



ENM TUTORIALS

How to Use Chipster for Bioinformatics Analysis of Nanomaterial-Based Omics Data

Deliverable Dx-y

RELEASE DATE:	24 th May, 2016
USE:	How to Use Chipster for Bioinformatics Analysis of Nanomaterial-Based Omics Data
VERSION:	V.1.0.
MAIN AUTHOR:	Penny Nymark
PARTNER:	UM
CONTACT DETAILS:	penny.nymark@ki.se
AUTHORS:	Penny Nymark, Friederike Ehrhart, Egon Willighagen, Eija Korpelainen, Pekka Kohonen, Roland Grafström
LICENCE:	CC-BY 4.0



TABLE OF CONTENTS

[1. INTRODUCTION](#)

[2. APPLICATION DETAILS](#)

[Example Data set](#)

[Open Chipster](#)

[Quality control](#)

[Normalization](#)

[Phenodata \(Metadata\)](#)

[Pre-processing and Filtering](#)

[Clustering and Visualization](#)

[Statistical analysis on a Subset of the Samples](#)

[Pathway and Gene Ontology Enrichment analysis](#)

[Other useful functions](#)

[Generate a Data Matrix and Exporting data from Chipster](#)

[Further Recommendations and examples of Analyses in PathVisio](#)

[A Selection of Databases and Tools Relevant for Nanosafety-Related Bioinformatics](#)

[4. ACKNOWLEDGMENTS](#)

[5. REFERENCES](#)

[6. KEYWORDS](#)

1. INTRODUCTION

This tutorial will get you started on the basic use of the bioinformatics tool Chipster, an alternative to ArrayAnalysis.org described in other [eNM tutorials](#). It will guide you through the import of microarray data (Affymetrix .CEL files), quality checks, normalization, pre-processing, basic statistical analyses methods and visualization of your data. Chipster is not specific to the nanosafety field, but this tutorial will be focused on the specific needs of nanotoxicological research.

Chipster is an open source user-friendly analysis software for high-throughput data analysis. It offers over 350 bioinformatics tools and it is constantly updated according to the latest state-of-the-art tools and scripts. Users can analyse and visualize data interactively, and share complete analysis sessions and automatic workflows with colleagues. Chipster supports the analysis of over 120 different microarray platforms, including the most common types by Affymetrix, Agilent and Illumina. In addition, Chipster has extensive functions and tools for the analysis of NGS data.

For additional and deeper guidance on how to use Chipster, beyond this tutorial, see the [website](#) with manuals and tutorials or contact the [Chipster team](#). The website also includes guidance on how to analyze next-generation sequencing (NGS) data in Chipster.

Chipster's client software uses Java Web Start to install itself automatically and connects to computing servers to perform the actual analyses. Chipster has been installed on several servers around the world. If you would like to use Chipster, you can ask the IT people in your institute to set up a Chipster server for you. This is easy because Chipster is packaged as a [virtual machine image \(link\)](#). If your institute doesn't have the required computer hardware, Chipster server can be also installed (free of charge) in the [EGI infrastructure](#) for European researchers. EGI and ELIXIR are currently planning to set up also a ready-made Chipster server for end users, so in the future you will not need a local admin person to help you. You can also contact one of the following:

- **Finnish users:** [CSC](#)
- **Swedish users:** [UPPMAX](#)
- **Dutch users:** the [Dutch TechCentre for Life Sciences](#)
- **German users:** [DKFZ](#)
- **International users:** [EGI](#)

A demo username and password is available for testing e.g. [pre-made sessions](#) (see also below), but this account does not allow running analyses, (username and password: *guest*).

Courses, manuals and tutorials are available through:

- <http://chipster.csc.fi/>
- <https://www.youtube.com/channel/UCnL-Lx5gGIW01OkskZL7JEQ>

- <http://www.elixir-finland.org/next-generation-sequencing-data-analysis-with-chipster/>

Please also note that the program itself provides specific descriptions of each tool through the “More help” button in the upper right corner within the program itself.

2. APPLICATION DETAILS

EXAMPLE DATA SET

mRNA data (transcriptomics)

- Raw mRNA data – Affymetrix .CEL files
- Array type: Affymetrix HT HG-U133a 2.0 (=hgu133av2 in Chipster)
- GEO accession number: [GSE42067](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE42067)

Experiment set-up

- Human small airway epithelial cells (SAE) exposed to 10 and 100 µg/ml titanium dioxide nanobelts (TiO₂) and Cheaptubes multi-walled carbon nanotubes (MWCNT) for 1h and 24h (3 replicates)
 - o 24 samples
- Unexposed SAE cells (3 replicates)
 - o 6 samples

A [pre-made Chipster analysis](#) session based on the data presented above is available for exploration in Chipster under the ‘Open example session’ option ([microarray_Affymetrix_Nanosafety_Basel_Workshop.zip](#)) and a specific tutorial for the pre-made analysis session can be found [here](#).

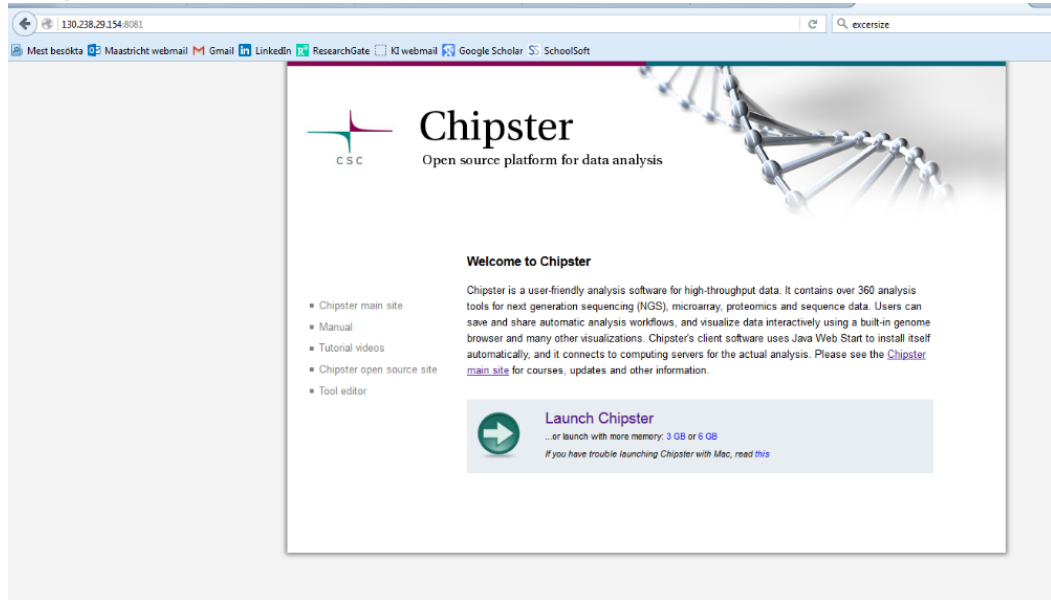
If you have access to Chipster on one of the European servers with your own username and password (and thus are able to run analysis tools), then download the data described above to your computer. Download the .CEL files and the Series Matrix File containing the descriptions of the data.

Physico-chemical data for the nanomaterials in the example data set (as described in the original [publication](#)) as well as links to the omics data will soon be available in the [eNanoMapper database](#).

