http://www.bigcat.unimaas.nl/images/logo_05-header.jpg 

# Practical session 2: Analysis of microarray data – preprocessing, QC and statistics

Pathway analysis helps you to interpret the data in a biological meaningful context. In this session we will continue exploring the analysis of microarray gene expression data.

1. Perform data preprocessing and quality control using online webtool [www.ArrayAnalysis.org](http://www.ArrayAnalysis.org)

2. Perform statistical analysis of pre-processed data from step 1

In this practical, we will use a gene expression dataset that studies the effects of silver nanoparticles on human intestine cell line Caco-2. The paper of the original experiment can be found here:

Molecular mechanism of silver nanoparticles in human intestinal cells, Linda Böhmert, Birgit Niemann, Dajana Lichtenstein, Sabine Juling & Alfonso Lampen, Nanotoxicology, 2015; 9(7): 852–860, [Link](http://www.ncbi.nlm.nih.gov/pubmed/25997095)

Check the GEO page for more information: <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE62253>

## Part 1: Data quality control and pre-processing

[ArrayAnalysis.org](http://www.arrayanalysis.org/) is an open source, free to use online platform for analysis of microarray data - and an alternative program for [Chipster](http://chipster.csc.fi/manual/supported-chips.html). ArrayAnalysis is a webtool, so there is no need for download or access code, and it provides more extensive quality control than Chipster but it is limited to two microarray formats: Affymetrix and Illumina. The exact microarray type (e.g. Affy-1) is automatically recognised. This tutorial shows how to use the Affymetrix quality control (affyQC) module which is designed for doing quality control and preprocessing of microarray data from Affymetrix microchips. All source code has been written in R and is open-source, available under the [Apache License version 2.0](http://www.apache.org/licenses/LICENSE-2.0). It is available on our [Download](http://www.arrayanalysis.org/download.php) page.

affyQC can be run :

* on-line via the [arrayanalysis.org](http://www.arrayanalysis.org) webportal (follow["Get started"](http://www.arrayanalysis.org/getStarted.php)).
* locally as an automated R workflow provided via a wrapper function

The main functions of affyQC are:

* to compute array quality information;
* to plot images that allow identifying any aberrations present in the dataset;
* to return pre-processed data and QC reports.

**JavaScript** has to be enabled (activated) in your web browser. You will be warned if it is not the case. You can activate it at any time in the browser options (see [activatejavascript.org](http://activatejavascript.org/) if needed)

You don't need to log in; you just need to prepare a zipped file containing all your Affymetrix .CEL files and possibly a file describing your dataset, called the description file. A presentation of this description file is available in the fourth section, subsection ["Parameter description"](http://www.arrayanalysis.org/documentation.php#paramDesc). The on-line module contains three steps before the launch of the analysis:

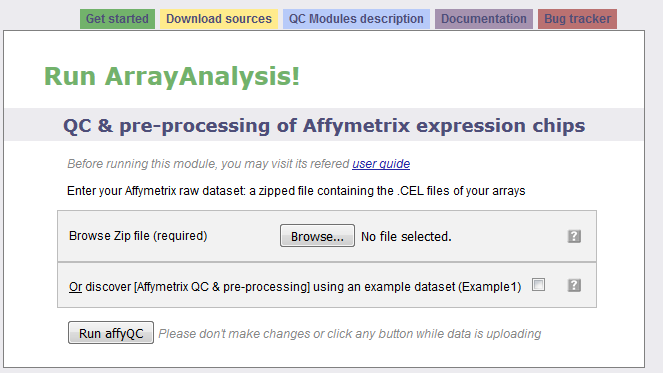
* Step1: First you load the archive of .CEL files
* Step2: Then you complete the description of the dataset
* Step3: And finally you choose the plots to be computed and their parameters.

Then:

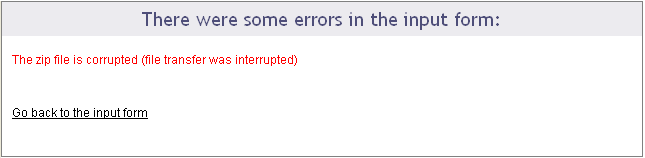
* Execution: The module is executed with the settings you choose
* Results: You get the results after the execution step, or by e-mail.

## First step: load the CEL files

The following picture shows the screen for the first step. Use **GSM1523726.CEL** in the folder “**Practical 2 data**”. This is a zipped folder.

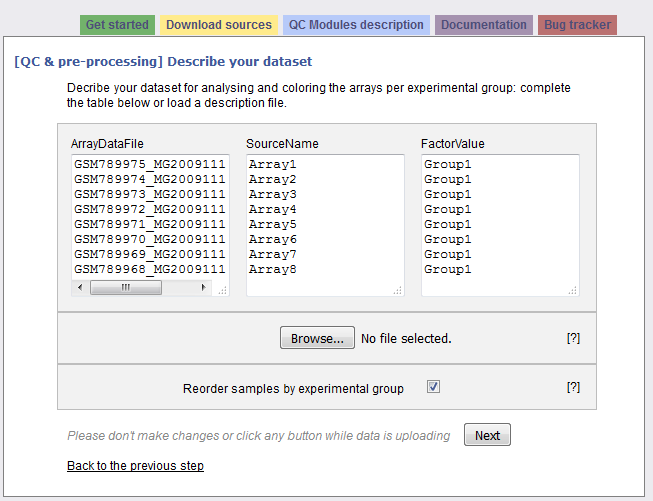


The interrogation mark button will help you by giving you a contextual help. Note that this feature is available when Javascript is activated and is not yet supported by Google Chrome and Safari browsers. Loading the zip file may take a while as .CEL files are often very big; don't click any button after clicking on the "Next" button otherwise the loading of the file may be compromised. When the file is loaded without error, you are automatically directed to the next step. Otherwise you get a message indicating the error encountered:



