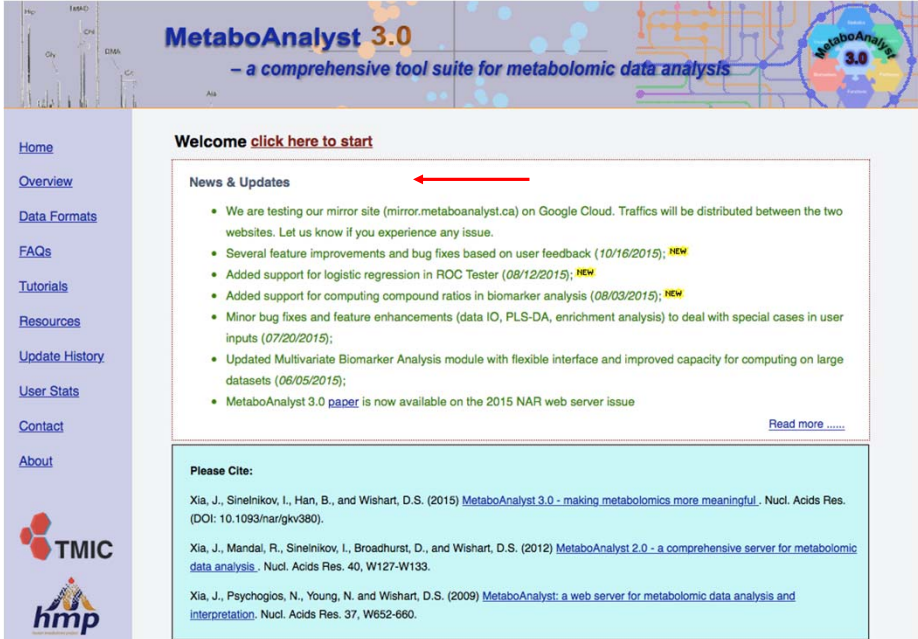


**UAB BLAZERS** *Knowledge that will change your world*

## Statistical analysis with MetaboAnalyst

Stephen Barnes and Xiangqin Cui  
University of Alabama at Birmingham  
[sbarnes@uab.edu](mailto:sbarnes@uab.edu) and [xcui@uab.edu](mailto:xcui@uab.edu)



**MetaboAnalyst 3.0**  
– a comprehensive tool suite for metabolomic data analysis

Welcome [click here to start](#)

**News & Updates**

- We are testing our mirror site ([mirror.metaboanalyst.ca](http://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 (08/05/2015);
- MetaboAnalyst 3.0 [paper](#) is now available on the 2015 NAR web server issue

[Read more .....](#)

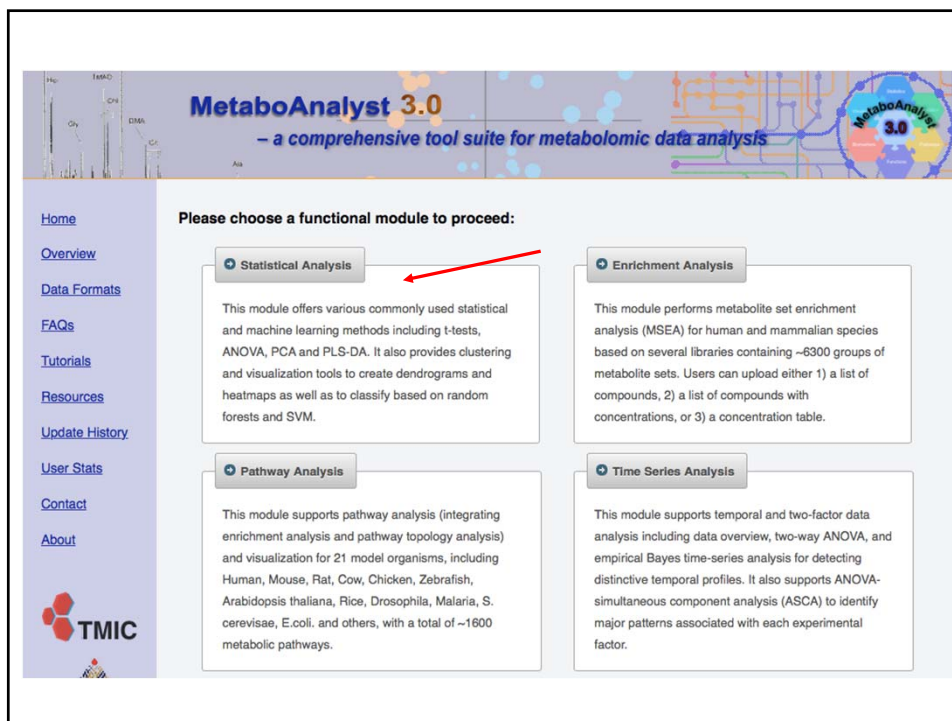
**Please Cite:**

Xia, J., Sinenikov, 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., Sinenikov, 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.

