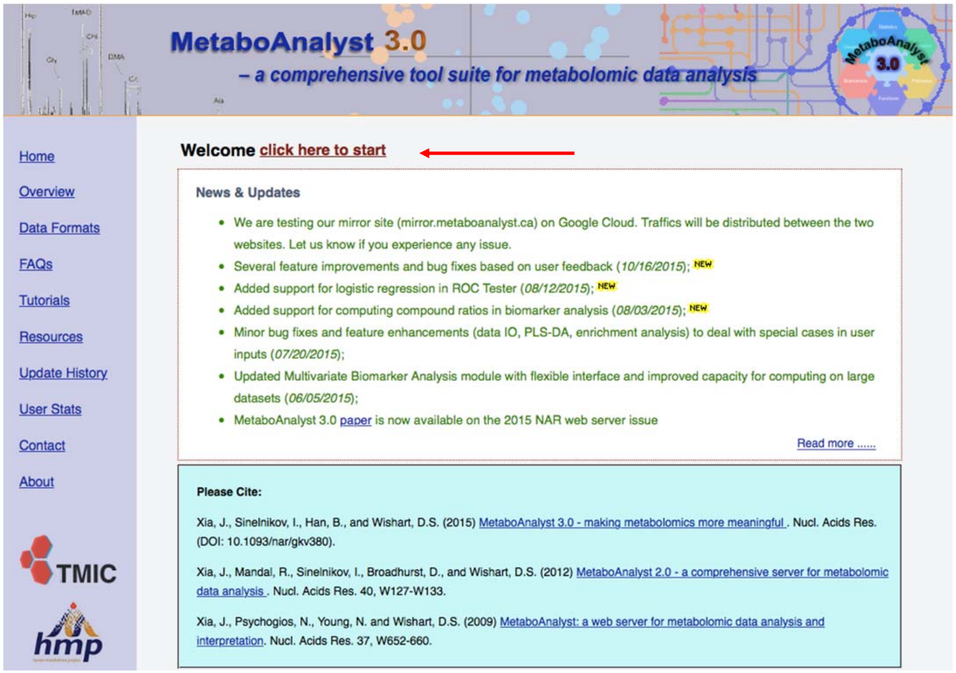
THE UNIVERSITY OF ALABAMA AT BIRMINGHAM

Knowledge that will change your world

GBS 724 02-08-16

Applying MetaboAnalyst

Xiangqin Cui, PhD



Welcome click here to start

News & Updates

- We are testing our mirror site (mirror.metaboanalyst.ca) on Google Cloud. Traffics will be distributed between the two websites. Let us know if you experience any issue.
- Several feature improvements and bug fixes based on user feedback (*10/16/2015*); NEW
- Added support for logistic regression in ROC Tester (*08/12/2015*); NEW
- Added support for computing compound ratios in biomarker analysis (*08/03/2015*); NEW
- Minor bug fixes and feature enhancements (data IO, PLS-DA, enrichment analysis) to deal with special cases in user inputs (*07/20/2015*);
- Updated Multivariate Biomarker Analysis module with flexible interface and improved capacity for computing on large datasets (*06/05/2015*);
- MetaboAnalyst 3.0 paper is now available on the 2015 NAR web server issue

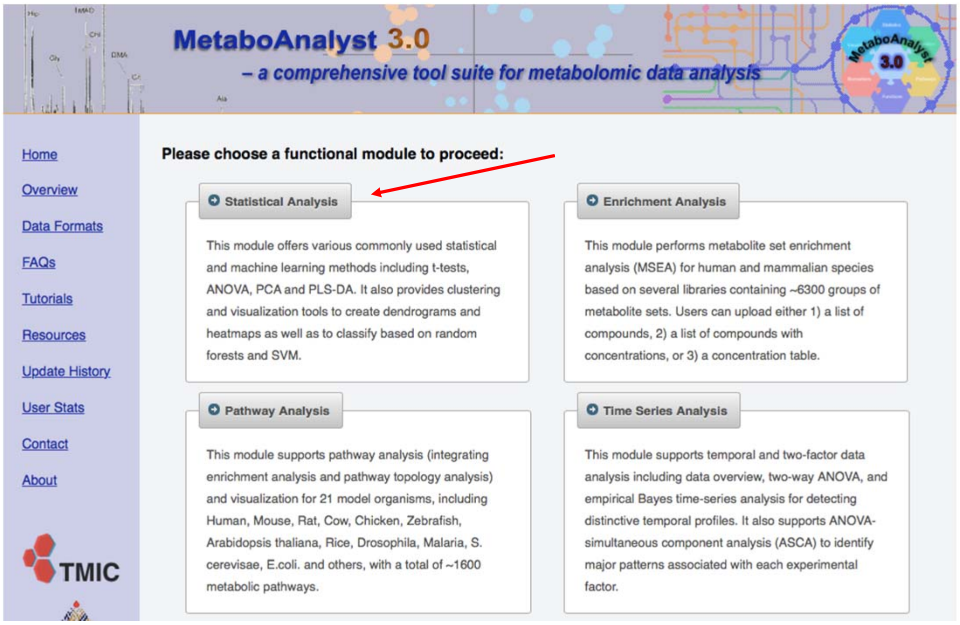
Read more ......

Please Cite:

Xia, J., Sinelnikov, I., Han, B., and Wishart, D.S. (2015) MetaboAnalyst 3.0 - making metabolomics more meaningful . Nucl. Acids Res. (DOI: 10.1093/nar/gkv380).

Xia, J., Mandal, R., Sinelnikov, I., Broadhurst, D., and Wishart, D.S. (2012) MetaboAnalyst 2.0 - a comprehensive server for metabolomic data analysis . Nucl. Acids Res. 40, W127-W133.

Xia, J., Psychogios, N., Young, N. and Wishart, D.S. (2009) MetaboAnalyst: a web server for metabolomic data analysis and interpretation. Nucl. Acids Res. 37, W652-660.



MetaboAnalyst 3.0

– a comprehensive tool suite for metabolomic data analysis

Please choose a functional module to proceed:

Statistical Analysis

This module offers various commonly used statistical and machine learning methods including t-tests, ANOVA, PCA and PLS-DA. It also provides clustering and visualization tools to create dendrograms and heatmaps as well as to classify based on random forests and SVM.

