

## Expression Profiler for Beginners – part 2

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 raw data
- How to normalize and transform the data
- How to identify differentially expressed genes using t-test
- How to identify differentially expressed genes using ordination based techniques

### *Contents:*

- 1 How to get started
- 2 How to upload raw data
- 3 How to normalize and transform the data
- 4 How to identify differentially expressed genes using t-test
- 5 How to identify differentially expressed genes using ordination based techniques
- 6 How to obtain a raw data for experiment E-MEXP-886



# 1 The basics of Expression Profiler - How to get started

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

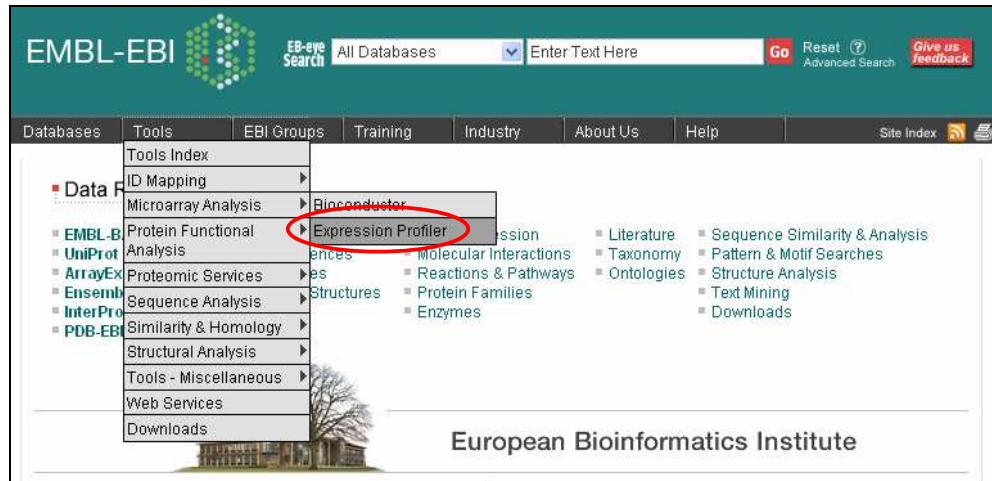


Fig. 1: Accessing Expression Profiler from the EBI homepage (<http://www.ebi.ac.uk/>)

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.

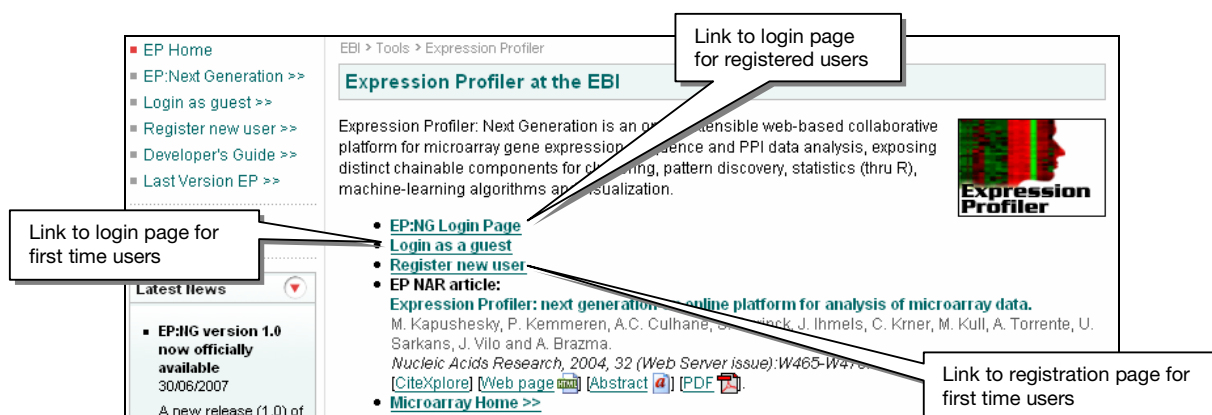


Fig. 2: Expression Profiler homepage (<http://www.ebi.ac.uk/expressionprofiler/>)

## 2 How to upload raw 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). The .CEL files can be uploaded by placing them into an archive (a .zip file, for instance) and then uploading the archive. The .zip file should contain only .CEL files from the same type of **Affymetrix arrays**.

For this tutorial, we will use the dataset E-MEXP-886 exported from the AE database in the additional exercise of the 'ArrayExpress for Beginners' Tutorial. All you need is the entire raw dataset, saved as .zip archive. At the end of this tutorial is a quick reminder on how to obtain this file.

In this study, transcription profiling of ataxin1-null vs. wild type mice was performed in order to investigate **spinocerebellar ataxia type 1**. Ten Affymetrix GeneChip Mouse Expression Arrays MOE430A were used, five hybridized with RNA extracted from ataxin1-null mice and 5 with RNA extracted from wild type mice [7].

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

The screenshot shows the 'Upload / Expression Data' form. It has three tabs: 'Tabular Data', 'Affymetrix', and 'ArrayExpress'. The 'Affymetrix' tab is selected. The form includes a text input for the file path (C:\E-MEXP-886.raw.zip) with a 'Browse...' button, a text input for the URL, a dropdown menu for 'Select data species' (set to 'Mus musculus'), a text input for the experiment name (E-MEXP-886), and an 'Execute' button. Callouts point to the 'Affymetrix' tab, the file path input, the species dropdown, and the experiment name input.