TMIC  
hmp



The image shows the homepage of MetaboAnalyst 3.0. At the top, there is a header with the text "MetaboAnalyst 3.0 - a comprehensive tool suite for metabolomic data analysis" and a logo on the right. Below the header is a navigation menu on the left with links for Home, Overview, Data Formats, FAQs, Tutorials, Resources, Update History, User Stats, Contact, and About. The main content area is titled "Please choose a functional module to proceed:" and contains four modules: Statistical Analysis, Enrichment Analysis, Pathway Analysis, and Time Series Analysis. Each module has a brief description of its capabilities. A red arrow points to the "Statistical Analysis" module.

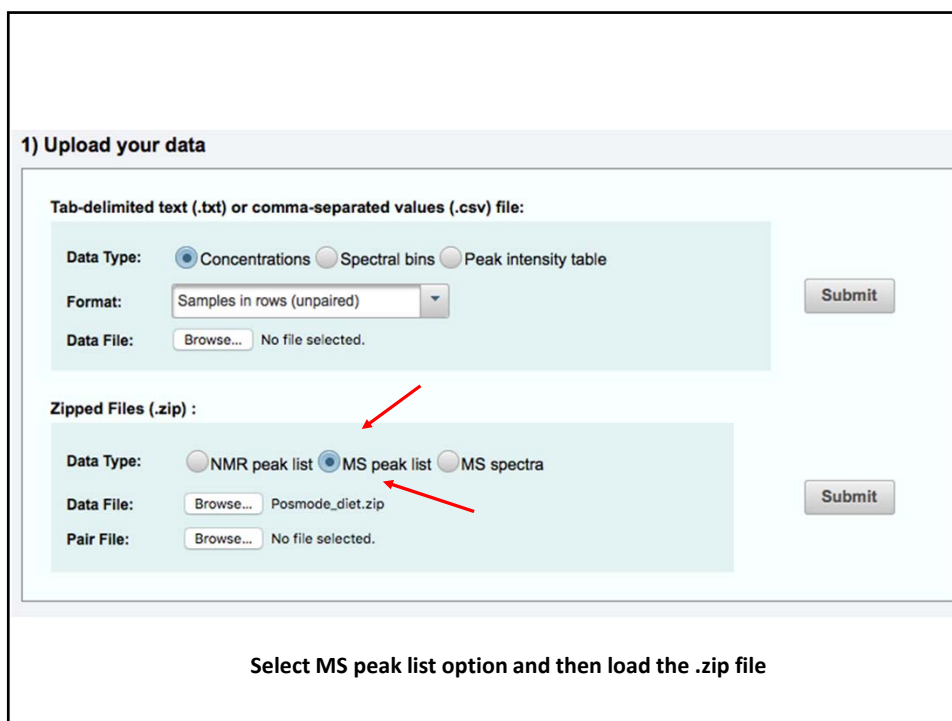
**MetaboAnalyst 3.0**  
– a comprehensive tool suite for metabolomic data analysis

[Home](#)  
[Overview](#)  
[Data Formats](#)  
[FAQs](#)  
[Tutorials](#)  
[Resources](#)  
[Update History](#)  
[User Stats](#)  
[Contact](#)  
[About](#)

**TMIC**

**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. cerevisiae, 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.



The image shows the "1) Upload your data" section of the MetaboAnalyst 3.0 interface. It is divided into two main sections: "Tab-delimited text (.txt) or comma-separated values (.csv) file:" and "Zipped Files (.zip) :". The first section has radio buttons for "Concentrations" (selected), "Spectral bins", and "Peak intensity table", a dropdown menu for "Format" set to "Samples in rows (unpaired)", and a "Data File" field with a "Browse..." button and "No file selected." text. The second section has radio buttons for "NMR peak list", "MS peak list" (selected), and "MS spectra", a "Data File" field with a "Browse..." button and "Posmode\_diet.zip" text, and a "Pair File" field with a "Browse..." button and "No file selected." text. Both sections have a "Submit" button. A red arrow points to the "MS peak list" radio button, and another red arrow points to the "Data File" field in the second section.

