



ENM TUTORIALS

How to use the statistics module of ArrayAnalysis.org for statistics analysis of microarray data

RELEASE DATE:	May 23rd 2016
USE:	How to use the statistics module of ArrayAnalysis.org for statistics analysis of microarray data
VERSION:	V.1.0.
MAIN AUTHOR:	Friederike Ehrhart
PARTNER:	UM
CONTACT DETAILS:	friederike.ehrhart@maastrichtuniversity.nl linda.rieswijk@maastrichtuniversity.nl egon.willighagen@maastrichtuniversity.nl +31(0)43-38 82913
AUTHORS:	Lars Eijssen, Anwasha Bohler, Linda Rieswijk, Egon Willighagen, Penny Nymark
LICENCE:	CC-BY 4.0



TABLE OF CONTENTS

[1. INTRODUCTION](#)

[2. APPLICATION DETAILS](#)

[First step: load the data file](#)

[Second step: describe the dataset](#)

[Third step: define your analysis](#)

[Execution](#)

[Results](#)

[3. ACKNOWLEDGMENTS](#)

[4. REFERENCES](#)

[5. KEYWORDS](#)

1. INTRODUCTION

ArrayAnalysis.org is an open source, free to use online platform for analysis of microarray data - and an alternative program for Chipster. 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.

How to use the documentation: As shown in the Table Of Content, you will find the separate sections :

- Using the on-line Stat module
- Interpreting the results provided

Bug tracking system: If you encounter an issue by using the code, you can report it at any moment on our internal tracking system : <http://trac.bigcat.unimaas.nl/arrayanalysis/newticket>. You can also use this system to post comments or feature suggestions. As an alternative you can contact the development team by email

Example data input file: An example dataset is available. When running the module, you can check a box to use this data set (Example1) in order to explore the functionality of the module.

2. APPLICATION DETAILS

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. A presentation of this description file is available in the fourth section, subsection [TO BE WRITTEN](#).

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

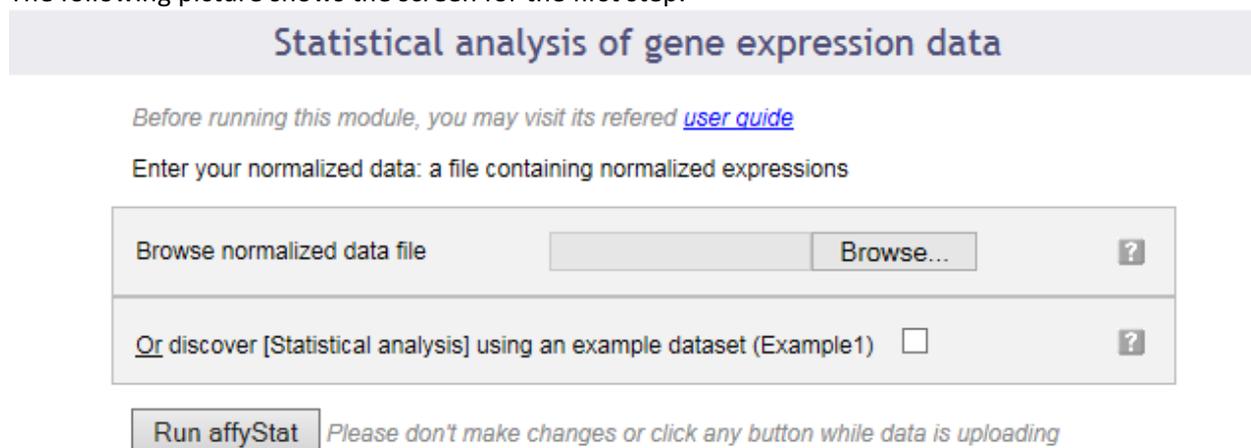
- [Step1](#): First you load the tab delimited text file containing the (cleaned) data;
- [Step2](#): Then you complete the description of the dataset;
- [Step3](#): And finally you choose the comparisons or contrasts to be computed and indicate output options.