OPEN CHIPSTER

1. Open Chipster through the preferred server (see details above).
 - a. Open a browser (*Chrome* or *Firefox*)

b. Navigate to the server URL



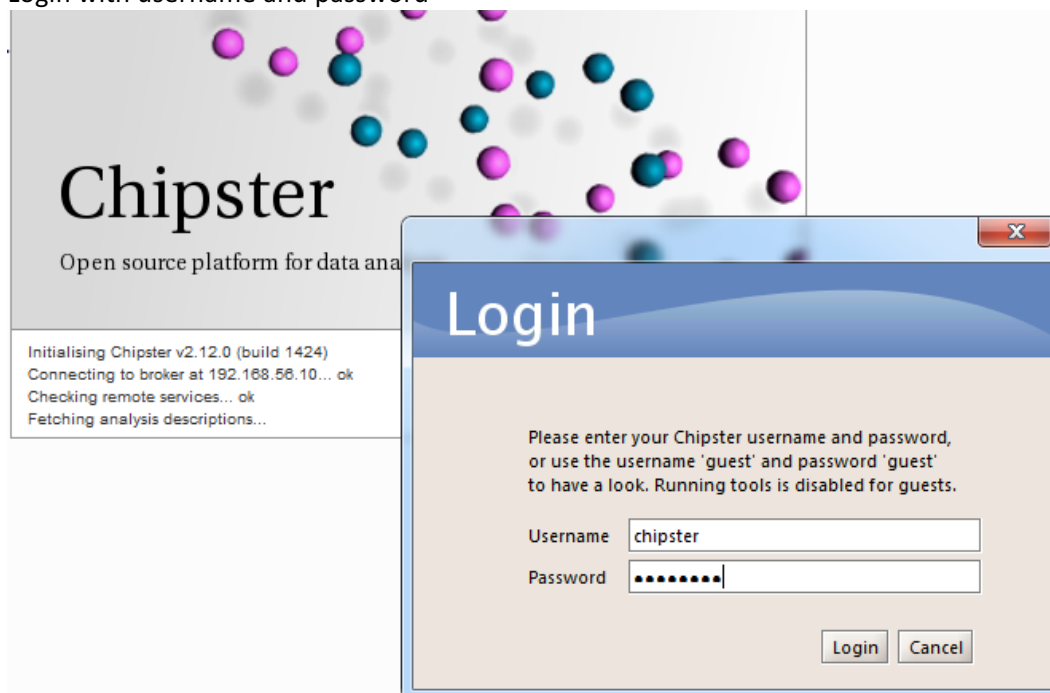
c. Click on **Launch Chipster**

d. The file **chipster.jnlp** begins to download. Keep this file and open it (Chrome) or open it directly (Firefox)

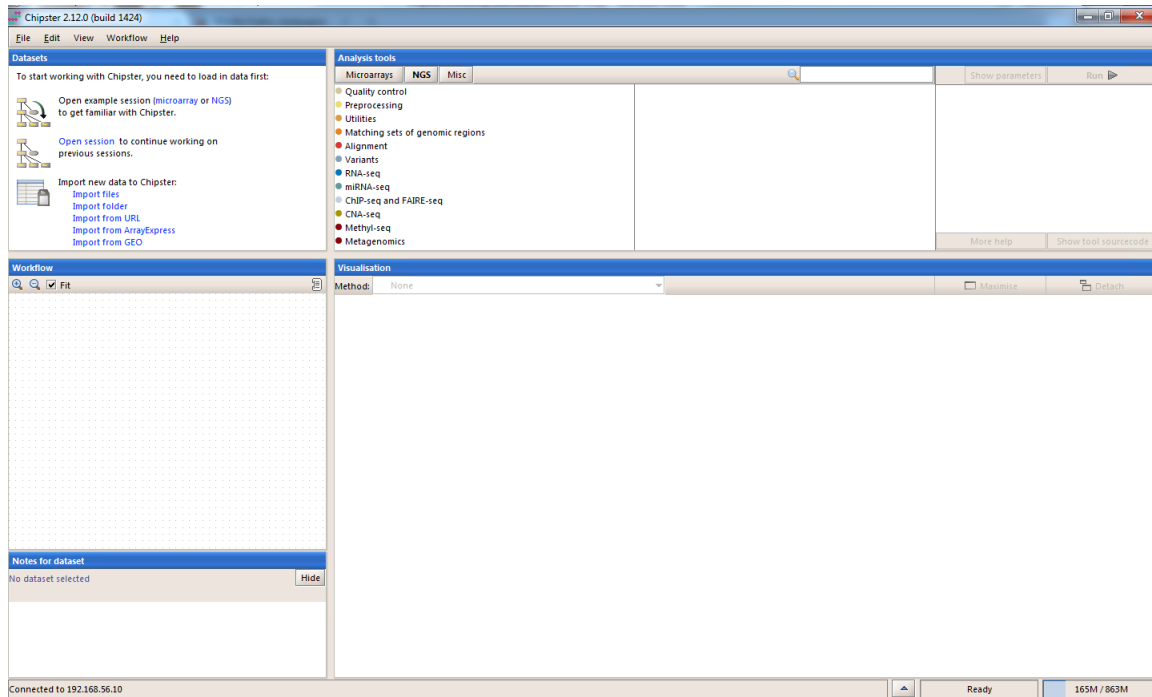
e. A security message appears. Click **Run**



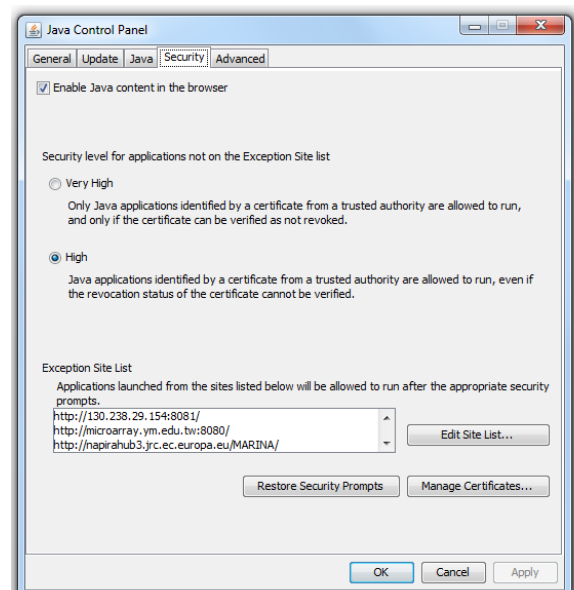
- f. Login with username and password



- g. The Chipster main window appears. There are 4 different sections which are used during analysis to keep track of what's being done.
- The **Datasets** section listing all the files in the analysis.
 - The **Analysis tools** section where different tools are listed and can be chosen for analysis.
 - The **Workflow** section which shows you the relationships between the different files.
 - The **Visualization** section in which analysis results can be visualized in various ways.



Note! If Java prompts a security message, the Chipster server URL needs to be added as an Exception Site in the Java security settings.



2. In the upper left section (Datasets section) click **Import files** and choose all the Affymetrix CEL files that you have downloaded to your computer. They will be imported automatically.

Tip! Data can also be retrieved directly from GEO or ArrayExpress under Utilities – Import data from GEO/ArrayExpress.

QUALITY CONTROL

1. Click once on the group of imported raw Affymetrix .CEL files in the lower left section (Workflow section).
2. Under the **Microarrays** tab in the upper right section (Analysis tools section) click on **Quality control**, then **Affymetrix basic** and then **Run** the tool (upper right corner).

Tip! For every tool in Chipster you can get more information on what the tool does in the box in the upper right corner of the Analysis tools section. In addition, you can click the “More help” button to get yet more information and references for the tools.

3. While the tool is running choose the tool **Affymetrix – using RLE and NUSE**, click **Show parameters** and choose for **Custom chip** the correct chip type (ending with ‘hgu133av2’)
4. Check the Quality Control (QC) files (RLE, NUSE, RNA-degradation, spike-in and simpleaffy plots)

Tip! Refer to the Chipster manual for [descriptions of QC plots](#). Also [ArrayAnalysis.org](#) provides a very extensive QC, please refer to the tutorial ‘[How to use the AffyQC web tool of ArrayAnalysis.org for quality control and pre-processing of Affymetrix microarray data](#)’

NORMALIZATION

1. Click once on the group of imported raw Affymetrix .CEL files in the Workflow section.
2. In the Analysis tools section, choose **Normalization** and then **Affymetrix**. Click **Show parameters** - leave the first two parameters at the default, but change the last one **Custom CDF annotation to be used** to the correct array type (hgu133av2). Click **Run**.

