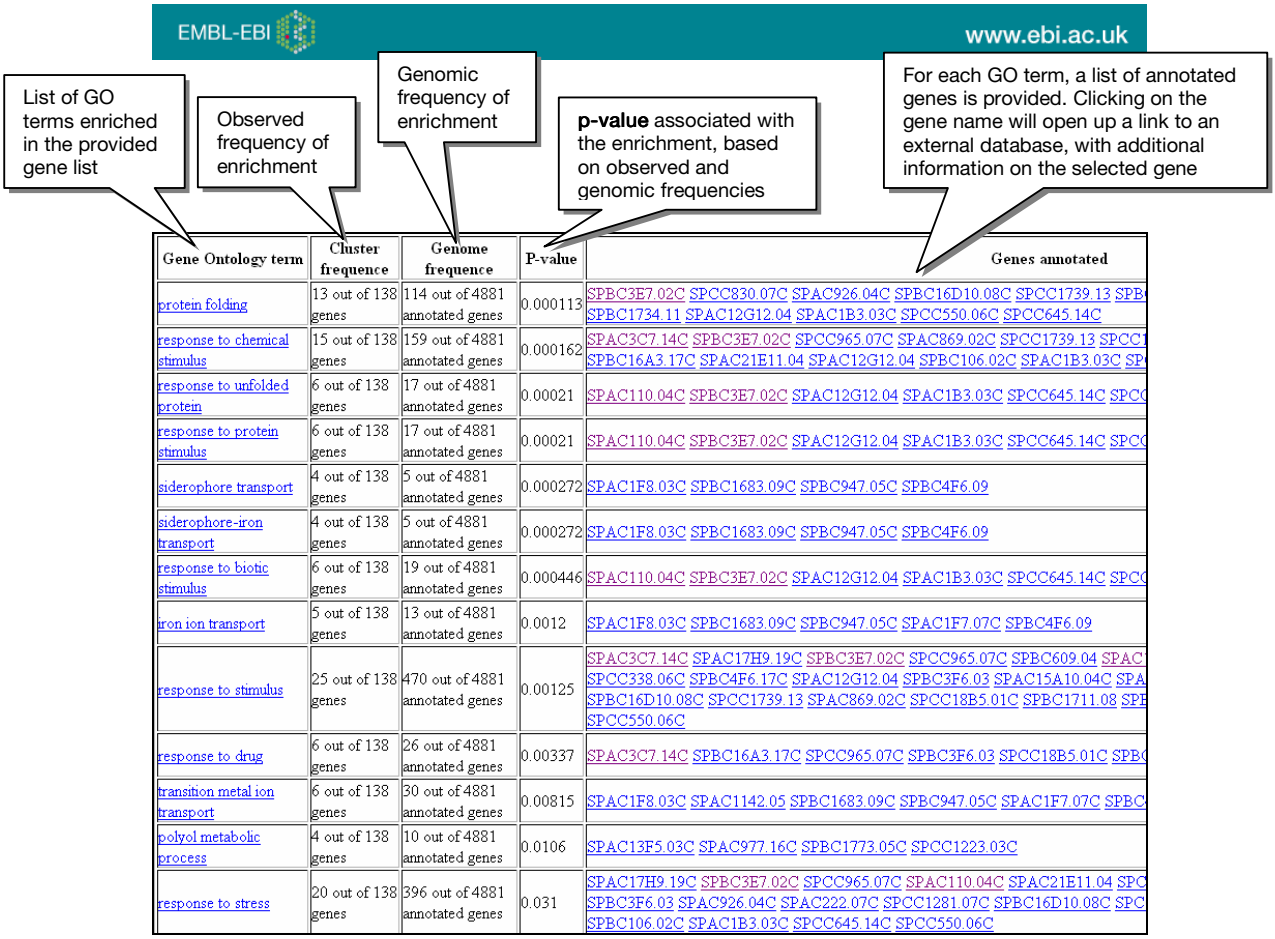


Fig. 16: 'Gene Ontology' menu option in EP

Results will be displayed as a tree view of GO terms and genes associated with each term.

