Analyzing data with Metaboanalyst 4.0

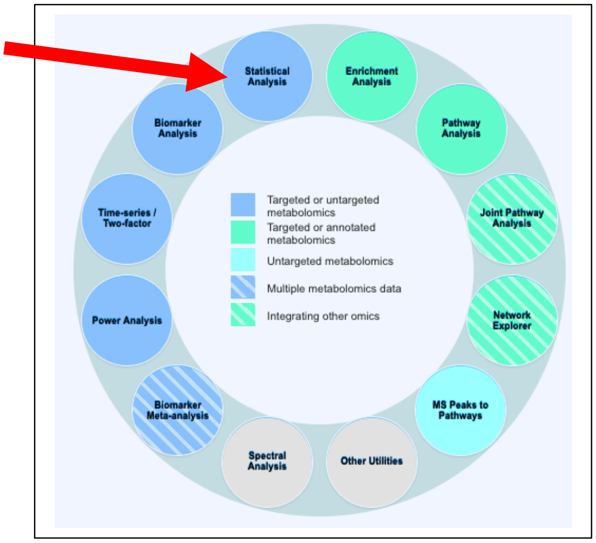
Ion features in liquid chromatography-mass spectrometry (LC-MS) datasets are aligned with regard to their retention time and the mass-to-charge (*m/z*) ratio. This analysis of datasets is carried out using MZmine (http://mzmine.github.io/) (as in Dr. Xiuxia Du's presentation) or by XCMS. The latter can be run in R, or (in many labs) using an online version of XCMS (https://xcmsonline.scripps.edu). Details of the features of XCMSonline can be obtained after setting up a registration account.

Information about the peaks identified using XCMS can be downloaded as a .zip file. Once unzipped, there is an Excel file that contains information on the *m/z* and retention time values of each ion feature, univariate statistical analysis (p-values and q-scores), and the area under the curve of each ion feature in each individual experimental sample.

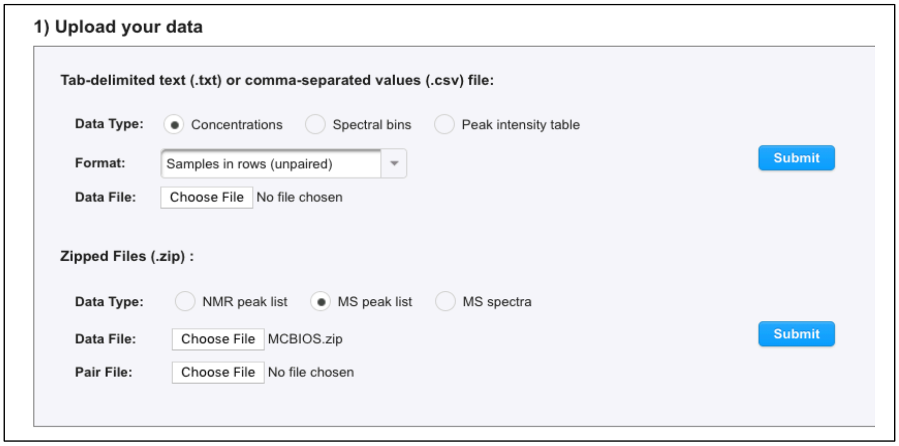
To prepare for Metaboanalyst, .csv files are created for each experimental sample. These contain three columns – the median *m/z* values, the median retention time values and the areas under the peaks. IMPORTANT: - the names used for each of the files should not contain spaces (an underscore should be used instead). The .csv files corresponding to a group are next placed in a folder (again, no spaces in the name used). This should be repeated for all groups. The folders are then zipped and the name given to the .zip file should again not contain a space. This is the file to be used by Metaboanalyst.

Uploading a .zip file to Metaboanalyst 4.0

Set your browser to https://www.metaboanalyst.ca/faces/ModuleView.xhtml



and then select the statistical analysis module – https://www.metaboanalyst.ca/faces/upload/StatUploadView.xhtml.



1) Upload your data

Tab-delimited text (.txt) or comma-separated values (.csv) file:

Data Type: ● Concentrations ○ Spectral bins ○ Peak intensity table

Format: Samples in rows (unpaired)

Data File: Choose File No file chosen

Submit

Zipped Files (.zip) :

Data Type: ○ NMR peak list ● MS peak list ○ MS spectra

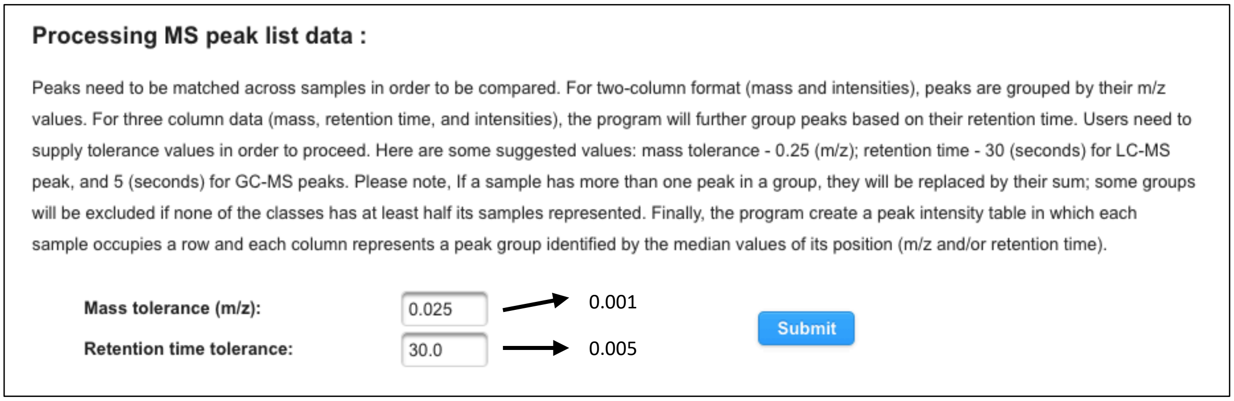
Data File: Choose File MCBIOS.zip

Pair File: Choose File No file chosen

Submit

Use the Zipped files option – first select “MS peak list” and then from “Choose file” browse to locate the .zip file. Then press **submit**.

Once loaded, entries are needed for mass and retention time tolerance. Since we set these to the same values for each .csv file, enter 0.001 for mass tolerance and 0.005 min for retention time tolerance (do not use zeros since these are not allowed entries). **Submit**.