Note! The resulting file contains the matrix of normalized data, which is useful for further analysis either in Chipster or in other tools. The file can be exported by right-clicking on the file either in the Datasets or in the Workflow section. The file is then saved as a .tsv file, which can be opened in Excel. Please note that Excel shifts the header one step “backwards”, because the first column does not have a header. So, when you open the file in Excel, please be sure to shift the header forward one step to correspond to the correct column.

The .tsv file can also be converted into a .txt file by either changing the extension manually or opening in Excel (correcting the header) and saving as a .txt file.

Tip! Similarly a text file (e.g. containing normalized data or a list of identifiers of interest) can be converted into a .tsv file readable by Chipster by changing the extension. However, please be sure to have the correct [format](#).

PHENODATA (METADATA)

1. Double click on the oval shaped **Phenodata file** in the Workflow section. Add data in the columns according to the information in GEO or in the downloaded Series Matrix File that you obtained from GEO.

In the group column add information as numerical values (e.g. control=1, exposed samples=2, 3, 4 etc.). Be sure to always use '1' for controls otherwise the analyses will not be performed correctly.

In the description column change the sample number (which was derived directly from the .CEL files) to a more descriptive name of the sample, e.g. from the Series Matrix File.

TIP! If you have only a few samples you can just add the information in the phenodata columns manually. If you have a lot of samples with a lot of parameters it may be convenient to export the phenodata file (right-click the phenodata icon in the Workflow section and export) and add the data in Excel, e.g. through VLOOKUP from another descriptor file related to the data. This file can then be imported again into Chipster and linked to your data set (right-click your data file and choose 'link to phenodata xx.tsv'). Another way is to just copy paste data from an Excel file (e.g. the Series Matrix File from GEO) – but then you should be completely sure the samples are in the same order!!

2. Add columns with additional parameters to your phenodata file by clicking on the right of the Visualization section. Add columns for NM, Time, Dose and Replicate according to the following figure if you are using the example data.

Phenodata editor									
sample	original_name	chiptype	group	description	NM	Time	Dose	Repl	
microarray001.cel	GSM1031905...	hgu133a2hs...	1	SAE_CON_0_1hr-1	cntrl	1	0	1	
microarray002.cel	GSM1031906...	hgu133a2hs...	1	SAE_CON_0_1hr-2	cntrl	1	0	2	
microarray003.cel	GSM1031907...	hgu133a2hs...	1	SAE_CON_0_1hr-3	cntrl	1	0	3	
microarray004.cel	GSM1031908...	hgu133a2hs...	2	SAE_MWCNT_10_1hr-1	MWCNT	1	10	1	
microarray005.cel	GSM1031909...	hgu133a2hs...	2	SAE_MWCNT_10_1hr-2	MWCNT	1	10	2	
microarray006.cel	GSM1031910...	hgu133a2hs...	2	SAE_MWCNT_10_1hr-3	MWCNT	1	10	3	
microarray007.cel	GSM1031911...	hgu133a2hs...	2	SAE_MWCNT_100_1hr-1	MWCNT	1	100	1	
microarray008.cel	GSM1031912...	hgu133a2hs...	2	SAE_MWCNT_100_1hr-2	MWCNT	1	100	2	
microarray009.cel	GSM1031913...	hgu133a2hs...	2	SAE_MWCNT_100_1hr-3	MWCNT	1	100	3	
microarray010.cel	GSM1031914...	hgu133a2hs...	3	SAE_TiO2_10_1hr-1	TiO2	1	10	1	
microarray011.cel	GSM1031915...	hgu133a2hs...	3	SAE_TiO2_10_1hr-2	TiO2	1	10	2	
microarray012.cel	GSM1031916...	hgu133a2hs...	3	SAE_TiO2_10_1hr-3	TiO2	1	10	3	
microarray013.cel	GSM1031917...	hgu133a2hs...	3	SAE_TiO2_100_1hr-1	TiO2	1	100	1	
microarray014.cel	GSM1031918...	hgu133a2hs...	3	SAE_TiO2_100_1hr-2	TiO2	1	100	2	
microarray015.cel	GSM1031919...	hgu133a2hs...	3	SAE_TiO2_100_1hr-3	TiO2	1	100	3	
microarray016.cel	GSM1031920...	hgu133a2hs...	1	SAE_CON_0_24hr-1	cntrl	24	0	1	
microarray017.cel	GSM1031921...	hgu133a2hs...	1	SAE_CON_0_24hr-2	cntrl	24	0	2	
microarray018.cel	GSM1031922...	hgu133a2hs...	1	SAE_CON_0_24hr-3	cntrl	24	0	3	
microarray019.cel	GSM1031923...	hgu133a2hs...	2	SAE_MWCNT_10_24hr-1	MWCNT	24	10	1	
microarray020.cel	GSM1031924...	hgu133a2hs...	2	SAE_MWCNT_10_24hr-2	MWCNT	24	10	2	
microarray021.cel	GSM1031925...	hgu133a2hs...	2	SAE_MWCNT_10_24hr-3	MWCNT	24	10	3	
microarray022.cel	GSM1031926...	hgu133a2hs...	2	SAE_MWCNT_100_24hr-1	MWCNT	24	100	1	
microarray023.cel	GSM1031927...	hgu133a2hs...	2	SAE_MWCNT_100_24hr-2	MWCNT	24	100	2	
microarray024.cel	GSM1031928...	hgu133a2hs...	2	SAE_MWCNT_100_24hr-3	MWCNT	24	100	3	
microarray025.cel	GSM1031929...	hgu133a2hs...	3	SAE_TiO2_10_24hr-1	TiO2	24	10	1	
microarray026.cel	GSM1031930...	hgu133a2hs...	3	SAE_TiO2_10_24hr-2	TiO2	24	10	2	
microarray027.cel	GSM1031931...	hgu133a2hs...	3	SAE_TiO2_10_24hr-3	TiO2	24	10	3	
microarray028.cel	GSM1031932...	hgu133a2hs...	3	SAE_TiO2_100_24hr-1	TiO2	24	100	1	
microarray029.cel	GSM1031933...	hgu133a2hs...	3	SAE_TiO2_100_24hr-2	TiO2	24	100	2	
microarray030.cel	GSM1031934...	hgu133a2hs...	3	SAE_TiO2_100_24hr-3	TiO2	24	100	3	

- In addition, add columns for information on which samples belong to certain groups that are later to be analysed together, e.g. specific controls vs. their exposed counterparts. Add columns for MWCNT-samples, TiO₂-samples, Dose100-samples, Dose10-samples etc. Give the samples that belong to a group the number '1' and the other samples '0' in each column, as shown in the figure below.

MWCNT	TiO2	Dose10	Dose100	1h	24h
1	1	1	1	1	0
1	1	1	1	1	0
1	1	1	1	1	0
1	0	1	0	1	0
1	0	1	0	1	0
1	0	1	0	1	0
1	0	0	1	1	0
1	0	0	1	1	0
1	0	0	1	1	0
0	1	1	0	1	0
0	1	1	0	1	0
0	1	1	0	1	0
0	1	0	1	1	0
0	1	0	1	1	0
0	1	0	1	1	0
1	1	1	1	0	1
1	1	1	1	0	1
1	1	1	1	0	1
1	0	1	0	0	1
1	0	1	0	0	1
1	0	1	0	0	1
1	0	0	1	0	1
1	0	0	1	0	1
1	0	0	1	0	1
0	1	1	0	0	1
0	1	1	0	0	1
0	1	1	0	0	1
0	1	0	1	0	1
0	1	0	1	0	1
0	1	0	1	0	1

At this point it is wise to save you session: File - save session!

PRE-PROCESSING AND FILTERING

Microarray data contains a lot of noise of both biological and technical background and before performing statistical analyses to identify differentially expressed genes or pathways, it is necessary to pre-process and filter the data, leaving out e.g. bad quality and invariant information, such as unaffected genes across all samples. Chipster provides several advanced tools for pre-processing and filtering and the advantage is that the tools can be run in parallel and results can be compared to find the best approach.