In addition, a table will summarize the results (Fig. 17). Enriched GO terms are ranked based on ***p-values***, in ascending order. The ***p-value*** of enrichment for each GO term is based on its frequency in the gene list provided and its frequency at the genomic level.

In this case, it is not surprising to observe an enrichment of GO terms such as 'response to chemical stimulus, 'response to stress' as well as 'siderophore transport' and 'iron ion transport'. All these categories include genes encoding for critical components of the oxidative stress response, which is the biological phenomenon being investigated in this study [9].



Gene Ontology term	Cluster frequence	Genome frequence	P-value	Genes annotated
protein folding	13 out of 138 genes	114 out of 4881 annotated genes	0.000113	SPBC3E7.02C SPCC830.07C SPAC926.04C SPBC16D10.08C SPCC1739.13 SPB SPBC1734.11 SPAC12G12.04 SPAC1B3.03C SPCC550.06C SPCC645.14C
response to chemical stimulus	15 out of 138 genes	159 out of 4881 annotated genes	0.000162	SPAC3C7.14C SPBC3E7.02C SPCC965.07C SPAC869.02C SPCC1739.13 SPCC SPBC16A3.17C SPAC21E11.04 SPAC12G12.04 SPBC106.02C SPAC1B3.03C SP
response to unfolded protein	6 out of 138 genes	17 out of 4881 annotated genes	0.00021	SPAC110.04C SPBC3E7.02C SPAC12G12.04 SPAC1B3.03C SPCC645.14C SPC
response to protein stimulus	6 out of 138 genes	17 out of 4881 annotated genes	0.00021	SPAC110.04C SPBC3E7.02C SPAC12G12.04 SPAC1B3.03C SPCC645.14C SPC
siderophore transport	4 out of 138 genes	5 out of 4881 annotated genes	0.000272	SPAC1F8.03C SPBC1683.09C SPBC947.05C SPBC4F6.09
siderophore-iron transport	4 out of 138 genes	5 out of 4881 annotated genes	0.000272	SPAC1F8.03C SPBC1683.09C SPBC947.05C SPBC4F6.09
response to biotic stimulus	6 out of 138 genes	19 out of 4881 annotated genes	0.000446	SPAC110.04C SPBC3E7.02C SPAC12G12.04 SPAC1B3.03C SPCC645.14C SPC
iron ion transport	5 out of 138 genes	13 out of 4881 annotated genes	0.0012	SPAC1F8.03C SPBC1683.09C SPBC947.05C SPAC1F7.07C SPBC4F6.09
response to stimulus	25 out of 138 genes	470 out of 4881 annotated genes	0.00125	SPAC3C7.14C SPAC17H9.19C SPBC3E7.02C SPCC965.07C SPBC609.04 SPAC SPCC338.06C SPBC4F6.17C SPAC12G12.04 SPBC3F6.03 SPAC15A10.04C SPA SPBC16D10.08C SPCC1739.13 SPAC869.02C SPCC18B5.01C SPBC1711.08 SPB SPCC550.06C
response to drug	6 out of 138 genes	26 out of 4881 annotated genes	0.00337	SPAC3C7.14C SPBC16A3.17C SPCC965.07C SPBC3F6.03 SPCC18B5.01C SPBC
transition metal ion transport	6 out of 138 genes	30 out of 4881 annotated genes	0.00815	SPAC1F8.03C SPAC1142.05 SPBC1683.09C SPBC947.05C SPAC1F7.07C SPBC
polyol metabolic process	4 out of 138 genes	10 out of 4881 annotated genes	0.0106	SPAC13F5.03C SPAC977.16C SPBC1773.05C SPCC1223.03C
response to stress	20 out of 138 genes	396 out of 4881 annotated genes	0.031	SPAC17H9.19C SPBC3E7.02C SPCC965.07C SPAC110.04C SPAC21E11.04 SPC SPBC3F6.03 SPAC926.04C SPAC222.07C SPCC1281.07C SPBC16D10.08C SPC SPBC106.02C SPAC1B3.03C SPCC645.14C SPCC550.06C

Fig. 17: GO term enrichment analysis table output

7 Identification of differentially expressed genes using t-test analysis

In the EP main menu, click on 't-test analysis', under the 'Statistics' menu (Fig. 4.1).