Processing MS peak list data :

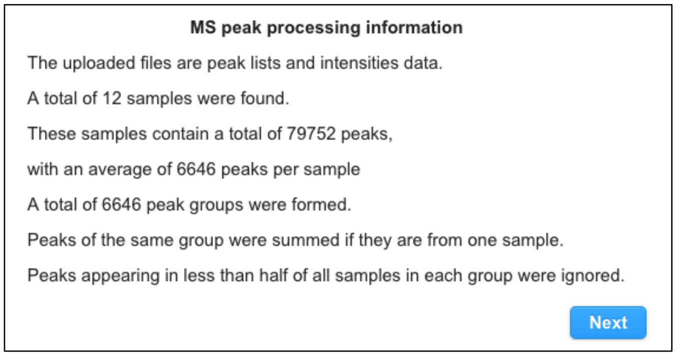
Peaks need to be matched across samples in order to be compared. For two-column format (mass and intensities), peaks are grouped by their m/z values. For three column data (mass, retention time, and intensities), the program will further group peaks based on their retention time. Users need to supply tolerance values in order to proceed. Here are some suggested values: mass tolerance - 0.25 (m/z); retention time - 30 (seconds) for LC-MS peak, and 5 (seconds) for GC-MS peaks. Please note, If a sample has more than one peak in a group, they will be replaced by their sum; some groups will be excluded if none of the classes has at least half its samples represented. Finally, the program create a peak intensity table in which each sample occupies a row and each column represents a peak group identified by the median values of its position (m/z and/or retention time).

Mass tolerance (m/z): 0.025 → 0.001

Retention time tolerance: 30.0 → 0.005

Submit

The summary will indicate that 12 sample files were loaded with 6,646 peak groups per file. A check is made for missing values – XCMS always puts a number into the Excel output file, so there is no need to do anything and therefore press SKIP.



MS peak processing information

The uploaded files are peak lists and intensities data.

A total of 12 samples were found.

These samples contain a total of 79752 peaks,

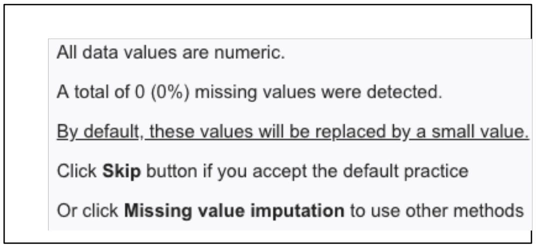
with an average of 6646 peaks per sample

A total of 6646 peak groups were formed.

Peaks of the same group were summed if they are from one sample.

Peaks appearing in less than half of all samples in each group were ignored.

Next



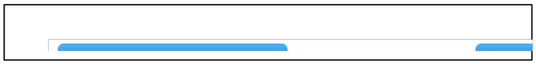
All data values are numeric.

A total of 0 (0%) missing values were detected.

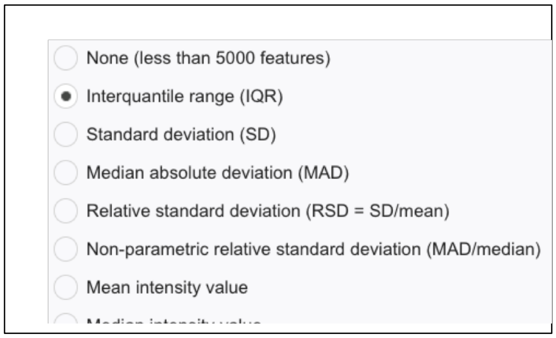
By default, these values will be replaced by a small value.

Click **Skip** button if you accept the default practice

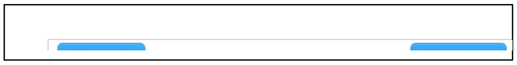
Or click **Missing value imputation** to use other methods



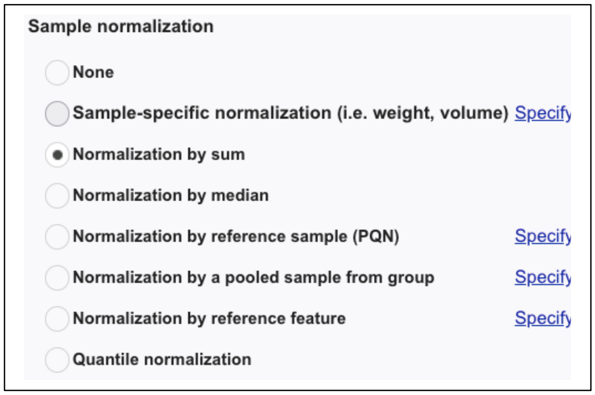
Since the number of features exceeds 5000, interquartile range (IQR) filtering is necessary. After selecting the IQR button, press SUBMIT. Then press PROCEED.



○ None (less than 5000 features)
● Interquantile range (IQR)
○ Standard deviation (SD)
○ Median absolute deviation (MAD)
○ Relative standard deviation (RSD = SD/mean)
○ Non-parametric relative standard deviation (MAD/median)
○ Mean intensity value
○ Median intensity value

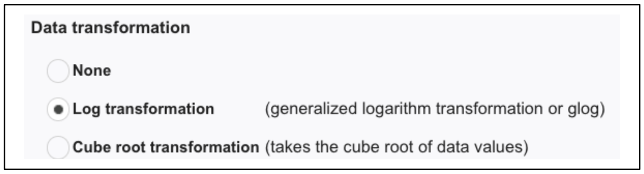


The next steps involve data normalization. First use "normalization by sum". Second, transform the data using "log transformation". Thirdly, use "Pareto scaling". These three selections deal with (1) variation (in this case) with urine volume differences between animals, (2) a need to ensure the data have a normal distribution (required for statistical analysis) and (3) to mean center the data to overcome large differences between individual ion features.



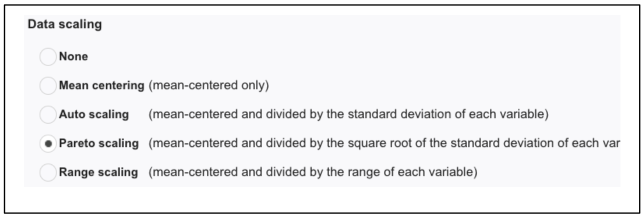
Sample normalization

○ None
○ Sample-specific normalization (i.e. weight, volume) Specify
● Normalization by sum
○ Normalization by median
○ Normalization by reference sample (PQN) Specify
○ Normalization by a pooled sample from group Specify
○ Normalization by reference feature Specify
○ Quantile normalization



