ENM TUTORIALS

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

RELEASE DATE:	05/31/2016
USE:	How to use the Pathway module of ArrayAnalysis.org for pathway 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, Anwesha Bohler, Linda Rieswijk, Egon Willighagen, Penny Nymark
LICENCE:	CC-BY 4.0

TABLE OF CONTENTS

1. INTRODUCTION

PathVisio and ArrayAnalysis.org are open source, free to use online platforms for analysis of microarray data - and an alternative program for Chipster. This tutorial shows how to use the Path module (Pathway module, PathVisio webtool) of ArrayAnalysis which is designed for doing pathway analysis on microarray data. All source code has been written in R and is available at https://github.com/BiGCAT-UM/Path_Module.

This technical documentation has two main objectives:

- to guide you in the use of the Path module
- to give interpretative help on the outputs of the module

The Path module can be run :

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

The main functions of the Path module are:

- to import a dataset;
- to create a visualization;
- to calculate z-scores based on the criterion;
- to return a list of pathways sorted on the basis of z-scores.

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

- Using the on-line Path 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.

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

2. APPLICATION DETAILS

You can access the on-line module on arrayanalysis.org webportal: (follow "Get started" and choose "Statistical analysis"). You don't need to log in; you just need to prepare a gene level statistics file containing the statistical contrasts between the different groups of your Affymetrix .CEL files (you may also obtain the file by running the statistical analysis module).

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

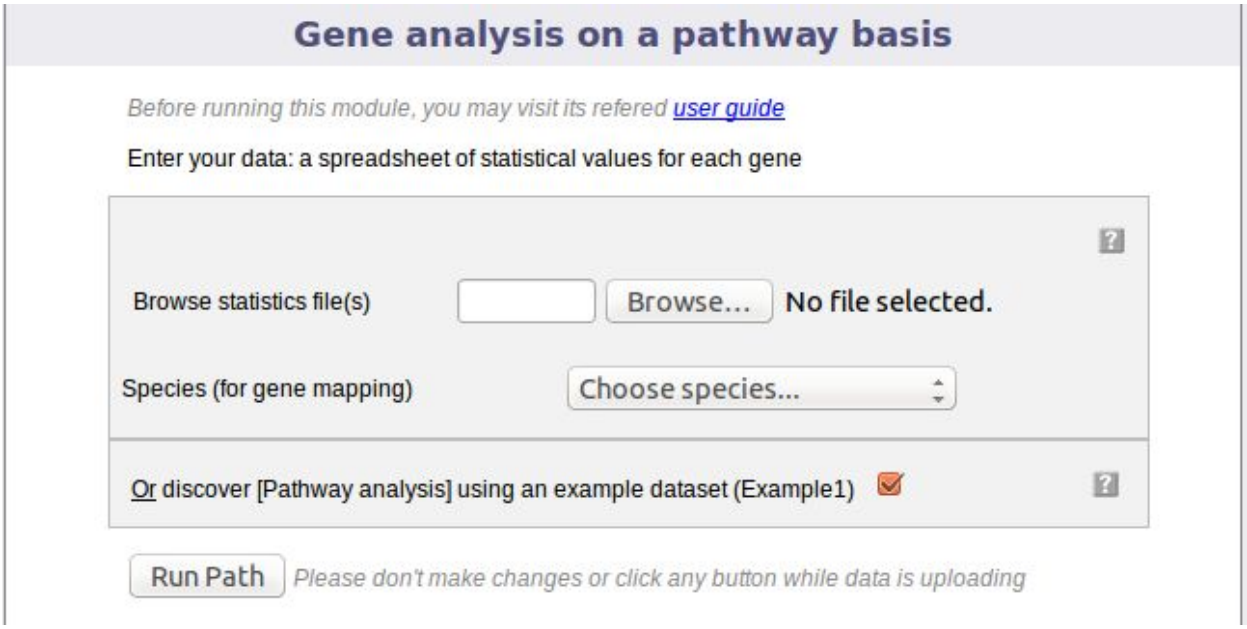
- Step1: First you load the gene level statistics file and select species. Alternatively, you might select the Example dataset for exploring the module. In that case you do not need to select the species. The dataset used is for Human. Click on Run Path, to proceed.
- Step2: Choose the column in the data file containing the identifiers and the database used for annotation if all the idenitifers are from the same database. If different identifier systems are used for annotating the dataset, the system code column has to be chosen. The system code specifies which database each idenitifier belongs to.
- Step3: Specify a criterion for calculating the z-scores.
- Step4: Select color criteria for visualization of the uploaded data on the pathways.