1. Click once on the normalized file (normalized.tsv) containing log₂ transformed intensity values.

Tip! You can change the name of the file by clicking on the name in the Visualization section (after having chosen the file in the Datasets or Workflow section). Be sure, however, to keep the .tsv extension and do not introduce spaces in the file name.

2. Under the **Preprocessing** tools, choose **Filter by expression**. Click **Show parameters** and change the **Underexpressed cut-off** to 6.0 (this is a commonly agreed upon gold standard cut-off often used for Affymetrix arrays). Leave the rest of the parameters at the default. Click **Run**.
Double click the resulting file and have a look at it to have an idea of what you are working with.

Tip! The resulting file is useful for example for Gene Set Enrichment Analysis (GSEA), which will not be discussed in this tutorial, but can be performed in Chipster under Pathways – Gene set test, for example for Gene Ontologies and KEGG pathways. Other GSEA tools for more custom built testing can be done for example with the [GSEA tool](#) developed at the Broad Institute.

CLUSTERING AND VISUALIZATION

Chipster offers a wide variety of different clustering and visualization tools, which are beneficial at different stages of the analysis workflow. They can be used for checking the quality of the filtered data and to compare the results of different filtering approaches. Different types of arrays require different types of filtering and it is often good to check whether the filtering approaches have been efficient in removing noise.

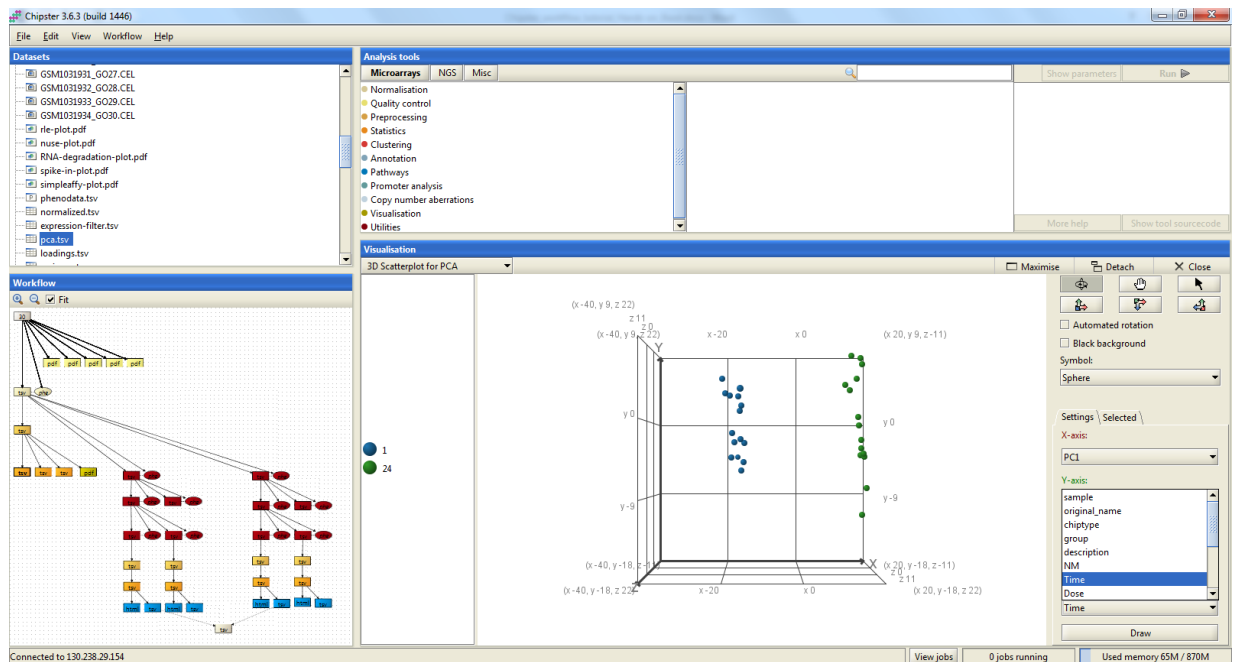
Furthermore, outlier samples are not necessarily noted during the initial quality check which is more focused on the technical aspects of the quality. Clustering and visualization of the normalized filtered data may reveal such outliers.

Finally clustering and visualization at this stage also allows for initial hypothesis generation and an overall feeling for the data.

Principle Component Analysis (PCA)

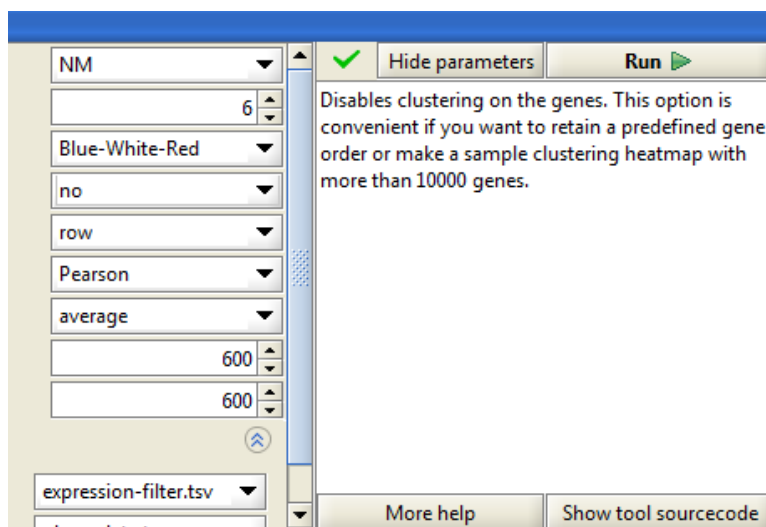
1. Click once on the filtered data file (expression-filter.tsv). Under **Statistics** in the Analysis tools section, choose **PCA**. With default parameters, click **Run**.
2. You will get three files containing information on the PCA analysis. Click once on the first file (pca.tsv). Then, in the Visualization section, open it with the **3D scatter plot for PCA** option.
3. Recolor the data according to different phenodata parameters (e.g. 'time') in the scroll-down menu on the right in the Visualization section (see example Figure below). Try out coloring

according to different phenodata information and draw some conclusions about the data based on what you see!



Hierarchical clustering

1. Click once on the filtered data (expression-filter.tsv). Under **Visualization** in the Analysis tools section, choose **Annotated heatmap**. Change the parameters according to the figure below. Click **Run**. This tool can be run at the same time as the PCA tool to save time.



2. Double click the resulting PDF to view the heatmap and see if you can draw any conclusions based on the way the samples/genes cluster.

Tip! The PCA plot and the hierarchical clustering can be used at different stages of the analysis to keep track of what is being done with the data and to see the quality of different analysis steps. Performing PCA or hierarchical clustering on unfiltered and filtered data, or on only differentially expressed genes, can provide different views of the data and offer a basis for drawing conclusions for further analysis.

At this point it is wise to save you session again: File - save session! The more you save in between the less you lose if the computer, the internet or the Chipster server crashes!

STATISTICAL ANALYSIS ON A SUBSET OF THE SAMPLES

To perform a statistical analysis on a subset of samples with e.g. control samples and corresponding treatment samples, these samples need to be extracted from the bulk of samples.