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

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

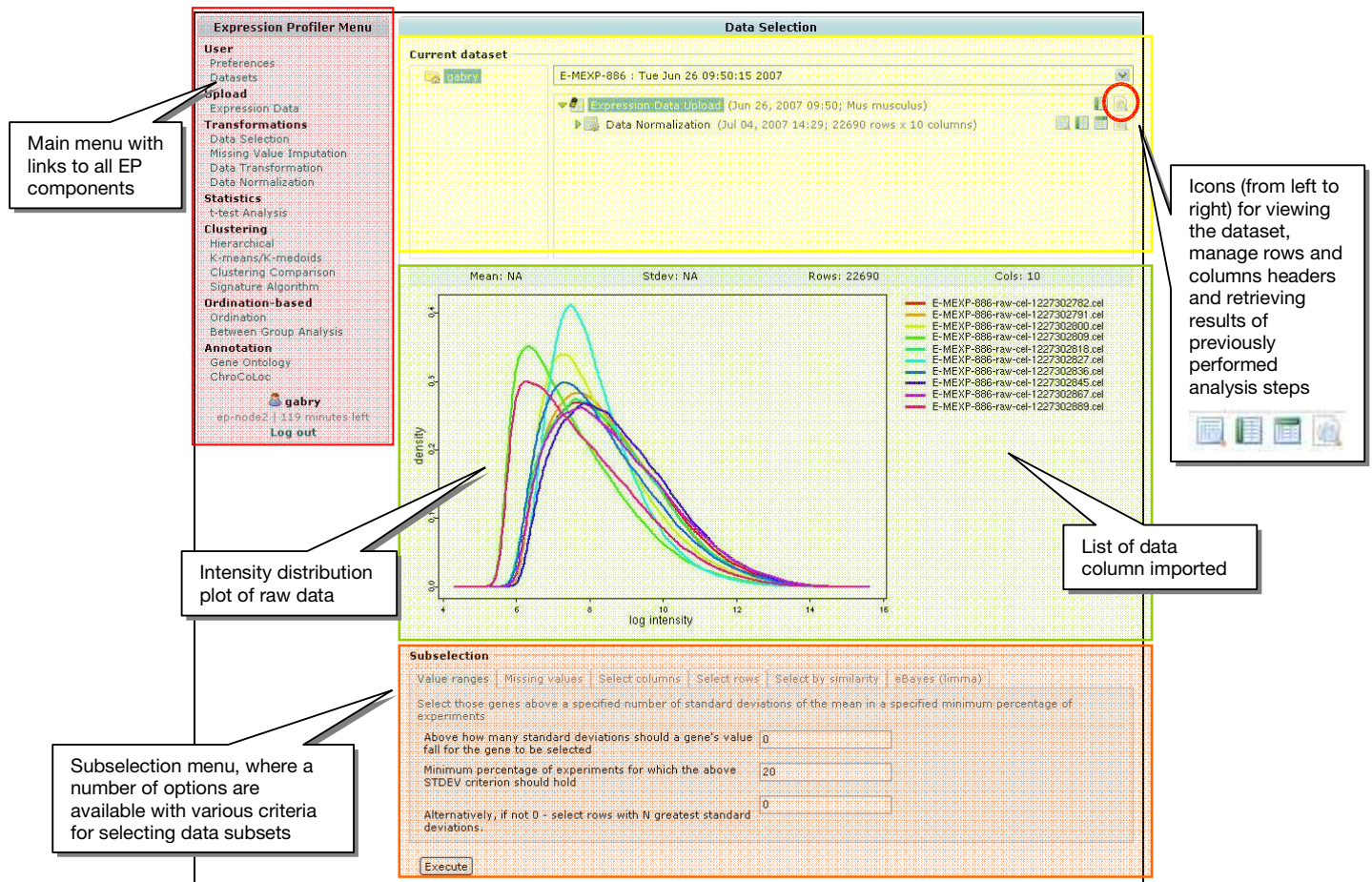
This view has three sections:

- '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 intensity distribution is shown in the 'Descriptive statistics' plot and the data is now available for further **pre-processing** and analysis such as **normalization**, **transformation**, **t-test** and **principal component analysis**.

Explore the intensity **distribution plot** of the raw data in the 'Descriptive statistics' graph. Each coloured line represents the **perfect match (PM) probe intensities** distribution for one array.



**Fig. 4:** EP Data selection view after uploading experiment E-MEXP-886. 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.

You can view a **box plot** distribution of the raw data by clicking on the 'View action output' icon to the right of the 'Expression data upload' link in the analysis history menu (Fig. 4, icon indicated by red circle). This will open a window with the thumbnails images of each .CEL file and the dataset **box plot** (Fig. 4.1). Notice the variability of the distribution of the 10 samples.

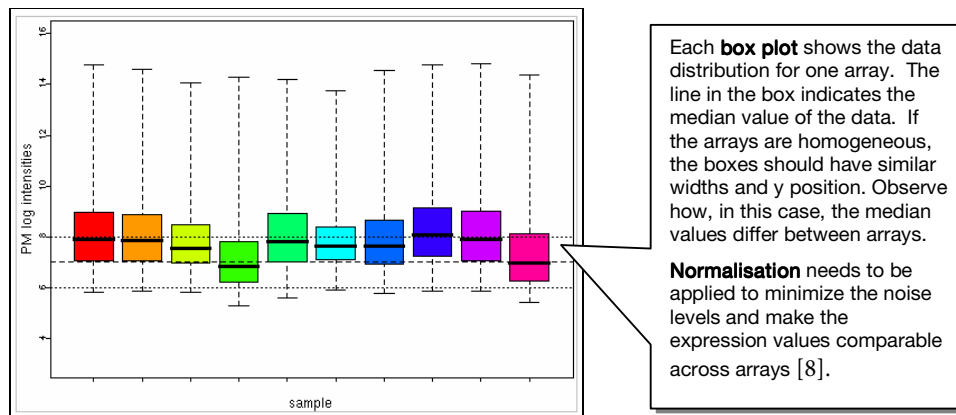


Fig. 4.1: Box plot distribution of E-MEXP-886 raw dataset.