Enrichment Analysis

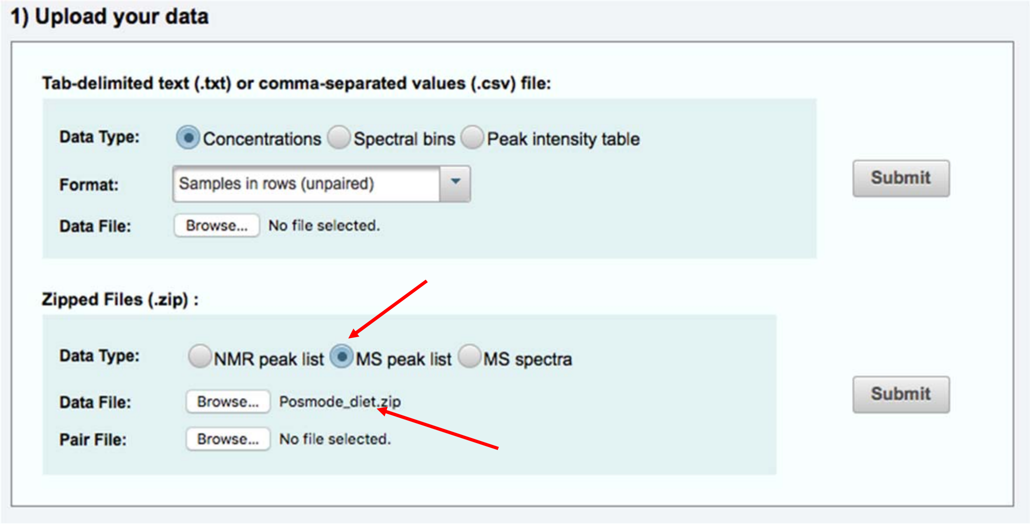
This module performs metabolite set enrichment analysis (MSEA) for human and mammalian species based on several libraries containing ~6300 groups of metabolite sets. Users can upload either 1) a list of compounds, 2) a list of compounds with concentrations, or 3) a concentration table.

Pathway Analysis

This module supports pathway analysis (integrating enrichment analysis and pathway topology analysis) and visualization for 21 model organisms, including Human, Mouse, Rat, Cow, Chicken, Zebrafish, Arabidopsis thaliana, Rice, Drosophila, Malaria, S. cerevisae, E.coli. and others, with a total of ~1600 metabolic pathways.

Time Series Analysis

This module supports temporal and two-factor data analysis including data overview, two-way ANOVA, and empirical Bayes time-series analysis for detecting distinctive temporal profiles. It also supports ANOVA-simultaneous component analysis (ASCA) to identify major patterns associated with each experimental factor.



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: Browse... No file selected.

Submit

Zipped Files (.zip) :

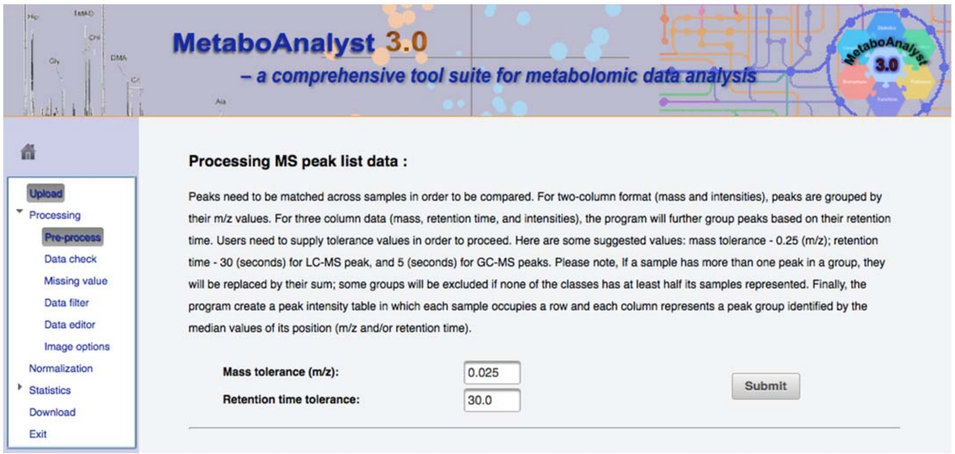
Data Type: ○ NMR peak list ● MS peak list ○ MS spectra

Data File: Browse... Posmode_diet.zip

Pair File: Browse... No file selected.

Submit

Select MS peak list option and then load the .zip file

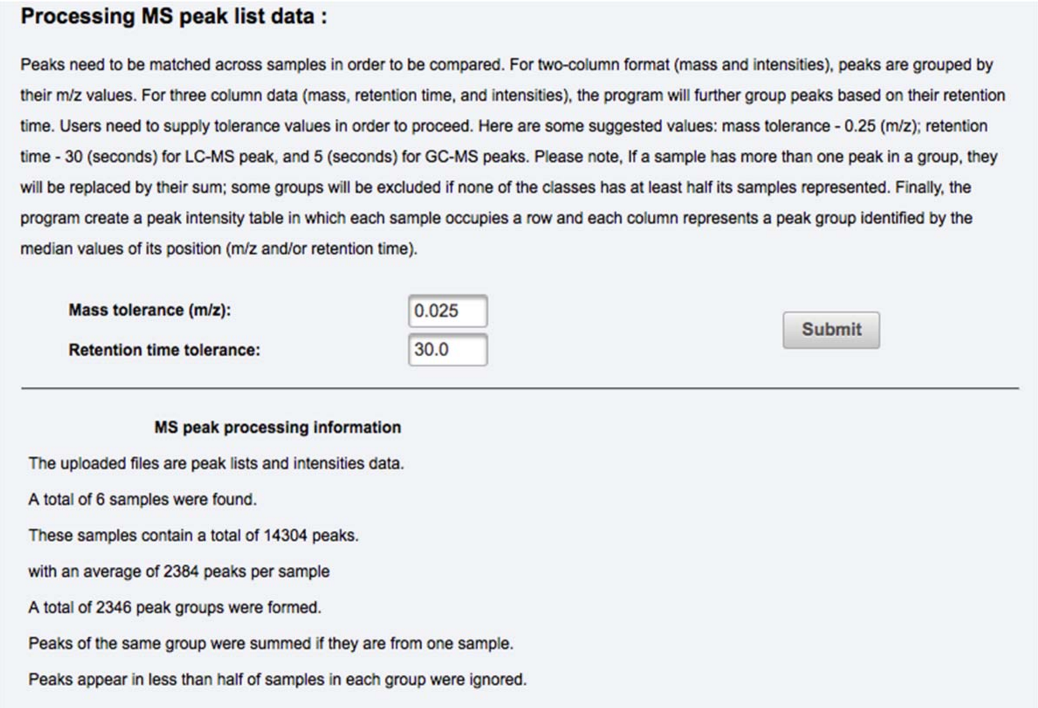


Processing MS peak list data :

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

Retention time tolerance: 30.0



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

Retention time tolerance: 30.0

MS peak processing information

The uploaded files are peak lists and intensities data.

