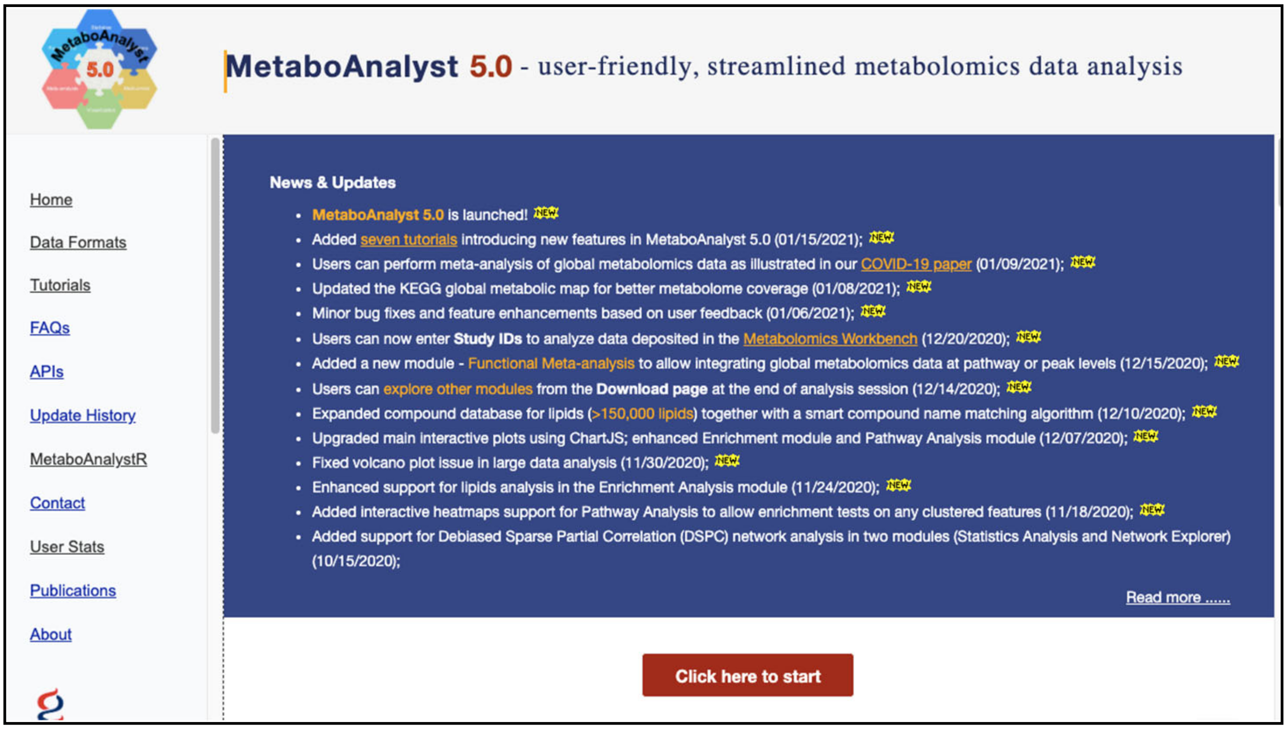
UAB BLAZERS

Knowledge that will change your world

Metaboanalyst

Stephen Barnes, PhD
BBRB 709; 205-934-7117
sbarnes@uab.edu

1



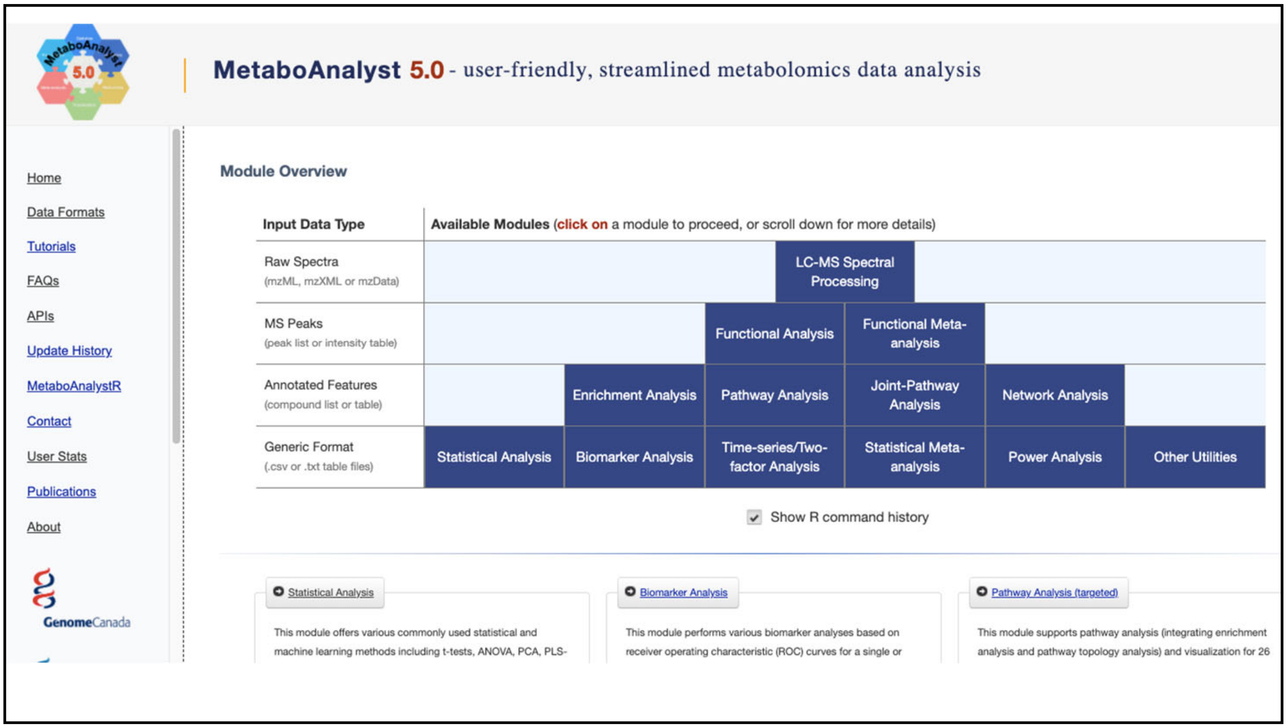
MetaboAnalyst 5.0 - user-friendly, streamlined metabolomics data analysis

News & Updates

- MetaboAnalyst 5.0 is launched! NEW
- Added seven tutorials introducing new features in MetaboAnalyst 5.0 (01/15/2021); NEW
- Users can perform meta-analysis of global metabolomics data as illustrated in our COVID-19 paper (01/09/2021); NEW
- Updated the KEGG global metabolic map for better metabolome coverage (01/08/2021); NEW
- Minor bug fixes and feature enhancements based on user feedback (01/06/2021); NEW
- Users can now enter **Study IDs** to analyze data deposited in the Metabolomics Workbench (12/20/2020); NEW
- Added a new module - Functional Meta-analysis to allow integrating global metabolomics data at pathway or peak levels (12/15/2020); NEW
- Users can explore other modules from the **Download page** at the end of analysis session (12/14/2020); NEW
- Expanded compound database for lipids (>150,000 lipids) together with a smart compound name matching algorithm (12/10/2020); NEW
- Upgraded main interactive plots using ChartJS; enhanced Enrichment module and Pathway Analysis module (12/07/2020); NEW
- Fixed volcano plot issue in large data analysis (11/30/2020); NEW
- Enhanced support for lipids analysis in the Enrichment Analysis module (11/24/2020); NEW
- Added interactive heatmaps support for Pathway Analysis to allow enrichment tests on any clustered features (11/18/2020); NEW
- Added support for Debiased Sparse Partial Correlation (DSPC) network analysis in two modules (Statistics Analysis and Network Explorer) (10/15/2020);

Read more ......