## Second step: describe the dataset

The following picture shows the screen obtained after completing the first step:



The interrogation mark buttons will help you by giving you a contextual help. Your dataset has been read and the following information is presented in a three columns table:

Column "ArrayDataFile" contains the .CEL file names of your N arrays found in the input zip file. You cannot edit this column.

Column "SourceName" is filled with Array1 - ArrayN. These names will be used for the analyses. Feel free to modify these names at the condition you use only unique names.

Column "FactorValue" is always set to "Group1". If you want your array groups to be represented in the analyses and plots, rename the factor groups. Please note, that the name must not start with a number or contain special characters.

Have a look at the excel file **data sample relationship.** This information is from ArrayExpress and describes the different arrays in the file. **Group the arrays as described in the file!**

You may also prefer to enter directly this information from a file you have prepared. If this is the case, browse your description file in the second section. If you enter such a file the information contained in the previous table will be skipped. You'll find a presentation of the description file on the fourth section of this documentation: ["Parameter description"](http://www.arrayanalysis.org/affyQC/doc_affyQC_web#paramDesc)

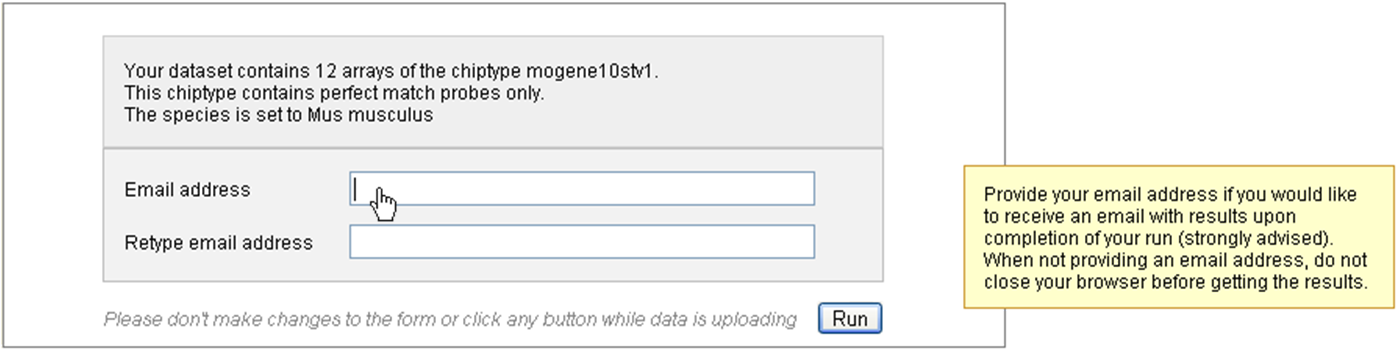
Clicking on the "Next" button will direct to the last step if no error has been detected.

## Third step: define your analysis

This last input form is divided into three main sections: the first part allows a quick launch, the second part defines in details the analysis parameters applied to the raw data and the last part is dedicated to the pre-processing (parameters for the normalization and re-annotation) and its evaluation (definition of the analysis parameters applied to the normalized data).

First part of the input form

The following picture presents the first part; it recalls briefly what your dataset contains and asks you to enter an e-mail. This is optional: if you don't enter your e-mail, you will need to keep the browser opened and not close the page before the end of the calculation. On the contrary, if you enter your e-mail address - which is recommended - you can close the windows as soon as the next page appears and you will be inform of the end of the analysis by e-mail. You would just have to follow the links to the result files given in the e-mail.



You may launch the analysis with the "Run" button right after this first section. In this case default parameters will be used.

Second part of the input form

This part contains four frames representing the four families of analysis applied to your raw data: 1) Sample quality, 2) Hybridization and overall signal quality 3) Signal comparability and bias diagnostic and 4) Array correlation.

Most of the parameters are checkboxes that you would tick or untick to indicate whether a certain plot or table has to be computed or not. The analyses and plots are described in the [module description](http://www.arrayanalysis.org/affyQC/moduleDesc_affyQC.php) page, which is reachable also from the left vertical menu (we recommend you to open the pages in a new tab to not lose the information entered in the input form you are filling).

Some analyses or plot construction, such as the MA-plot and the hierarchical clustering, need particular parameters. You may modify the default values.