A total of 6 samples were found.

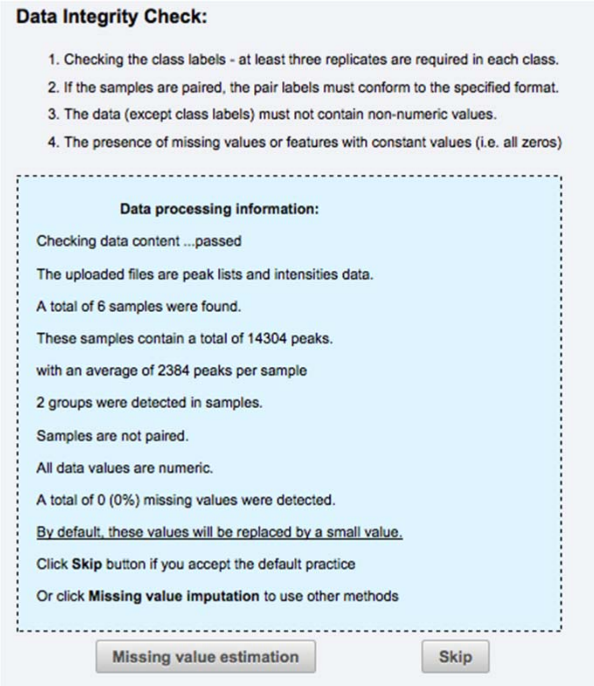
These samples contain a total of 14304 peaks.

with an average of 2384 peaks per sample

A total of 2346 peak groups were formed.

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

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



Data Integrity Check:

1. Checking the class labels - at least three replicates are required in each class.
2. If the samples are paired, the pair labels must conform to the specified format.
3. The data (except class labels) must not contain non-numeric values.
4. The presence of missing values or features with constant values (i.e. all zeros)

Data processing information:

Checking data content ...passed

The uploaded files are peak lists and intensities data.

A total of 6 samples were found.

These samples contain a total of 14304 peaks.

with an average of 2384 peaks per sample

2 groups were detected in samples.

Samples are not paired.

All data values are numeric.

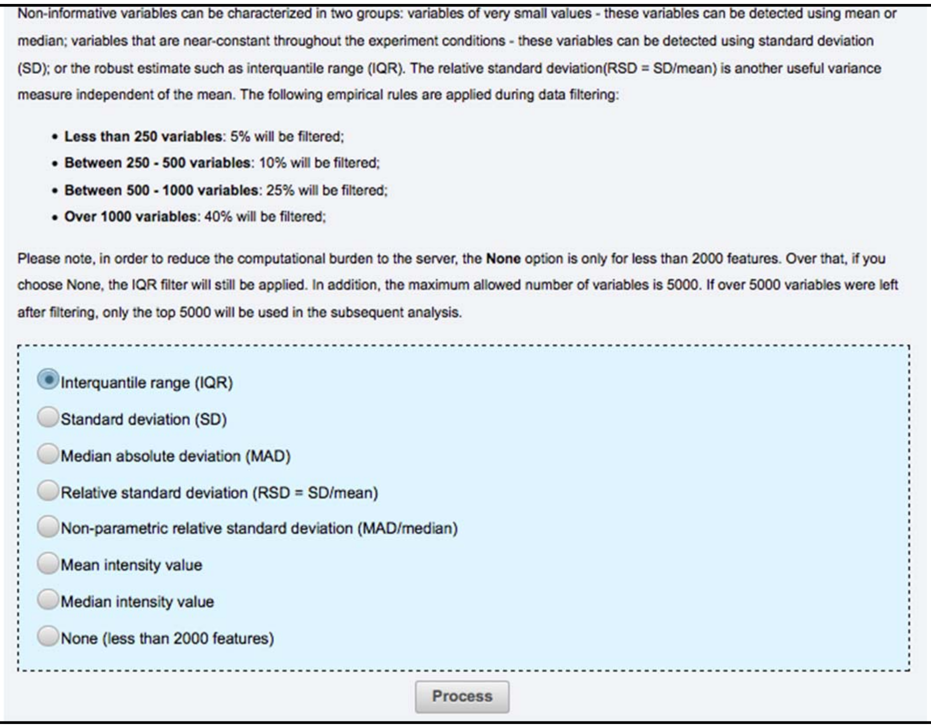
A total of 0 (0%) missing values were detected.

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

Note that XCMSonline filled in peaks

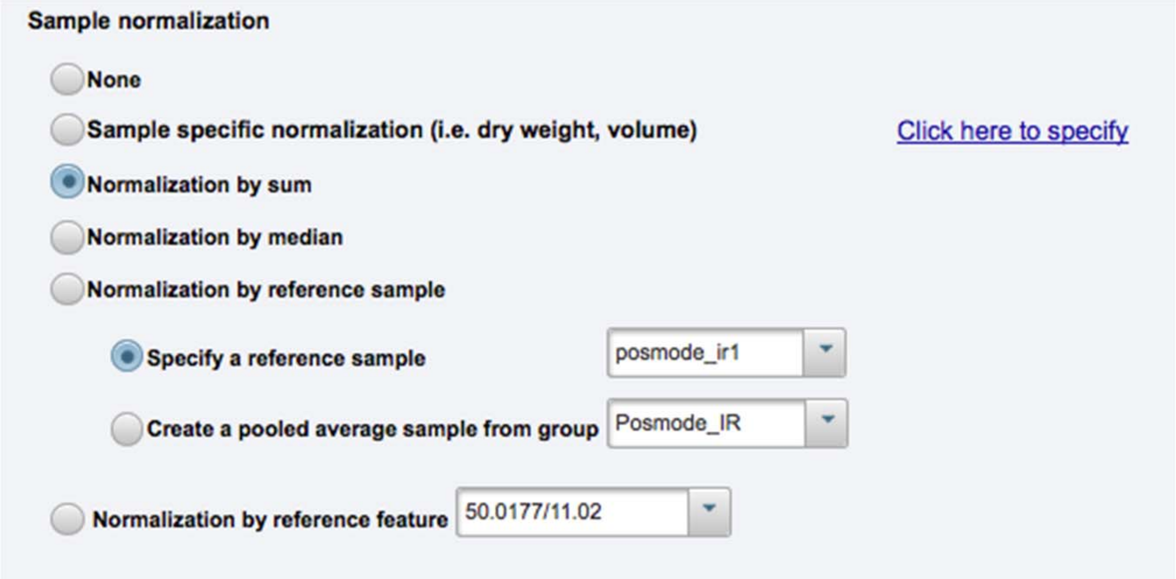


Non-informative variables can be characterized in two groups: variables of very small values - these variables can be detected using mean or median; variables that are near-constant throughout the experiment conditions - these variables can be detected using standard deviation (SD); or the robust estimate such as interquantile range (IQR). The relative standard deviation(RSD = SD/mean) is another useful variance measure independent of the mean. The following empirical rules are applied during data filtering:

- **Less than 250 variables**: 5% will be filtered;
- **Between 250 - 500 variables**: 10% will be filtered;
- **Between 500 - 1000 variables**: 25% will be filtered;
- **Over 1000 variables**: 40% will be filtered;

Please note, in order to reduce the computational burden to the server, the **None** option is only for less than 2000 features. Over that, if you choose None, the IQR filter will still be applied. In addition, the maximum allowed number of variables is 5000. If over 5000 variables were left after filtering, only the top 5000 will be used in the subsequent analysis.