The **t-test** component provides a way to apply this basic statistics test for comparing the means from 2 distributions in the following **differentially expressed** gene identification situations: looking for genes expressed significantly above background/control, or looking for genes expressed differentially between 2 sets of conditions. In the first case ('one class' option), the user specifies either the background level to compare against, or selects the genes in the dataset that are to be used as controls. In the second case ('two classes in one dataset' option), the user specifies which columns in the dataset represent the first group of conditions and which represent the second group. The user will now try an example of the latter case.

Click on 'Two classes in one dataset' tab, type in 5-8 for Class 1 (hydrogen peroxide treated samples) and 1-4 for Class 2 (untreated samples) and click 'Execute' (Fig. 18).

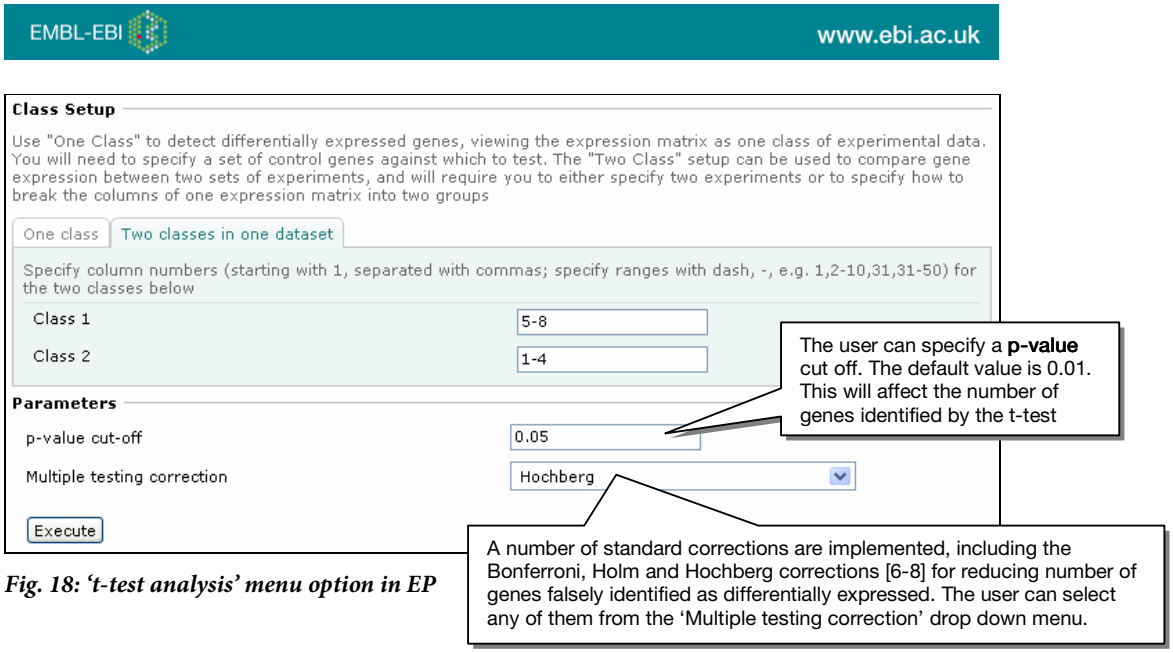


Fig. 18: 't-test analysis' menu option in EP

Upon execution, the **t-test** involves, for each gene, the calculation of the mean in both groups being tested (when testing against controls, the mean over all control genes is taken as the second group mean), and comparing the difference between the two means to a theoretical t-statistic [14]. A table of gene names, ***p-values*** and **confidence intervals** is output (Fig. 19), as well as a plot of the top 15 genes found (Fig. 20).