Data transformation

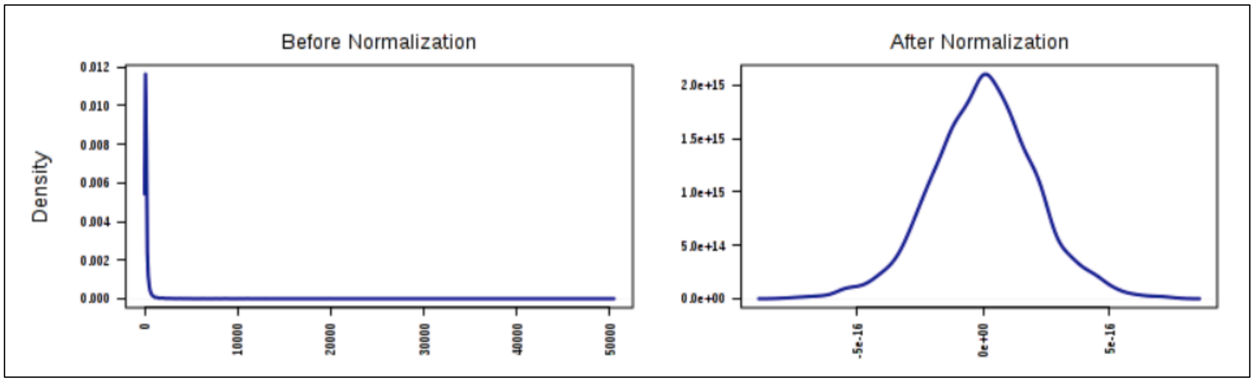
○ None
● Log transformation (generalized logarithm transformation or glog)
○ Cube root transformation (takes the cube root of data values)

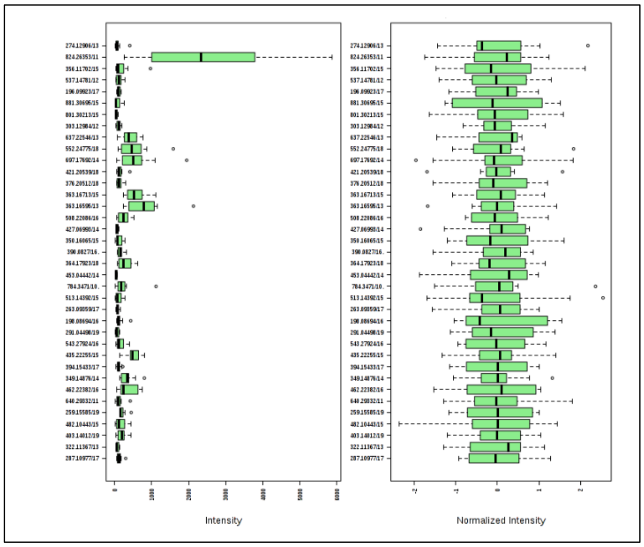


Data scaling

○ None
○ Mean centering (mean-centered only)
○ Auto scaling (mean-centered and divided by the standard deviation of each variable)
● Pareto scaling (mean-centered and divided by the square root of the standard deviation of each var
○ Range scaling (mean-centered and divided by the range of each variable)

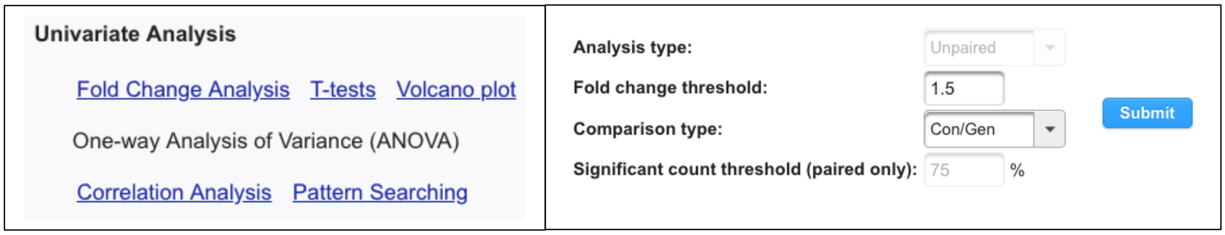
Now press **NORMALIZE**, **VIEW RESULT** and **PROCEED**.





Clicking onto PROCEED brings you into the statistical analysis section of Metaboanalyst.

Univariate Analysis



Univariate Analysis

Fold Change Analysis T-tests Volcano plot

One-way Analysis of Variance (ANOVA)