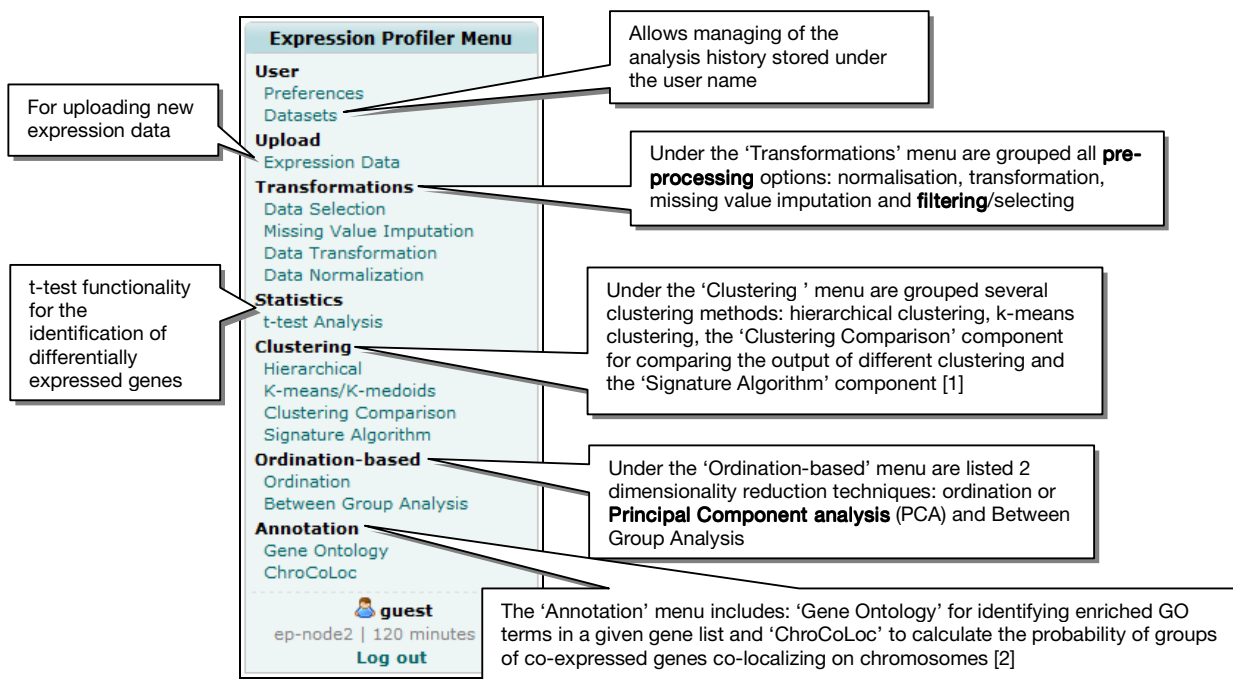


Fig. 4.2: EP main menu with links to all EP data analysis components.

### 3 How to normalize and transform the data

In the EP main menu, click on 'Data normalisation', under the 'Transformations' menu (Fig. 4.2).

When dealing with raw data, it is important to perform a **normalisation** step to minimize the noise levels and make the expression values comparable across arrays [8]. There are many approaches to normalizing expression levels. EP provides a graphical interface to four data **normalisation** routines: RMA, GCRMA, 'Li & Wong' and VSN [9-12]. Of the four methods, RMA, GCRMA and 'Li & Wong' can only be applied to Affymetrix .CEL file imports, while VSN can be applied to all types of data.

Select the RMA tab and click 'Execute'.

**Normalization Methods**

Several methods are available for data normalization. They are arranged (left-to-right) in order of increasing complexity, effectiveness, and slowness.

RMA GCRMA Li & Wong VSN

Robust Multi-Array Average expression measure

Execute

Fig. 5: Normalization menu in EP – the ‘RMA’ option is selected

In the output page, the normalized results are visualized as a dataset **heatmap** and a **box plot** (Fig. 6 – right panel). Observe the effect of **normalisation** on the data distribution by comparing this **box plot** with the one in Fig. 4.1. The 10 **box plots** now have nearly identical median values and homogenous distributions. Explore the changes in the log intensity **distribution plot** by going back to the EP Data Selection view (Fig. 6 – left panel).