Depending on the number of samples in each group (which corresponds to the number of biological replicates), the test's reliability is reflected in the **confidence intervals** of the ***p-values*** that are produced (in this case, the likelihood that the two means are significantly different, i.e., that the gene is **differentially expressed**).

Table of all differentially expressed genes (above the 0.05 **p-value** cut off) identified by the t-test, between wild type untreated and hydrogen peroxide treated samples.

Gene	(adjusted) p-value ↑	mean/difference of means	confidence interval
SP[illegible]C1F8.03C	8.44e-05	-4.09	(-4.48, -3.7)
[illegible]AC1F8.03C.1	8.44e-05	-4.09	(-4.48, -3.7)
[illegible]PAC222.07C	0.00136	2.38	(1.98, 2.77)
SPAC222.07C.1	0.00136	2.38	(1.98, 2.77)
SPCC338.06C	0.00718	2.5	(2.04, 2.95)
SPCC338.06C.1	0.00718	2.5	(2.04, 2.95)
SPBC19C7.04C	0.011	3.59	(2.85, 4.34)
SPBC19C7.04C.1	0.011	3.59	(2.85, 4.34)
SPBC16A3.17C	0.011	-2.83	(-3.23, -2.43)
SPBC16A3.17C.1	0.011	-2.83	(-3.23, -2.43)
SPAC1039.02	0.0129	-2.39	(-2.87, -1.9)
SPAC1039.02.1	0.0129	-2.39	(-2.87, -1.9)
SPAC1F7.07C	0.0162	-2.58	(-3.13, -2.03)
SPAC1F7.07C.1	0.0162	-2.58	(-3.13, -2.03)

Fig. 19: 't-test analysis' table output

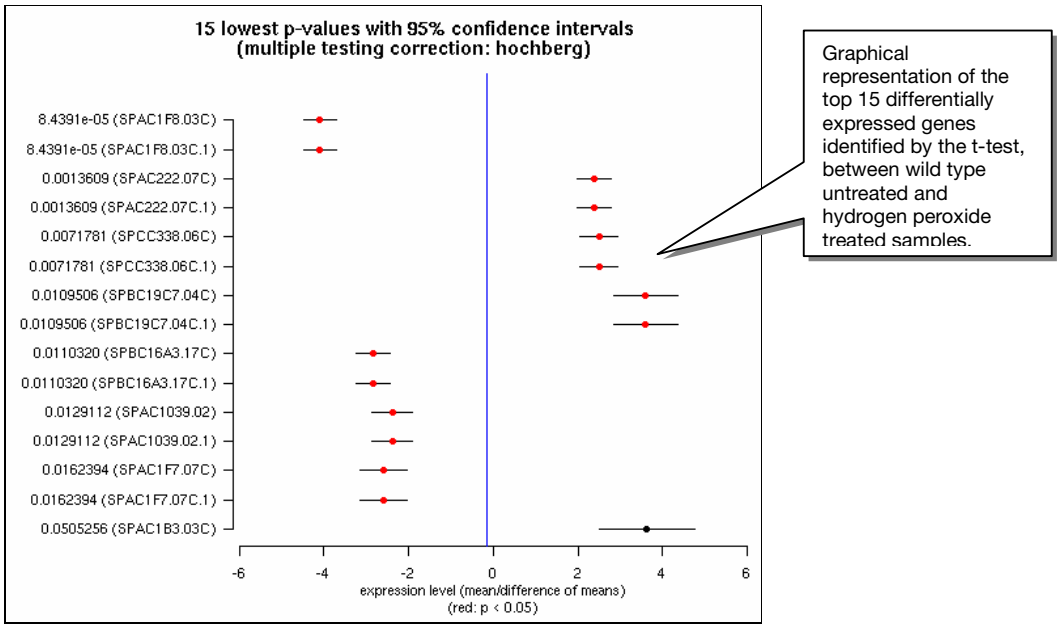


Fig. 20: 't-test analysis' graphical output – plot showing the top 15 differentially expressed genes. For each gene the expression level (mean/difference of means) is plotted with 95% confidence intervals. The expression level is shown as red dot when the corresponding p-value is below the chosen cut off (in this case 0.05).