The following picture presents you this part of the input form, which defines the graphs built from the raw data:



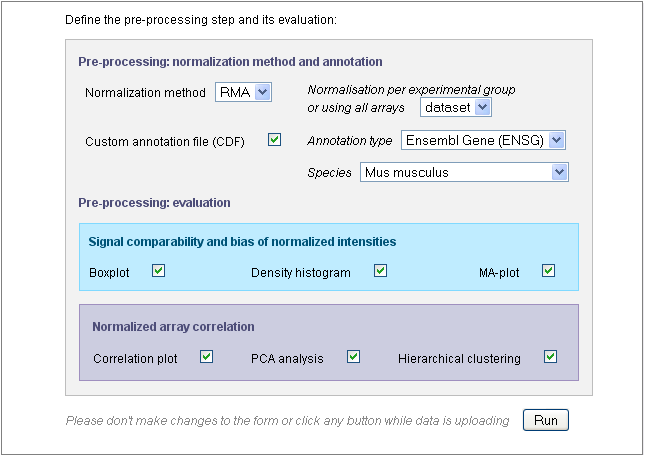
You may note that all the plots are not selected by default; you may select all of them with the first checkbox: [toggle select all].

You may also note that some plots cannot be selected, such as the "Sample prep controls", the "Background intensity" or the "Scale factors". This is because the dataset used for this example (public dataset available on ArrayExpress: [E-GEOD-13278](http://www.ebi.ac.uk/arrayexpress/browse.html?keywords=E-GEOD-13278)), was built with PM-only arrays and the construction of these particular graphs uses the [MAS5 algorithm](http://media.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf) which cannot be applied to PM-only arrays.

Be aware that the generation of 2D PLM-based images for spatial biases are highly time-consuming; the generation of the complete set of images (4 different images representing the raw data, the PLM weights, residuals and residual signs) is not computed by default. See examples of these images on the [description page](http://www.arrayanalysis.org/affyQC/moduleDesc_affyQC.php#SO) or on [Bolstad PLM page](http://plmimagegallery.bmbolstad.com/).

Third part of the input form

The following picture presents the part of the input form concerning the pre-processing step and its evaluation:



Use the "Normalization method" drop-down menu to define the pre-processing step. You may chose "*none*" and keep the raw data. In this case, further parameters will be skipped. By default, the **GC-RMA** is applied to arrays containing both PM and MM probes and RMA is applied to PM-only arrays.

If the species could have been deduced from the CEL files in the previous steps, the "*Species*" field is already filled, as shown in this example. Otherwise, you would need to fill this field yourself or to untick the "Custom annotation" checkbox. **Make sure that you have chosen the correct species as indicated in the sample-data relationship file!**

After defining the pre-processing, you chose the analyses you want to apply to the normalized data. Only six graphs are proposed (other graphs are not meaningful on normalized data) and the parameters entered for the MA-plot and hierarchical clustering applied on raw data will be also used for the normalized data.

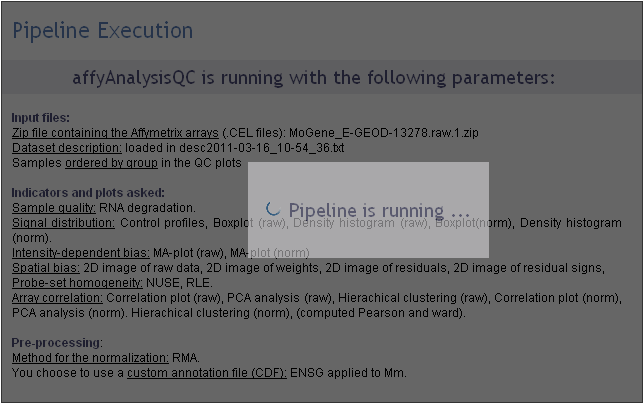
Once the input form is completely filled, you can launch the analysis with the "Run" button. Don't click any button after clicking on the "Run" button and before being automatically redirected to the execution page, otherwise you may compromise your analysis.

## Excecution step

After the third step, affyQC has all it needs to launch the analysis. The page becomes grey with a message telling you that the analysis is running. If you entered your e-mail address in the previous step, you can now close the window.

You will find on this page a recalling of the choices you made for this analysis: which files were loaded or created, which plots you decided to create for raw and normalized data and how you managed the pre-processing step.

The following picture shows the screen for the execution step:



## Getting the results

If you entered your e-mail address during the third step, you will receive an e-mail such as the one presented on the following picture:



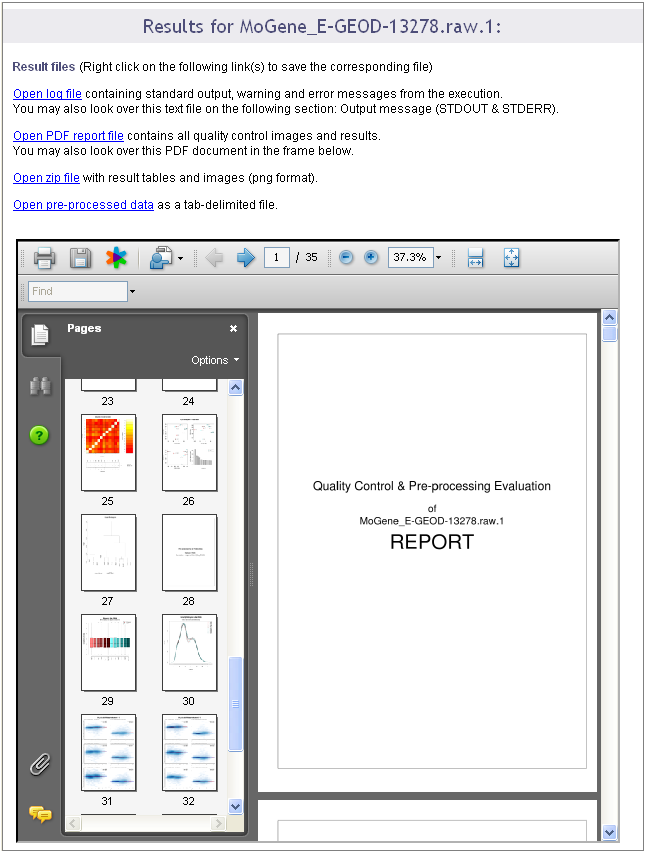
The e-mail contains direct links to the log file, PDF report, ZIP file containing the resulting files (png images, usable for your presentations, and result files such as the PMA table) and normalized dataset (presented as a tab-delimited file). If you closed the browser once you analysis was launched, you can only reach these result files through the links given in the e-mail. You cannot access your results from the arrayanalysis.org portal anymore.

On the contrary, if you did not close the browser, the result page presented in the following pictures shows up when the calculation are ended.

A first section gives you the same links to the result files than the e-mail: we recommend you either to save these links or to save the result files because if you did not enter your e-mail, once you close this result page, you will not be able to reach them again.

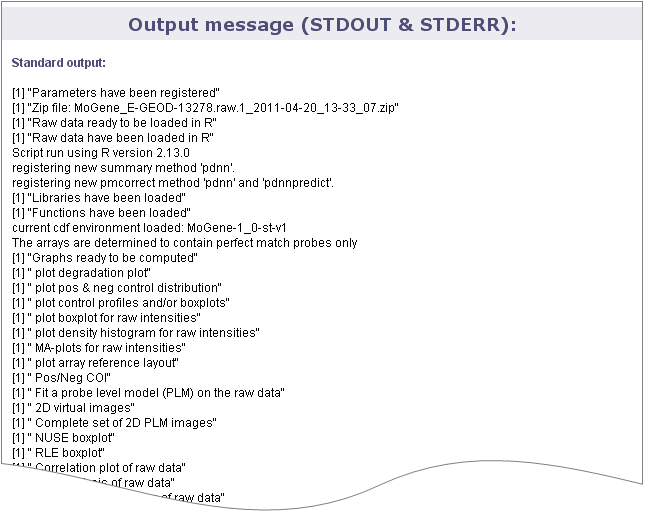
You can download your result files during one week from the links given by e-mail or by the result page. Make sure you download the files before this period elapses.

This section ends with a frame in which the PDF report is opened. You can visualize the document and save it from this frame.



A second section of the result page shows the log file content. This information is important when you encountered a bug in the execution: you can report the bug in our [internal tracking system](http://trac.bigcat.unimaas.nl/arrayanalysis/report) or [by email](http://www.arrayanalysis.org/affyQC/doc_affyQC_web.php). If you do so, please send us the log information by:

* either saving the log file on your computer (see previous links) and attach it to the ticket/e-mail
* or copy and paste the text from the screen.



## Part 2: Perform statistical analysis of pre-processed data from step 1

This tutorial shows how to use the statistics module of ArrayAnalysis which is designed for doing statistics on pre-processed (quality controlled, normalized data - e.g. from the previous step using AffyQC module) microarray data. All source code has been written in R and is available at<https://github.com/BiGCAT-UM/Stat_Module>.

**This technical documentation has two main objectives:**

* to guide you in the use of the Stat module
* to give interpretative help on the outputs of the module

**AnalysisStat can be run:**

* on-line via the<http://www.arrayanalysis.org> webportal (follow "Get started" and choose "Statistical analysis")
* or as an automated R workflow from a local computer

**The main functions of AnalysisStat are:**

* to perform statistical analysis on a table of (cleaned) data;
* to allow easy specification of experimental groups to be compared;
* to return tables containing (log) fold changes and P values for each measured element.;
* to plot diagnostic fold change and p value histograms and summary tables.

You can access the on-line module on the<http://www.arrayanalysis.org> webportal (follow "Get started" and choose "Statistical analysis"). You don't need to log in; you just need to have a tab delimited data file containing the (cleaned) data of your raw data files (you may also obtain such a file by running the affyAnalysisQC workflow) and possibly a file describing your dataset, called the description file.

The on-line module contains three steps before the launch of the analysis:

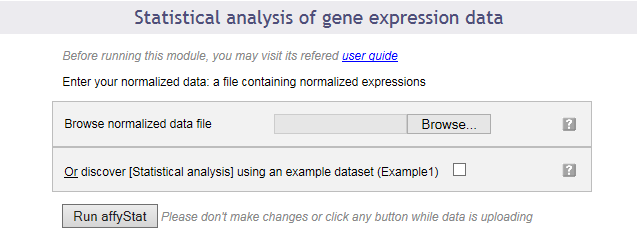
* Step1: First you load the tab delimited text file containing the (cleaned) data;
* Step2: Then you complete the description of the dataset;
* Step3: And finally you choose the comparisons or contrasts to be computed and indicate output options.