Correlation Analysis Pattern Searching

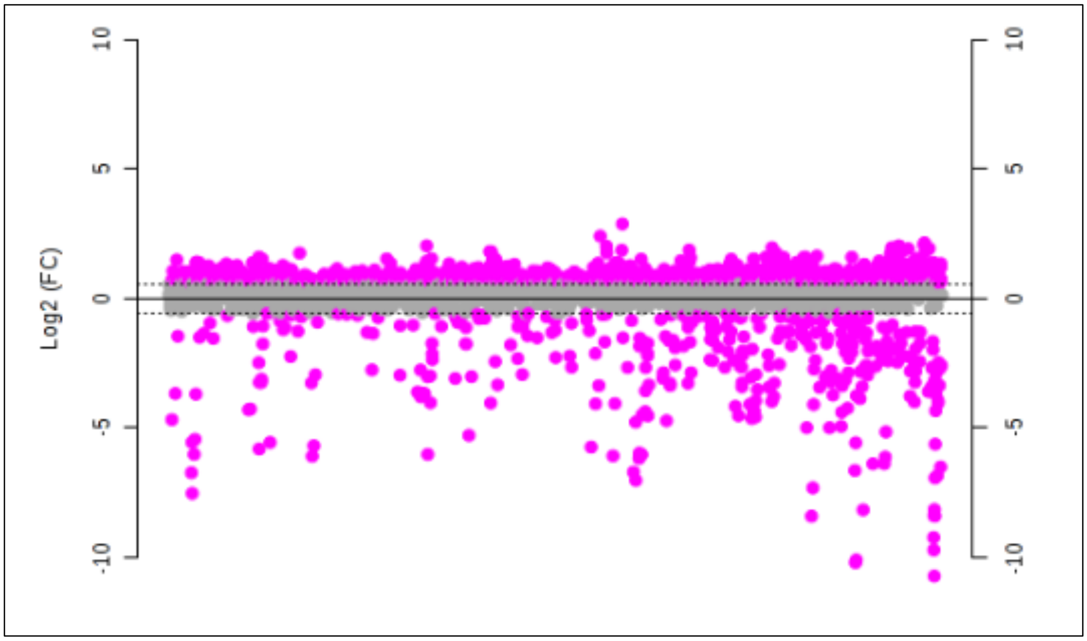
Analysis type: Unpaired

Fold change threshold: 1.5

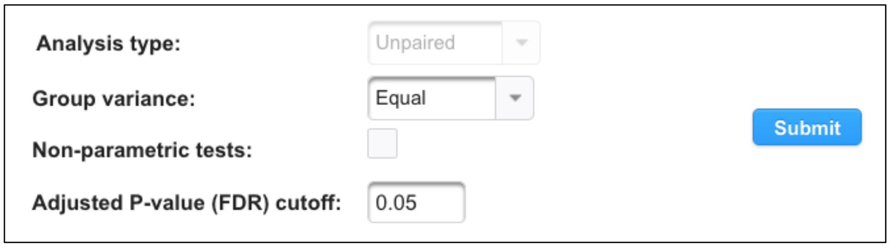
Comparison type: Con/Gen

Significant count threshold (paired only): 75 %

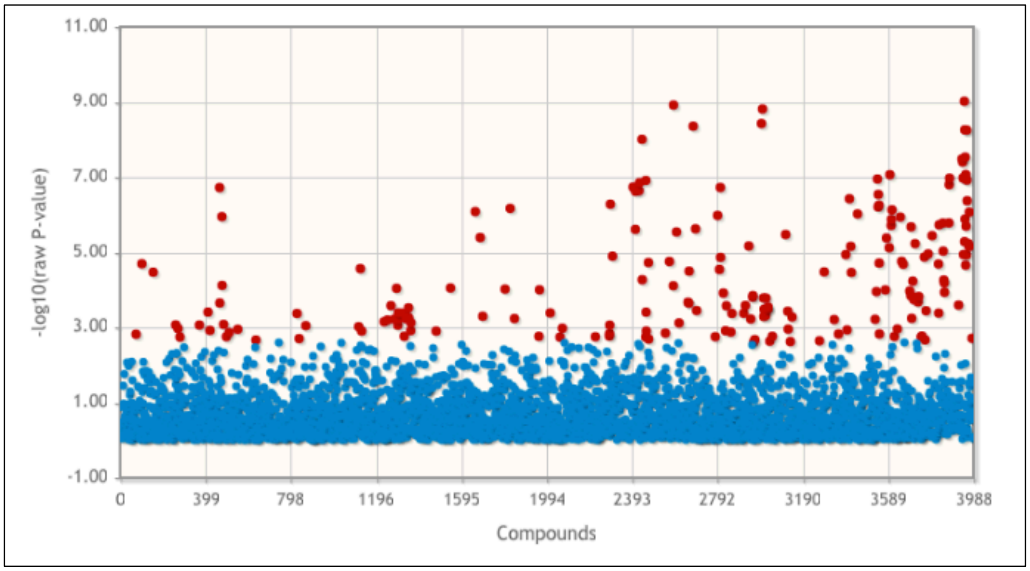
Fold change – update fold change threshold to 1.5 and **SUBMIT**.



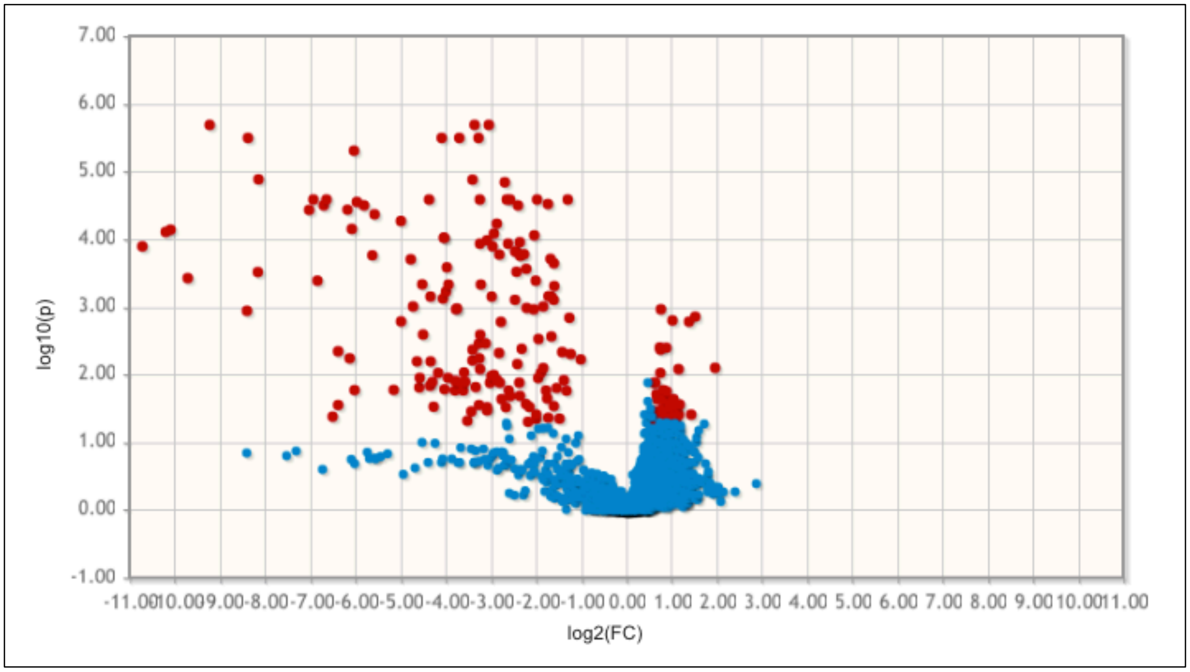
T-test – the assumption is that group variance is equal and adjusted P-value (for multiple testing) is 0.05.



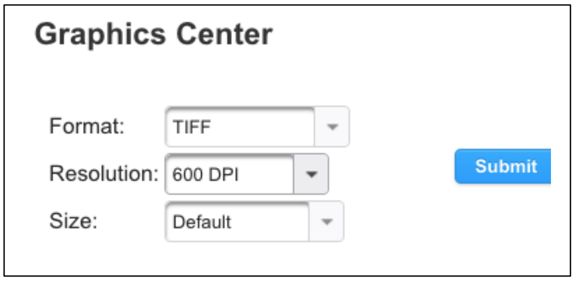
Analysis type:	Unpaired
Group variance:	Equal
Non-parametric tests:	□
Adjusted P-value (FDR) cutoff:	0.05



Volcano plot – reset to fold-change to 1.5 and p-value to 0.05. Then **Submit**.



A publication quality figure can be generated by clicking on the palette on the upper right of the Volcano plot. Set format to TIFF and resolution to 600 dpi – press **SUBMIT**. Click on volcano_1_dpi600.tiff to download figure.