2



MetaboAnalyst 5.0 - user-friendly, streamlined metabolomics data analysis

Module Overview

Input Data Type	Available Modules (click on a module to proceed, or scroll down for more details)					
Raw Spectra (mzML, mzXML or mzData)			LC-MS Spectral Processing			
MS Peaks (peak list or intensity table)			Functional Analysis	Functional Meta-analysis		
Annotated Features (compound list or table)		Enrichment Analysis	Pathway Analysis	Joint-Pathway Analysis	Network Analysis	
Generic Format (.csv or .txt table files)	Statistical Analysis	Biomarker Analysis	Time-series/Two-factor Analysis	Statistical Meta-analysis	Power Analysis	Other Utilities

☑ Show R command history

Statistical Analysis

This module offers various commonly used statistical and machine learning methods including t-tests, ANOVA, PCA, PLS-

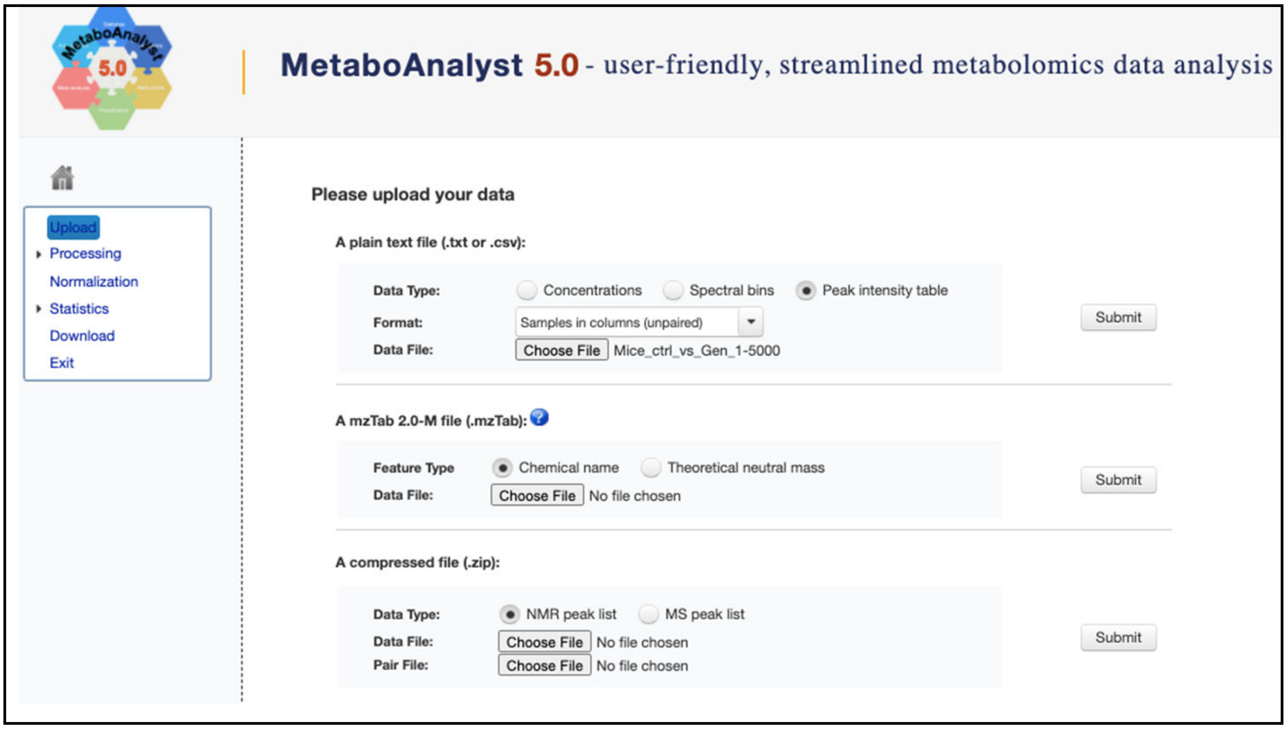
Biomarker Analysis

This module performs various biomarker analyses based on receiver operating characteristic (ROC) curves for a single or

Pathway Analysis (targeted)

This module supports pathway analysis (integrating enrichment analysis and pathway topology analysis) and visualization for 26

3



MetaboAnalyst 5.0 - user-friendly, streamlined metabolomics data analysis

Please upload your data

A plain text file (.txt or .csv):

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

Format: Samples in columns (unpaired)

Data File: Choose File Mice_ctrl_vs_Gen_1-5000

A mzTab 2.0-M file (.mzTab):

Feature Type ● Chemical name ○ Theoretical neutral mass

Data File: Choose File No file chosen

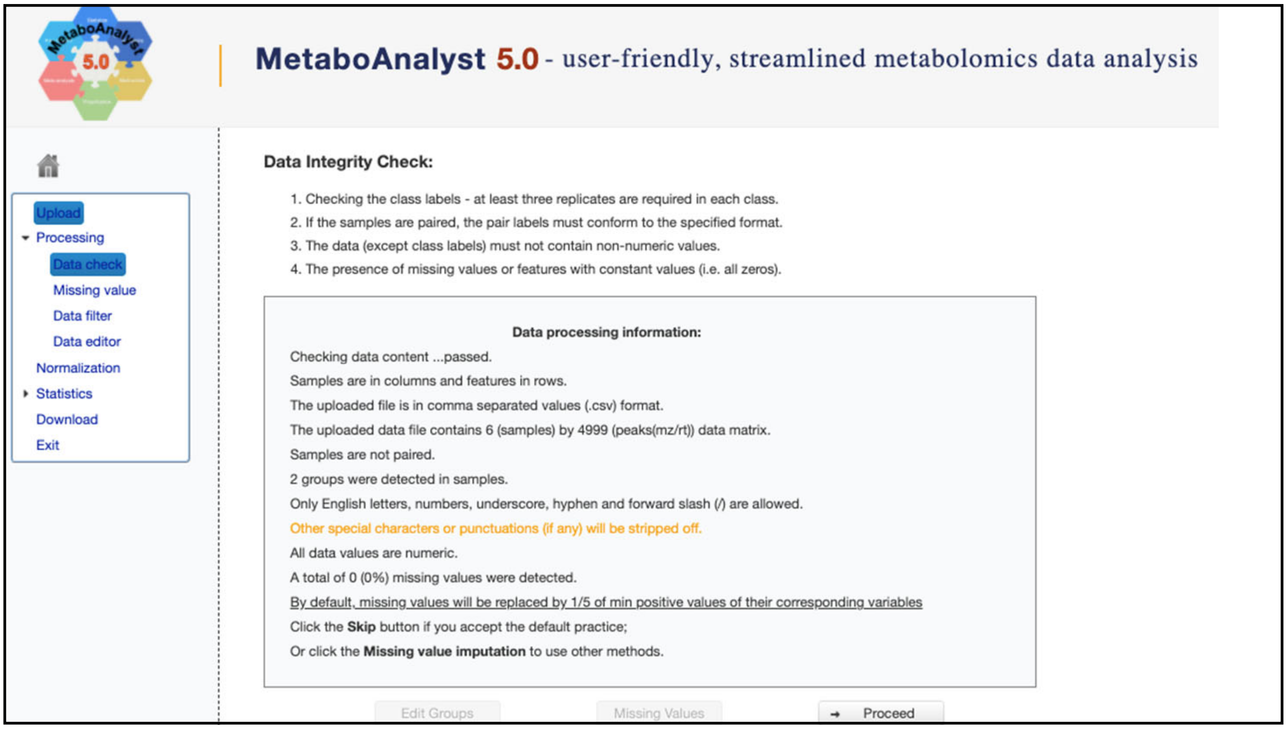
A compressed file (.zip):

Data Type: ● NMR peak list ○ MS peak list

Data File: Choose File No file chosen

Pair File: Choose File No file chosen

4



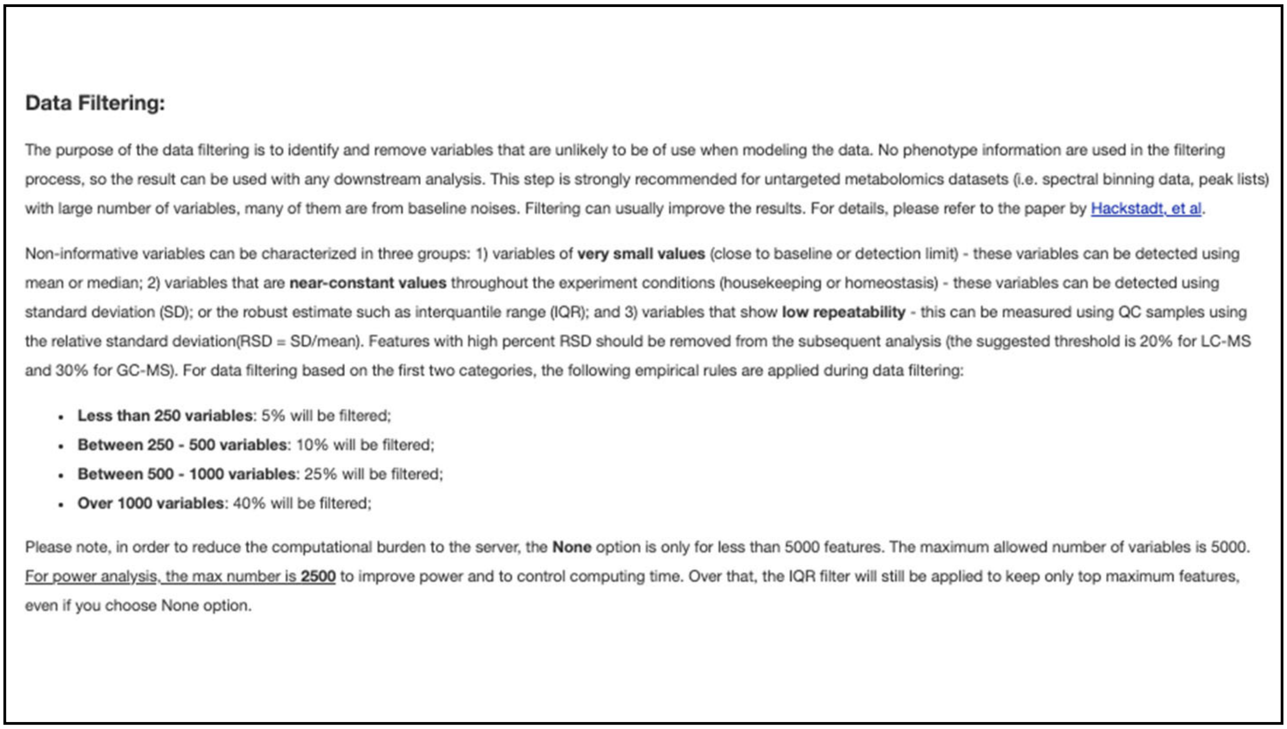
Data Integrity Check:

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.
Samples are in columns and features in rows.
The uploaded file is in comma separated values (.csv) format.
The uploaded data file contains 6 (samples) by 4999 (peaks(mz/rt)) data matrix.
Samples are not paired.
2 groups were detected in samples.
Only English letters, numbers, underscore, hyphen and forward slash (/) are allowed.
Other special characters or punctuations (if any) will be stripped off.
All data values are numeric.
A total of 0 (0%) missing values were detected.
By default, missing values will be replaced by 1/5 of min positive values of their corresponding variables
Click the **Skip** button if you accept the default practice;
Or click the **Missing value imputation** to use other methods.

5



Data Filtering:

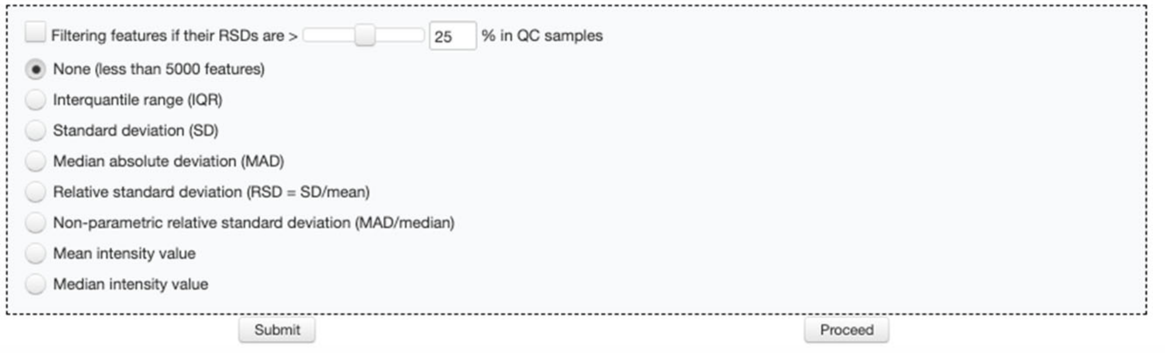
The purpose of the data filtering is to identify and remove variables that are unlikely to be of use when modeling the data. No phenotype information are used in the filtering process, so the result can be used with any downstream analysis. This step is strongly recommended for untargeted metabolomics datasets (i.e. spectral binning data, peak lists) with large number of variables, many of them are from baseline noises. Filtering can usually improve the results. For details, please refer to the paper by Hackstadt, et al.

Non-informative variables can be characterized in three groups: 1) variables of **very small values** (close to baseline or detection limit) - these variables can be detected using mean or median; 2) variables that are **near-constant values** throughout the experiment conditions (housekeeping or homeostasis) - these variables can be detected using standard deviation (SD); or the robust estimate such as interquantile range (IQR); and 3) variables that show **low repeatability** - this can be measured using QC samples using the relative standard deviation(RSD = SD/mean). Features with high percent RSD should be removed from the subsequent analysis (the suggested threshold is 20% for LC-MS and 30% for GC-MS). For data filtering based on the first two categories, 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 5000 features. The maximum allowed number of variables is 5000. For power analysis, the max number is **2500** to improve power and to control computing time. Over that, the IQR filter will still be applied to keep only top maximum features, even if you choose None option.

6



□ Filtering features if their RSDs are > 25 % in QC samples

● None (less than 5000 features)

○ Interquantile range (IQR)

○ Standard deviation (SD)

○ Median absolute deviation (MAD)

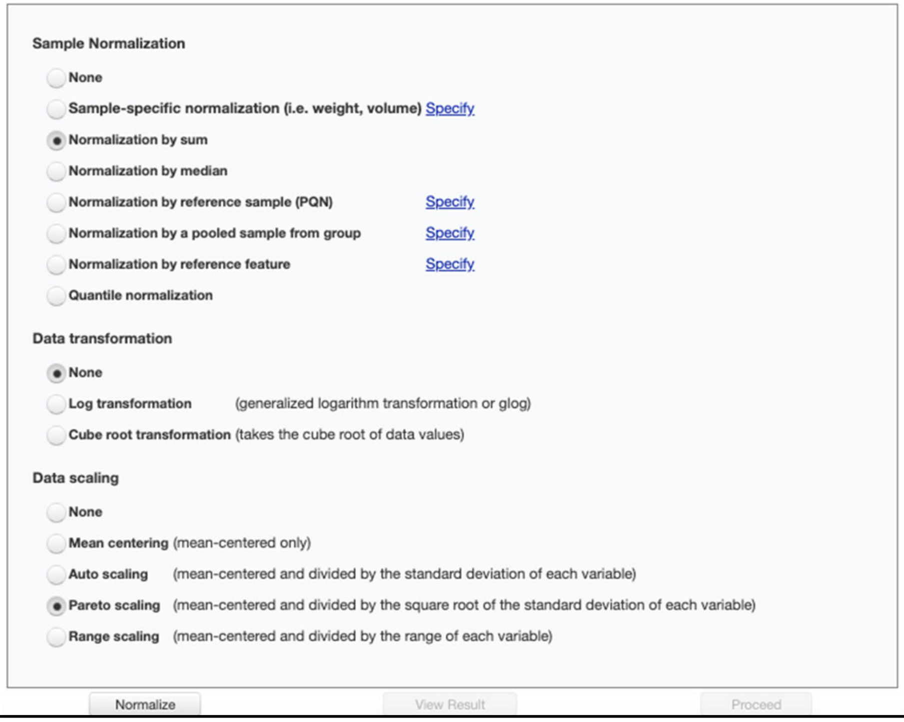
○ Relative standard deviation (RSD = SD/mean)

○ Non-parametric relative standard deviation (MAD/median)

○ Mean intensity value

○ Median intensity value

7



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**

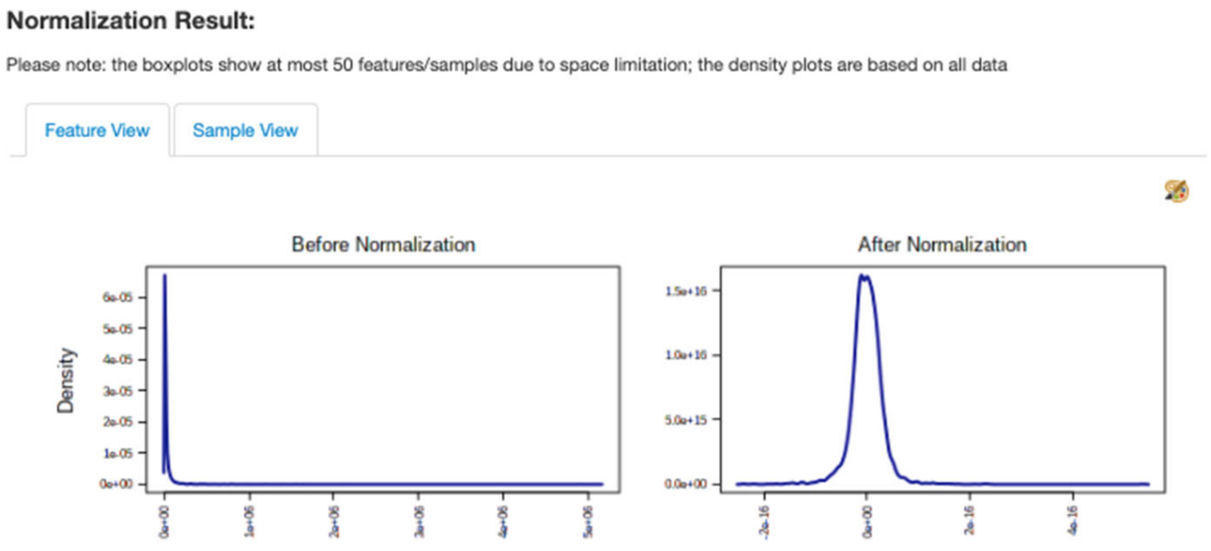
○ **Mean centering** (mean-centered only)