**1) Upload your data**

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

Data Type:  Concentrations  Spectral bins  Peak intensity table

Format:

Data File:  No file selected.

**Zipped Files (.zip) :**

Data Type:  NMR peak list  MS peak list  MS spectra

Data File:  Posmode\_diet.zip

Pair File:  No file selected.

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

**MetaboAnalyst 3.0**  
– a comprehensive tool suite for metabolomic data analysis

**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 ( $m/z$ ):

Retention time tolerance:

**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 ( $m/z$ ):

Retention time tolerance:

---

**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**

Missing value estimation

Skip

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 interquartile 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.

- Interquartile 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)

Process

**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

Create a pooled average sample from group

Normalization by reference feature

## Data options before stats analysis

**Data transformation**

None

Log transformation (generalized logarithm transformation or glog)

Cube root transformation (take cube root of data values)

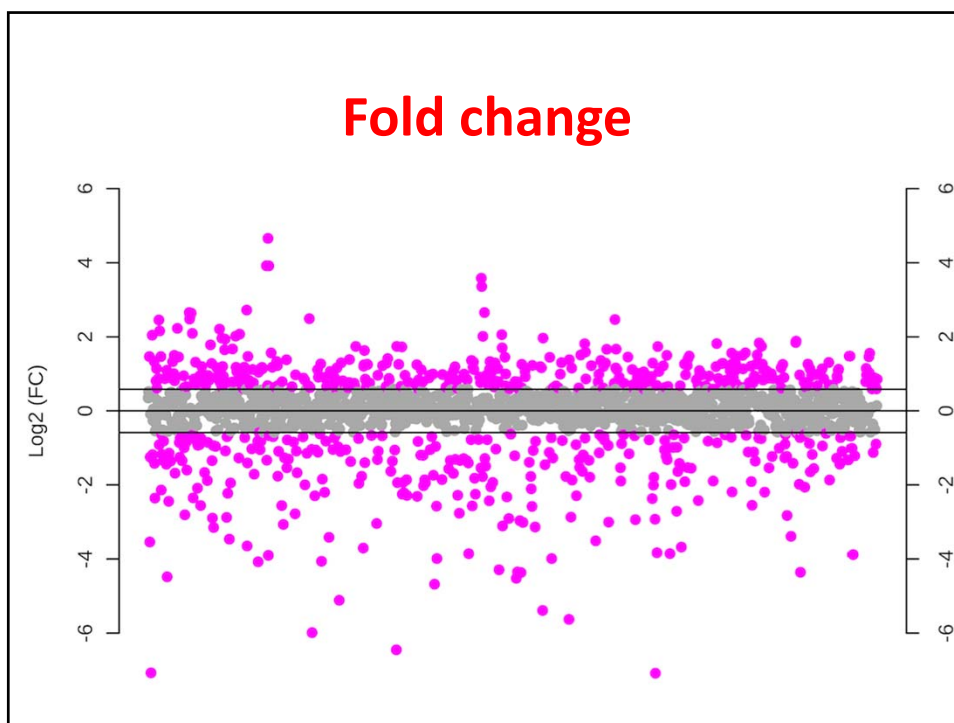
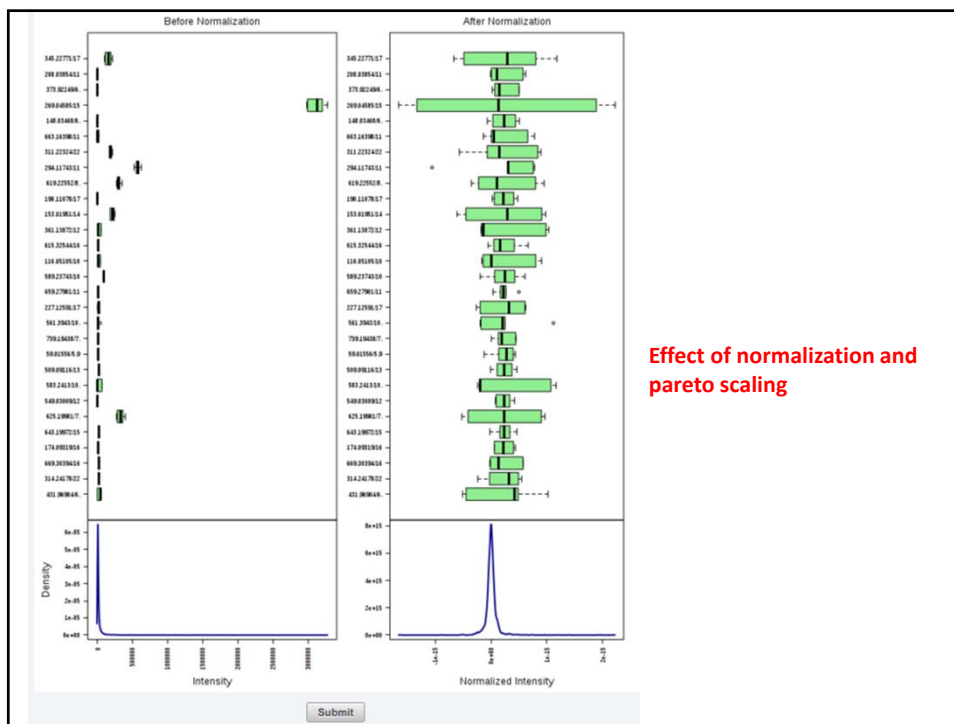
**Data scaling**