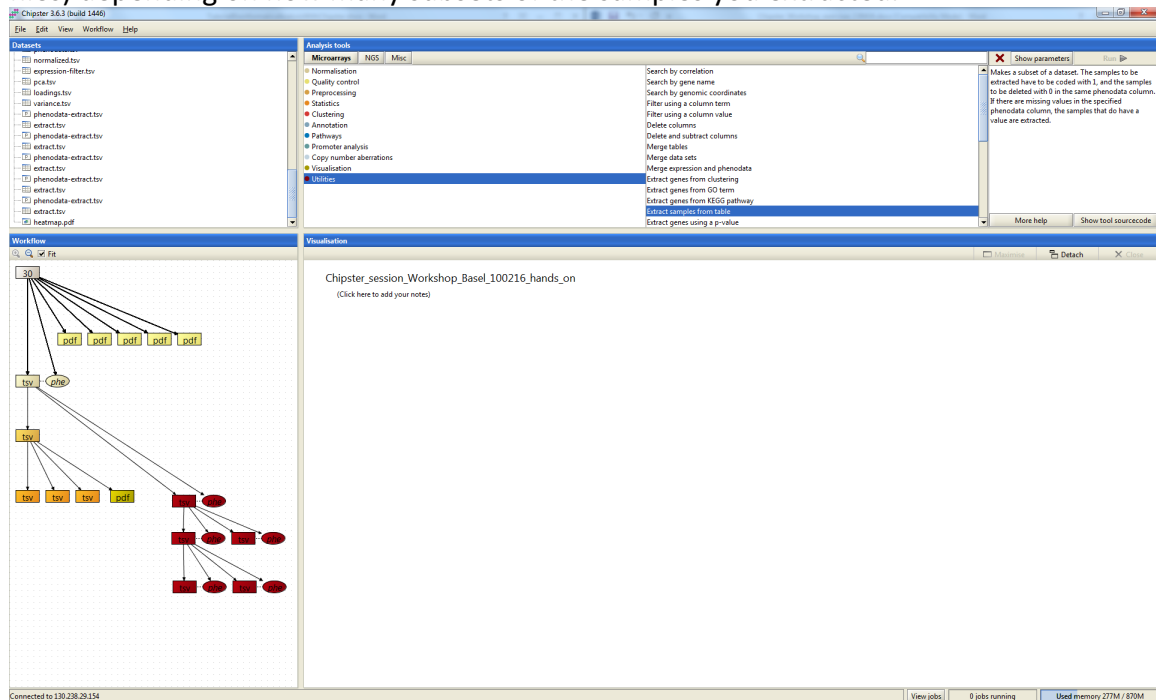
1. Click once on the unfiltered normalized file (normalized.tsv). Under **Utilities** in the Analysis tools section, choose **Extract samples from table**. Under **Show parameters** define which column contains the extraction info (e.g. TiO₂). You will get a new normalized file (with an associated phenodata file), which contains only the TiO₂ samples and their controls. I.e. all the samples marked with '1' in the column chosen in the parameters section. The samples with '0' will be excluded from the newly generated dataset.
2. Redo the sample extraction step on the new TiO₂-file to get samples from only one dose (e.g. Dose10). Again you will get a new normalized file with associated phenodata file.

Tip! You can check the new phenodata file to see that the correct samples have been extracted.

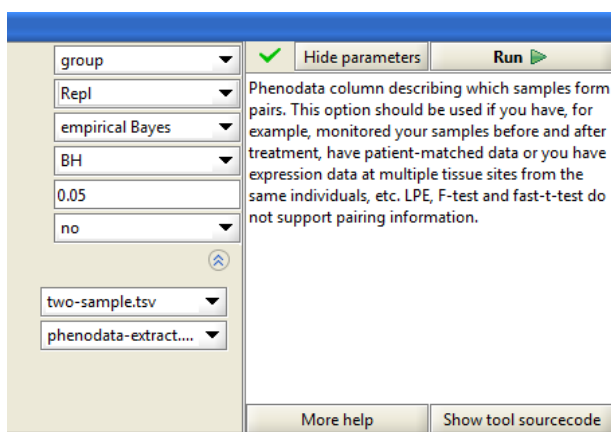
3. Redo the sample extraction once more to get only the samples for one time point (e.g. 1h). This can be done over and over again to get all the different subsets of samples, i.e. TiO₂-Dose10-1h, TiO₂-Dose10-24h, TiO₂-Dose100-1h, TiO₂-Dose100-24h.

Note! The exact parameters for each file are shown in the Visualization section when clicking once on a file in the Dataset or the Workflow section. This will help you keep track of the files with the same name. When extracting files it is, however, recommended that they are renamed according to the content and analysis.

If you have completed all the steps above, the session will now look something like the figure below. You will have different numbers of normalized-phenodata pairs (dark red files) depending on how many subsets of the samples you extracted.



- Click once on one of the normalized files with a subset of 6 samples (3 controls and 3 treatments, e.g. TiO₂-Dose10-1h). Preprocess the normalized file according to the Preprocessing and filtering steps described above.
- Under **Statistics** in the Analysis tools section, choose **Two groups tests**. Set the **parameters** according to the figure below.



Note! In this case multiple testing is being performed. In general, it is recommended to use multiple testing correction (different methods are provided under the ‘p-value adjustment method’ parameter). This study, however used low doses, resulting in low level changes, which are not detected by the stringent criteria of multiple testing correction in some of the sample subsets, e.g. MWCNT-Dose10-1h and other subsets as well. In these cases the multiple testing correction parameter can be set to ‘none’. The unadjusted p-value can be set to a lower threshold, e.g. 0.01 to be more confident about the results. Results obtained with unadjusted p-values can still be significant and relevant in the context of other results and on pathway-level analyses. This is the case in the Nanosafety example session provided within Chipster and [here](#). Thus, the other sample subsets can be analysed without multiple testing and the results from different subsets can be compared. However, when comparing results, it is highly recommended to use the same parameter settings for all sample subsets. I.e. if one subset is tested without multiple testing correction then all subsets should be tested that way.

6. The resulting file contains statistically differentially expressed genes (DEGs) between the control cells and the treated cells. In the case of the subset TiO₂-Dose10-1h there should be 54 DEGs when using the parameters in the figure above.

Tip! The different lists of DEGs resulting from statistical analysis of the different sample subsets can be compared in Venn diagrams by choosing several files in the Workflow section (hold down the Ctrl-key) and then choosing the **Venn-diagram icon** in the Visualization section.

PATHWAY AND GENE ONTOLOGY ENRICHMENT ANALYSIS

1. To perform a basic pathway or Gene Ontology (GO) enrichment analysis click once on the file with the DEGs identified during the statistical analysis. Under **Pathways** in the Analysis tools section, choose **Hypergeometric test for GO**. Under **Show parameters** choose which sub-ontology to use (you can also choose all). Leave the other parameters at default. Click **Run**.

2. While the GO tool is running, click once on the same DEGs file as before and choose the tool **Hypergeometric test for ConsensusPathDB** (with default parameters). Click **Run**.
3. Double click on the HTML file from the GO analysis. This file provides you with links to the enriched pathways. The hypergeo.tsv file contains the same information, but in a sortable and exportable format.