Go to 'Data selection', under the 'Subselection' menu and click on 'select rows' (Fig. 21). Cut and paste (with the help of a text editor) the list of top 15 **differentially expressed** genes into the text box and click 'Execute'. By doing this, you will retrieve the expression profiles for the selected genes which will be shown as **heatmap** on a new window (Fig. 21). This provides a visual confirmation of the **t-test** results. The expression profiles show differential expression of the selected genes in the 2 conditions (Fig. 21).

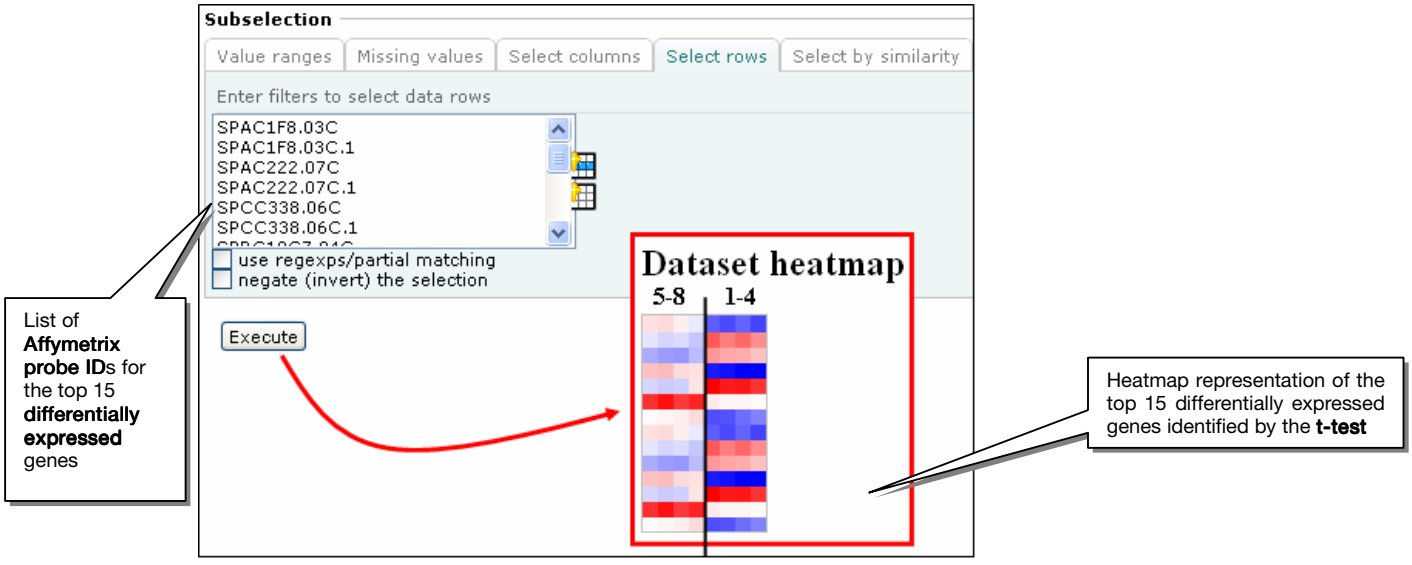


Fig. 21 The 'select rows' function in the Subselection menu was used to retrieve the heatmap representation of the top 15 differentially expressed genes identified by the t-test. The 15 genes

show a different behaviour in the 2 conditions studied: hydrogen peroxide treated samples (5-8) and untreated samples (1-4). Red indicates induced genes, blue repressed genes.