● 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
○ None (less than 2000 features)



Sample normalization

○ None

○ Sample specific normalization (i.e. dry weight, volume) — Click here to specify

● Normalization by sum

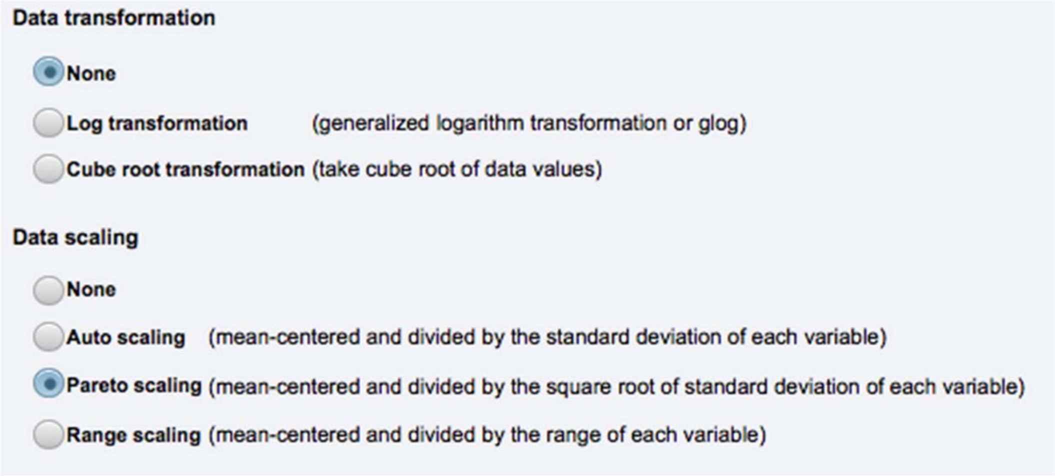
○ Normalization by median

○ Normalization by reference sample

- ● Specify a reference sample: posmode_ir1
- ○ Create a pooled average sample from group: Posmode_IR

○ Normalization by reference feature: 50.0177/11.02

Data options before stats analysis



Data transformation

● None

○ Log transformation (generalized logarithm transformation or glog)

○ Cube root transformation (take cube root of data values)

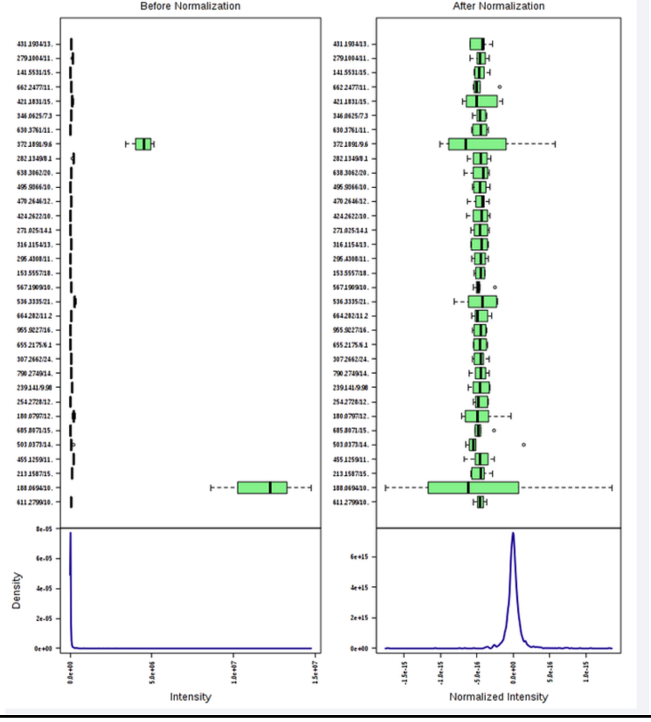
Data scaling

○ None

○ Auto scaling (mean-centered and divided by the standard deviation of each variable)

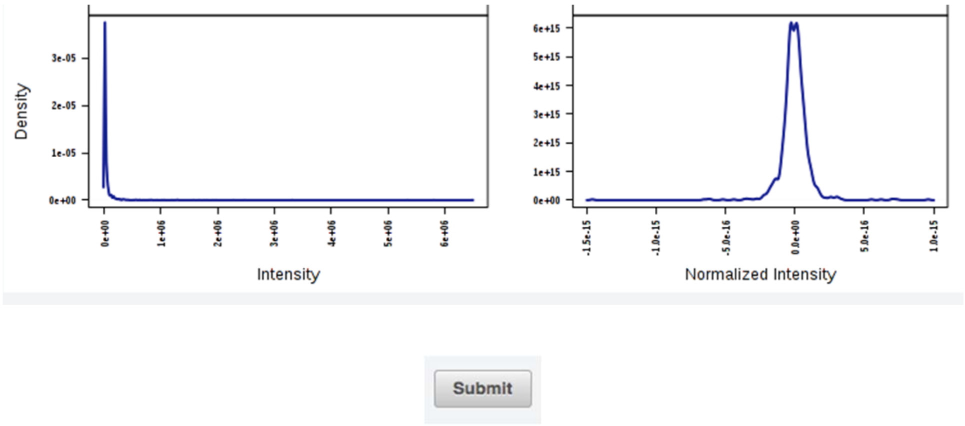
● Pareto scaling (mean-centered and divided by the square root of standard deviation of each variable)