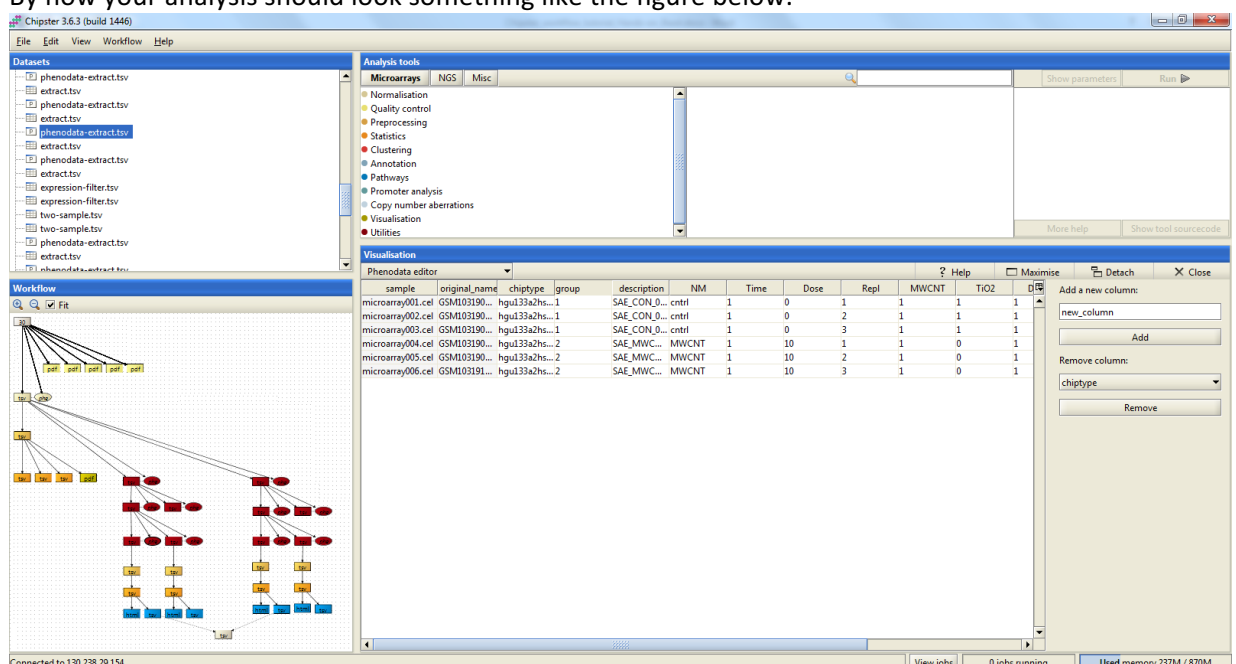
Tip! The .tsv file listing the enriched GO terms can be compared with other GO enrichment results from other comparisons and visualized in a Venn diagram, similarly as gene lists as described above.

4. From the ConsensusPathDB analysis you will get three files. Double click on the HTML file to inspect which pathways are enriched and for links to information about the pathways. The cpdb-pathways.tsv file lists the same information in a sortable and exportable format.

Note! ConsensusPathDB (CPDB) combines knowledge from several different pathway tools to avoid the need for running analyses for each database separately.

Tip! When comparing different lists using the Venn diagram, you can click on the different regions of the diagram to identify the genes in that region (on the far right in the Visualization section, choose the **Selected** tab and view the genes. By clicking the button on the bottom **Create data set from selected** you can create a new file containing only those genes.

By now your analysis should look something like the figure below.



The screenshot shows the Chipster 3.6.3 (build 1446) interface. On the left, a workflow diagram shows a sequence of steps: 'Fit' followed by 'extract' and 'phenodata-extract' steps. The main window displays the 'Phenodata editor' with a table of data. The table has columns for sample, original_name, chiptype, group, description, NM, Time, Dose, Rep, MWCNT, TiO2, and a status column. The data rows represent different microarray samples and their associated parameters.

sample	original_name	chiptype	group	description	NM	Time	Dose	Rep	MWCNT	TiO2	
microarray001.cel	GSM103190...	hgu133a2hs...	1	SAE_CON_0..._cntrl	1	0	1	1	1	1	1
microarray002.cel	GSM103190...	hgu133a2hs...	1	SAE_CON_0..._cntrl	1	0	2	1	1	1	1
microarray003.cel	GSM103190...	hgu133a2hs...	1	SAE_CON_0..._cntrl	1	0	3	1	1	1	1
microarray004.cel	GSM103190...	hgu133a2hs...	2	SAE_MWVC..._MWCNT	1	10	1	1	0	1	1
microarray005.cel	GSM103190...	hgu133a2hs...	2	SAE_MWVC..._MWCNT	1	10	2	1	0	1	1
microarray006.cel	GSM103191...	hgu133a2hs...	2	SAE_MWVC..._MWCNT	1	10	3	1	0	1	1

OTHER USEFUL FUNCTIONS

1. Under **Utilities** you can find several useful functions for manipulating your files. E.g.
 - **Extract** genes based on different criteria
 - **Delete** columns
 - **Search** for specific items
 - **Merge** tables with different samples
 - **Sort** information in files
 - And many many more...
2. Under **Visualization** you can visualize genes in a list according to their chromosomal position (**Idiogram** or **Chromosomal position**).
3. Under **Annotation** you can add information to your data, e.g. alternative identifiers, genomic location, GO annotations etc.
4. There is a wide variety of other tools that can and should be explored, e.g. many other statistical tools (including limma – **Linear modelling** under **Statistics**) other than the **Two group test** used in this tutorial. Limma, for example, provides the possibility to build more complex statistical models for testing dose-response relationships and time-dependent effects.

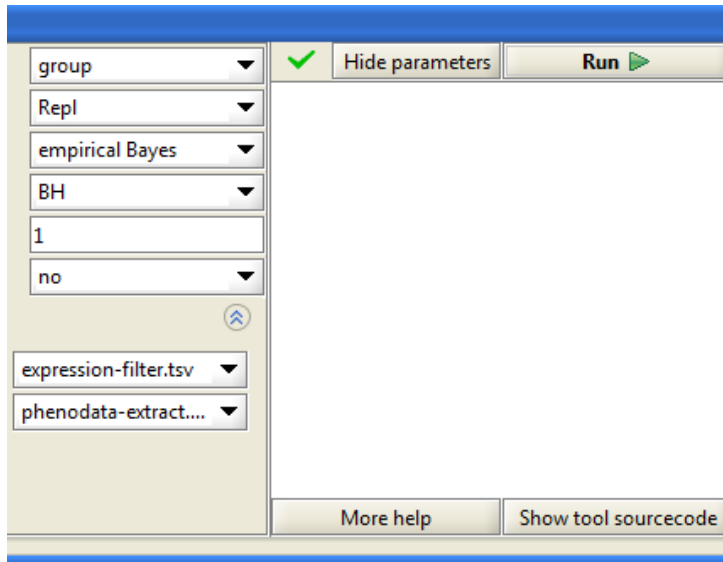
GENERATING A DATA MATRIX AND EXPORTING DATA FROM CHIPSTER

Some online bioinformatics tools require a data matrix with statistically tested fold changes for each treatment condition. Such a file can be generated through a few different ways in Chipster. One approach is described below.

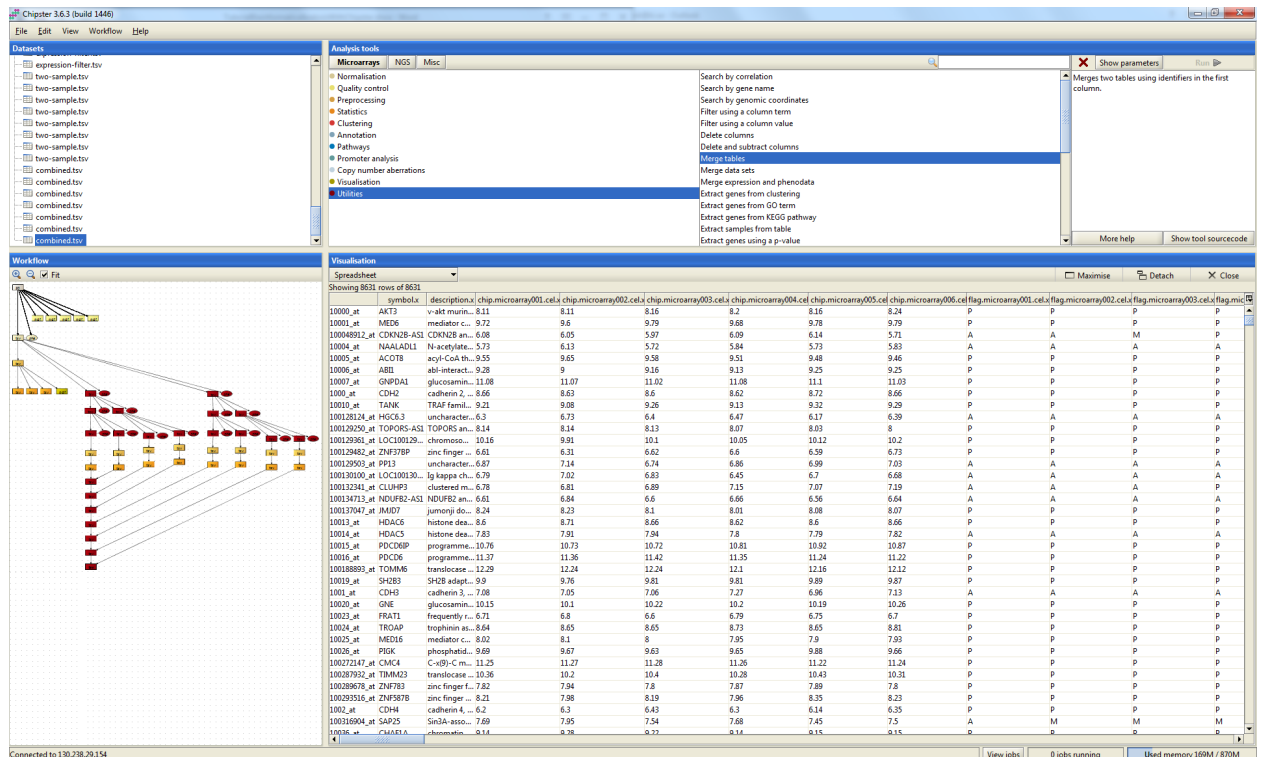
1. Extract all the different sample subsets to be compared, in the case of the sample data 8 different subsets with 6 samples each.
2. Preprocess each normalized file for each sample subset according to the Preprocessing and filtering steps above.