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 and select species

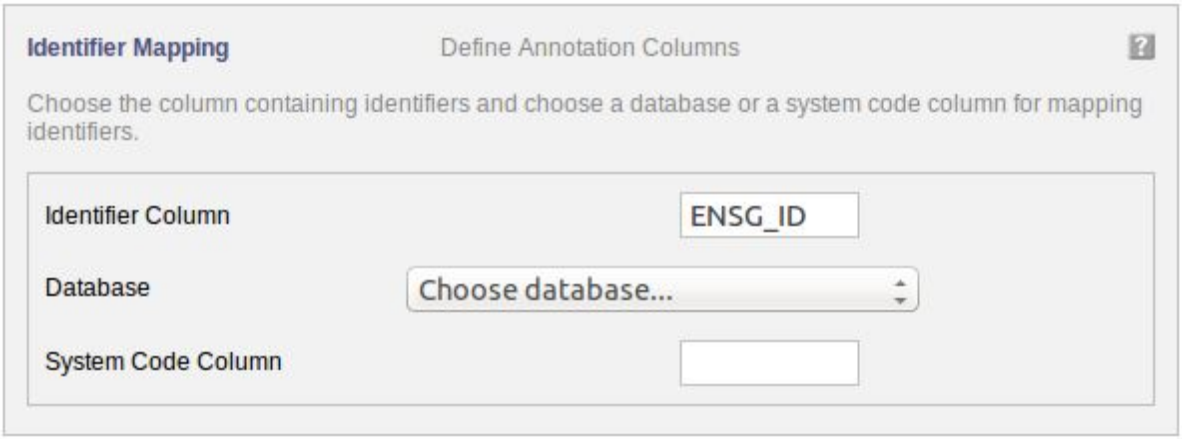
The following picture shows the screen for the first step:



This dialog allows you to upload a tab-delimited text file with (gene level statistics) data and choose the relevant species. Alternatively, the module can be run with an example data set, by ticking the checkbox presented. The interrogation mark button will give you contextual help.

Second step: Identifier Mapping

The following part of the online form is used for the second step:



Your dataset has been uploaded. For mapping the uploaded data to the pathways, the annotation information needs to be filled in.

"Identifier Column" Choose the column in the uploaded data file containing the identifiers used for annotation.

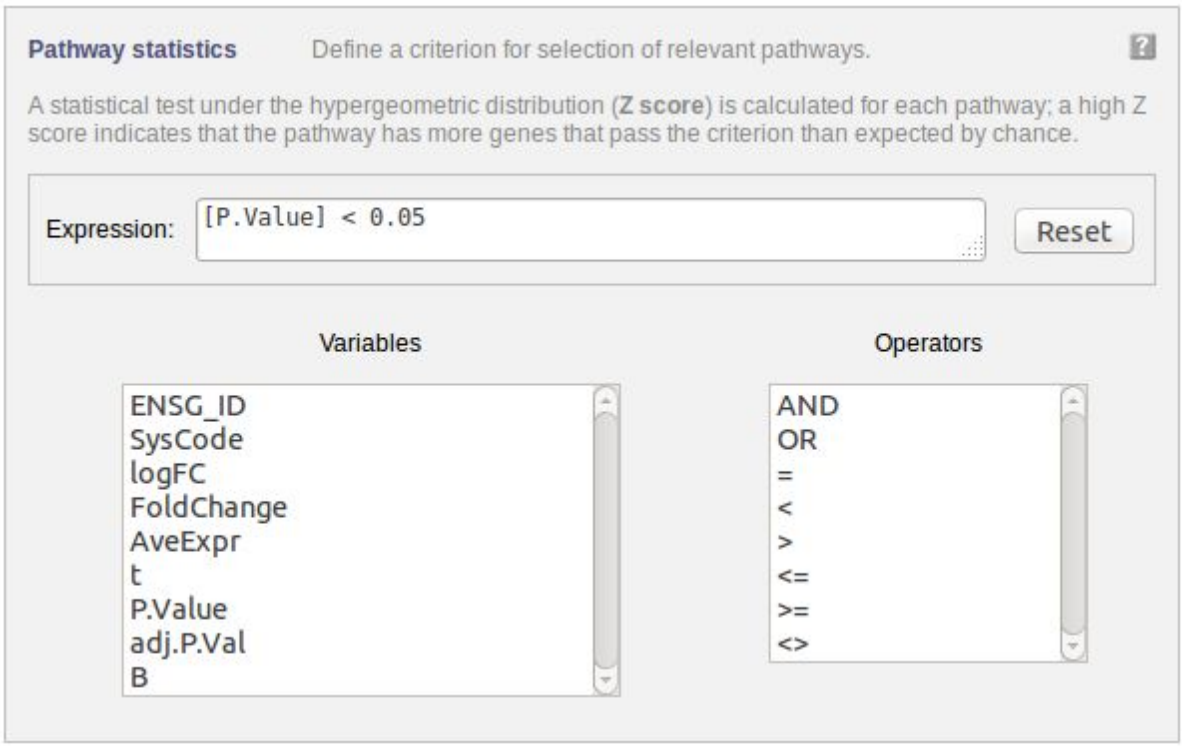
"Database" If the identifiers used for annotation are all from the same database, then select the database.