Then:

* Execution: The module is executed with the settings you choose
* Results: You get the results after the execution step, at the website or by e-mail.

## First step: load the data file

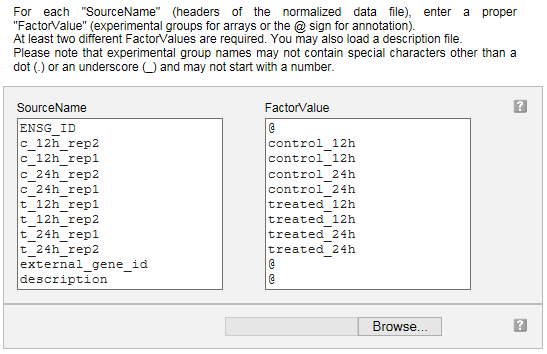
The following picture shows the screen for the first step:



This dialog allows you to upload a tab-delimited text file with (cleaned) data. Alternatively, the module can be run with an example data set, by ticking the checkbox presented. For this practical please use the text file **“normalized data” in your “practical 2 data” folder.**

## Second step: describe the dataset

The following picture shows the screen obtained after completing the first step:



Your dataset has been read and the following information is presented in a two columns table:

Column "SourceName" is filled with the columns headers from your data file. These names will be used for the analyses.

Column "FactorValue" is to be completed. Add an *at sign* (@) for each columns that does not represent sample measurements (e.g. annotation columns), and add the appropriate desired experimental group name for each other column.

You may also prefer to enter directly this information from a file you have prepared. If this is the case, browse your description file in the second section. If you enter such a file the information contained in the previous table will be ignored. Please use the information from **the excel file “data sample relationship**”.

Note that in case you reach the module directly from the AffyAnalysisQC module, you will find this table already filled with the array names and groups you entered in that module. You can modify the groups here if you wish.

Clicking on the "Next" button will direct to the last input form.

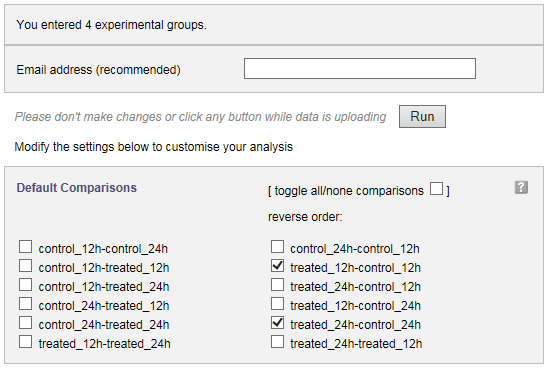
### Third step: define your analysis

The input form is divided into three main sections: the first part allows a quick launch using direct group comparisons; the second part gives you a chance of defining advanced contrasts to be computed. The third part allows customization of the required output options.

First part of the input form

The following image presents the first part; it recalls briefly what your dataset contains and asks you to enter an e-mail address. This is optional: if you don't enter your e-mail address, you will need to keep the browser opened and not close the page before the end of the calculation. On the contrary, if you enter your e-mail address - which is recommended - you can close the windows as soon as the next page appears and you will be informed of the completion of the analysis by e-mail. You will be presented with links to the result files in the e-mail.

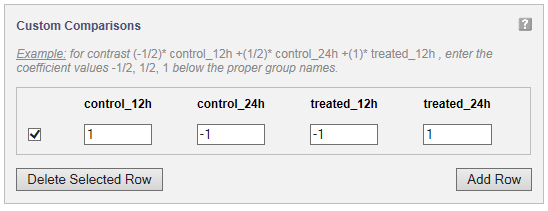
You may launch the analysis with the "Run" button right after this first section. In this case pairwise statistical comparisons will be done between each of the groups of data by default (option available only in the case of 4 or less experimental groups).



You can also choose the groups between which you require pairwise (e.g. experimental group - control group for each condition) statistical analysis to be done, by simply checking and unchecking boxes before you launch the analysis, as shown in the image above.

Second part of the input form

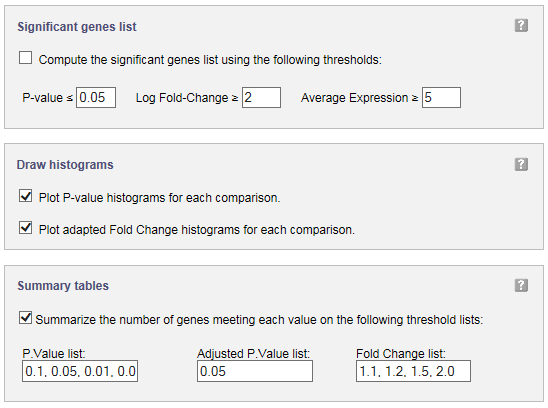
This part contains a text input box in which you can enter any custom contrast to be computed. When adding contrasts as well as predefined group comparisons (see above), both of them will be computed.



A statistical contrast can be any linear combination of the experimental groups. For instance, when for the example study presented in the image, you want to compute the difference over time for the treated samples, corrected for the difference over time for the control samples, you would compute (treated\_24h - treated\_12h) - (control\_24h - control\_12h). Simple arithmetic tells you this is control\_12h - control\_24h - treated\_12h + treated\_24h, which corresponds to the entries given in the figure. It is advisable to only use this option if you know about statistical contrasts or after consulting a statistician.