○ **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 variable)

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

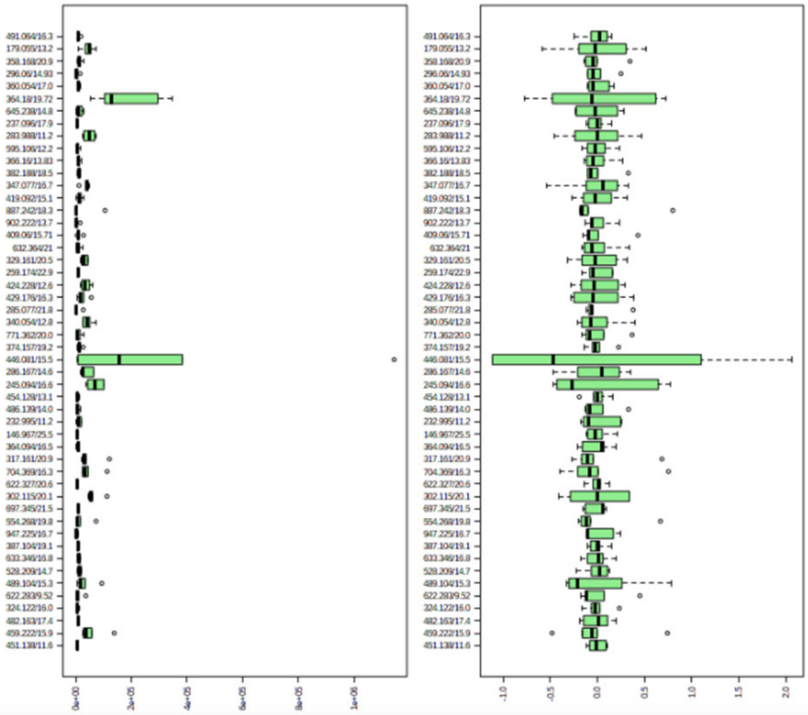
8



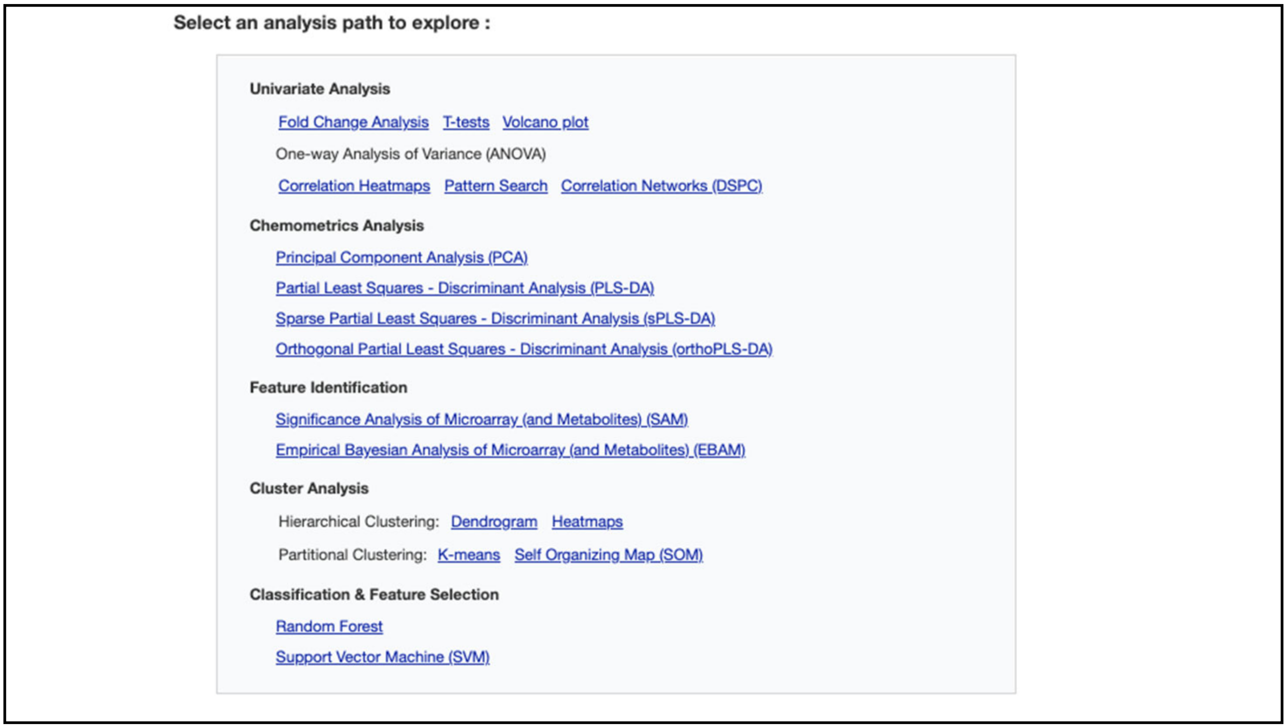
Normalization Result:

Please note: the boxplots show at most 50 features/samples due to space limitation; the density plots are based on all data

9



10



Select an analysis path to explore :

Univariate Analysis

Fold Change Analysis T-tests Volcano plot

One-way Analysis of Variance (ANOVA)

Correlation Heatmaps Pattern Search Correlation Networks (DSPC)

Chemometrics Analysis

Principal Component Analysis (PCA)

Partial Least Squares - Discriminant Analysis (PLS-DA)

Sparse Partial Least Squares - Discriminant Analysis (sPLS-DA)

Orthogonal Partial Least Squares - Discriminant Analysis (orthoPLS-DA)

Feature Identification

Significance Analysis of Microarray (and Metabolites) (SAM)

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

Cluster Analysis

Hierarchical Clustering: Dendrogram Heatmaps

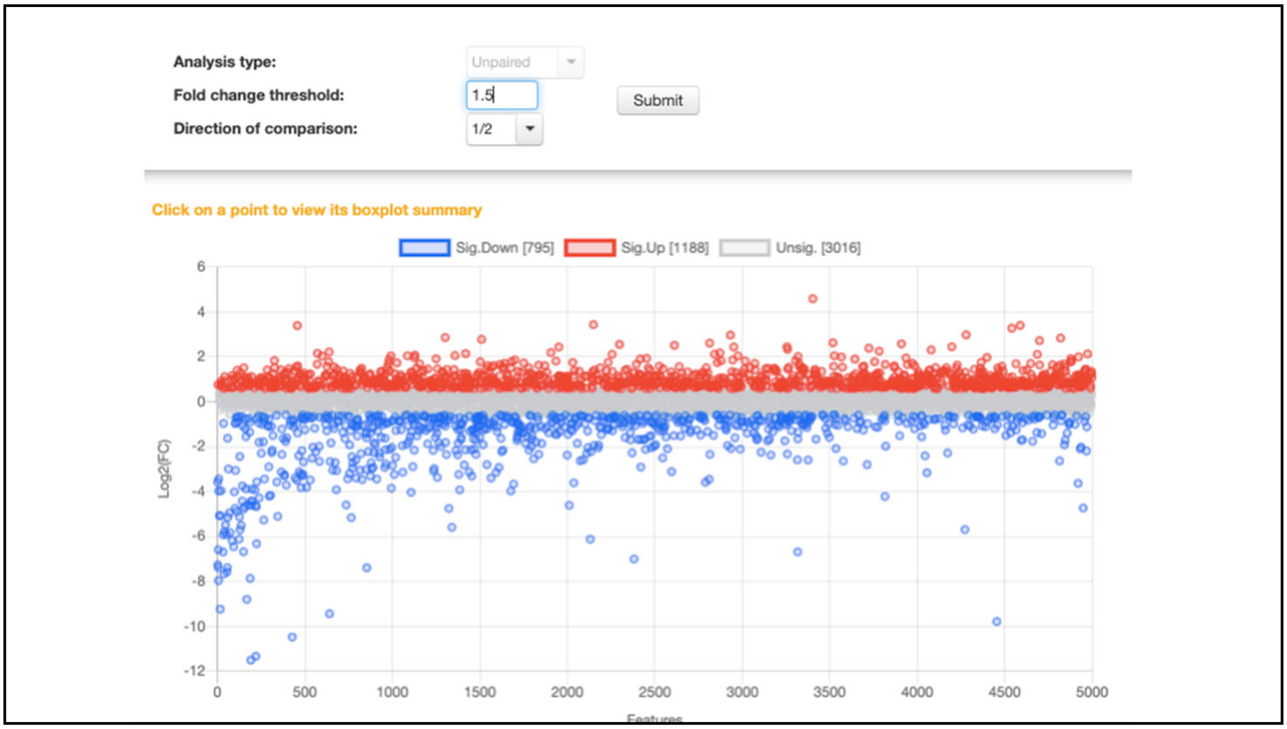
Partitional Clustering: K-means Self Organizing Map (SOM)