***Tip!** This can be done for all files simultaneously by choosing all normalized files (hold down the Ctrl-key) and then the tool. Remember to set the parameters.*

3. Perform a **Two groups test** on each of the filtered files for each samples subset using the same parameters. Set the p-value threshold to 1 to include all genes in the results file. Like the figure below.



4. Choose (hold down Ctrl) the two first filtered files. Under **Utilities** choose the tool **Merge tables** and with default parameters click **Run**. Then choose the resulting file from the merge and the next filtered file from the third samples subset. Click **Run** for the **Merge tables** tool. Again choose the resulting file and the next filtered file from the fourth sample subset. Click run for the **Merge tables** tool. Your session will then look like this.



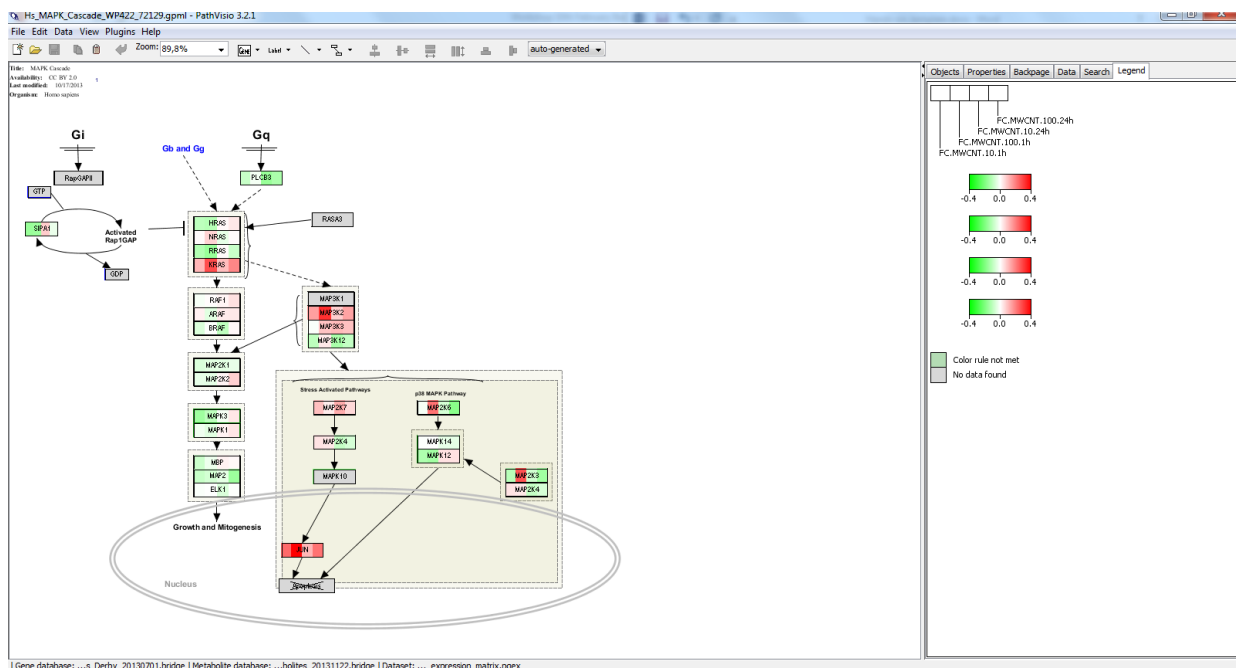
- To export the data file right click the mouse, choose **Export...** and save the file on the Desktop. The file is in .tsv format, which is basically a tab-delimited text file. Open the file in Excel. Be sure to open it correctly (Excel tends to automatically reformat values if the settings are off) and shift the headers one step to the right. Save the file as a tab-delimited text file (.txt). The file can be used and processed to include information required for analysis in different online tools.

FURTHER RECOMMENDATIONS AND EXAMPLES OF ANALYSES IN PATHVISIO

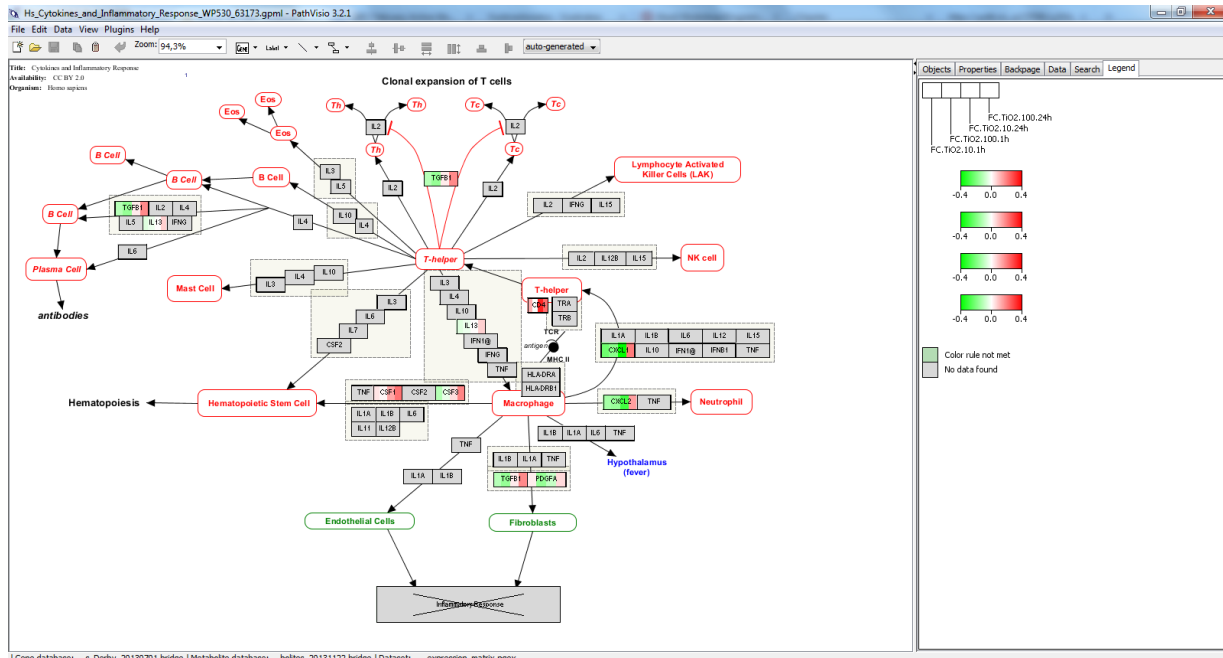
[PathVisio](#) is a free open source biological pathway analysis tool, which provides detailed exploration and visualization of biological and toxicological pathways. The tool allows you to draw, edit and analyze specific biological pathways of interest. A [nanomaterials-specific WikiPathways portal](#) has been established and lists nano-related pathways relevant for the nanosafety community, though it excludes general pathways that can be affected by nanomaterials. Nevertheless, in contrast to more general scientific studies, a nano-focused curated subset of pathways may benefit nanosafety risk analysis/assessment activities. These pathways can be explored in PathVisio or through the Pathway module of [ArrayAnalysis.org](#).