The list of **differentially expressed** genes identified with the **t-test** can now be further analyzed using **clustering** and GO term enrichment analysis as previously shown.

8 How to obtain a data matrix for experiment E-MEXP-29

1. Go to the AE main homepage, at http://www.ebi.ac.uk/arrayexpress/
2. In the 'Experiments' box, on the left-hand side of the page, type in the accession number E-MEXP-29 and click 'Query'
3. Expand the experiment view by clicking on the plus sign next to E-MEXP-29
4. Click on the 'View detailed data retrieval page' link
5. Go to the 'Processed Data Group 1'
6. In the 'Experimental conditions' table select the following 8 conditions: MEXP-763, 764, 767, 768 (wild type untreated) and MEXP-1346, 1347, 1348, 1349 (wild type treated with hydrogen peroxide)
7. Scroll down to the 'Quantitation type' and the 'Array Annotation' tables
8. Select 'Quantitation type: normalized' from the 'Quantitation type' table and select 'Database DB:genedb' from the 'Array annotation' table
9. Click on 'Export data'
10. Once computed, save the **data matrix** onto your computer as .csv file

Glossary

Affymetrix arrays

Affymetrix is a leading manufacturer of oligonucleotide arrays (http://www.Affymetrix.com/). Affymetrix expression arrays use a set of features (often referred to as “spots”) designed to recognize each molecule of interest. Each feature consists of millions of identical single-stranded 25-mer nucleotide probes, each designed to hybridize to a specific transcript. On a gene-level array, each of these Perfect Match (PM) features is accompanied by an adjacent Mis-Match (MM) feature in which the middle residue is changed. Hybridization conditions are designed to maximise binding to the PM features while minimizing binding to the MM ones. Each MM feature can therefore be used to provide a measure of probe specific background for its PM partner. Multiple PM/MM pairs are used for each transcript. On most gene-level arrays, 11 PM/MM pairs are used per transcript, and the complete set of 22 features is referred to as a probeset.

Affymetrix .CEL data files

The Cell Intensity (.CEL) file contains fluorescence intensities for each cell (feature) on the microarray. A single intensity value is stored per cell.

Atf1

Transcription factor involved in *S. pombe* stress response mechanisms. More information for this gene can be found at http://www.genedb.org/genedb/Dispatcher?formType=navBar&organism=pombe&name=atf1&desc=yes&submit=Search.

Average Linkage

Average Linkage is a type of **hierarchical clustering** in which the distance between one cluster and another cluster is considered to be equal to the average distance from any member of one cluster to any member of the other cluster.

Clustering

Clustering aims at grouping objects, such as genes, together, according to some measure of similarity, so that objects within one group or cluster are more similar to each other than to objects in other groups. It is a mean to visualize patterns of gene expression in the data.

Correlation-based distance or Pearson correlation

Pearson correlation measures the similarity in shape between 2 profiles. It will return a score of how correlated the expression profiles of the 2 genes are, where 1= similar, 0 = no similarity, -1 = dissimilar.