Classification & Feature Selection

Random Forest

Support Vector Machine (SVM)

11

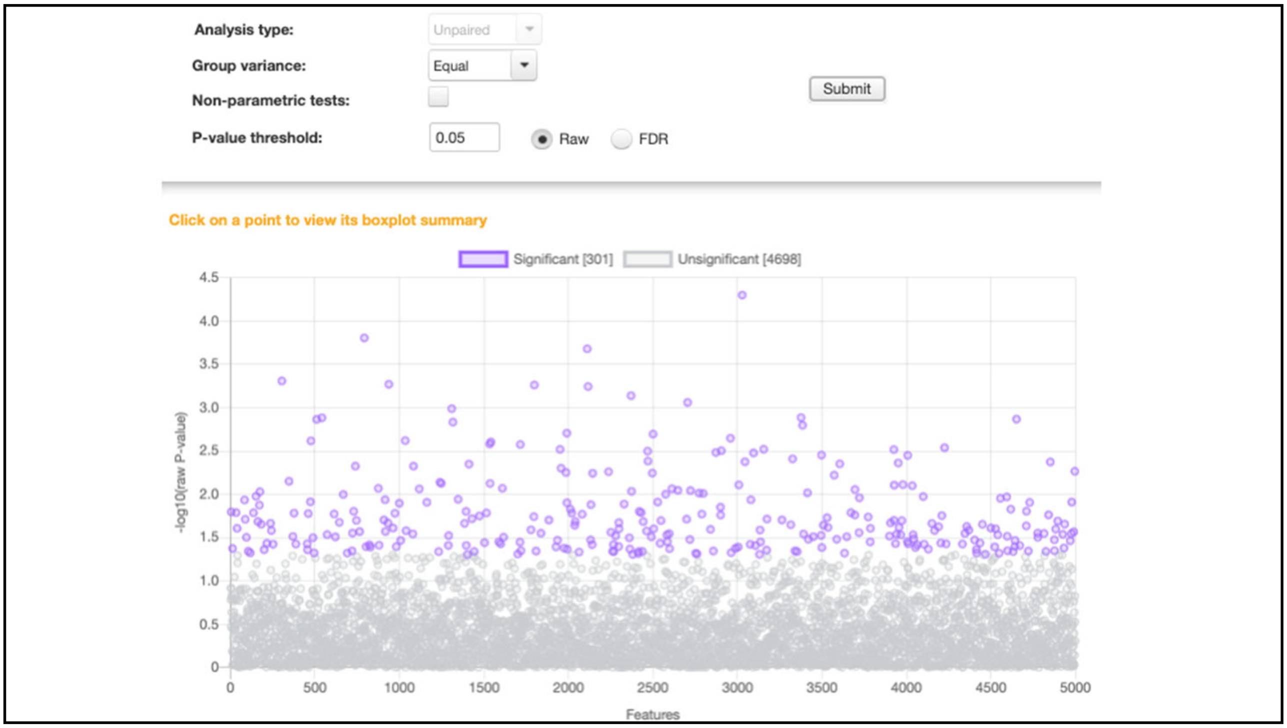


Analysis type: Unpaired

Fold change threshold: 1.5

Direction of comparison: 1/2

12



Analysis type: Unpaired

Group variance: Equal

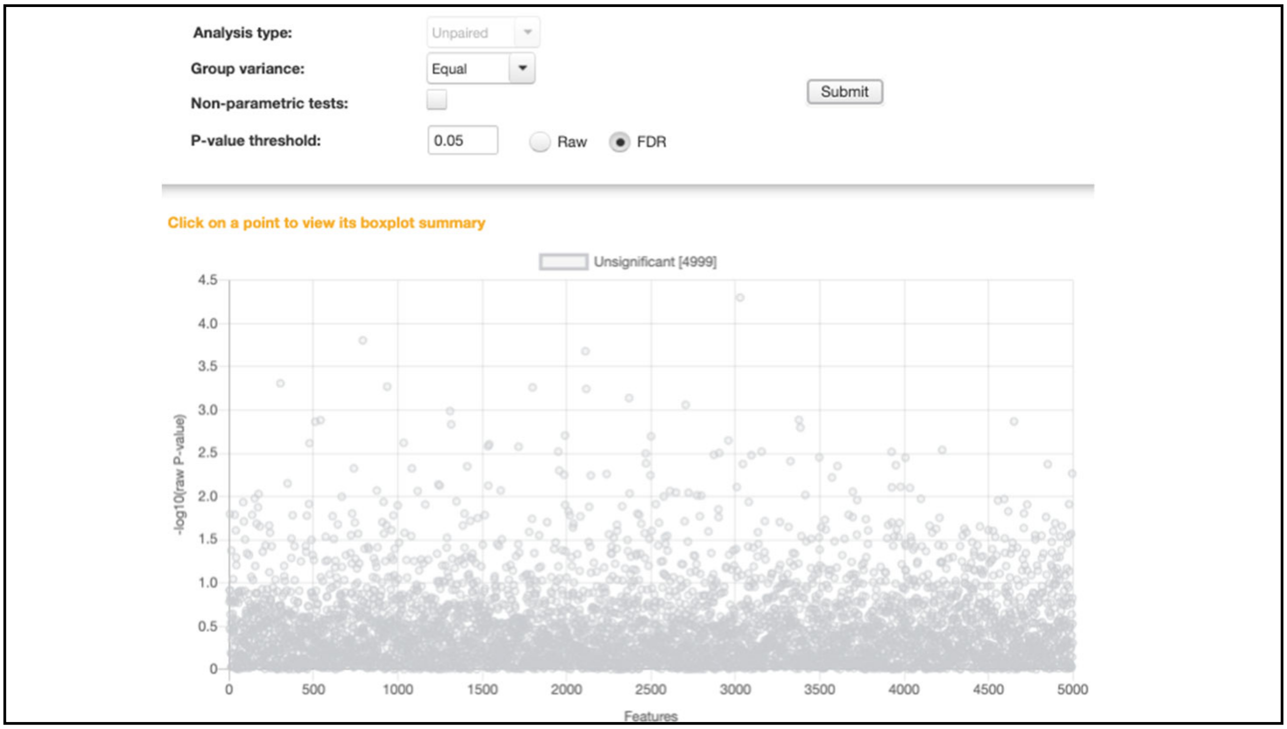
Non-parametric tests: □

P-value threshold: 0.05 ● Raw ○ FDR

Submit

Click on a point to view its boxplot summary

13



Analysis type: Unpaired

Group variance: Equal

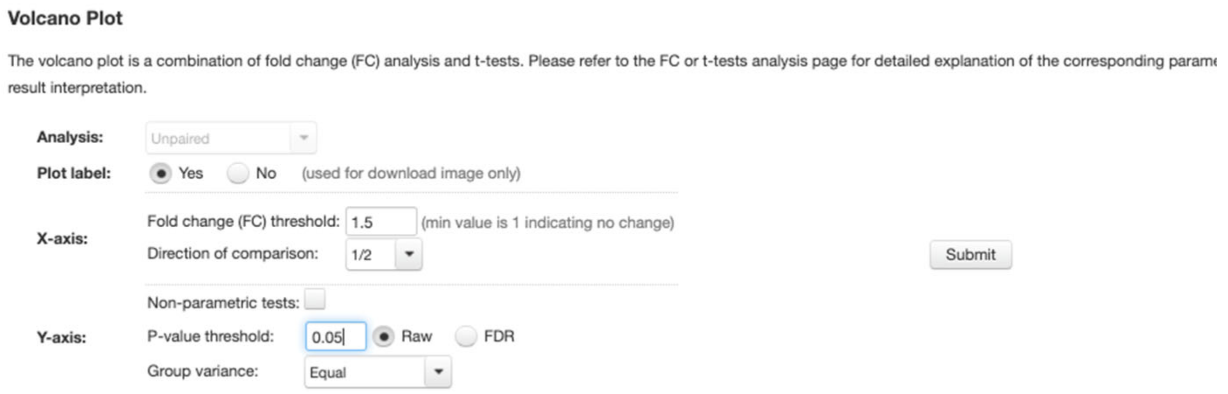
Non-parametric tests: □

P-value threshold: 0.05 ○ Raw ● FDR

Submit

Click on a point to view its boxplot summary

14



Volcano Plot

The volcano plot is a combination of fold change (FC) analysis and t-tests. Please refer to the FC or t-tests analysis page for detailed explanation of the corresponding parame result interpretation.