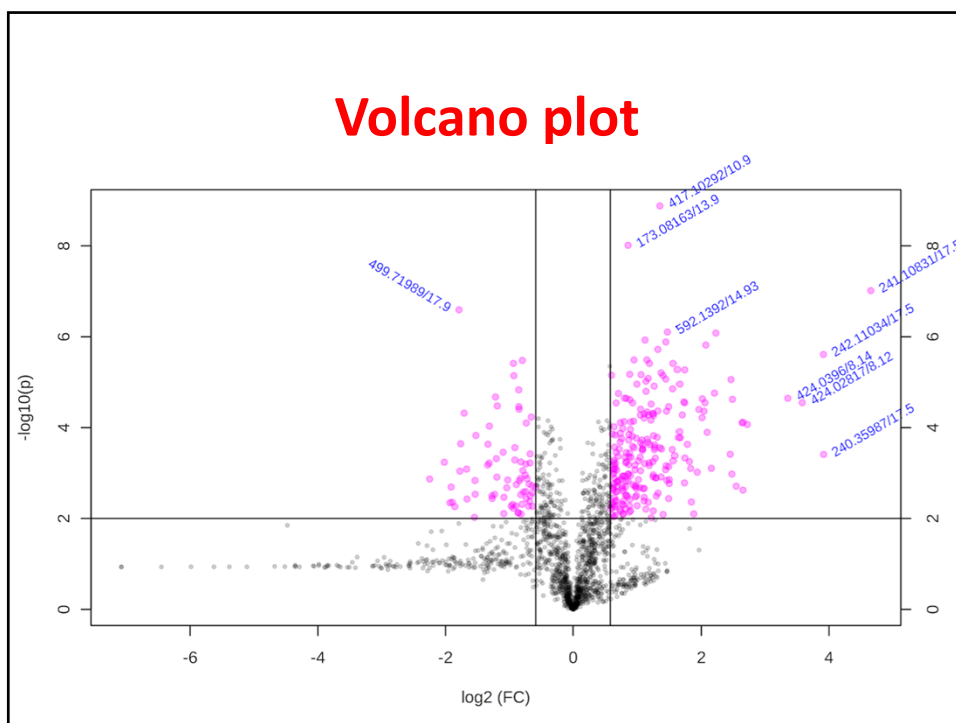
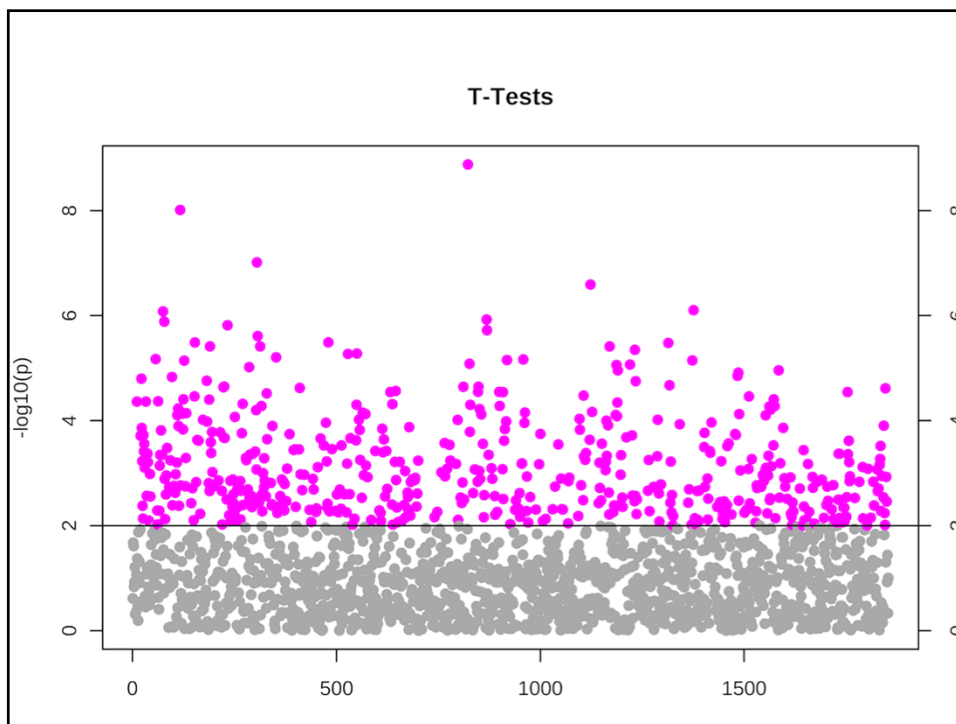
None