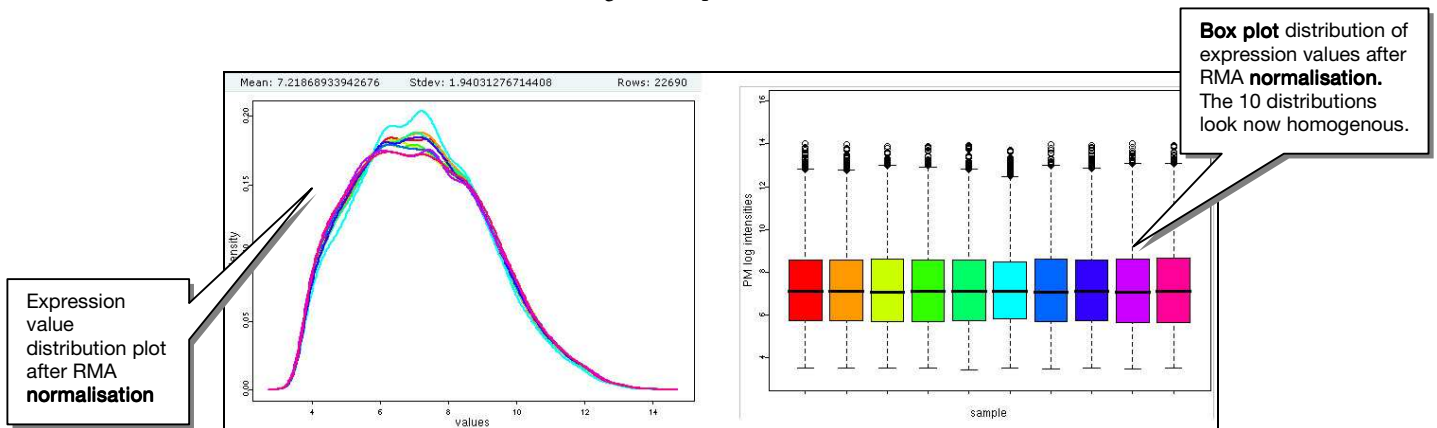


Fig. 6: RMA normalization output graphs. The results of data normalization can be view as a box plot of Perfect Match (PM) log intensities distribution (right) or as a line graph in the ‘Descriptive statistics’ view (left). Above the line graph, the post-normalisation mean and standard deviation values are displayed

After **normalisation**, data might need to be transformed. In the EP main menu, click on ‘Data transformation’, under the ‘Transformations’ menu (Fig. 4.2). Several types of **transformations** are available (Fig. 7):

**Transformation**

Intensity → (Log N) Ratio Ratio → Log N Ratio Average row identifiers KNN imputation Transpose Abs → Rel

Mean-center Compute relative expression ratios from available columns

Compute ratios relative to ... gene's average value

use regexps/partial matching

What log to take (if any) No log: these are Log 2 data (post-RMA)

Execute

Fig. 7: Transformation menu in EP – ‘Absolute to relative’ is selected

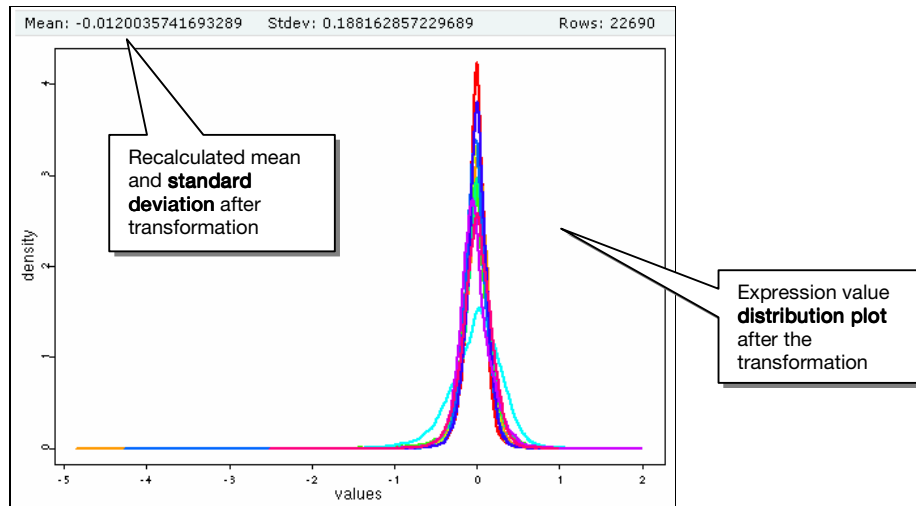
The description of all **transformation** available, from left to right, is as follows:

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 [13]
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)

When working with a set of **Affymetrix .CEL files**, it may be desired to look for genes whose expression varies relative to a reference sample, i.e., to one of the imported CEL files. If there is no reference sample, as in this case, relative values can still be calculated with respect to the gene's average mean. This will allow comparison between arrays in the absence of a common reference.

We will now apply an 'Absolute to relative' **transformation** to the RMA normalised E-MEXP-886 dataset. Make sure the normalised dataset is selected, in the 'current dataset' menu. In the Transformation menu, click on the 'Abs-to-Rel' tab and select 'gene's average value' from the 'compute ratios relative to' drop down menu (Fig. 7). In the 'What log to take' drop down menu, select 'No log: these are log<sub>2</sub> data (post-RMA)' since the data has already been log transformed by the RMA algorithm. Then click 'Execute'.

The result of data **transformation** will be displayed in a new window as a dataset **heatmap**. Explore the changes in the log intensity **distribution plot** after **transformation** by going back to the EP Data Selection view (Fig. 8). Observe that the log intensity distribution is now centred on 0 (mean = -0.01). This ensures that both repressed and induced genes are equally represented allowing us to perform further analysis.



*Fig. 8: Data transformation output graph. The transformed data is now shown in the descriptive statistic view. At the top of the graph, the post-transformation mean and standard deviation values are displayed.*

Following **normalization** and **transformation**, the data can be analyzed using any of the tools available in EP. We will now use two different methods for identifying **differentially expressed** genes in the E-MEXP-886 dataset.