Third part of the input form

The following image presents the part of the input form concerning the plotting of the p-value and fold change histograms and computing the significant genes table.



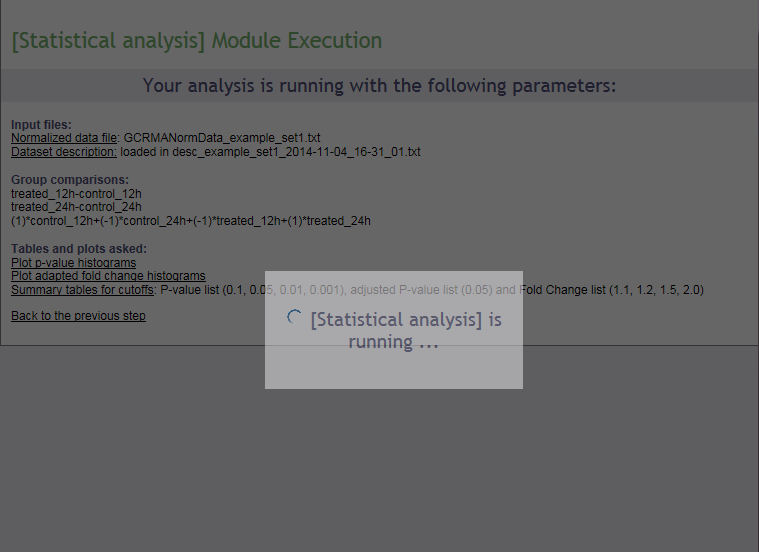
In significant genes list, you can define filters to select the genes to be added to a table of significantly changed genes and their statistical results. A table with p values and (log) fold changes for **all** genes is computed by default, and requires no ticking of a checkbox.

You can obtain histograms for the comparisons you have chosen to perform by ticking the checkboxes in the input form.

Significant genes summary tables can be obtained by entering a list of p value, adjusted p.value, and fold change cut-offs. For each of these cut-offs the number of genes meeting them will be computed and presented in a table.

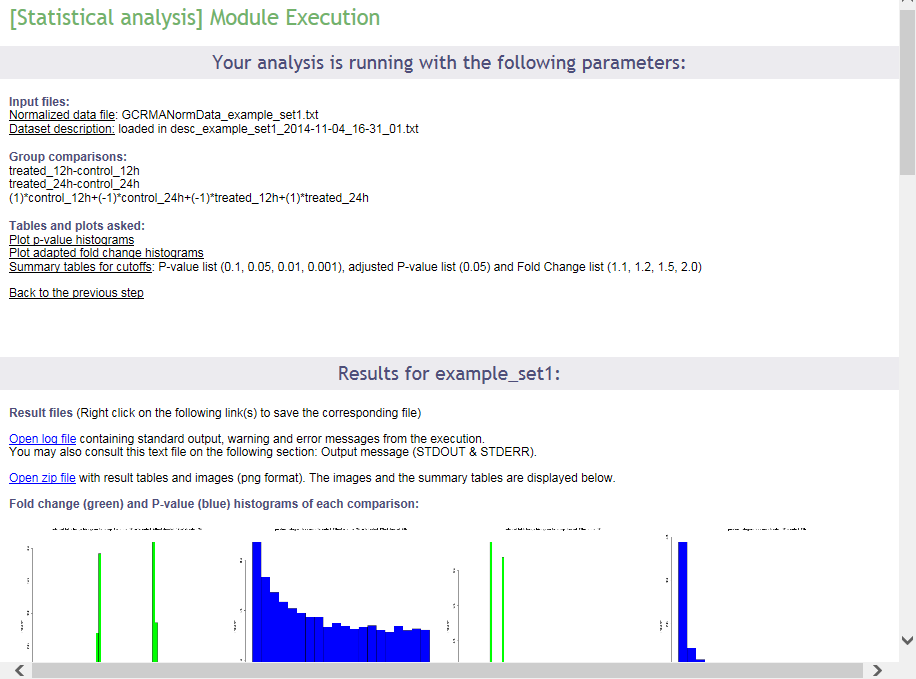
## Execution

After clicking 'Run' the module is executed.