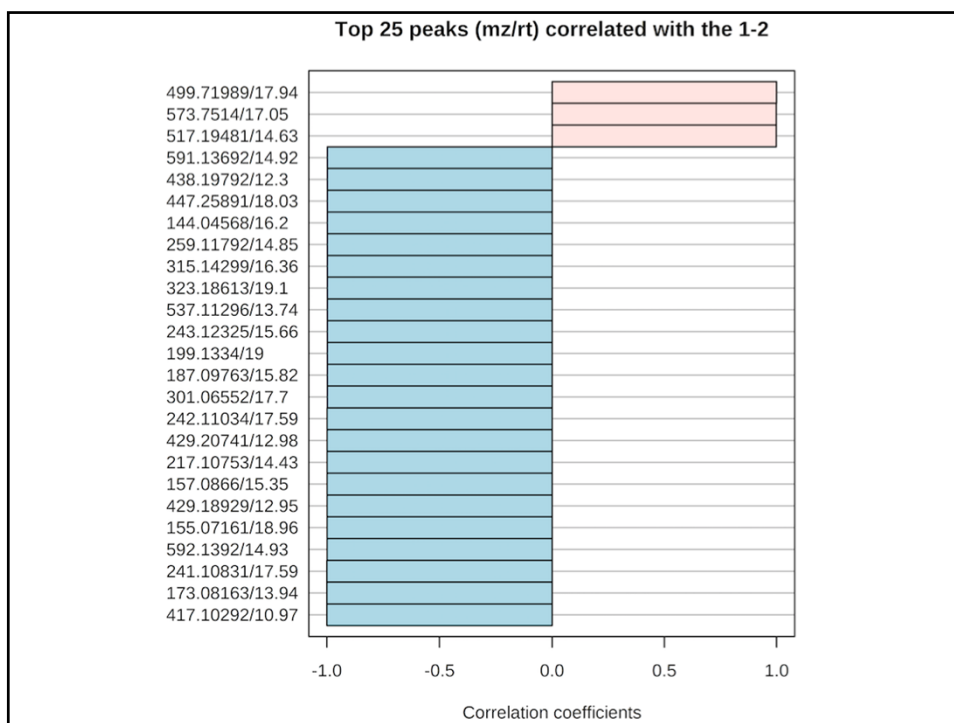
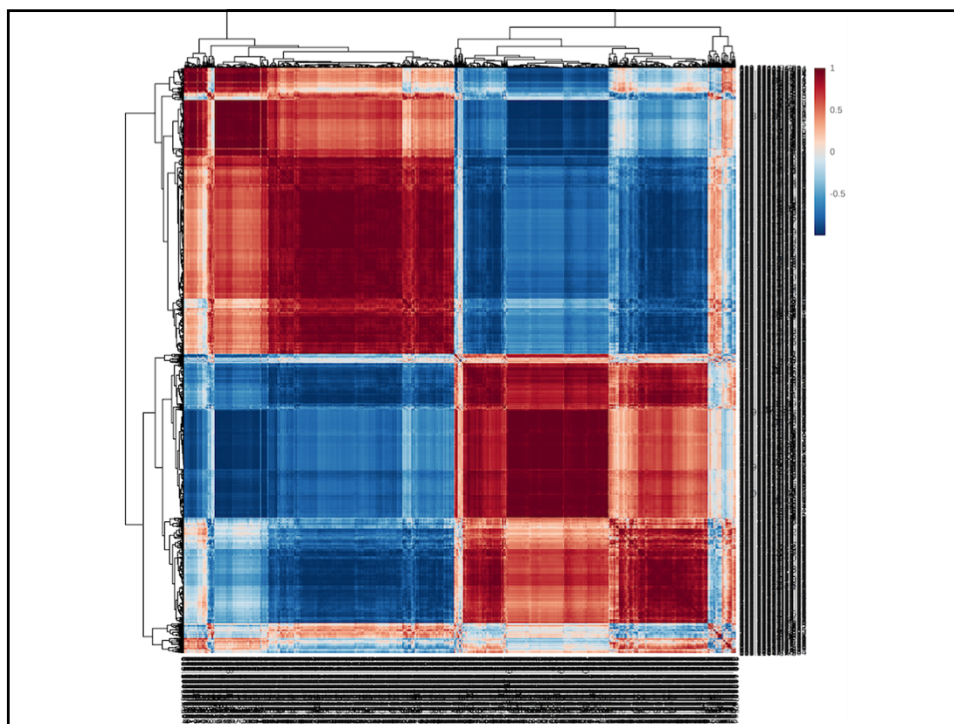
Auto scaling (mean-centered and divided by the standard deviation of each variable)