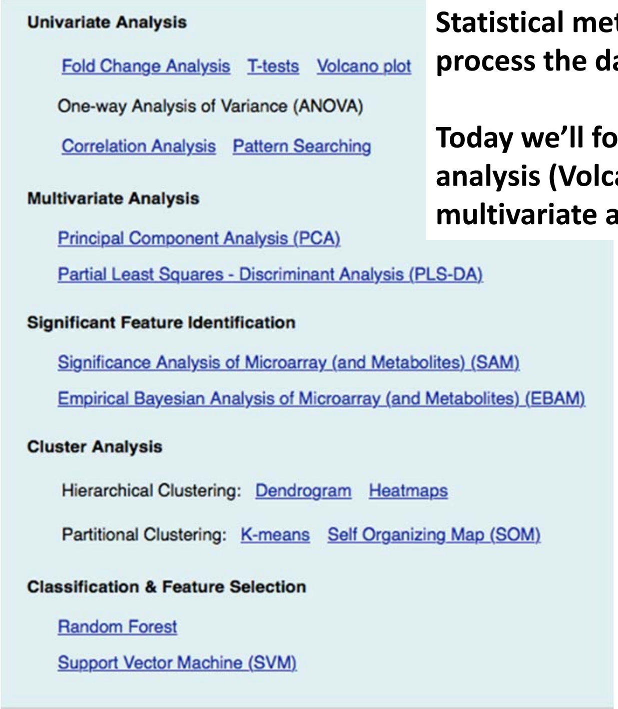
○ Range scaling (mean-centered and divided by the range of each variable)



Now (on the right) the data are mean-centered and variance is reduced by dividing by the standard error of the mean

Effect of normalization, mean centering and Pareto scaling





Univariate Analysis

Fold Change Analysis   T-tests   Volcano plot

One-way Analysis of Variance (ANOVA)

Correlation Analysis   Pattern Searching

Multivariate Analysis

Principal Component Analysis (PCA)

Partial Least Squares - Discriminant Analysis (PLS-DA)

Significant Feature Identification

Significance Analysis of Microarray (and Metabolites) (SAM)

Empirical Bayesian Analysis of Microarray (and Metabolites) (EBAM)

Cluster Analysis

Hierarchical Clustering: Dendrogram   Heatmaps

Partitional Clustering: K-means   Self Organizing Map (SOM)

Classification & Feature Selection

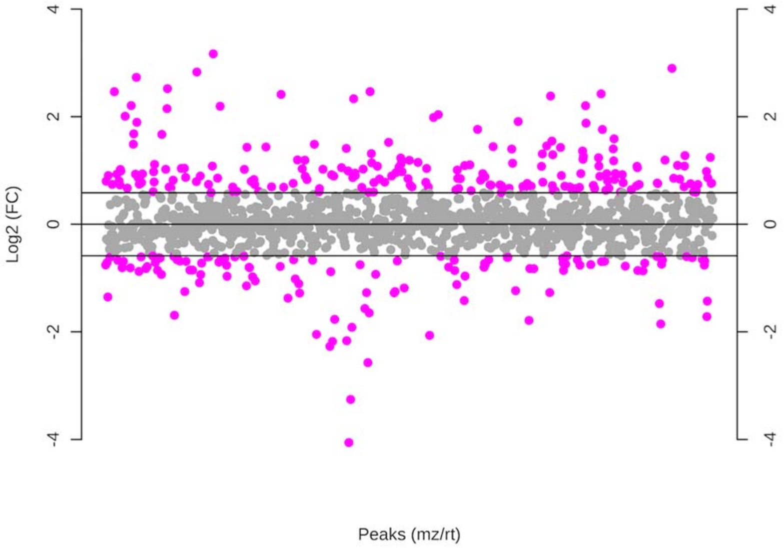
Random Forest

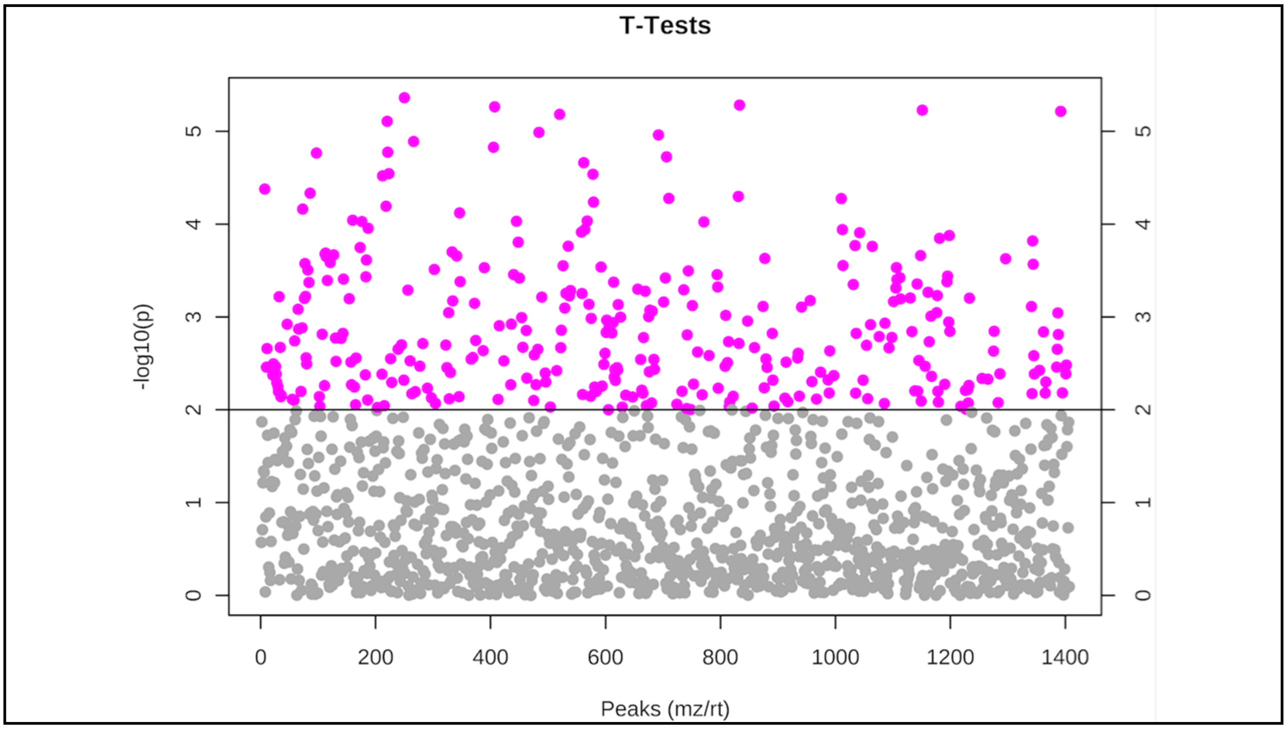
Support Vector Machine (SVM)

Statistical methods available to process the data on MetaboAnalyst

Today we'll focus on univariate analysis (Volcano plots) and multivariate analysis (PCA and PLS-DA)