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:



Statistical analysis of gene expression data

Before running this module, you may visit its referred [user guide](#)

Enter your normalized data: a file containing normalized expressions

Browse normalized data file Browse... ?

Or discover [Statistical analysis] using an example dataset (Example1) ?

Run affyStat *Please don't make changes or click any button while data is uploading*

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.

The interrogation mark button on all dialog forms will help you by giving you a contextual help.

SECOND STEP: DESCRIBE THE DATASET

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

For each "SourceName" (headers of the normalized data file), enter a proper "FactorValue" (experimental groups for arrays or the @ sign for annotation). At least two different FactorValues are required. You may also load a description file. Please note that experimental group names may not contain special characters other than a dot (.) or an underscore (_) and may not start with a number.

SourceName	FactorValue
ENSG_ID	@
c_12h_rep2	control_12h
c_12h_rep1	control_12h
c_24h_rep2	control_24h
c_24h_rep1	control_24h
t_12h_rep1	treated_12h
t_12h_rep2	treated_12h
t_24h_rep1	treated_24h
t_24h_rep2	treated_24h
external_gene_id	@
description	@

Browse... ?

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.

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 customisation 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 entered 4 experimental groups.

Email address (recommended)

Please don't make changes or click any button while data is uploading

Modify the settings below to customise your analysis

Default Comparisons [toggle all/none comparisons] [?](#)

reverse order:

<input type="checkbox"/> control_12h-control_24h	<input type="checkbox"/> control_24h-control_12h
<input type="checkbox"/> control_12h-treated_12h	<input checked="" type="checkbox"/> treated_12h-control_12h
<input type="checkbox"/> control_12h-treated_24h	<input type="checkbox"/> treated_24h-control_12h
<input type="checkbox"/> control_24h-treated_12h	<input type="checkbox"/> treated_12h-control_24h
<input type="checkbox"/> control_24h-treated_24h	<input checked="" type="checkbox"/> treated_24h-control_24h
<input type="checkbox"/> treated_12h-treated_24h	<input type="checkbox"/> treated_24h-treated_12h

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.

Custom Comparisons ?

Example: for contrast $(-1/2) \times \text{control_12h} + (1/2) \times \text{control_24h} + (1) \times \text{treated_12h}$, enter the coefficient values -1/2, 1/2, 1 below the proper group names.

	control_12h	control_24h	treated_12h	treated_24h
<input checked="" type="checkbox"/>	<input type="text" value="1"/>	<input type="text" value="-1"/>	<input type="text" value="-1"/>	<input type="text" value="1"/>

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 $(\text{treated_24h} - \text{treated_12h}) - (\text{control_24h} - \text{control_12h})$. Simple arithmetic tells you this is $\text{control_12h} - \text{control_24h} - \text{treated_12h} + \text{treated_24h}$, 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.

Significant genes list ?

Compute the significant genes list using the following thresholds:

P-value \leq Log Fold-Change \geq Average Expression \geq

Draw histograms ?

Plot P-value histograms for each comparison.

Plot adapted Fold Change histograms for each comparison.

Summary tables ?

Summarize the number of genes meeting each value on the following threshold lists:

P.Value list: Adjusted P.Value list: Fold Change list:

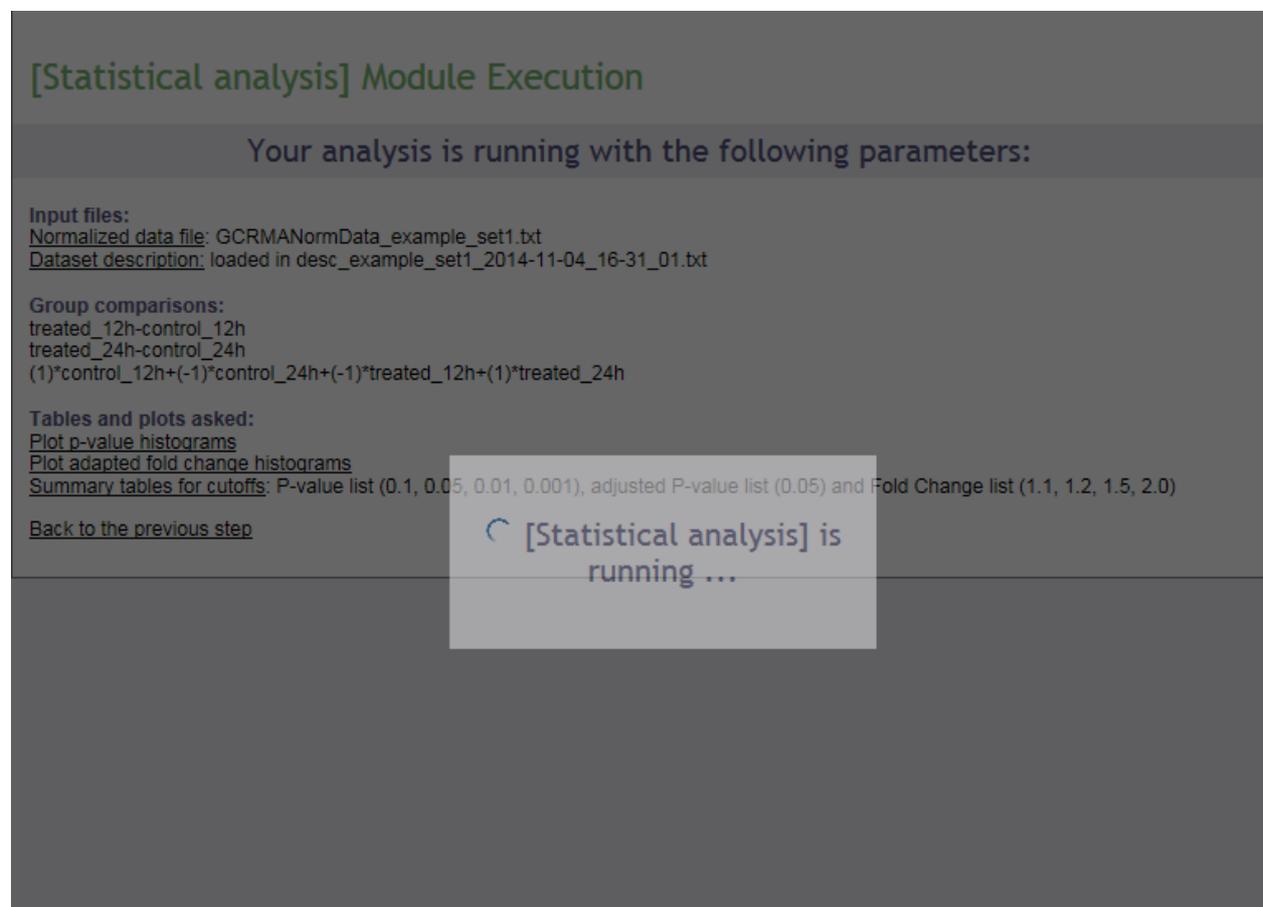
In significant genes list, you can define filters to select the genes to be added to a table of significantly changed genes and their statistical results. A table with p values and (log) fold changes for **all** genes is computed by default, and requires no ticking of a checkbox.

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.



[Statistical analysis] Module Execution

Your analysis is running with the following parameters:

Input files:
Normalized data file: GCRMANormData_example_set1.txt
Dataset description: loaded in desc_example_set1_2014-11-04_16-31_01.txt

Group comparisons:
treated_12h-control_12h
treated_24h-control_24h
(1)*control_12h+(-1)*control_24h+(-1)*treated_12h+(1)*treated_24h

Tables and plots asked:
Plot p-value histograms
Plot adapted fold change histograms
Summary tables for cutoffs: P-value list (0.1, 0.05, 0.01, 0.001), adjusted P-value list (0.05) and Fold Change list (1.1, 1.2, 1.5, 2.0)

[Back to the previous step](#)

[Statistical analysis] is running ...