OR "System Code" If identifiers from different databases are used for annotation then the a column containing the system code of the databse needs to be selected.

The interrogation mark button will give you contextual help.

THIRD STEP: SET CRITERION FOR Z-SCORE CALCULATION

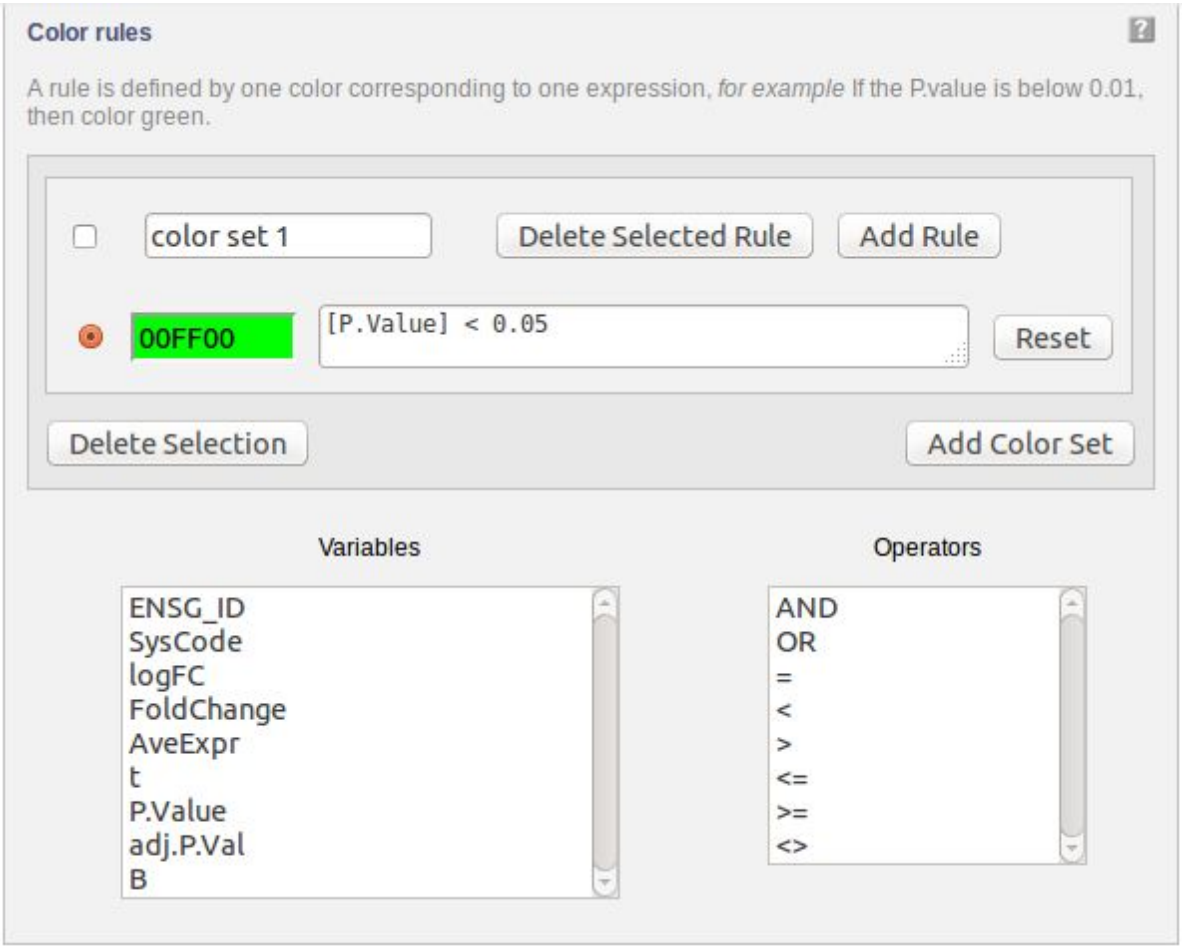
The following part of the online form is used for the third step:

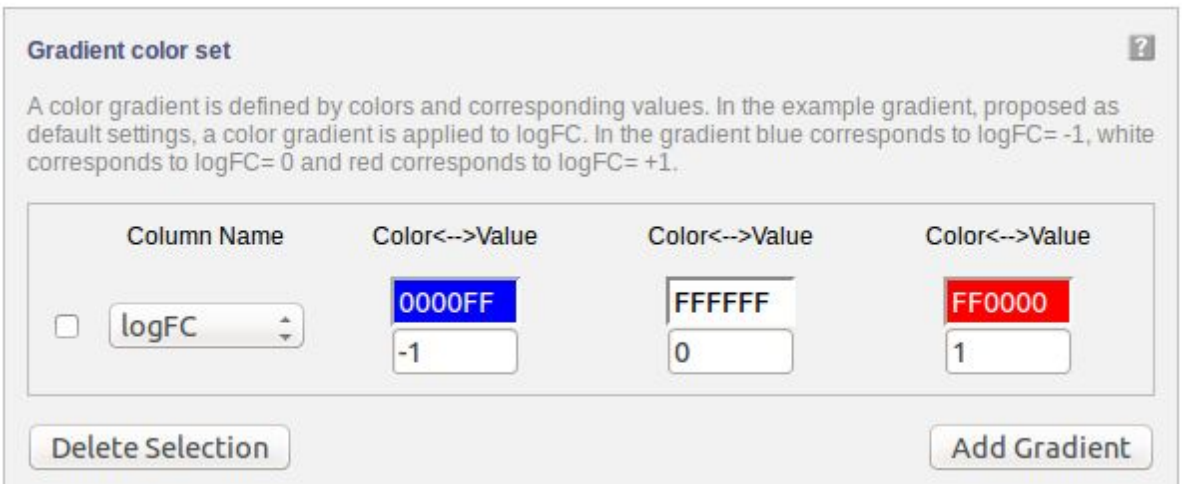


Select a criterion for calculating the z-score. You could, e.g. specify a criterion based on a fold change threshold. You can either type the expression in the "Expression" field or you can use the available parameters and operators listed by clicking on them.

Fourth step: Creating a visualization

The following part of the online form is used for the fourth step:

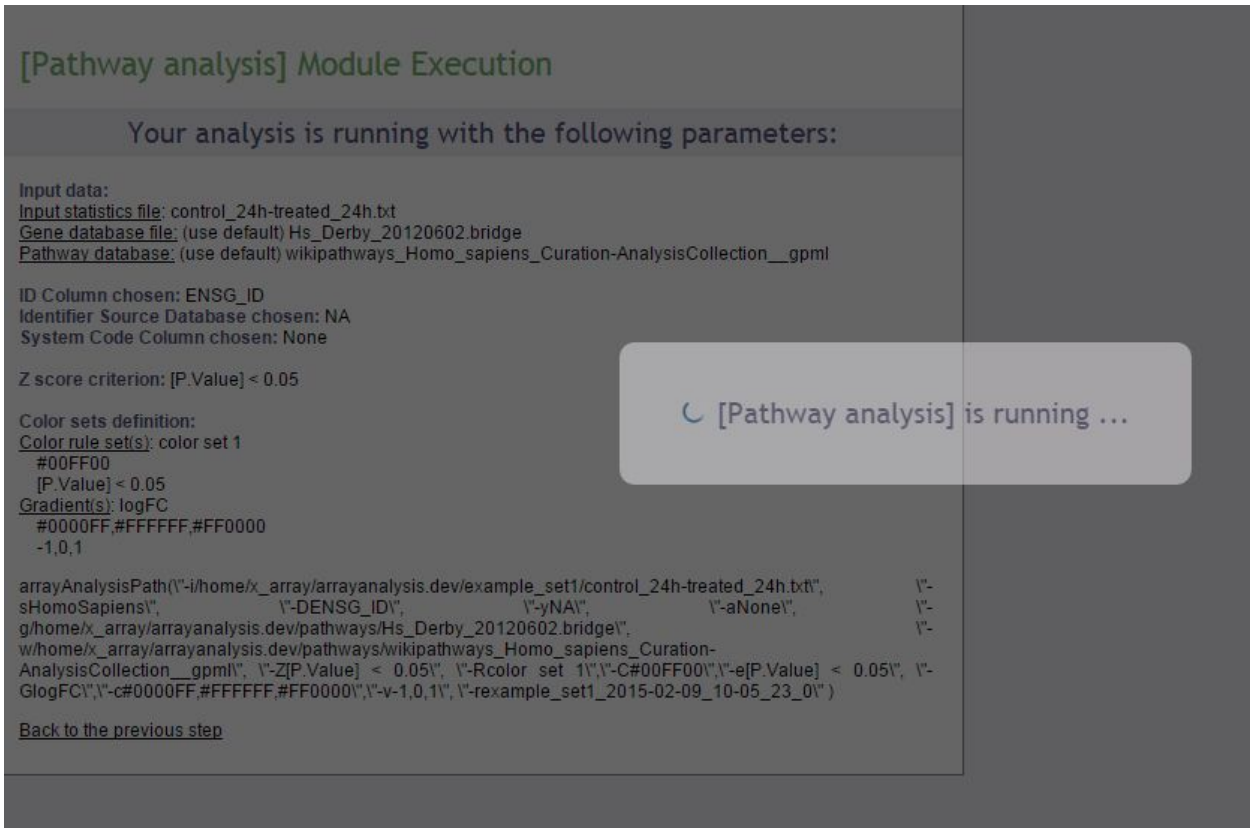




Data can be visualized on pathways using colours. A gradient colouring scheme can be used to visualize a range of data on a gene (e.g. fold change) while a rule can be applied for certain criteria allowing only the genes which qualify to be coloured (e.g. P Value "<" 0.05)

EXECUTION

After clicking 'Run' the module is executed.



RESULTS

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

[Pathway analysis] Module Execution

Your analysis is running with the following parameters:

Input data:
Input statistics file: control_24h-treated_24h.txt
Gene database file: (use default) Hs_Derby_20120602.bridge
Pathway database: (use default) wikipathways_Homo_sapiens_Curation-AnalysisCollection__gpml

ID Column chosen: ENSG_ID
Identifier Source Database chosen: En
System Code Column chosen: None

Z score criterion: [P.Value] < 0.05

Color sets definition:
Color rule set(s): color set 1
#00FF00
[P.Value] < 0.05
Gradient(s): logFC
#0000FF,#FFFFFF,#FF0000
-1,0,1

arrayAnalysisPath(\"-i/home/x_array/arrayanalysis.dev/example_set1/control_24h-treated_24h.txt\", \"-sHomoSapiens\", \"-DENSG_ID\", \"-yEn\", \"-aNone\", \"-g/home/x_array/arrayanalysis.dev/pathways/Hs_Derby_20120602.bridge\", \"-w/home/x_array/arrayanalysis.dev/pathways/wikipathways_Homo_sapiens_Curation-AnalysisCollection__gpml\", \"-Z[P.Value] < 0.05\", \"-Rcolor set 1\",\"-C#00FF00\",\"-e[P.Value] < 0.05\", \"-GlogFC\",\"-c#0000FF,#FFFFFF,#FF0000\",\"-v-1,0,1\", \"-rexample_set1_2015-02-09_10-10_44_0\")