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)

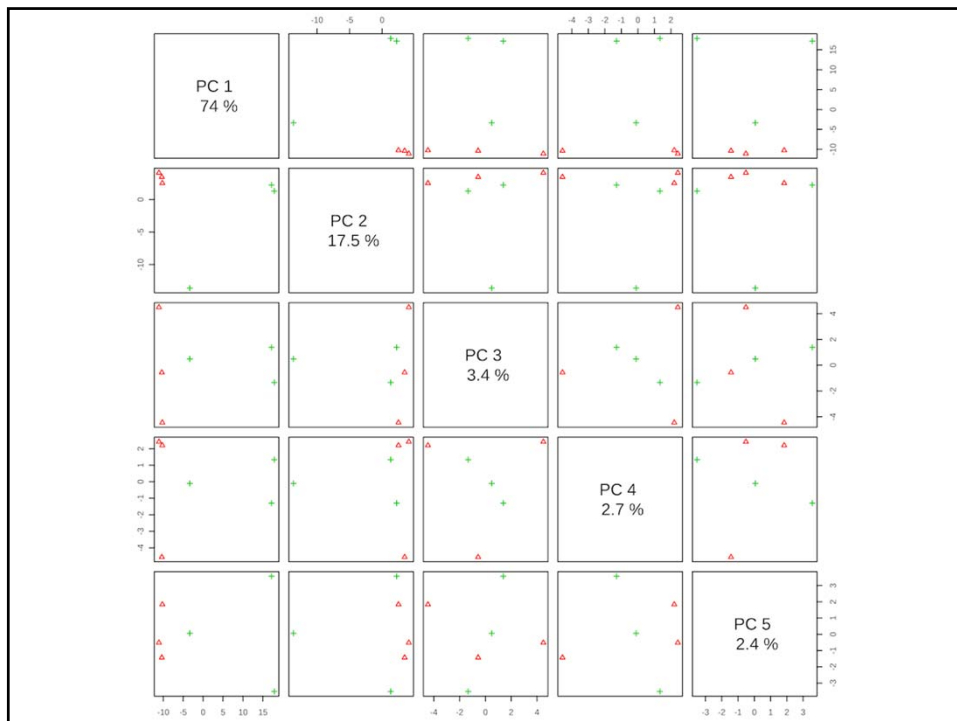