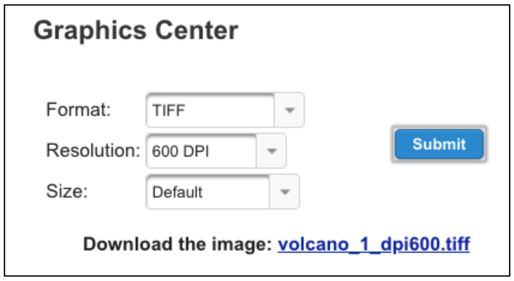


Graphics Center

Format: TIFF

Resolution: 600 DPI

Size: Default



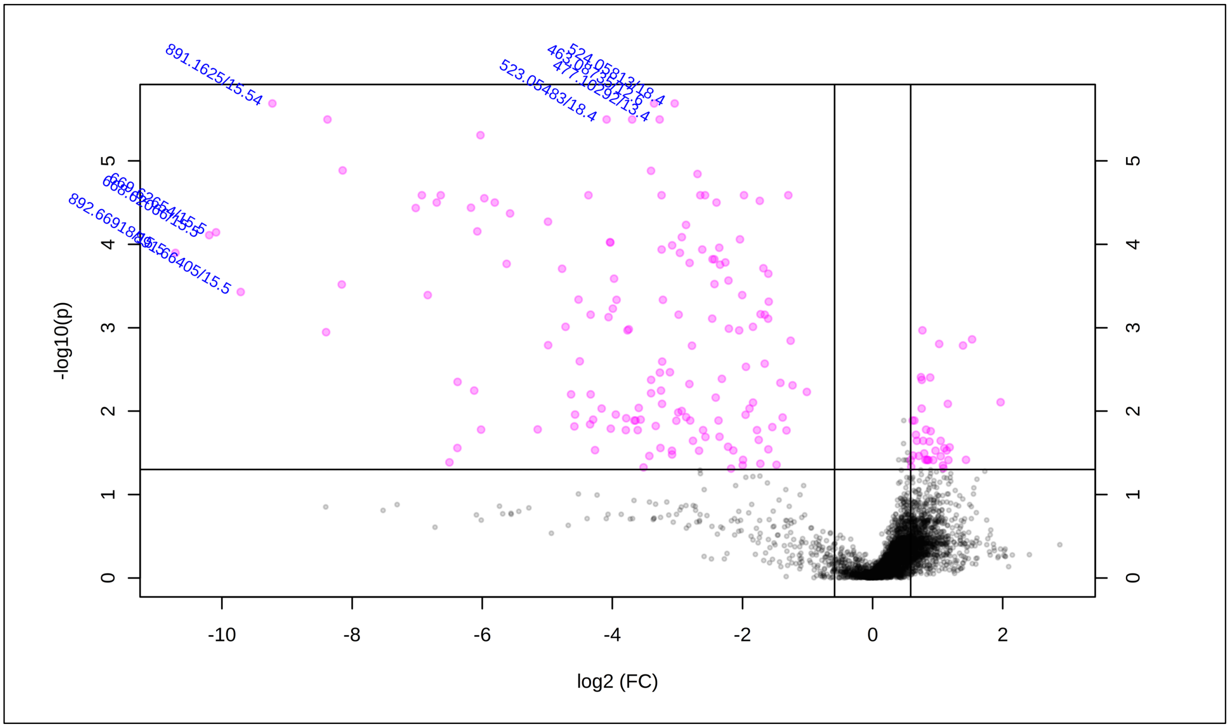
Graphics Center

Format: TIFF

Resolution: 600 DPI

Size: Default

Download the image: volcano_1_dpi600.tiff

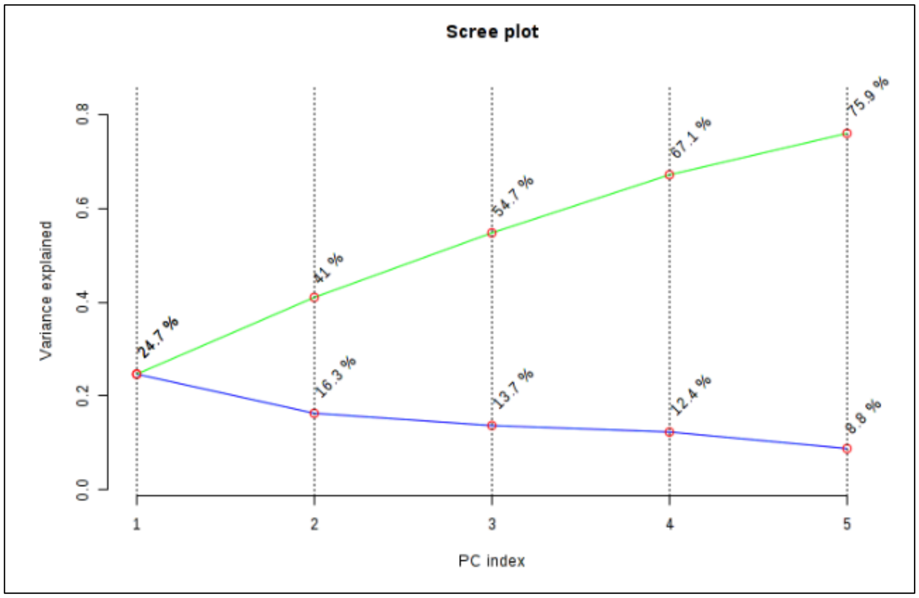


Multivariate Analysis

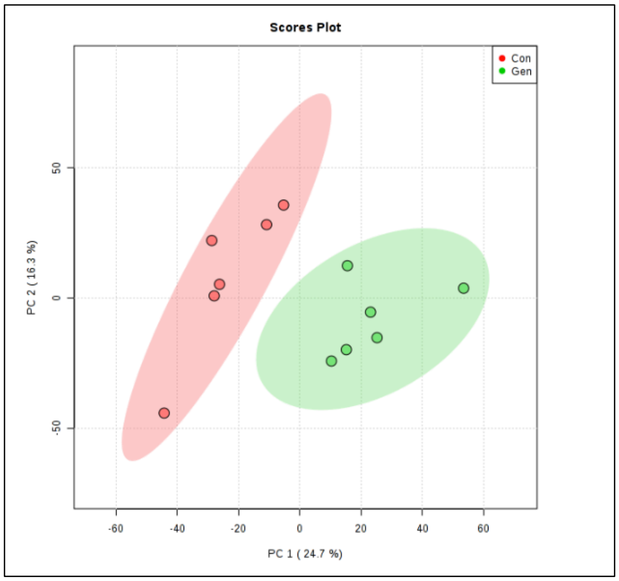
In univariate analysis, the ion features are considered to be independent. However, this is very unlikely in a biological system and therefore multivariate analysis is needed.

The data are first analyzed by *Principal Components Analysis* (PCA). This is an unsupervised analysis – taking the data from both groups as if they were one, principal component 1 (PC1) comes from a line that bisects the data to produce to the maximum error differences between the line and the data. Subsequent principal components represent lines drawn through the data orthogonal to PC1.