Back to the previous step

```
cd /home/x_array/arrayanalysis.dev//Path_Module/
 Rscript /home/x_array/arrayanalysis.dev/temp/launchFilePath_example_set1_2015-02-09_10-10_44_0.R

cd /home/x_array/arrayanalysis.dev/temp/example_set1_2015-02-09_10-10_44_0_Path/
 zip -r pathways_results_example_set1_2015-02-09_10-10_44_0.zip contents index.html
```

Results for example_set1_2015-02-09_10-10_44_0:

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).

Download zip file with result html pages.

Result page Open the complete result page in a new tab or window

Summary tables for relevant pathways:

Powered by PathVisio

Gene/Protein found in pathways (n): 3252
Gene/Protein meeting criterion (r): 892

Twenty more significant pathways	n	r	z
Apoptosis Modulation and Signaling	76	35	3.68
Serotonin Receptor 2 -> ELK-SRF/GATA4 signaling	15	9	2.83
Cholesterol Biosynthesis	13	8	2.76
Serotonin Receptor 4/6/7 -> NR3C signaling	17	9	2.36
mRNA processing	116	42	2.16
Proteasome Degradation	58	23	2.11
FAS pathway and Stress induction of HSP regulation	35	15	2.06
Eukaryotic Transcription Initiation	35	15	2.06
MAPK signaling pathway	152	52	1.92
Fatty Acid Biosynthesis	22	10	1.90
IL-6 Signaling Pathway	92	33	1.84
Translation Factors	40	16	1.79
Apoptosis	80	29	1.79

IL-6 Signaling Pathway	92	33	1.84
Translation Factors	40	16	1.79
Apoptosis	80	29	1.79
Pentose Phosphate Pathway	7	4	1.76
Acetylcholine Synthesis	7	4	1.76
Keap1-Nrf2	12	6	1.76
Cell cycle	84	30	1.72
Oxidative Stress	26	11	1.71
Folic Acid Network	26	11	1.71
Fluoropyrimidine Activity	29	12	1.69

more pathways...

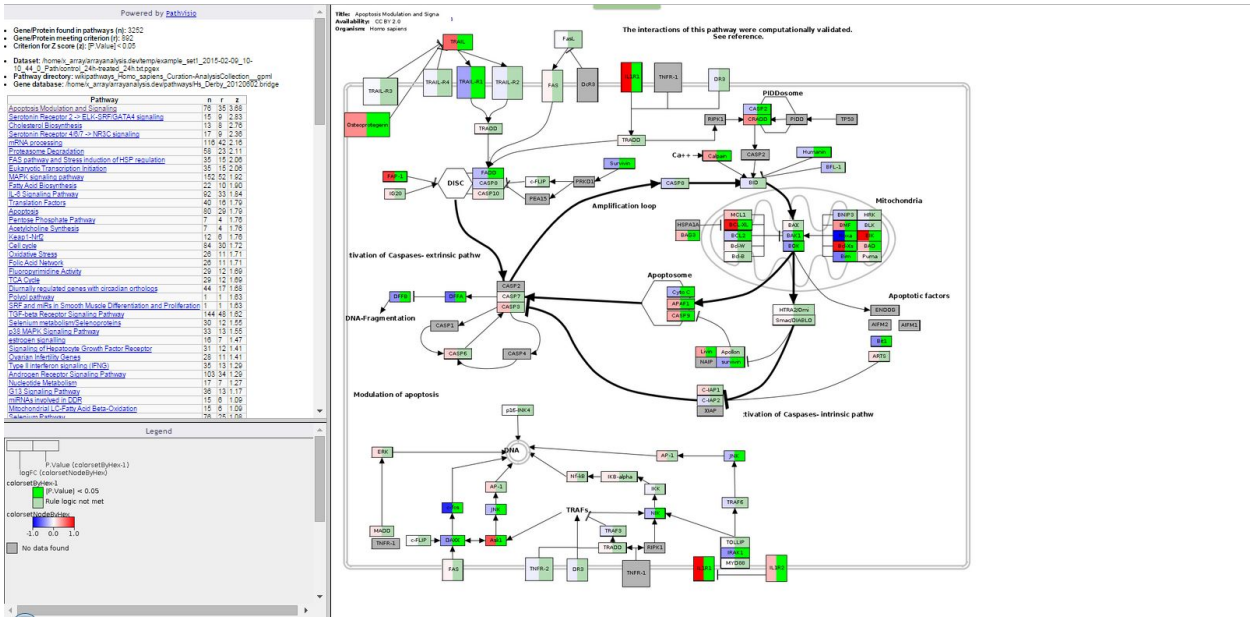
Output message (STDOUT & STDERR):