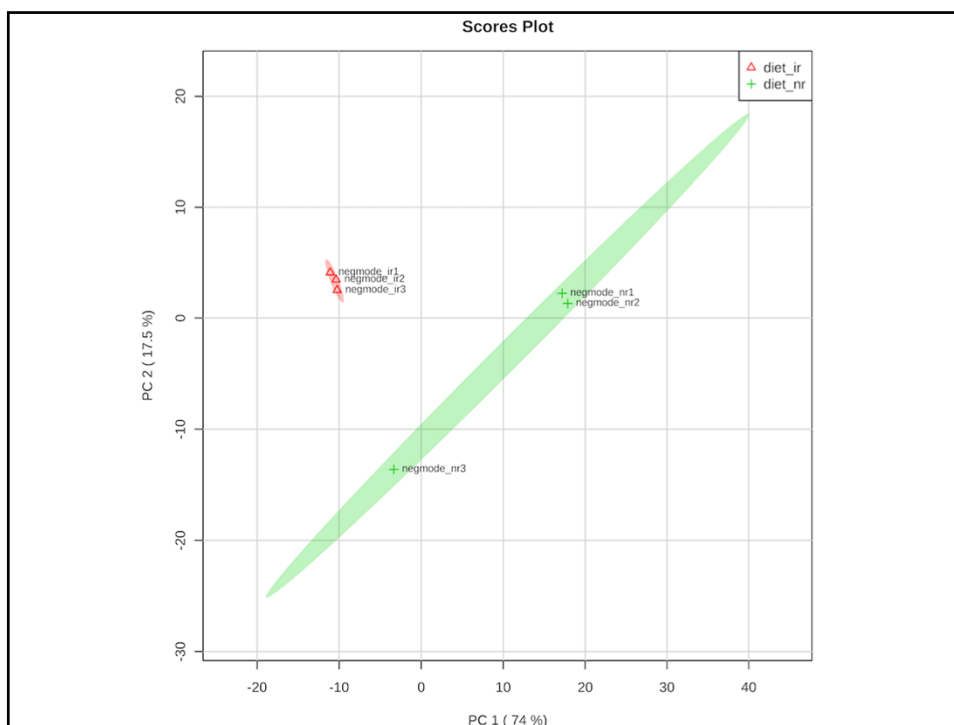
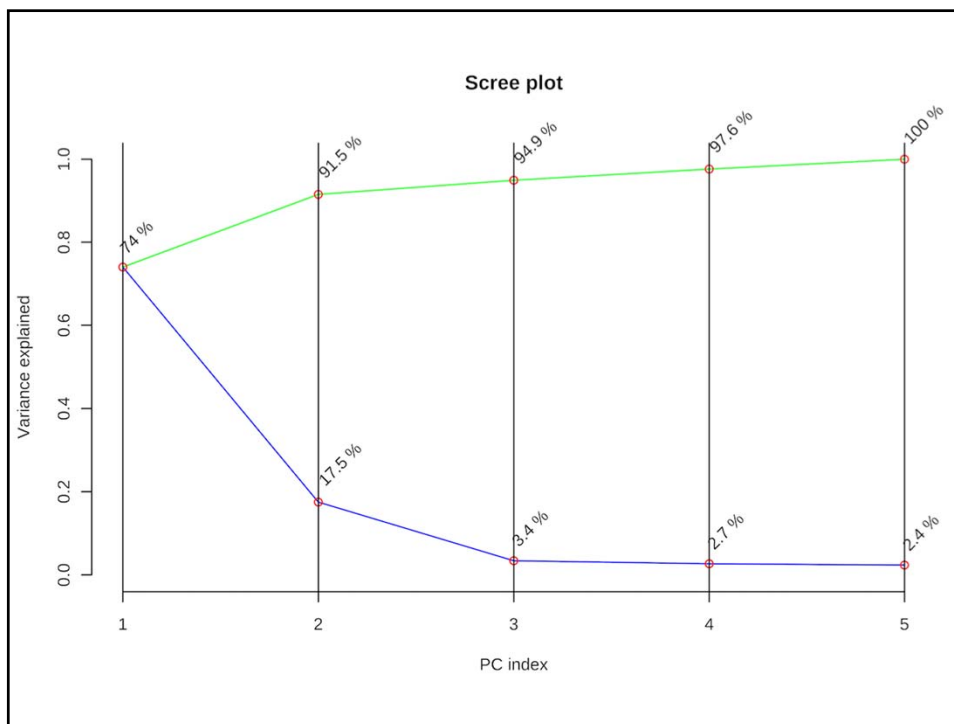