RESULTS

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

[Statistical analysis] Module Execution

Your analysis is running with the following parameters:

Input files:

[Normalized data file](#): GCRMANormData_example_set1.txt
[Dataset description](#): loaded in desc_example_set1_2014-11-04_16-31_01.txt

Group comparisons:

treated_12h-control_12h
 treated_24h-control_24h
 (1)*control_12h+(-1)*control_24h+(-1)*treated_12h+(1)*treated_24h

Tables and plots asked:

[Plot p-value histograms](#)
[Plot adapted fold change histograms](#)
[Summary tables for cutoffs](#): P-value list (0.1, 0.05, 0.01, 0.001), adjusted P-value list (0.05) and Fold Change list (1.1, 1.2, 1.5, 2.0)

[Back to the previous step](#)

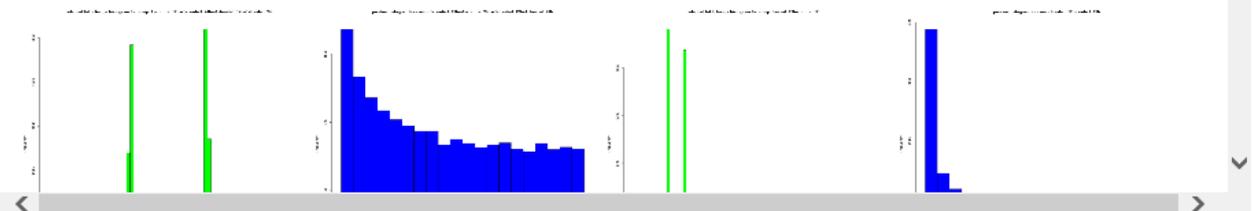
Results for example_set1:

Result files (Right click on the following link(s) to save the corresponding file)

[Open log file](#) containing standard output, warning and error messages from the execution.
 You may also consult this text file on the following section: Output message (STDOUT & STDERR).

[Open zip file](#) with result tables and images (png format). The images and the summary tables are displayed below.

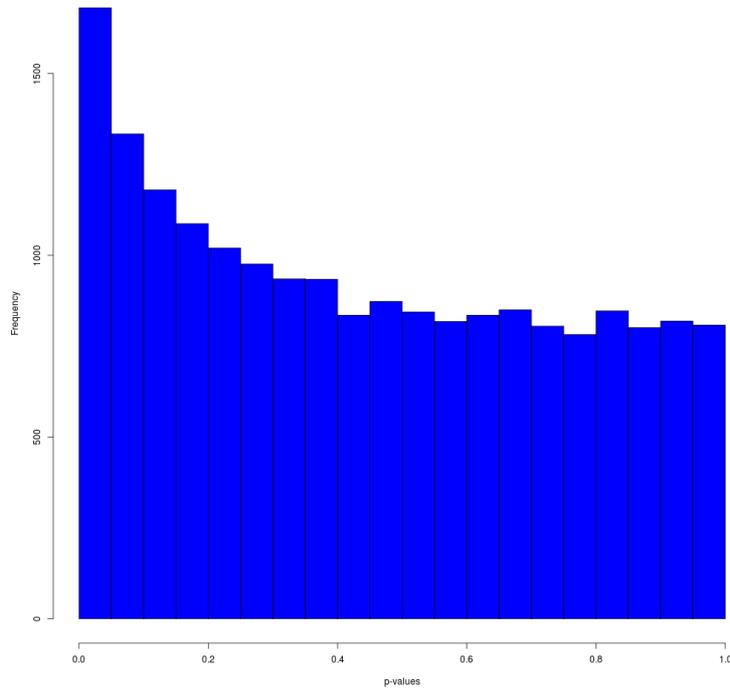
Fold change (green) and P-value (blue) histograms of each comparison:



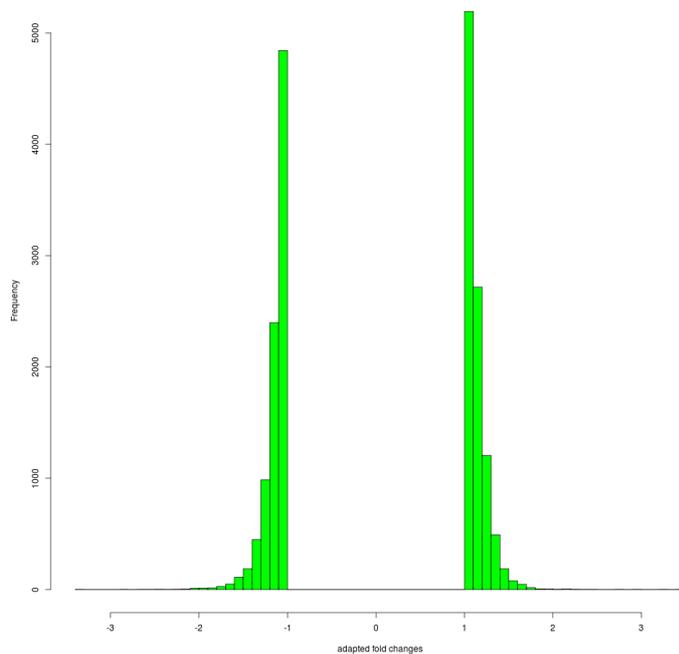
In the first part of the screen, your settings are recalled. Then links to the log file of the run and to the zip file containing all results (tables and images) are presented. Then, the p value and fold change histograms for each computed comparison are shown (clickable to enlarge, see below), if they 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 an 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.

No_of_genes: 19064

P.Values	Comparisons	pVal< 0.1 tot	pVal< 0.1 up	pVal< 0.1 down	pVal< 0.05 tot	pVal< 0.05 up	pVal< 0.05 down	pVal< 0.01 tot	pVal< 0.01 up	pVal< 0.01 down	pVal< 0.001 tot	pVal< 0.001 up	pVal< 0.001 down
Expected	Expected NoOfGenes	1906	953	953	953	476.5	476.5	191	95.5	95.5	19	9.5	9.5
	comp_treated_12h-control_12h	5321	2670	2651	3876	2020	1856	1947	1046	901	756	433	323
	comp_treated_24h-control_24h	6109	3241	2868	4577	2444	2133	2413	1285	1128	969	475	494
	comp_1xcontrol_12h+-1xcontrol_	3015	1539	1476	1681	858	823	438	180	258	51	24	27

Adj.pValues.html	Comparisons	Adj_pVal< 0.05 tot	Adj_pVal< 0.05 up	Adj_pVal< 0.05 down
	comp_treated_12h-control_12h	1184	663	521
	comp_treated_24h-control_24h	1819	961	858
	comp_1xcontrol_12h+-1xcontrol_	0	0	0

Fold.Changes.html	Comparison	FC >= 1.1 tot	FC >= 1.1 up	FC >= 1.1 down	FC >= 1.2 tot	FC >= 1.2 up	FC >= 1.2 down	FC >= 1.5 tot	FC >= 1.5 up	FC >= 1.5 down	FC >= 2 tot	FC >= 2 up	FC >= 2 down
	comp_treated_12h-control_12h	8415	4052	4363	3725	1903	1822	708	384	324	156	93	63
	comp_treated_24h-control_24h	9153	4742	4411	4450	2331	2119	892	433	459	218	101	117
	comp_1xcontrol_12h+-1xcontrol_	9030	4771	4259	3915	2054	1861	411	172	239	40	16	24

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.