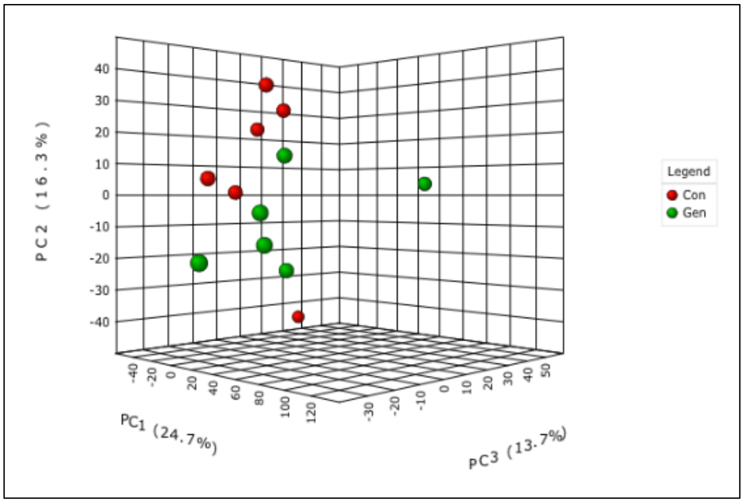
A Scree plot summarizes the accumulation of explained error.



The 2D Scores plot plots PC2 versus PC1. Individual samples are color-coded to represent the two groups and a 95% confidence limit is drawn around the data in each group.



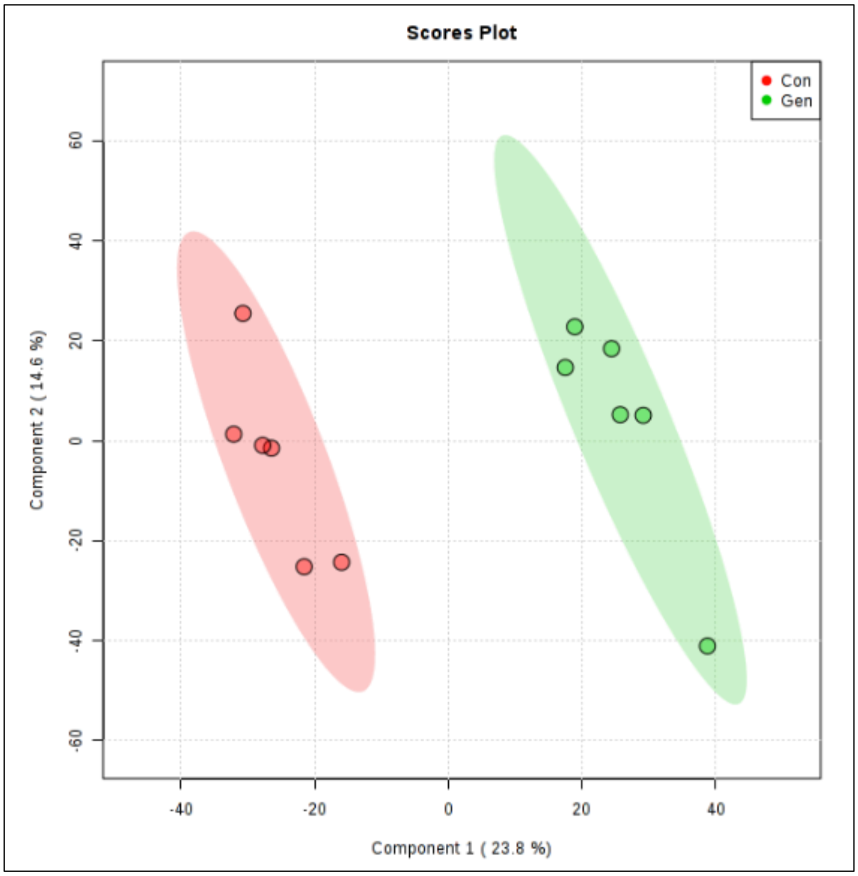
Use the palette to make a publication quality figure. Press **SUBMIT**.



A 3D-view of the PCA plot using PC1, PC2 and PC3 can be viewed – it's rotatable! However, it cannot be saved into a publication quality figure.

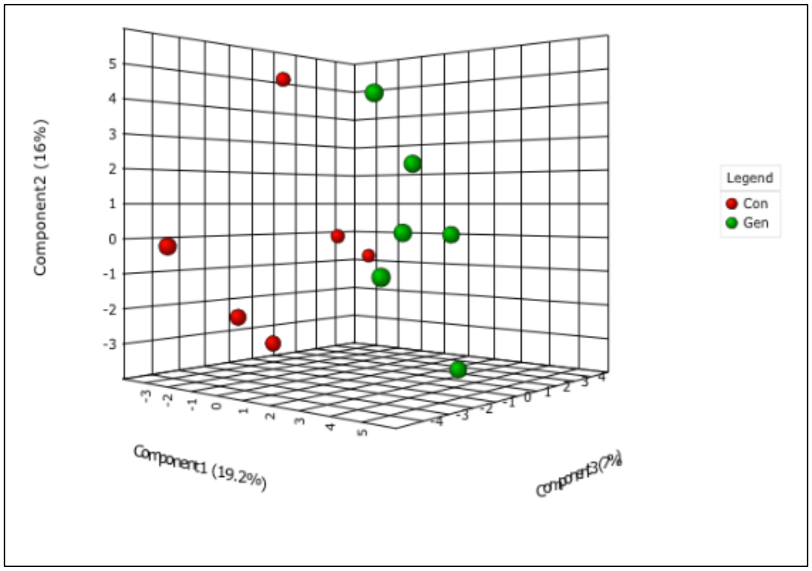
In the Loadings plot, most ions are centered around 0,0. Clicking on ions that are most away from 0,0 produces box-and-whisker plots for individual ions.

Partial least squares-discriminant analysis (PLSDA) is a supervised analysis that includes information that there are (in this case) two groups.

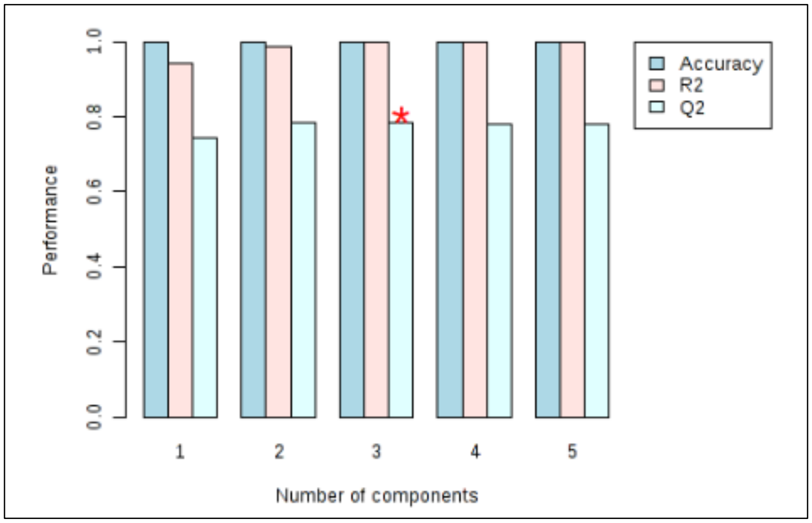


The *2D-scores plot* shows a complete separation of the two groups.