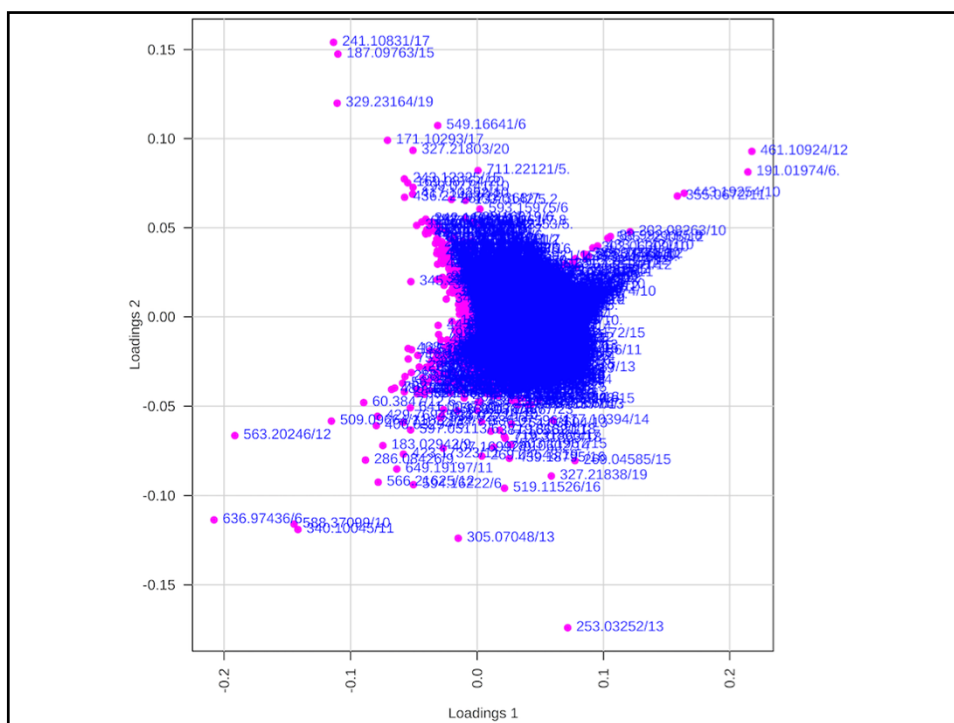
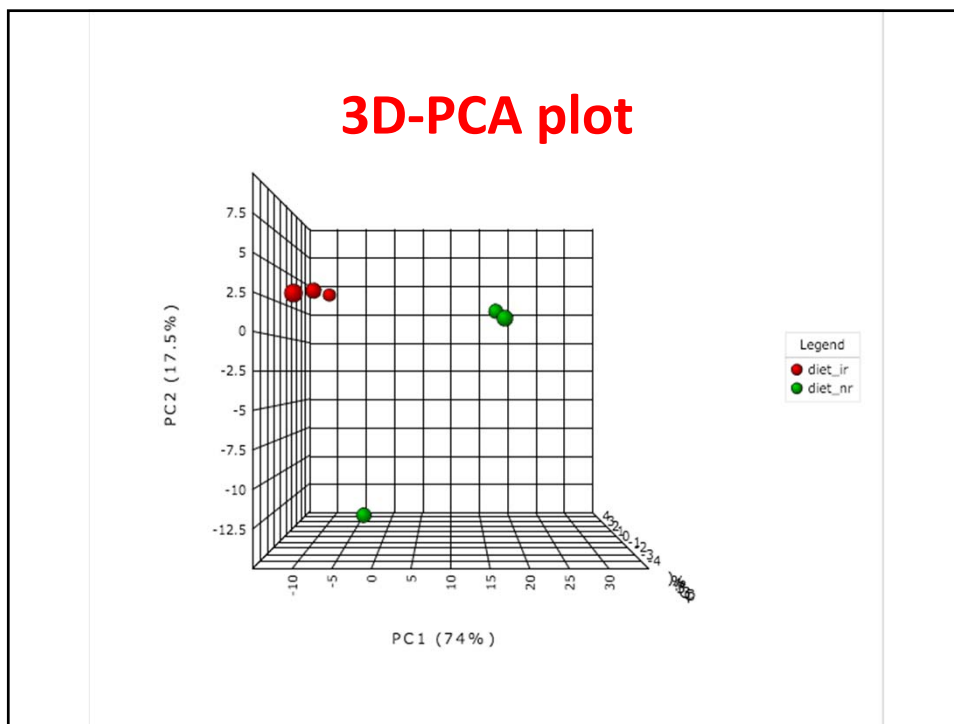




# Principal components analysis







## Partial least squares discriminant analysis