Fold-change - pink = >1.5

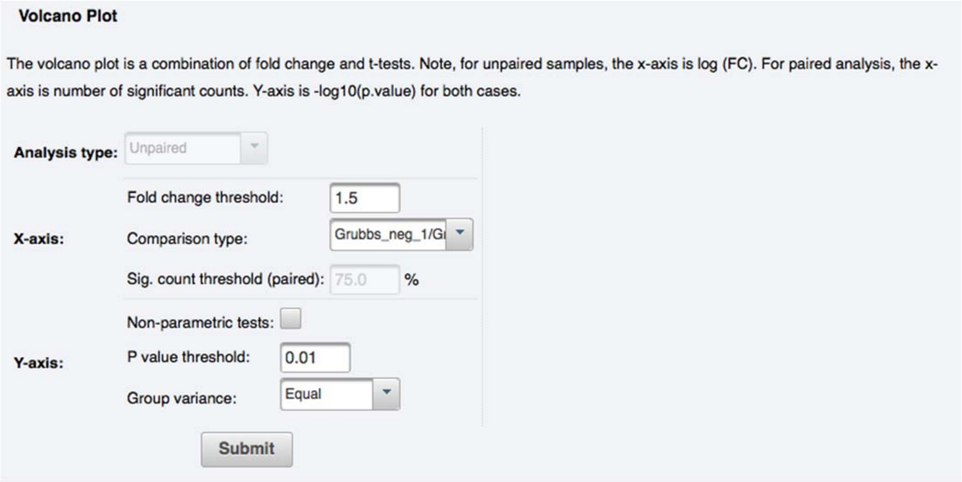




Univariate analysis – Volcano plot

- This is a useful measure
- The data points represent comparisons made one metabolite at a time
- The plot identifies metabolites with p-values <0.01 and a fold change >1.5

Volcano plot set up



Volcano Plot

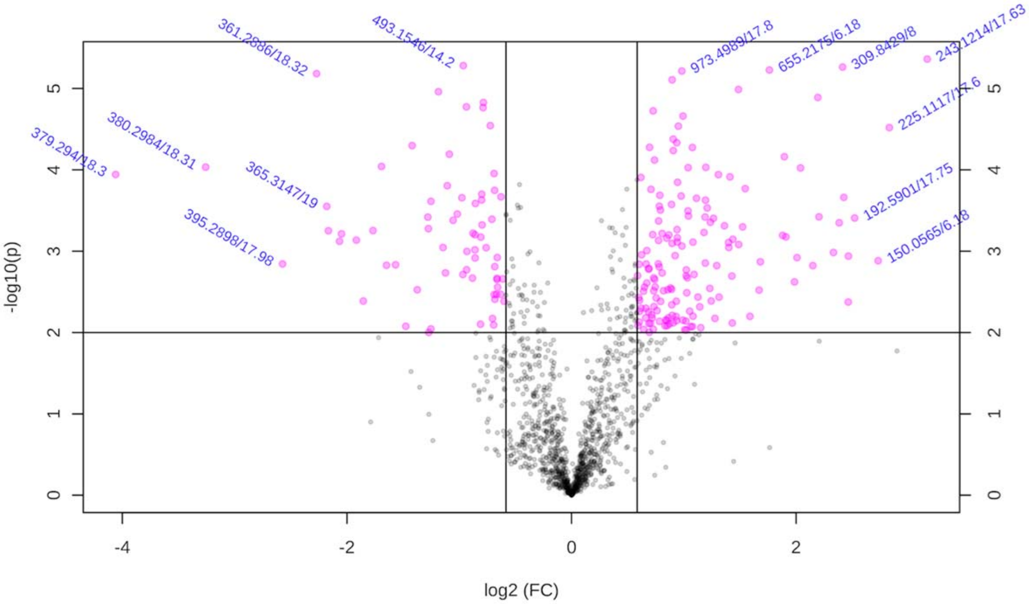
The volcano plot is a combination of fold change and t-tests. Note, for unpaired samples, the x-axis is log (FC). For paired analysis, the x-axis is number of significant counts. Y-axis is -log10(p.value) for both cases.