## Results

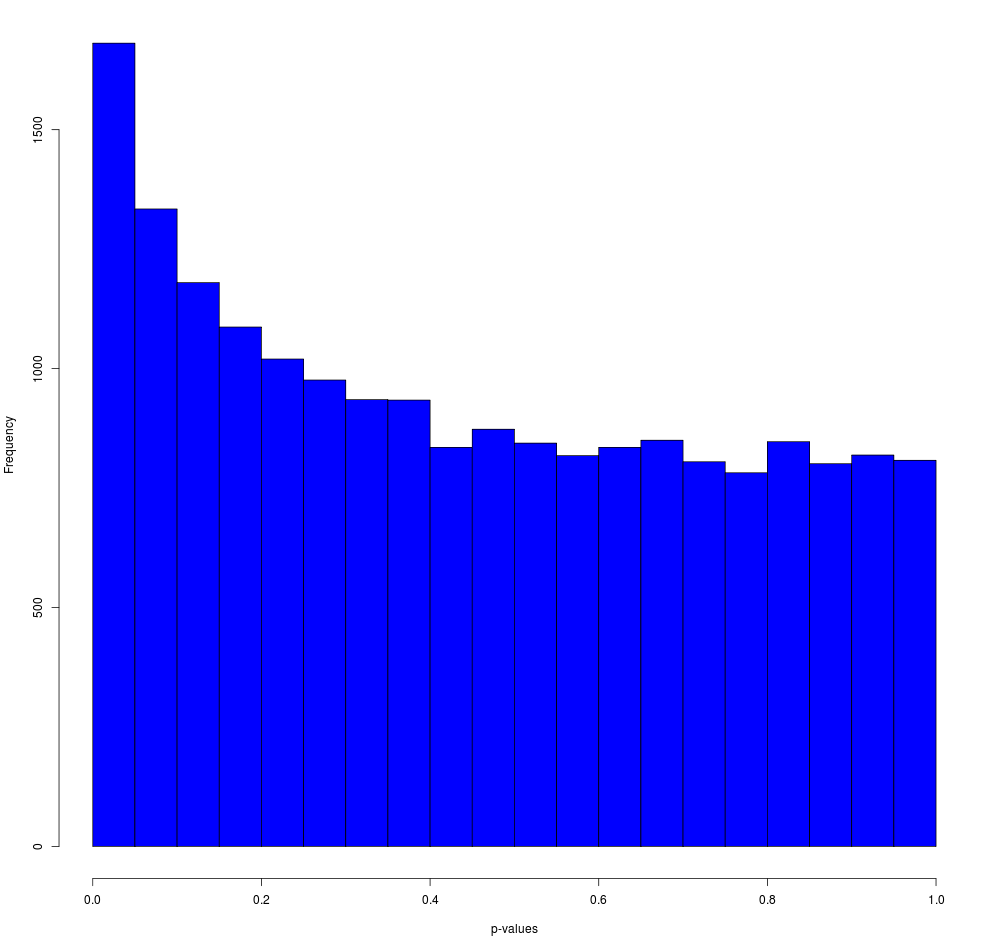
Upon completion a page of results is displayed on your screen.



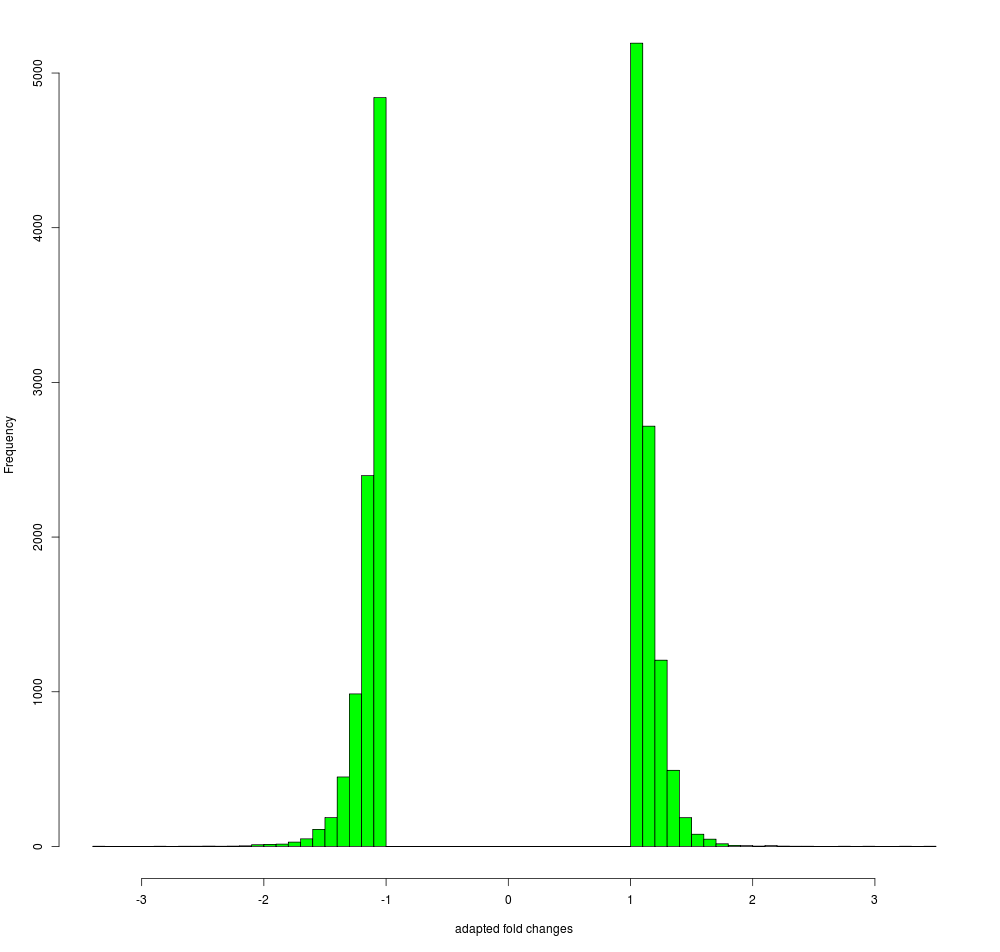
In the first part of the screen, your settings are recalled. Then links to the log file of the run and to the zip file containing all results (tables and images) are presented. Then, the p value and fold change histograms for each computed comparison are shown (clickable to enlarge, see below), if they have been chosen to be computed. These images are also part of the zip archive. The results will be described in the next section of this documentation.