Complete Linkage

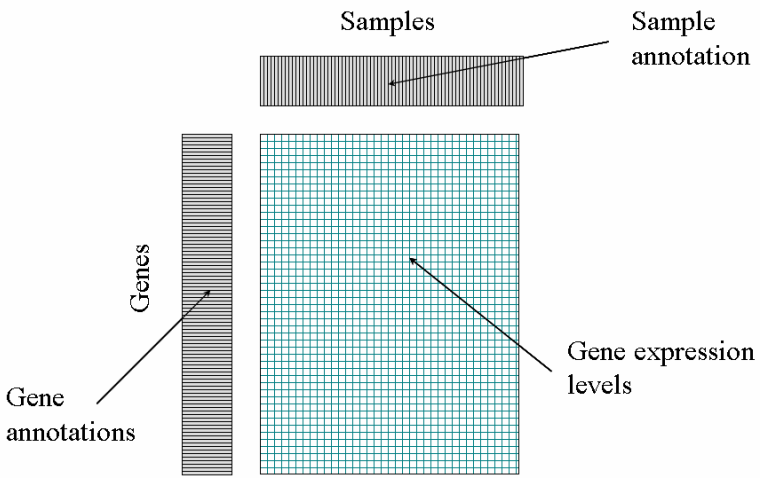
Complete linkage is a type of **hierarchical clustering** in which the distance between one cluster and another cluster is considered to be equal to the longest distance from any member of one cluster to any member of the other cluster.

Confidence intervals

A **confidence interval** for the difference between two means specifies a range of values within which the difference between the means of the two populations lies.

Data matrix

In a gene expression **data matrix**, each row represents a gene and each column represents an experimental sample or array. An entry in the **data matrix** usually represents the expression level or expression ratio of a gene in a given sample or array. In addition to numerical values, the matrix can also contain additional columns for gene annotation or additional rows for **sample** annotation.



Dendrogram

A **dendrogram** is a tree diagram used to illustrate the arrangement of the clusters produced by a **clustering** algorithm.

Density histogram

A graphical representation of a single dataset, tallied into classes. The graph consists of a series of rectangles whose widths are defined by the limits of the classes, and whose heights are calculated by dividing relative frequency by class width. Resulting rectangle heights are called densities; the vertical scale is called density scale.

Distance measure

Aims to quantify to what degree two expression profiles are similar. Such measure of similarity is called distance; the more distant two expression profiles are, in the multidimensional space, the more dissimilar they are. Distance can be measured in many different ways.

Differentially expressed

A gene is **differentially expressed** when its expression values under two or more conditions are statistically significantly different.

Distribution plot

Chart used to graphically characterize the distribution of measurements.

Euclidean distance

Euclidean distance measures the shortest distance between 2 points.

Filtering

Filtering is a process of data transformation where any observations that do not fulfil a pre-formulated condition are excluded from the data.

Gene ontology (GO)

GO is a controlled vocabulary used to describe the biology of a gene product in any organism. There are 3 independent sets of vocabularies, or ontologies, that describe: the molecular function of a gene product, the biological process in which the gene product participates and the cellular component where the gene product can be found (http://www.geneontology.org).

Heatmap

A **heatmap** is a graphical representation of data where the values taken by a variable in a two-dimensional map are represented as colours. Heat maps are typically used in microarray data analysis to represent gene expression levels across several conditions.

Hierarchical Clustering

Hierarchical clustering is a type of **clustering** which can be agglomerative ("bottom-up") or divisive ("top-down"). Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters. It merges or splits until a required degree of similarity holds for the elements of the clusters.

Normalisation

Normalisation is a fundamental pre-processing step in microarray data analysis. It aims to compensate for systematic technical differences between arrays, to see more clearly the systematic biological differences between samples.

One- and two-channel experiments

Two-channel or two-colour hybridisation experiments aim to compare the relative transcript abundance in two mRNA or DNA samples (for example a 'test' cell state and a 'reference' cell state) which are labelled using two different fluorescent dyes (say, a red dye for the test and a green dye for the reference), mixed and then hybridized to the arrayed DNA spots.

Affymetrix arrays use instead a one-channel or single-color labeling strategy where experimental mRNA is enzymatically amplified, biotin-labeled for detection, hybridized to the array, and detected through the binding of a fluorescent compound [15].

Partitioning clustering

Partitioning **clustering** is a type of **clustering** which attempts to cluster a set of genes directly, in a manner that depends on predefined parameters. These parameters are then adjusted to optimally satisfy a chosen criterion of separation and compactness of clusters. K-means **clustering** is a type of **partitioning clustering** in which the number of clusters, K, needs to be provided.