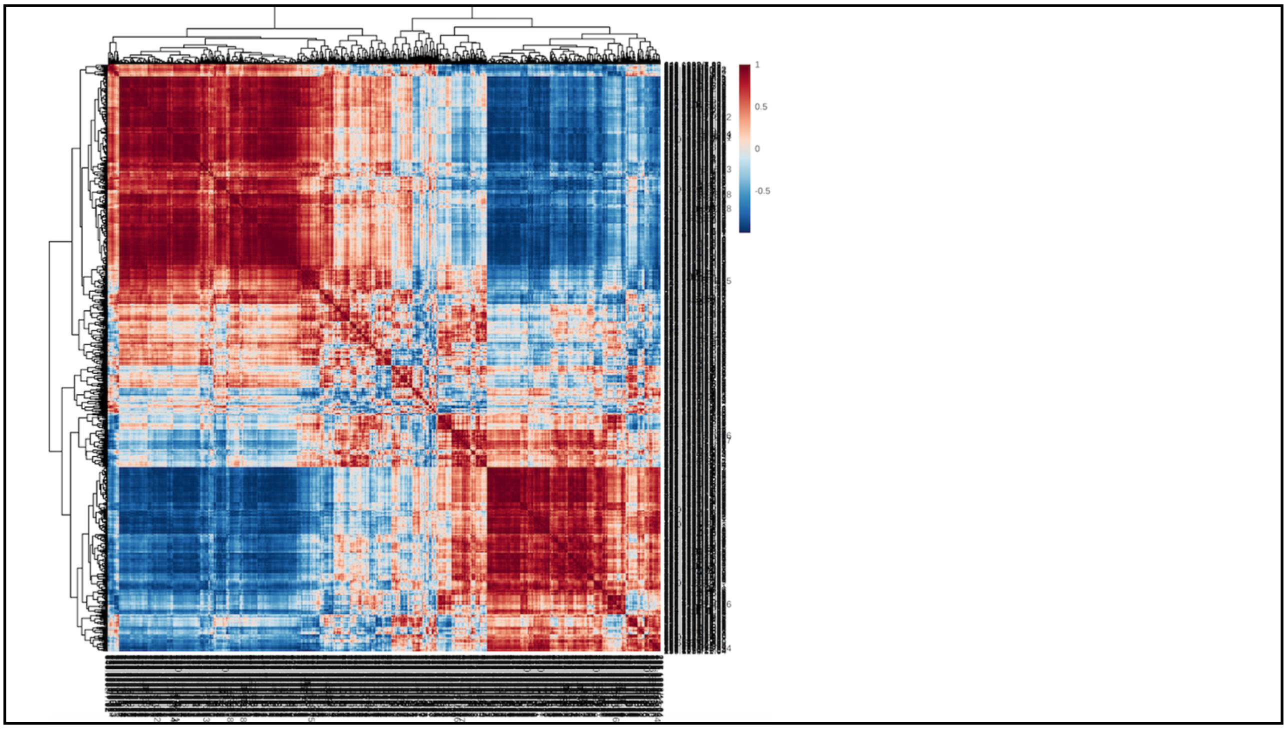
Analysis type: Unpaired

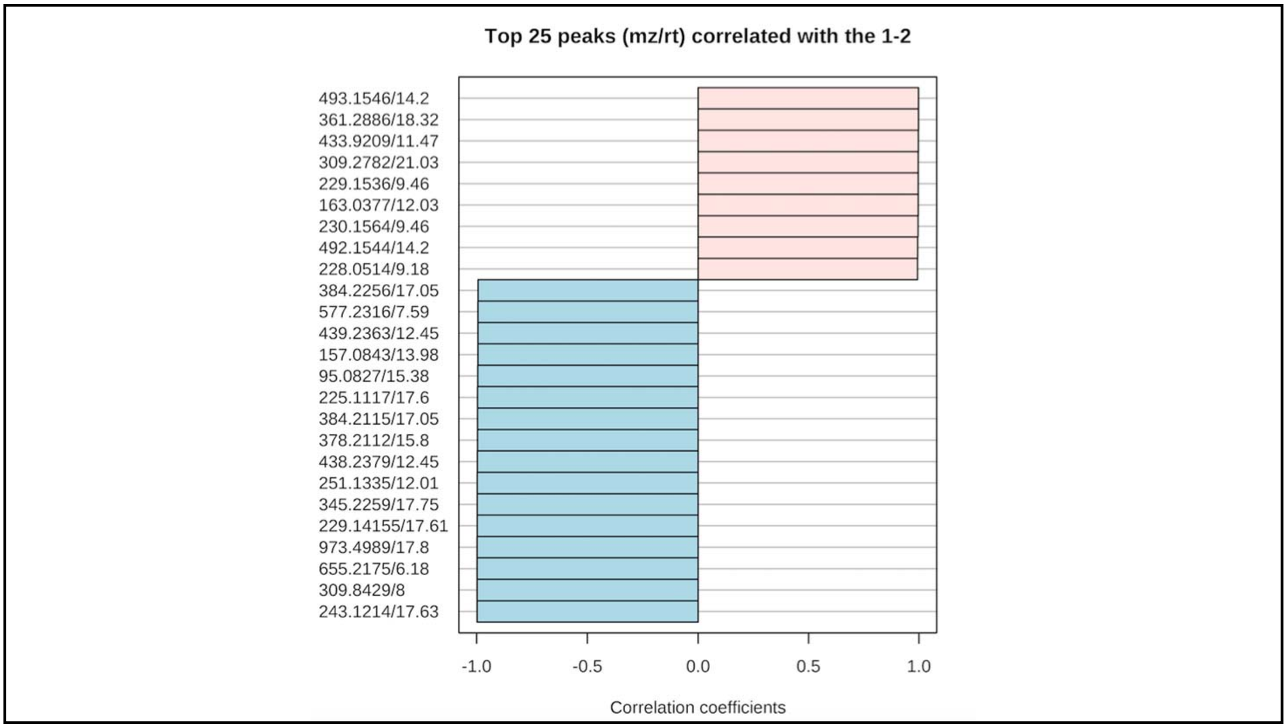
X-axis:
Fold change threshold: 1.5
Comparison type: Grubbs_neg_1/Gı
Sig. count threshold (paired): 75.0 %