Use the palette to make a publication quality figure. Press **SUBMIT**.



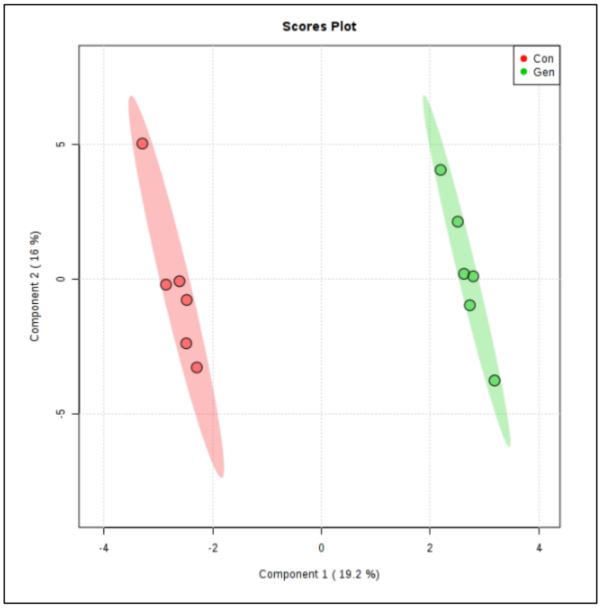
The *3D-PLSDA scores plot* shows improved resolution over the 3D-PCA plot.

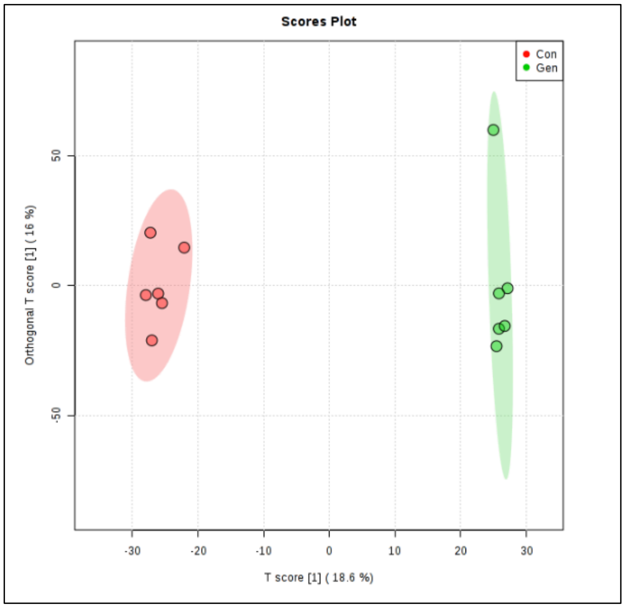


Cross validation reveals that there are 3 principal components and that 3D-plots are the best representation of the data.

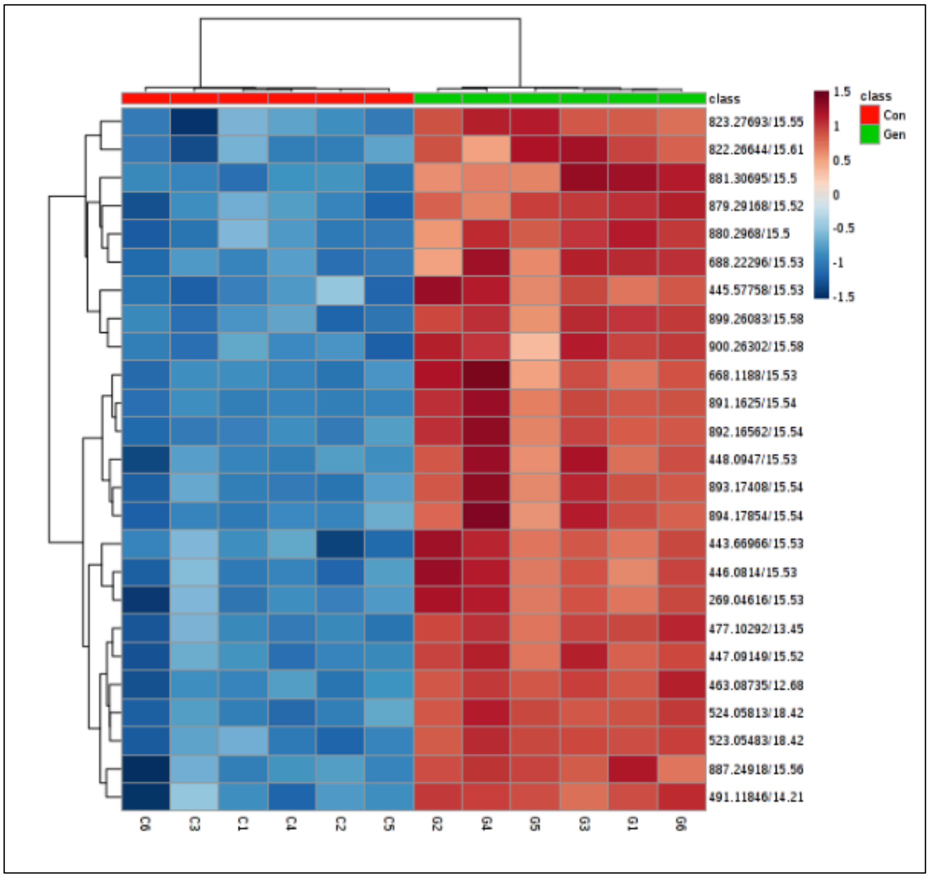
To identify the ion features most contributing to the differences between the two groups, the *Variable importance in projection* values are calculated (VIP scores). Usually these are selected from component 1. Increase the number to 25 shown over the default number (15) and press UPDATE. Use the palette to make a publication quality figure. Press **SUBMIT**.

Two other multivariate analyses that are frequently found in publications on metabolomics are sparse PLS-DA and orthogonal PLSDA. These are both available in Metaboanalyst 4.0.