Perfect Match (PM)

A probe that is an exact complementary to the transcript of interest. See the glossary term **Affymetrix arrays** for more details.

Pre-processing

Data pre-processing includes data normalization, transformation and **filtering**. It aims to prepare the data for the following analysis steps.

Principal Component Analysis (PCA)

Principal component analysis is a data transformation process for simplifying a dataset, by reducing multidimensional dataset to lower dimensions for analysis.

Probe intensity

It is the florescent intensity value that is detected by the scanner for each probe on the array.

p-value

The **p-value** measures the probability that a difference between two experimental conditions happened by chance. The lower the **p-value**, the more likely it is that the difference between the two conditions is a true reflection of the biological process being studied either than a random phenomenon.

Single Linkage

Single linkage is a type of **hierarchical clustering** in which the distance between one cluster and another is considered to be equal to the shortest distance from any member of one cluster to any member of the other cluster.

Standard deviation

The **standard deviation** measures the spread of the data about the mean value.

Sty1

S. pombe MAP Kinase involved in stress activated signal transduction pathways. More information for this gene can be found at http://www.genedb.org/genedb/Dispatcher?formType=navBar&organism=pombe&desc=yes&wildcard=yes&name=sty1&ohmr=.&ohmr2=.

t-test

The **t-test** is a common statistical test that is used to find out if there is a significant difference between the means (averages) of two different groups.

Transformation

Data transformation is the conversion of data from one format to another.

Further reading

1. Ihmels, J., et al., *Revealing modular organization in the yeast transcriptional network.* Nat Genet, 2002. **31**(4): p. 370-7.
2. Blake, J., et al., *ChroCoLoc: an application for calculating the probability of co-localization of microarray gene expression.* Bioinformatics, 2006. **22**(6): p. 765-767.
3. Ihaka, R. and R. Gentleman, *R: a language for data analysis and graphics.* J. Comput. Graph. Stat., 1996. **5**: p. 299-314.
4. Quackenbush, J., *Computational analysis of microarray data.* Nat Rev Genet, 2001. **2**(6): p. 418-27.
5. Torrente, A., M. Kapushesky, and A. Brazma, *A new algorithm for comparing and visualizing relationships between hierarchical and flat gene expression data clusterings.* Bioinformatics, 2005. **21**(21): p. 3993-9.
6. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society B, 1995. **57**(289-300).
7. Hochberg, Y., *A sharper Bonferroni procedure for multiple tests of significance.* Biometrika, 1988. **75**: p. 800-803.
8. Holm, S., *A Simple Sequentially Rejective Bonferroni Test Procedure.* Scandinavian Journal of Statistics, 1979. **6**: p. 65-70.
9. Chen, D., et al., *Global transcriptional responses of fission yeast to environmental stress.* Mol Biol Cell, 2003. **14**(1): p. 214-29.
10. Troyanskaya, O., et al., *Missing value estimation methods for DNA microarrays.* Bioinformatics, 2001. **17**(6): p. 520-5.
11. Smyth, G.K., *Linear models and empirical bayes methods for assessing differential expression in microarray experiments.* Stat Appl Genet Mol Biol, 2004. **3**(12): p. Article3.
12. Johansson, P. and J. Hakkinen, *Improving missing value imputation of microarray data by using spot quality weights.* BMC Bioinformatics, 2006. **7**(306): p. 306.
13. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-29.
14. Manly, K.F., D. Nettleton, and J.T. Hwang, *Genomics, prior probability, and statistical tests of multiple hypotheses.* Genome Res, 2004. **14**(6): p. 997-1001.
15. Chee, M., et al., *Accessing genetic information with high-density DNA arrays.* Science, 1996. **274**(5287): p. 610-4.

What to do next

Once you have read this tutorial, you might want to test your understanding by trying the related online quiz or reflective tasks. Please see the EBI moodle at www.ebi.ac.uk/training/ for these and other eLearning resources.