## 4 How to identify differentially expressed genes using t-test

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

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 1-5 for Class 1 (wild type mice samples) and 6-10 for Class 2 (ataxin-null mice samples) and click 'Execute' (Fig. 9).

**Class Setup**

Use "One Class" to detect differentially expressed genes, viewing the expression matrix as one class of experimental data. You will need to specify a set of control genes against which to test. The "Two Class" setup can be used to compare gene expression between two sets of experiments, and will require you to either specify two experiments or to specify how to break the columns of one expression matrix into two groups

One class **Two classes in one dataset**

Specify column numbers (starting with 1, separated with commas; specify ranges with dash, -, e.g. 1,2-10,31,31-50) for the two classes below

Class 1

Class 2

**Parameters**

p-value cut-off

Multiple testing correction

The user can specify a p-value cut off. The default value is 0.01. This will affect the number of genes identified by the t-test

A number of standard corrections are implemented, including the Bonferroni, Holm and Hochberg corrections [4-6] for reducing number of genes falsely identified as differentially expressed. The user can select any of them from the 'Multiple testing correction' drop down menu.

Fig. 9: '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. 10 – right panel), as well as a plot of the top 15 genes found (Fig. 10 – left panel).

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**).

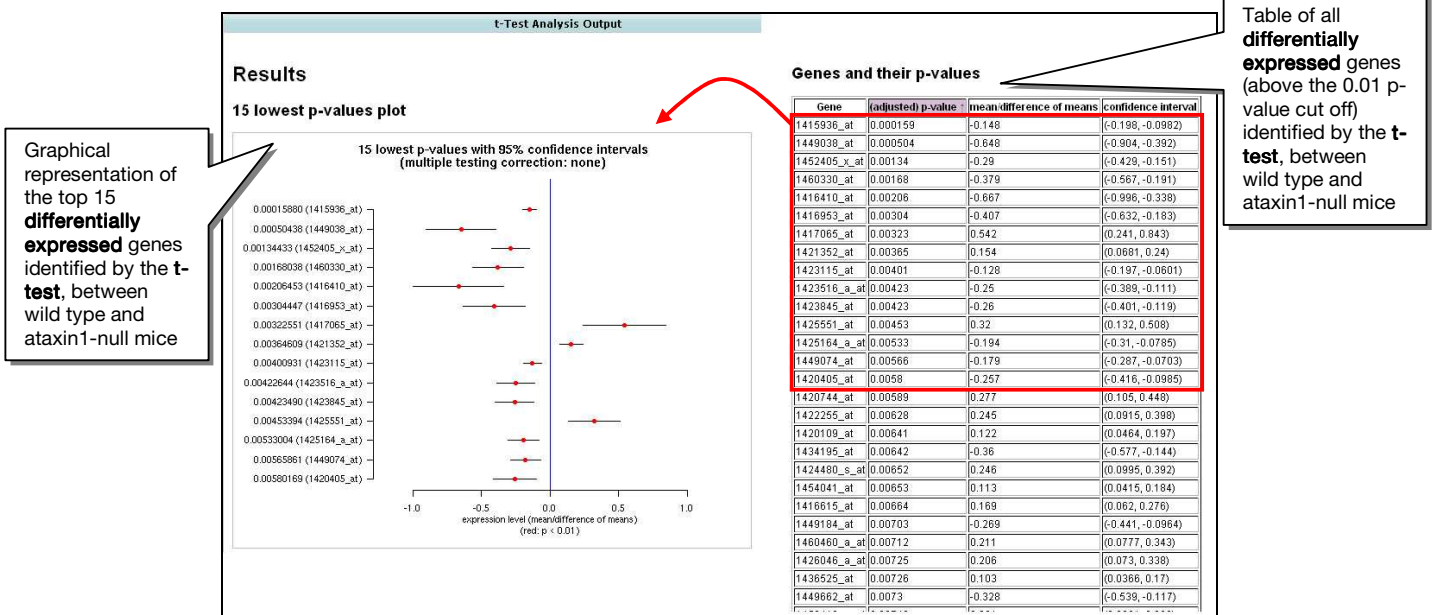
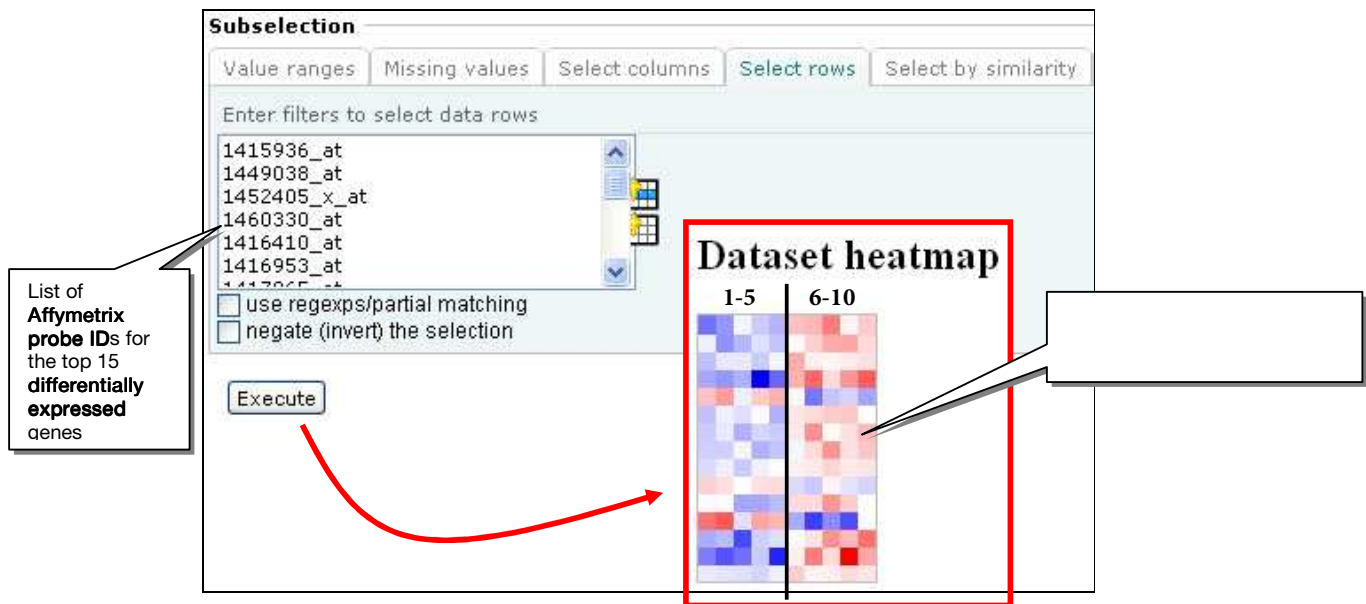