Analysis: Unpaired

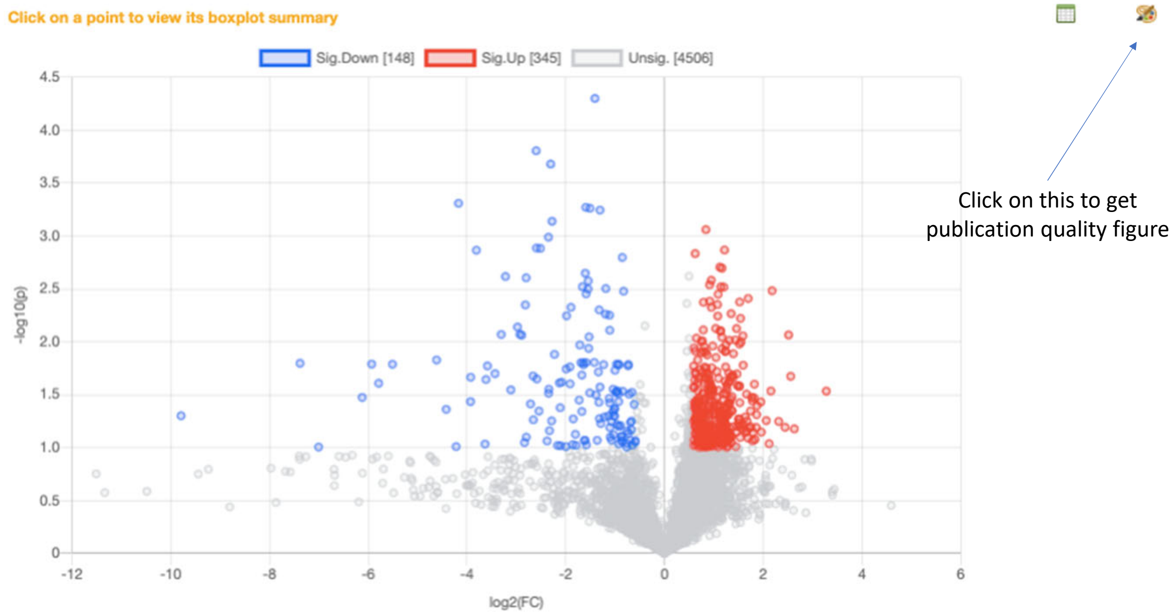
Plot label: ● Yes ○ No (used for download image only)

X-axis:
Fold change (FC) threshold: 1.5 (min value is 1 indicating no change)
Direction of comparison: 1/2

Y-axis:
Non-parametric tests: □
P-value threshold: 0.05 ● Raw ○ FDR
Group variance: Equal