Standard output:

```
[1] "Parameters have been registered"
[1] "Libraries are loaded"
[1] "Importing Data ...."
[1] "/home/x_array/arrayanalysis.dev/temp/example_set1_2015-02-09_10-10_44_0_Path/control_24h-treated_24h.txt.pgexfile created!"
[1] "Creating visualisation..."
[1] "/home/x_array/arrayanalysis.dev/temp/example_set1_2015-02-09_10-10_44_0_Path/control_24h-treated_24h.txt.pgex.xml-visualization file created!"
[1] "Calculating Pathway Statistics and Exporting results..."
[1] "Pathway Collection used: /home/x_array/arrayanalysis.dev/pathways/wikipathways_Homo_sapiens_Curation-AnalysisCollection__gpml/"
[1] "Gene Identifier Mapping Database Used: /home/x_array/arrayanalysis.dev/pathways/Hs_Derby_20120602.bridge"
[1] "Dataset Imported: /home/x_array/arrayanalysis.dev/temp/example_set1_2015-02-09_10-10_44_0_Path/control_24h-treated_24h.txt.pgex"
[1] "Z score criterion: [P.Value] < 0.05"
[1] "Result Directory on Calculation server /home/x_array/arrayanalysis.dev/temp/example_set1_2015-02-09_10-10_44_0_Path"
[1] "Run completed"
```

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 (index file, pathway images, and related backpages) are presented. The results will be described in the next section of this documentation.

INTERPRET THE RESULTS OF THE PATH MODULE



The output consists of :
An index file in html format, which contains:

- The number of genes/proteins in the dat aset found in the pathways of the pathway collection.
- The number of genes/proteins out of the above that meet the user defined criterion for z-score calculation.
- The criterio used for z-score calculation.
- The dataset used.
- The pathway directory used.
- The identifier mapping database used.
- A clickable list of all the pathways on which the genes of interest have been visualized sorted on the basis of z-scores.
- The file also contains the legend to better understand the visualization.

A contents folder: contains the backpages for all the pathways containing gene expression information, the legend file, the statistics file and all the colour coded pathway images.

WikiPathways provides a portal for nanomaterial relevant pathway information:
http://www.wikipathways.org/index.php/Portal:Nanomaterials

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

4. REFERENCES

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 Anwesha Bohler.

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 doi: 10.1093/nar/gkt293

Adding automated Statistical Analysis and Biological Evaluation modules to www.arrayanalysis.org. Master's thesis of Anwesha Bohler (Dutta) Advisor : Lars M.T. Eijssen, doi: 10.13140/2.1.5150.6244
PathVisio 3: An Extendable Pathway Analysis Toolbox, Kutmon M, van Iersel MP, Bohler A, Kelder T, Nunes N, Pico AR, and Evelo, CT, PLoS Computational Biology (2015) 11(2): e1004085, doi: 10.1371/journal.pcbi.1004085, pubmed: 25706687

Presenting and exploring biological pathways with PathVisio, van Iersel MP, Kelder T, Pico AR, Hanspers K, Coort S, Conklin BR, and Evelo C., BMC Bioinformatics. 2008 Sep 25;9:399. doi: 10.1186/1471-2105-9-399, pubmed: 18817533

5. KEYWORDS

Microarray data analysis

Statistics

Systems biology

Pathway and network analyis