Fig. 10: t-test analysis output graphs. The t-test analysis results are summarized in a table, where the genes are ranked according to the p-value, with the most significant genes at the top (right). The top 15 genes are also plotted in a graph (left). The expression level is shown as red dot when the corresponding p-value is below the chosen cut off (in this case 0.01).

Go to 'Data selection', under the 'Subselection' menu and click on 'Select rows' (Fig. 11). Cut and paste (with the help of a text editor) the list of **Affymetrix probe IDs** for the top 15 **differentially expressed** genes into the text box and click on '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. 11). 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. 11).



*Fig. 11: 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: wild type mice samples (1-5) and ataxin-null mice samples (6-10). Red indicates induced genes, blue repressed genes.*

The user can retrieve more information about any of the top 15 **Affymetrix probe IDs** just identified by the **t-test** analysis.

Go to 'Data selection', under the 'Subselection' menu and click on 'select by row' (Fig. 12). Click on the small top table icon, type in the text box an Affymetrix ID and click search. The corresponding gene symbol, gene description and chromosome location will be returned in the result window (Fig. 12).

The screenshot shows the EBI Affymetrix data interface. At the top, there is a 'Subselection' menu with options for 'Value ranges', 'Missing values', 'Select columns', and 'Select rows'. Below this menu is a search box containing '1415936\_at' and a 'search' button. A red arrow points from this search box to a search box below the main table. The main table lists Affymetrix identifiers, gene symbols, descriptions, and feature IDs. The identifier '1415936\_at' is highlighted in the table. Below the main table is a search box with '1415936\_at' and a 'search' button. Below this search box is a table with additional information for the selected identifier.

affymetrix	geneSymbol	Description	feature_id
1415670_at	Copg	coatamer protein complex, subunit gamma [Source:MarkerSymbol,Acc:MGI:1858696]	1415670_at
1415671_at	Atp6v0d1	ATPase, H+ transporting, lysosomal V0 subunit D1 [Source:MarkerSymbol,Acc:MGI:1201778]	1415671_at
1415672_at	Golga7	golgi autoantigen, golgin subfamily a, 7 [Source:MarkerSymbol,Acc:MGI:1931029]	1415672_at
1415673_at	Pspk	phosphoserine phosphatase [Source:MarkerSymbol,Acc:MGI:97788]	1415673_at
1415674_a_at	Trappc4	trafficking protein particle complex 4 [Source:MarkerSymbol,Acc:MGI:1926211]	1415674_a_at
1415675_at	Dpm2	dolichol-phosphate (beta-D) mannosyltransferase 2 [Source:MarkerSymbol,Acc:MGI:1330238]	1415675_at
1415676_a_at	Psmb5	proteasome (prosome, macropain) subunit, beta type 5 [Source:MarkerSymbol,Acc:MGI:1194513]	1415676_a_at

affymetrix	geneSymbol	Description	feature_id	chromosome_location	composite_feature_id
1415936_at	Bcar3	breast cancer anti-estrogen resistance 3 [Source:MarkerSymbol,Acc:MGI:1352501]	1415936_at	3G3	Bcar3

**Fig. 12:** The 'select rows' function in the Subselection menu was used to additional information on the Affymetrix identifier 1415936\_at. The corresponding gene symbol (Brca3), gene description (breast cancer anti-estrogen resistant 3) and chromosome location were retrieved.

This example represents the first step towards reconstructing a biological network of genes which might be involved in determining **spinocerebellar ataxia type 1**.

## 5 How to identify differentially expressed genes using ordination based techniques

The Between Group Analysis (BGA) component, under the 'Ordination-based' menu, provides a statistically rigorous framework for a more comprehensive multi-group analysis of microarray data. BGA simply is a dimensionality reduction technique which is carried out on defined groups of samples rather than individual samples [15]. The ordination step involved in BGA, as implemented in EP, can be either **Principal Components Analysis (PCA)**, or **Correspondence Analysis (COA)**, both standard statistical tools for reducing the dimensionality of a dataset by calculating an ordered set of values that correspond to greatest sources of variation in the data and using these values to "reorder" the genes and samples of the matrix. BGA combined with COA is especially powerful, because it provides a simultaneous view of the grouped samples and the genes that most facilitate the discrimination between them. The BGA component's algorithms are provided through an interface of the BioConductor package made4 [16], which, in turn, refers to the R multivariate data analysis package ade4.