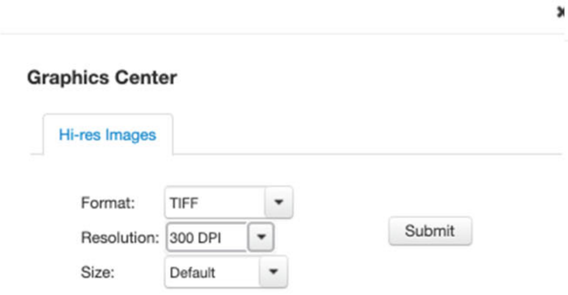
Submit

15



Click on this to get publication quality figure

16

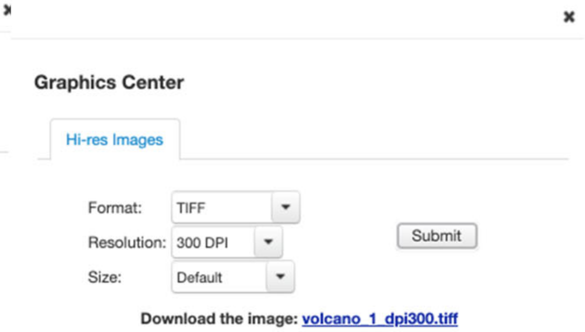


Graphics Center

Format: TIFF

Resolution: 300 DPI

Size: Default



Graphics Center

Format: TIFF

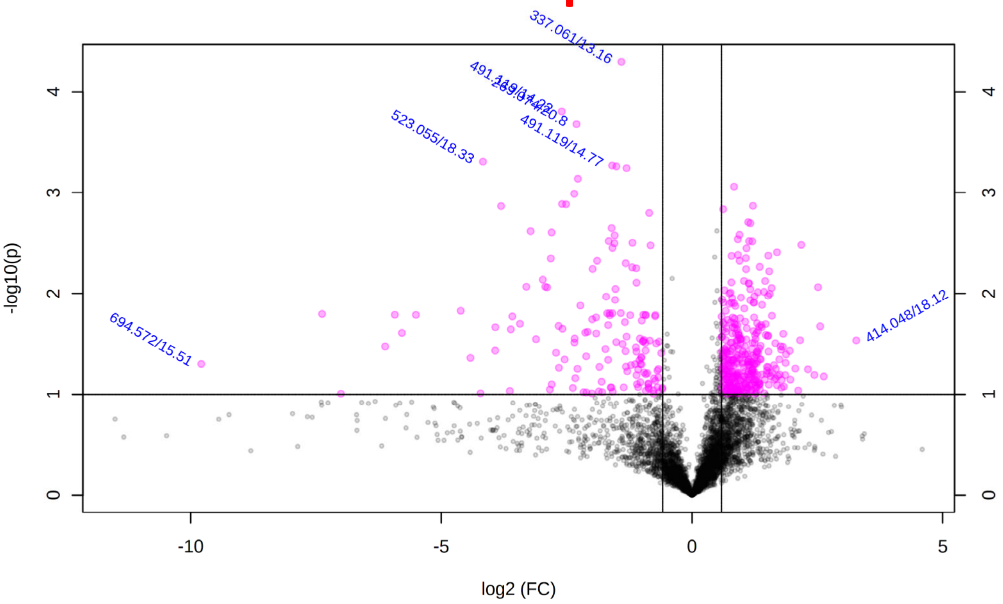
Resolution: 300 DPI

Size: Default

Download the image: volcano_1_dpi300.tiff

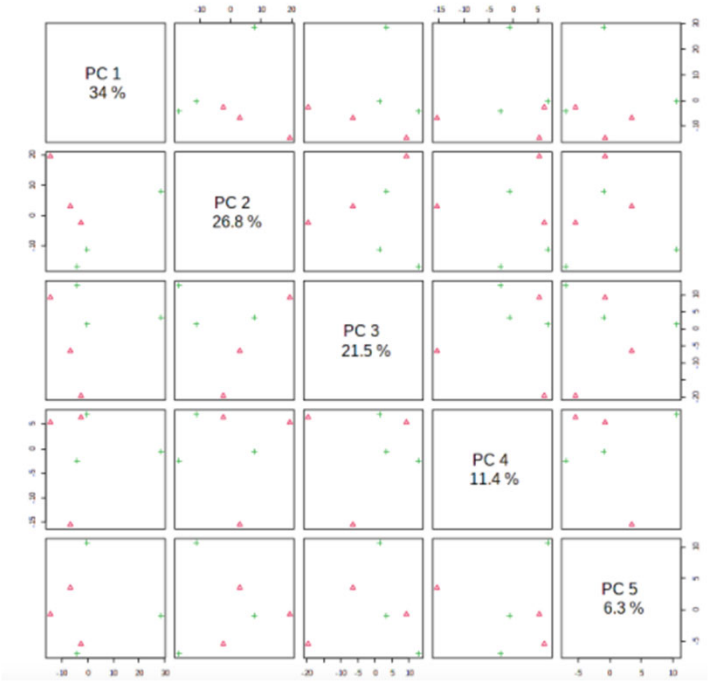
17

Volcano plot

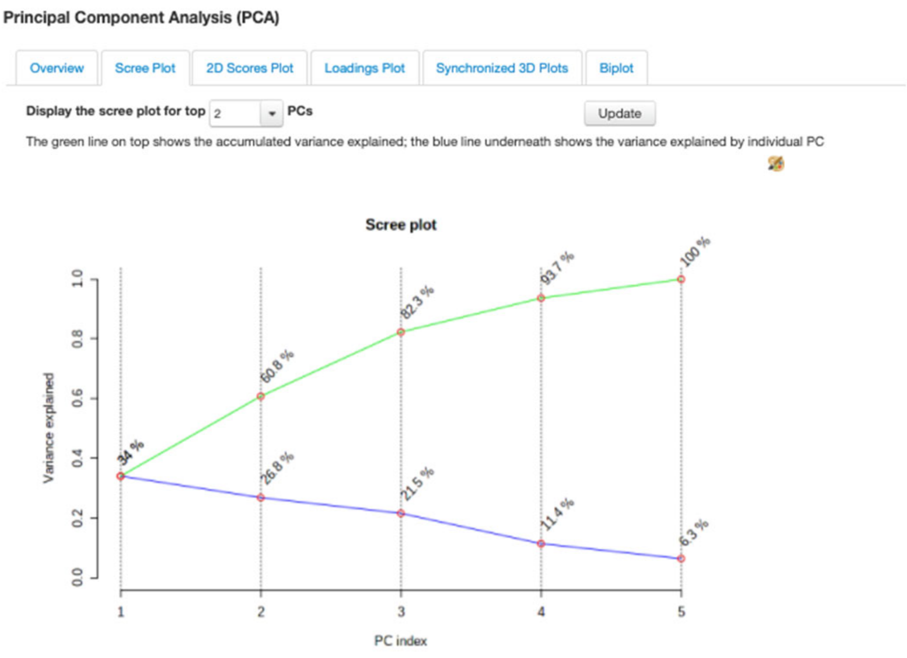


18

Principal Component Analysis



19



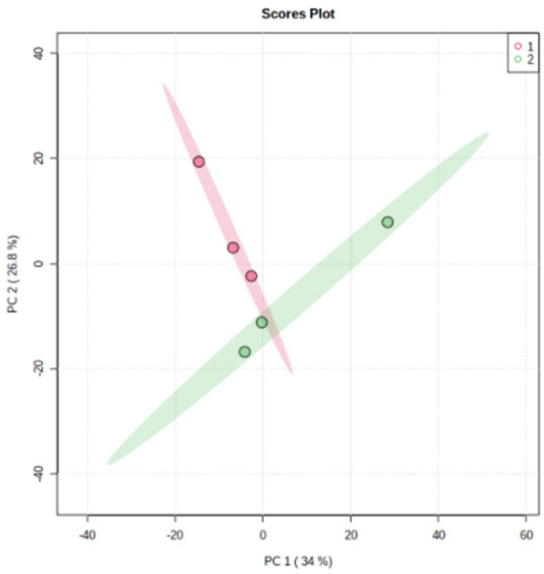
Principal Component Analysis (PCA)

Display the scree plot for top 2 **PCs**

The green line on top shows the accumulated variance explained; the blue line underneath shows the variance explained by individual PC

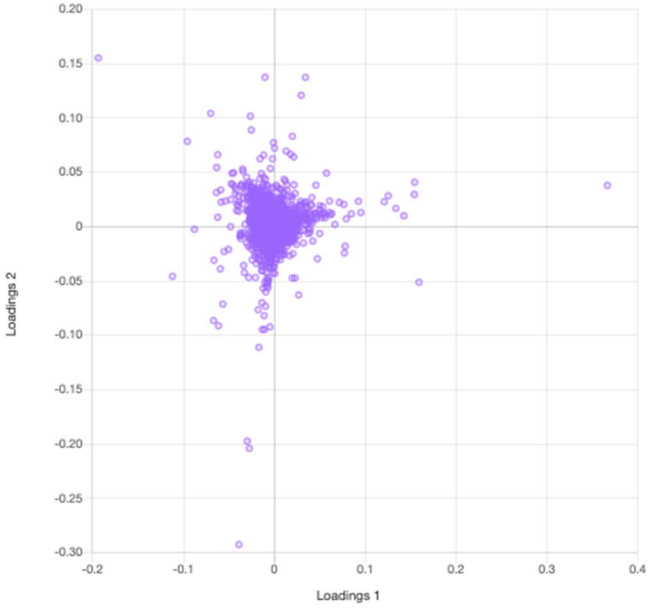
20

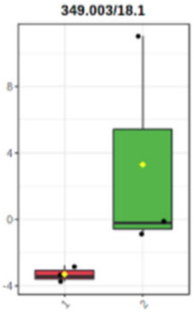
2D-PCA analysis

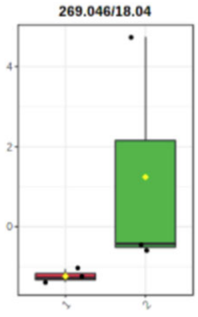


21

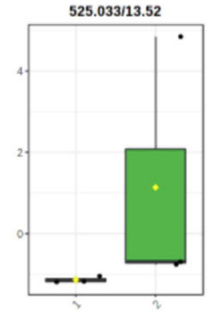
Click on a point to view its boxplot summary







The groups are different, but there is large variation in the genistein group. Discuss.

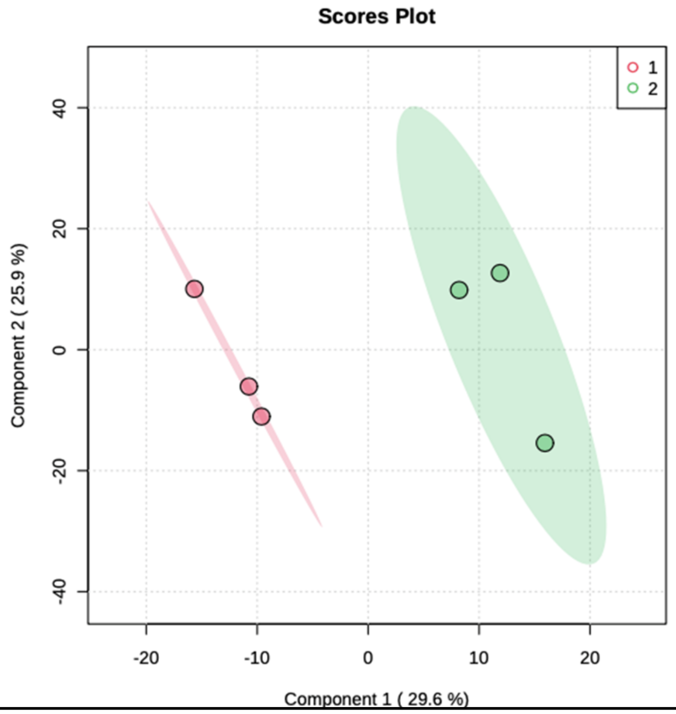


22

Partial least square discriminant analysis (PLSDA)