Y-axis:
Non-parametric tests: □
P value threshold: 0.01
Group variance: Equal

Volcano plot with fold change=1.5 and p <0.01



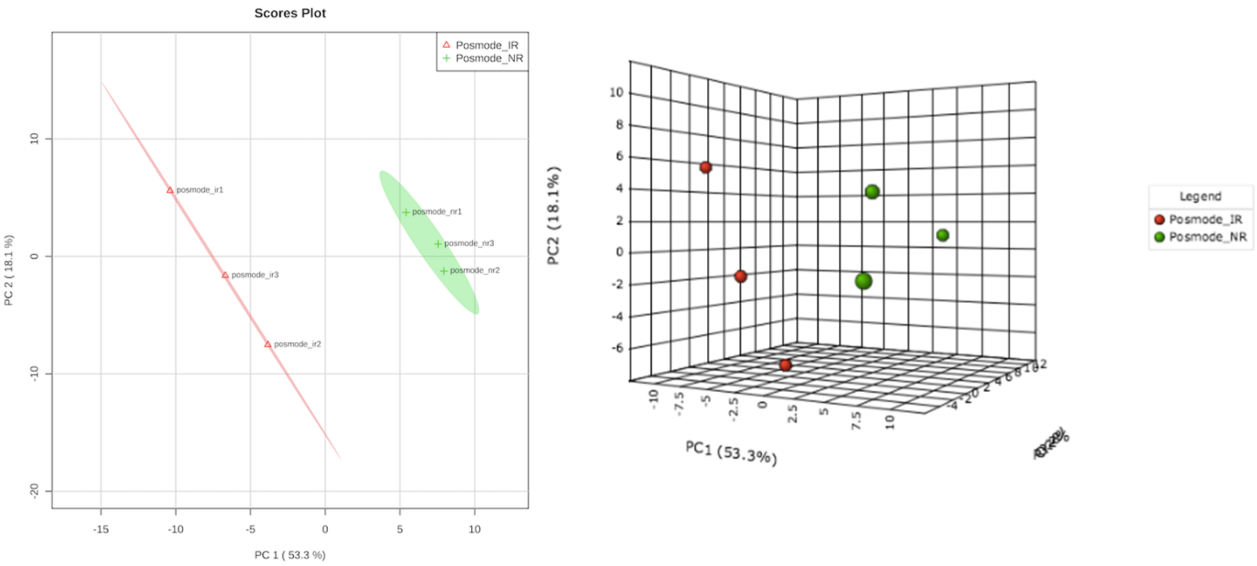



Top 25 peaks (mz/rt) correlated with the 1-2

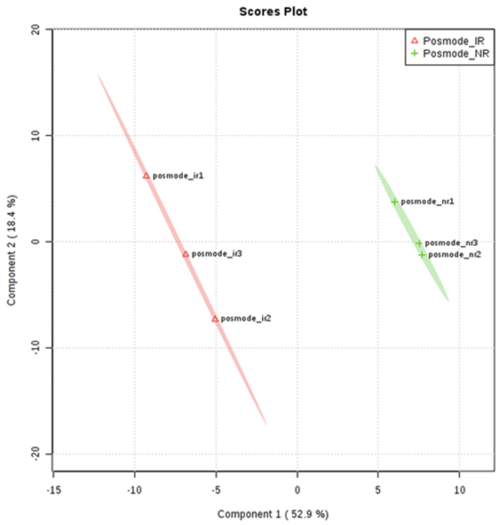
Multivariate Analysis

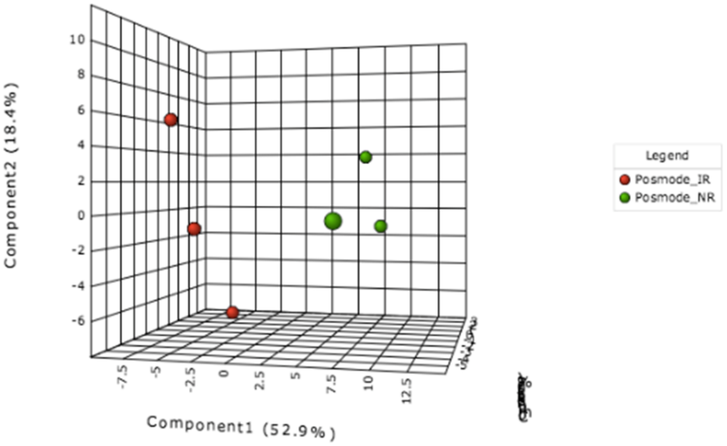
- In this type of analysis, principal components are identified.
- Contributions from all the metabolites to each principal component are calculated, reducing >1000 values to one number for each component
- When this done in an unsupervised way, the group information is added after principal component analysis (PCA)
- Partial least squares discriminant analysis (PLS-DA) is a supervised procedure and group information is included in the analysis

2D- and 3D-PCA plots

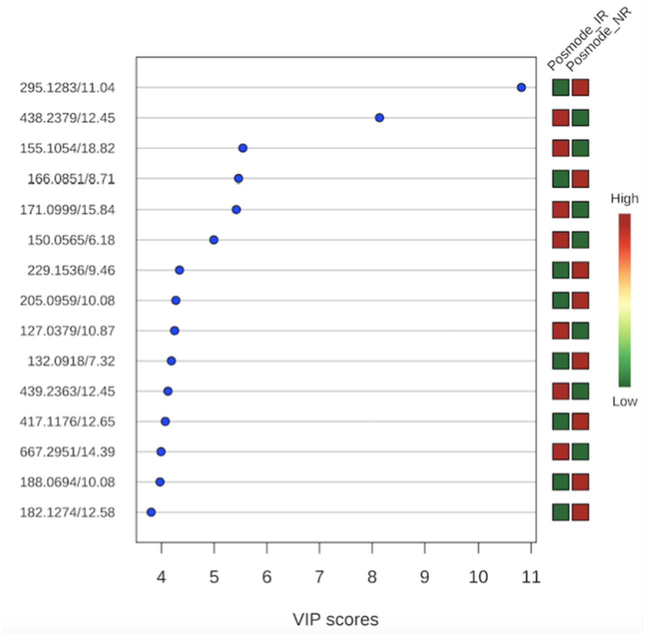


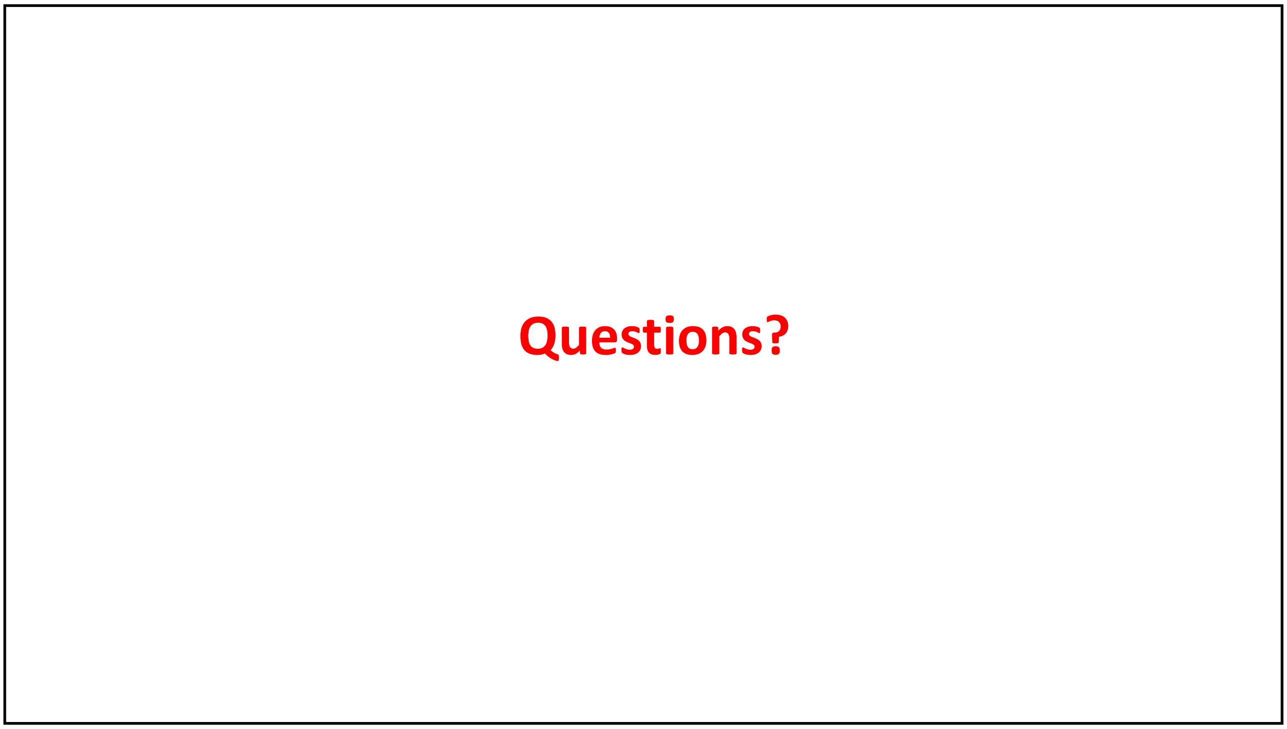
2D- and 3D-PLS-DA plots





From download file – plsda_vip.csv file





Questions?