Examples of analysis of DEG lists obtained from analyses of the example data in this tutorial are given in the following.

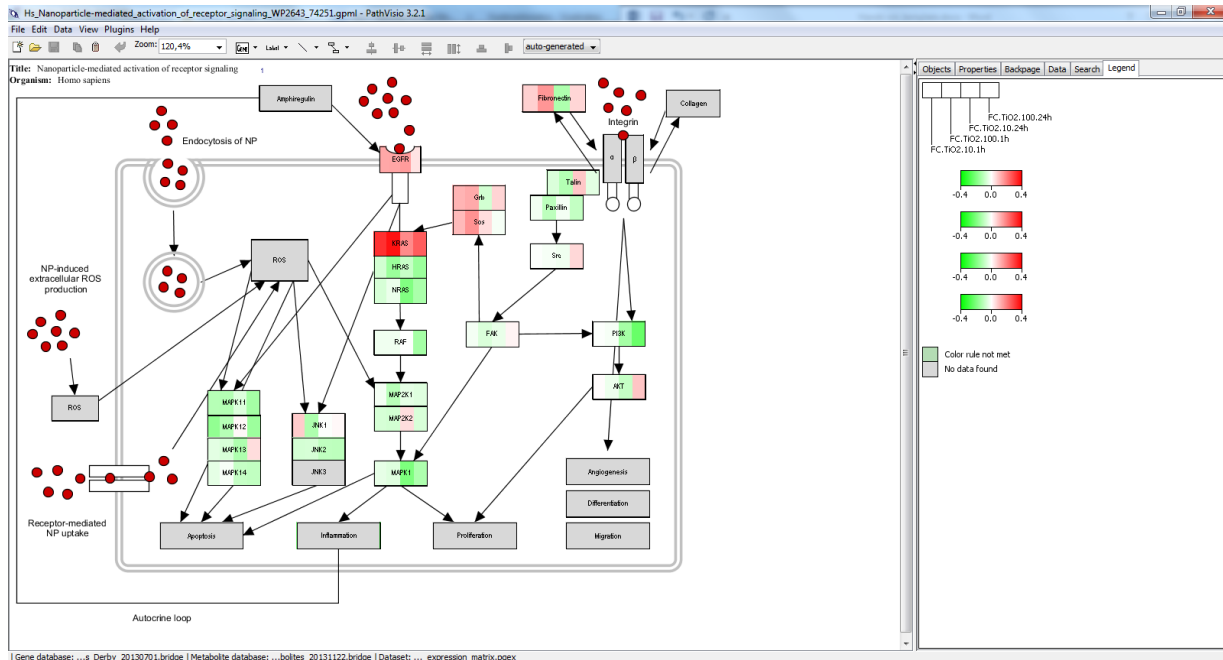
1. A matrix with DEGs found to be statistically significant in at least one MWCNT comparison was used to visualize the [MAPK Cascade](#) pathway in PathVisio (below). The results identified in the GO enrichment analysis in Chipster found MAPK cascade related GOs to be enriched at both 1h and 24h of the lower dose (Dose10), leading the researcher to the hypothesis that this pathway may be of interest with regard to MWCNT exposure in small airway epithelial cells.



2. Similarly a data matrix with DEGs found to be statistically significant in at least one TiO₂ comparison was used to visualize the pathway for [Cytokines and Inflammatory Response](#) in PathVisio (below). Immune response GO terms were enriched, especially at 24h following TiO₂ exposure. This is also in line with the original publication ([Tilton et al, 2014](#)).



3. Finally, the WikiPathway [Nanoparticle mediated activation of receptor signaling](#) listed in the [WikiPathways Nanoportal](#) was visualized using the TiO₂ DEGs matrix (below), showing the effects of TiO₂ exposure on small airway epithelial cells with regard to this specific pathway. Pathways such as this one, may be an interesting starting point for nanoparticle-related bioinformatics gene network analysis on a wider level.



A SELECTION OF DATABASES AND TOOLS RELEVANT FOR NANOSAFETY-RELATED BIOINFORMATICS

Databases

[ArrayExpress](#)

[GEO](#)

[NanoMiner](#)

[Comparative Toxicogenomics Database \(CTD\)](#)

[eNanoMapper Database](#) and the [interactive search interface](#)

[WikiPathways](#)

Tools

[PathVisio](#)

[Ingenuity Pathway Analysis](#) (commercial)

[ConsensusPathDB](#) (database with interactive tools)

[Reactome](#) (database with interactive tools)

[EnrichR](#)

[Gene Set Enrichment Analysis \(GSEA\)](#)

3. ACKNOWLEDGMENTS

We would like to express our gratitude to the developers of Chipster, especially Eija Korpelainen, who has provided extensive support on both the use of and access to Chipster. This tutorial is derived from <http://chipster.csc.fi/> documentation.

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

Kallio MA, Tuimala JT, Hupponen T, Klemelä P, Gentile M, Scheinin I, Koski M, Käki J, Korpelainen EI. *Chipster: user-friendly analysis software for microarray and other high-throughput data*. BMC Genomics. 2011 Oct 14;12:507. doi: 10.1186/1471-2164-12-507

Tilton SC, Karin NJ, Tolic A, Xie Y, Lai X, Hamilton RF Jr, Waters KM, Holian A, Witzmann FA, Orr G. *Three human cell types respond to multi-walled carbon nanotubes and titanium dioxide nanobelts with cell-specific transcriptomic and proteomic expression patterns*. Nanotoxicology. 2014 Aug;8(5):533-48. doi: 10.3109/17435390.2013.803624

Grafström RC, Nymark P, Hongisto V, Spjuth O, Ceder R, Willighagen E, Hardy B, Kaski S, Kohonen P. *Toward the Replacement of Animal Experiments through the Bioinformatics-driven Analysis of 'Omics' Data from Human Cell Cultures*. Altern Lab Anim. 2015 Nov;43(5):325-32

Nymark P, Wijshoff P, Cavill R, van Herwijnen M, Coonen ML, Claessen S, Catalán J, Norppa H, Kleinjans JC, Briedé JJ. *Extensive temporal transcriptome and microRNA analyses identify molecular mechanisms underlying mitochondrial dysfunction induced by multi-walled carbon nanotubes in human lung cells*. Nanotoxicology. 2015;9(5):624-35. doi: 10.3109/17435390.2015.1017022

Kohonen P, Ceder R, Smit I, Hongisto V, Myatt G, Hardy B, Spjuth O, Grafström R. *Cancer biology, toxicology and alternative methods development go hand-in-hand*. Basic Clin Pharmacol Toxicol. 2014 Jul;115(1):50-8. doi: 10.1111/bcpt.12257. Review.

Kohonen P, Benfenati E, Bower D, Ceder R, Crump M, Cross K, Grafström RC, Healy L, Helma C, Jeliaskova N, Jeliaskov V, Maggioni S, Miller S, Myatt G, Rautenberg M, Stacey G, Willighagen E, Wiseman J and Hardy B. *The ToxBank Data Warehouse: Supporting the Replacement of In Vivo Repeated Dose Systemic Toxicity Testing*. Mol. Inf., 32: 47–63. doi: 10.1002/minf.201200114

5. KEYWORDS

Microarray data analysis
Quality control
Data pre-processing and filtering
Pathway analysis
Systems biology
Nanomaterials