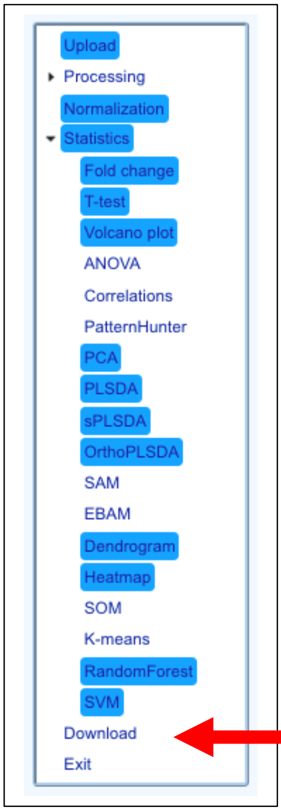




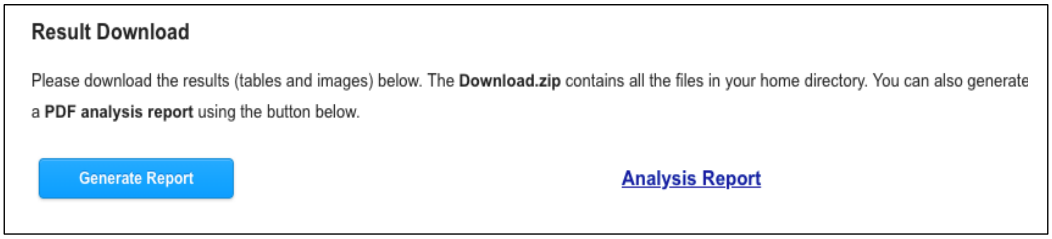
<u>Heatmaps</u>



Metaboanalyst allows for ion feature clustering and representation of the differences between the two groups in shades of blue and red. Select the option to show the top 25 ion features. These show the very clear differences between the two groups.

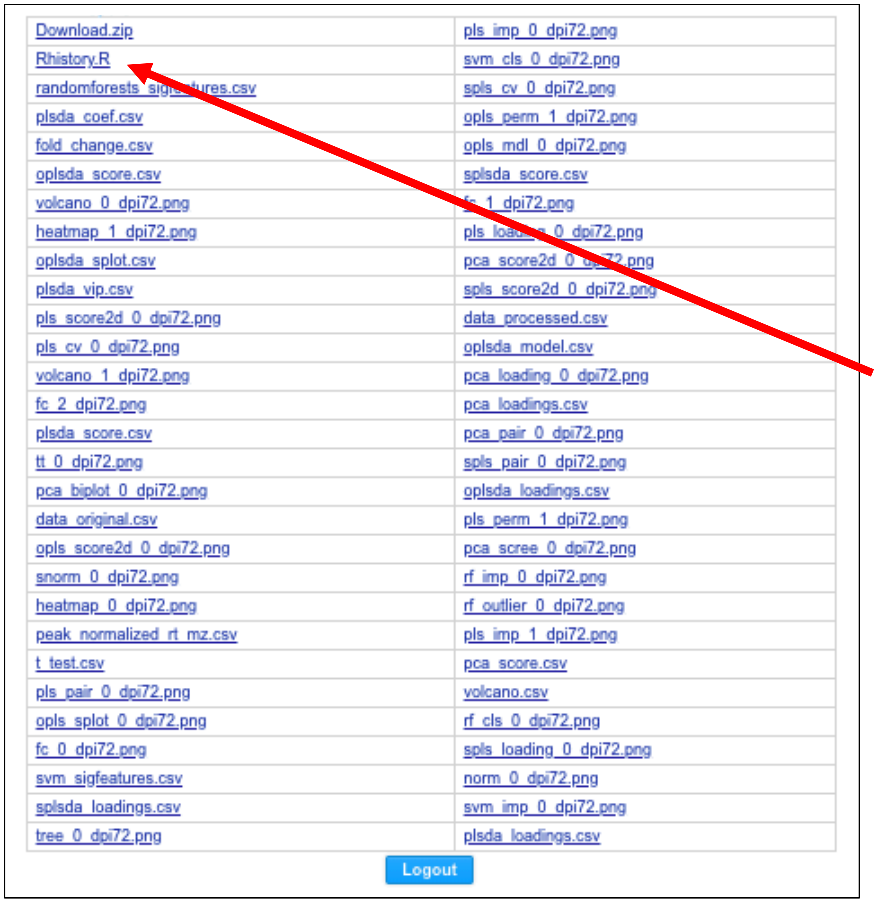


A nice feature of Metaboanalyst is that when you've finished the analyses you wanted to do, you're able to automatically compile them into a downloadable file. A written compiled report in PDF is obtained by pressing "Generate Report".



Result Download

Please download the results (tables and images) below. The **Download.zip** contains all the files in your home directory. You can also generate a **PDF analysis report** using the button below.



Download.zip	pls_imp_0_dpi72.png
Rhistory.R	svm_cls_0_dpi72.png
randomforests_sigfeatures.csv	spls_cv_0_dpi72.png
plsda_coef.csv	opls_perm_1_dpi72.png
fold_change.csv	opls_mdl_0_dpi72.png
oplsda_score.csv	splsda_score.csv
volcano_0_dpi72.png	fc_1_dpi72.png
heatmap_1_dpi72.png	pls_loading_0_dpi72.png
oplsda_splot.csv	pca_score2d_0_dpi72.png
plsda_vip.csv	spls_score2d_0_dpi72.png
pls_score2d_0_dpi72.png	data_processed.csv
pls_cv_0_dpi72.png	oplsda_model.csv
volcano_1_dpi72.png	pca_loading_0_dpi72.png
fc_2_dpi72.png	pca_loadings.csv
plsda_score.csv	pca_pair_0_dpi72.png
tt_0_dpi72.png	spls_pair_0_dpi72.png
pca_biplot_0_dpi72.png	oplsda_loadings.csv
data_original.csv	pls_perm_1_dpi72.png
opls_score2d_0_dpi72.png	pca_scree_0_dpi72.png
snorm_0_dpi72.png	rf_imp_0_dpi72.png
heatmap_0_dpi72.png	rf_outlier_0_dpi72.png
peak_normalized_rt_mz.csv	pls_imp_1_dpi72.png
t_test.csv	pca_score.csv
pls_pair_0_dpi72.png	volcano.csv
opls_splot_0_dpi72.png	rf_cls_0_dpi72.png
fc_0_dpi72.png	spls_loading_0_dpi72.png
svm_sigfeatures.csv	norm_0_dpi72.png
splsda_loadings.csv	svm_imp_0_dpi72.png
tree_0_dpi72.png	plsda_loadings.csv

The report also contains the R-commands that were invoked in the Metaboanalyst session.

Finally, remember to logout.