Go to 'Ordination-based' menu and click on 'Between Group Analysis' (Fig. 4.2). We first need to create a factor to discriminate between different groups of samples. Click on the 'Define new factors' icon under the 'Factors' text box (Fig. 13).

In the new window, click on the 'Add factor' button. In this example, we want to identify the genes which are **differentially expressed** between 2 conditions: wild type and ataxin1-null mice. The top 5 data files are wild type (WT) and the bottom 5 are ataxin1-null mice (KO). Select a name for the new experimental factor (e.g. WT/KO), fill the table as shown in Fig. 13 and click 'Save factor'. The

newly created experimental factor will now be showing in the 'Factors' box, in the BGA window and can now be selected as parameter for the analysis (Fig. 13).

Select the WT/KO factor. From the top dropdown menu select either COA or PCA (Fig. 12). Different output graphics can also be added, if needed. For this example, we will leave the default parameters. Click 'Execute'.

**Between Group Analysis**

**Factors**  
Select which factors determine the groups for this analysis:  
column\_id: 10 groups  
WT/KO: 2 groups

**Transformation options:**  
Correspondence Analysis (COA) or Principal Components Analysis (PCA)  
Impute data via row averaging

**Results to be displayed (optional parameters)**  
Graphs. By default, a graph of the eigenvalues and a plot of the arrays and genes on the first 2 axes will be shown. In addition a 3D graph of the arrays is shown. Optional: select additional output files: Default: Overall plot of Arrays, Genes  
Text Files. Output a summary and the following files. By default, a log of the analysis and co-ordinates of the arrays and the genes are given. This is sufficient generally. Optional: select additional output files:

**Submit**  
Execute

column number	factors	
	column_id	new factor: WT/KO
1	E-MEXP-886-raw-cel-1227302782.cel	new value: WT
2	E-MEXP-886-raw-cel-1227302791.cel	new value: WT
3	E-MEXP-886-raw-cel-1227302800.cel	new value: WT
4	E-MEXP-886-raw-cel-1227302809.cel	new value: WT
5	E-MEXP-886-raw-cel-1227302818.cel	new value: WT
6	E-MEXP-886-raw-cel-1227302827.cel	new value: KO
7	E-MEXP-886-raw-cel-1227302836.cel	new value: KO
8	E-MEXP-886-raw-cel-1227302845.cel	new value: KO
9	E-MEXP-886-raw-cel-1227302867.cel	new value: KO
10	E-MEXP-886-raw-cel-1227302889.cel	new value: KO

Save factor  
Cancel

Callouts:  
 - 'Define new factor' icon: Points to the plus icon in the Factors section.  
 - Factor box: Points to the list of factors in the Factors section.  
 - Once created, the new factor will appear in the factor box and can be selected for the following analysis.  
 - Different transformation options and output graphs can be selected.  
 - Select a name for the new experimental factor (e.g. WT/KO), and then assign to individual samples.  
 - Window used for creating a new factor: Points to the table below.

**Fig 13: Define new factor window.** When running an ordination-based technique, the user might need to create a new experimental factor in order to identify the genes differentially expressed between 2 conditions. In this example, the genotype is the discriminating factor (wild type vs. knock-out) and the new factor can be created filling the table as shown.

The 'overall plot' provides a graphical representation of the most discriminating arrays and/or genes. In addition to this plot, BGA produces two numerical tables, the table of genes coordinates (Coordinates of columns) and the table of array coordinates (Coordinate of rows) (Fig. 14). The gene coordinates table is of special interest, because it provides, for each gene, a measure of how variable that gene is in each of the identified strong sources of variation. The sources of variation (principal axes/components) are ordered from left to right. In this example we only have one main source of variation (Component 1). Thus genes that have the highest or lowest values in the first column of the gene coordinates table make up the likeliest candidates for differential expression.

Coordinates of rows		Coordinates of columns	
Name	Axis 1 ↑	Name	Component 1 ↑
E.MEXP.886.raw.cel.1227302889.cel	-56.6	X1416906_at	-6.72e-05
E.MEXP.886.raw.cel.1227302867.cel	-49.3	X1423774_a_at	9.73e-06
E.MEXP.886.raw.cel.1227302827.cel	-41.3	X1424391_at	-5.9e-05
E.MEXP.886.raw.cel.1227302836.cel	-35.7	X1426472_at	2.26e-05
E.MEXP.886.raw.cel.1227302845.cel	-20.1	X1429839_a_at	7e-04
E.MEXP.886.raw.cel.1227302800.cel	17.1	X1450081_x_at	-2.53e-05
E.MEXP.886.raw.cel.1227302809.cel	19.1	X1452070_at	8.47e-05
E.MEXP.886.raw.cel.1227302782.cel	40.9	X1415936_at	-0.926
E.MEXP.886.raw.cel.1227302791.cel	59.5	X1449038_at	-0.903
E.MEXP.886.raw.cel.1227302818.cel	66.5	X1416410_at	-0.863
		X1452405_x_at	-0.863
		X1423115_at	-0.856
		X1460330_at	-0.855
		X1423845_at	-0.849
		X1423516_a_at	-0.839
		X1416953_at	-0.829
		X1434195_at	-0.819
		X1427891_at	-0.817
		X1425164_a_at	-0.814

**Fig 14: BGA analysis output tables.** The BGA analysis results are summarized in a table of genes coordinates (right) and a table of array coordinates (left).