#### Interpret the results of the Stat module

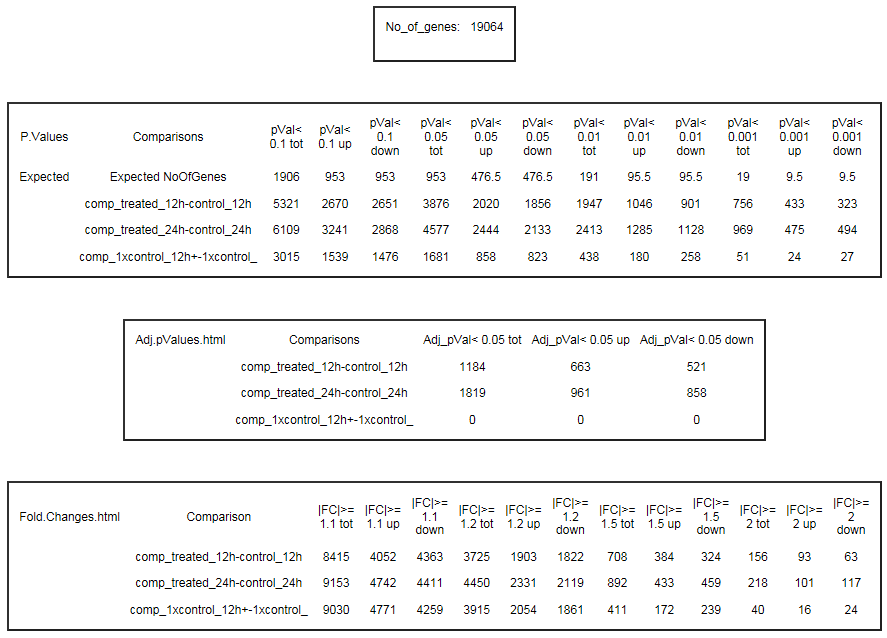
The p value histogram (see below) shows how often each interval of p values occurs. For random data, each p value is equally likely, and the histogram will be flat. For comparison of real different groups, the significant p values are expected to be overrepresented, and the histogram bars to be higher on the left side of the plot. For groups that are similar, non-significant p values are expected to be overrepresented, and the histogram bars to be higher on the right side of the plot.



The fold change histogram (see below) shows how often each fold change occurs, and gives and idea about the number of up- and downregulated genes, as well as the size of the differences. This can also be relevant to determine cut-offs for further analysis procedures. Note that the fold change is an adapted fold change: upregulated values are kept as they are, but downregulated values are represented by the negative value of their upregulated counterpart. So, for example, a 2 fold downregulated gene, does not have the value 0.5 (ordinary ratio), but -2 (minus the equivalent upregulated ratio).



Furthermore, as shown below, the summary tables of p values and fold changes are shown, if they have been chosen to be computed. These tables, in a tab-delimited version, are also part of the zip archive.

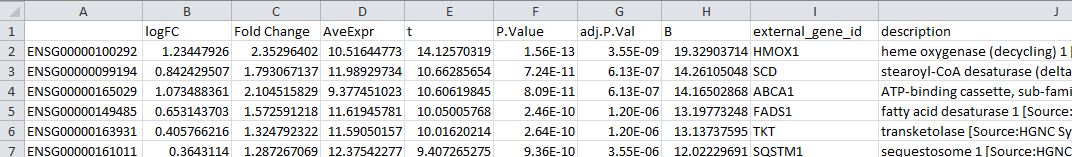


These tables indicate how many genes meet the chosen p value cut-offs and how many are expected to meet those purely by chance, how many genes meet the chosen adjusted p value cut-offs, and how many meet the chosen fold change cut-offs. The p value tables also indicate how many of the genes are up- and downregulated.

In addition, logging information is presented, as (partially) shown below.



Finally, this documentation describes the format of the statistical tables that are available from the zip archive, and contain the p values and (log) fold changes for all genes. The figure below shows a representative screenshot of an example table.



The table contains the following columns, as described below:

* identifier column - in the example containing Ensembl identifiers, but this depends on the data set that has been uploaded
* logFC - the log2 of the (regular) fold change, so the log2 of the ratio of the expression in both groups compared (or of the contrast outcome)
* Fold Change - an adapted version of the fold change, the ratio of the expression in both groups compared (or of the contrast outcome). For upregulated values the value is just the ordinary ratio; for negative values, the value is replaced by the negative of its upregulated counterpart
* AveExpr - the average expression over all samples in the experimental groups that have been compared
* t - the t-statistic of the limma adapted t-test
* P.Value - the p-value of the limma adapted t-test
* adj.P.Val - the Benjamini-Hochberg (FDR) corrected p-value
* B - the B-statistic
* external\_gene\_id - if available, an external name related to the identifier
* description - if available, the description belonging to the external name related to the identifier