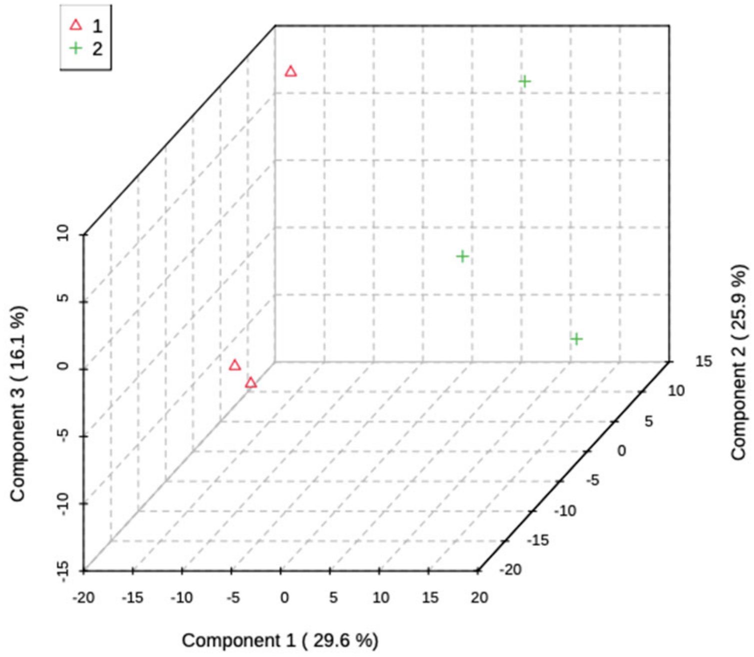
23

2D-PLSDA plot



24

3D-PLSDA plot



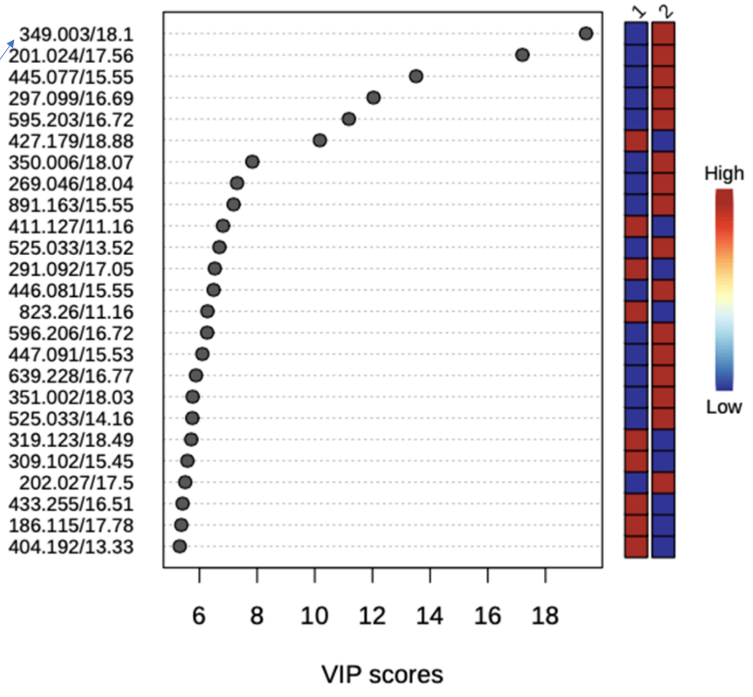
25

Ion features contributing most to the group separation

Genistein metabolites

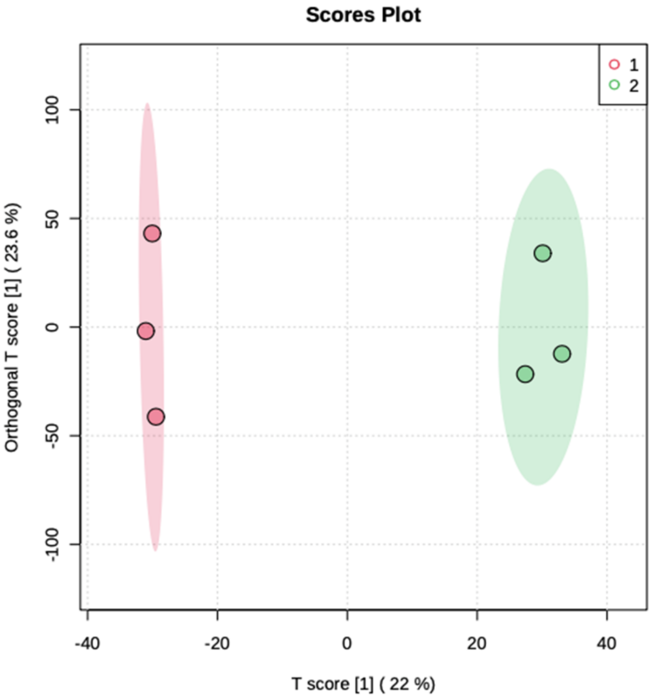
Isotopes?

Dimers/other adducts?



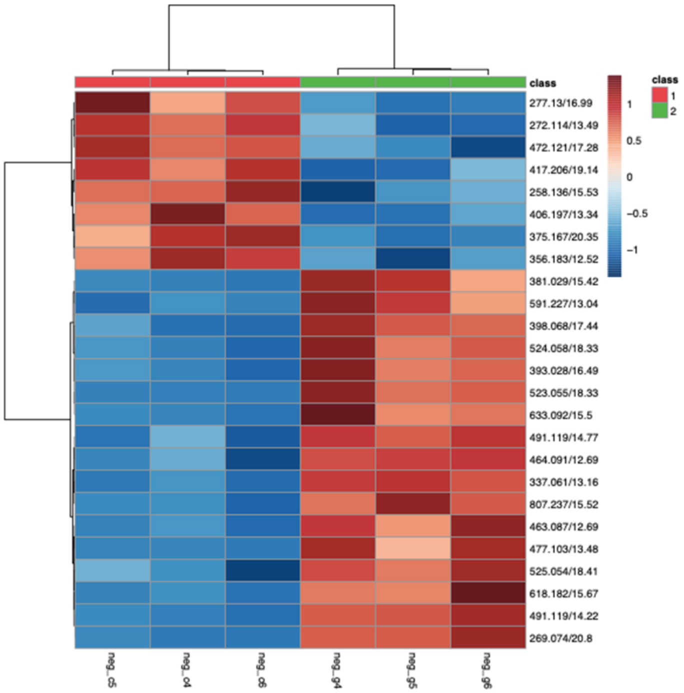
26

Orthogonal PLSDA

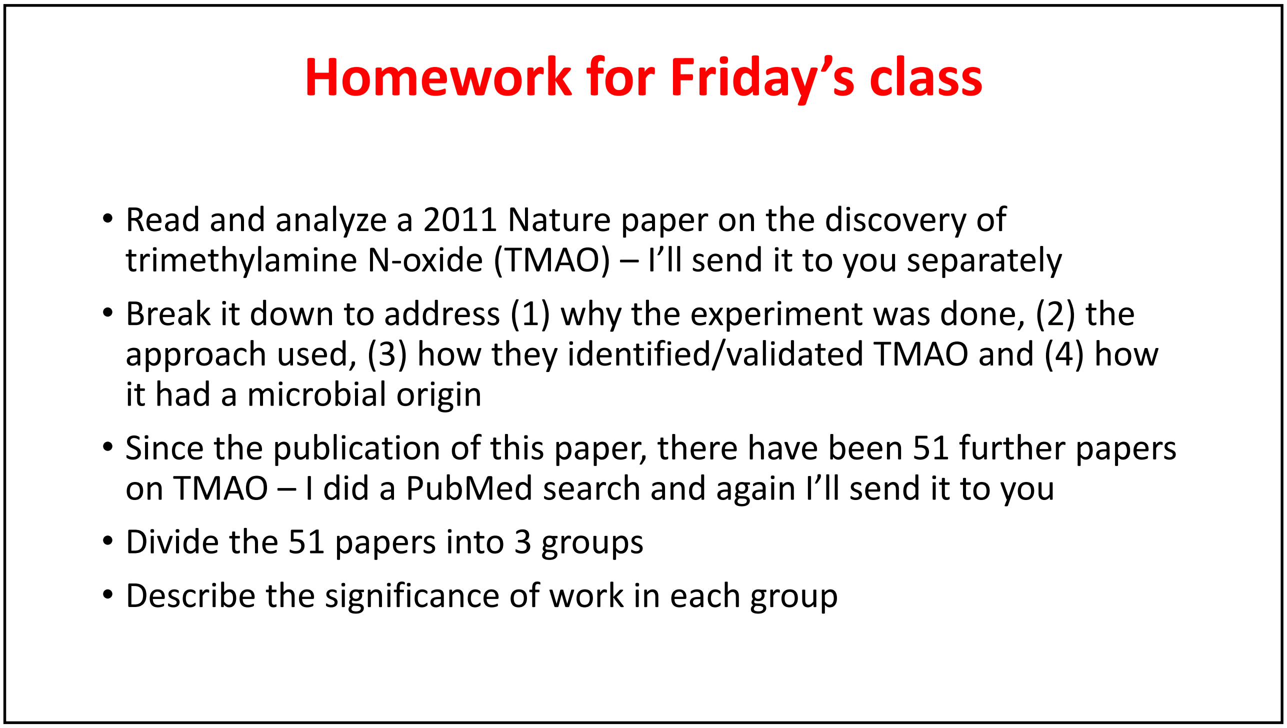


27

Hierarchal clustering



28



Homework for Friday’s class

- Read and analyze a 2011 Nature paper on the discovery of trimethylamine N-oxide (TMAO) – I’ll send it to you separately
- Break it down to address (1) why the experiment was done, (2) the approach used, (3) how they identified/validated TMAO and (4) how it had a microbial origin
- Since the publication of this paper, there have been 51 further papers on TMAO – I did a PubMed search and again I’ll send it to you
- Divide the 51 papers into 3 groups
- Describe the significance of work in each group

29