The identification of **differentially expressed** genes is only the first step in the analysis of microarray data. Once identified, the **differentially expressed** genes can be clustered in order to discover patterns of gene expression in the data. Ultimately, additional biological information (e.g. gene and sample annotation,...) should be brought in to obtain new insights into the biology of the system studied. For more examples, including clustering and the use of gene annotation, we suggest looking at the 'Expression Profiler for Beginners – part 1' tutorial.

## 6 How to obtain a raw data for experiment E-MEXP-886

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-886 and click 'Query'
3. Expand the experiment view by clicking on the plus sign next to E-MEXP-886
4. Click on the Affymetrix raw data icon on the title line to download onto your PC a zip archive containing all .CEL files related to this experiment.
5. Save the 'E-MEXP-886.raw.zip' file onto your computer

## 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.

### Affymetrix probe ID

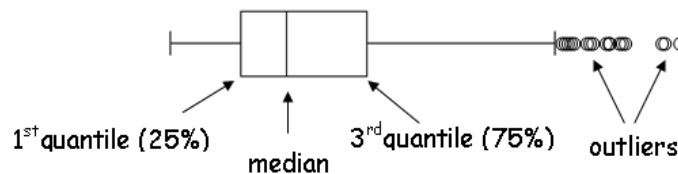
Affymetrix unique identifier assigned to each probe on the array

### Box plot

A **box plot**, or box and whisker diagram provides a simple graphical summary of a dataset. It is often used in exploratory data analysis to show the shape of the distribution, its central value, and its variability.

The **box plot** is interpreted as follows:

- The box itself contains the middle 50% of the data. The upper edge (3<sup>rd</sup> quantile) of the box indicates the 75th percentile of the data set, and the lower edge indicates the 25th percentile (1<sup>st</sup> quantile).
- The line in the box indicates the median value of the data. If the median line within the box is not equidistant from the hinges, then the data is skewed.
- The ends of the vertical lines or "whiskers" indicate the minimum and maximum data values, unless outliers are present in which case the whiskers extend to a maximum of 1.5 times the inter-quartile range.
- The points outside the ends of the whiskers are outliers or suspected outliers.



### 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.

### Correspondance analysis (COA)

Correspondence analysis is an explorative computational method for the study of associations between variables. Much like principal component analysis, it displays a low-dimensional

projection of the data, e.g., into a plane. It does this, though, for two variables simultaneously, thus revealing associations between them.

## 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.

## 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.

## GCRMA

GCRMA is a **normalisation** method using RMA with the help of probe sequence and with GC-content background correction

## 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.

## 'Li &Wong'

The 'Li & Wong' or Invariant Set Normalisation method takes a subset of Perfect Match (PM) probes with small within-subset rank difference in the two arrays as a basis for the **normalisation**. For more information on the smoothing process involved in Invariant Set Normalisation, please refer to Li and Wong, 2001.

## 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 [17].

## 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 **normalisation**, **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

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.

## RMA

Robust Multichip Average (RMA) is a three step **normalisation** procedure for Affymetrix data. The three steps consist of: background correction, quantile **normalisation** and summarization.

## Spinocerebellar ataxia type 1

Spinocerebellar ataxia type 1 (SCA1) is one specific type of ataxia among a group of inherited diseases of the central nervous system. In SCA1, genetic defects lead to impairment of specific nerve fibers carrying messages to and from the brain, resulting in degeneration of the cerebellum.

## Standard deviation

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

## 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.

## VSN

The Variance Stabilization and Normalisation (VSN) algorithm builds upon the fact that the variance of microarray data depends on the signal intensity and that a transformation can be found after which the variance is approximately constant.

## 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. 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).
5. Hochberg, Y., *A sharper Bonferroni procedure for multiple tests of significance*. Biometrika, 1988. **75**: p. 800-803.
6. Holm, S., *A Simple Sequentially Rejective Bonferroni Test Procedure*. Scandinavian Journal of Statistics, 1979. **6**: p. 65-70.
7. Goold, R., et al., *Down-regulation of the Dopamine Receptor D2 in mice lacking Ataxin 1*. Hum Mol Genet, 2007. **28**: p. 28.
8. Quackenbush, J., *Microarray data normalisation and transformation*. Nat Genet, 2002. **32 Suppl**: p. 496-501.
9. Huber, W., et al., *Variance stabilization applied to microarray data calibration and to the quantification of differential expression*. Bioinformatics, 2002. **18 Suppl 1**(1): p. S96-104.
10. Irizarry, R.A., et al., *Summaries of Affymetrix GeneChip probe level data*. Nucleic Acids Res, 2003. **31**(4): p. e15.
11. Li, C. and W.H. Wong, *Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection*. Proc Natl Acad Sci U S A, 2001. **98**(1): p. 31-6.
12. Wu, Z., et al., *A Model-Based Background Adjustment for Oligonucleotide Expression Arrays*. Journal of the American Statistical Association 2004. **99**(468): p. 909-917.
13. Troyanskaya, O., et al., *Missing value estimation methods for DNA microarrays*. Bioinformatics, 2001. **17**(6): p. 520-5.
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. Culhane, A.C., et al., *Between-group analysis of microarray data*. Bioinformatics, 2002. **18**(12): p. 1600-8.
16. Culhane, A.C., et al., *MADE4: an R package for multivariate analysis of gene expression data*. Bioinformatics, 2005. **21**(11): p. 2789-90.
17. 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/](http://www.ebi.ac.uk/training/) for these and other eLearning resources.