Output message (STDOUT & STDERR):

Standard output:

```
[1] "Parameters have been registered"
Script run using R version 2.15.3 and affyAnalysisStat version_1.0.0
[1] "Libraries has been loaded"
[1] "Functions have been loaded"
[1] "Statistical analysis with limma"
--[[ Saving table for contrast comp_treated_12h-control_12h ]]-
--[[ Saving table for contrast comp_treated_24h-control_24h ]]-
--[[ Saving table for contrast comp_1xcontrol_12h+-1xcontrol_24h+-1xtreated_12h+1xtreated_24h ]]-
[1] "p-value histograms"
--[[ Saving comp_treated_12h-control_12h.txt P.Value_hist.png ]]-
--[[ Saving comp_treated_24h-control_24h.txt P.Value_hist.png ]]-
--[[ Saving comp_1xcontrol_12h+-1xcontrol_24h+-1xtreated_12h+1xtreated_24h.txt P.Value_hist.png ]]-
[1] "adapted fold change histograms"
--[[ Saving comp_treated_12h-control_12h.txt FC_hist.png ]]-
--[[ Saving comp_treated_24h-control_24h.txt FC_hist.png ]]-
--[[ Saving comp_1xcontrol_12h+-1xcontrol_24h+-1xtreated_12h+1xtreated_24h.txt FC_hist.png ]]-
[1] "Creating table of significant genes"
--[[ Saving Summary_tables.txt ]]-
```

Warning and error messages:

Pre-loaded package(s):
- R.utils

Attaching package: 'methods'

The following object(s) are masked from 'package:R.oo':

getClasses, getMethods

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 belows shows a representative screenshot of an example table.

	A	B	C	D	E	F	G	H	I	J
1		logFC	Fold Change	AveExpr	t	P.Value	adj.P.Val	B	external_gene_id	description
2	ENSG00000100292	1.23447926	2.35296402	10.51644773	14.12570319	1.56E-13	3.55E-09	19.32903714	HMOX1	heme oxygenase (decycling) 1
3	ENSG00000099194	0.842429507	1.793067137	11.98929734	10.66285654	7.24E-11	6.13E-07	14.26105048	SCD	stearoyl-CoA desaturase (delta)
4	ENSG00000165029	1.073488361	2.104515829	9.377451023	10.60619845	8.09E-11	6.13E-07	14.16502868	ABCA1	ATP-binding cassette, sub-fami
5	ENSG00000149485	0.653143703	1.572591218	11.61945781	10.05005768	2.46E-10	1.20E-06	13.19773248	FADS1	fatty acid desaturase 1 [Source:
6	ENSG00000163931	0.405766216	1.324792322	11.59050157	10.01620214	2.64E-10	1.20E-06	13.13737595	TKT	transketolase [Source:HGNC Sy
7	ENSG00000161011	0.3643114	1.287267069	12.37542277	9.407265275	9.36E-10	3.55E-06	12.02229691	SQSTM1	sequestosome 1 [Source:HGNC

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 \log_2 of the (regular) fold change, so the \log_2 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

3. ACKNOWLEDGMENTS

We would like to express our gratitude for using the open-access applications of ArrayAnalysis.org. This tutorial is derived from <http://www.arrayanalysis.org/> documentation originally written by Lars Eijssen and Anwasha Bohler.

The eNanoMapper project is funded by the European Union's Seventh Framework Program for research, technological development and demonstration (FP7-NMP-2013-SMALL-7) under grant agreement no. 604134.

4. REFERENCES

ArrayAnalysis homepage and web tools: <http://www.arrayanalysis.org/>

User-friendly solutions for microarray quality control and pre-processing on ArrayAnalysis.org. Eijssen LM, Jaillard M, Adriaens ME, Gaj S, de Groot PJ, Müller M, Evelo CT. Nucleic Acids Res. 2013 Apr 24. PMID: [23620278](https://pubmed.ncbi.nlm.nih.gov/23620278/) doi: 10.1093/nar/gkt293

Adding automated Statistical Analysis and Biological Evaluation modules to www.arrayanalysis.org. Master's thesis of Anwasha Bohler (Dutta) Advisor : Lars M.T. Eijssen, doi: 10.13140/2.1.5150.6244

5. KEYWORDS

Microarray data analysis

Statistics

Systems biology

Pathway and network analysis