



*FEBRUARY 9, 2017*

1st NuGO ECN Online Webinar

# AN INTRODUCTION TO METABONALYST

A web-based freely accessible tool for -omics data  
analysis and interpretation

---

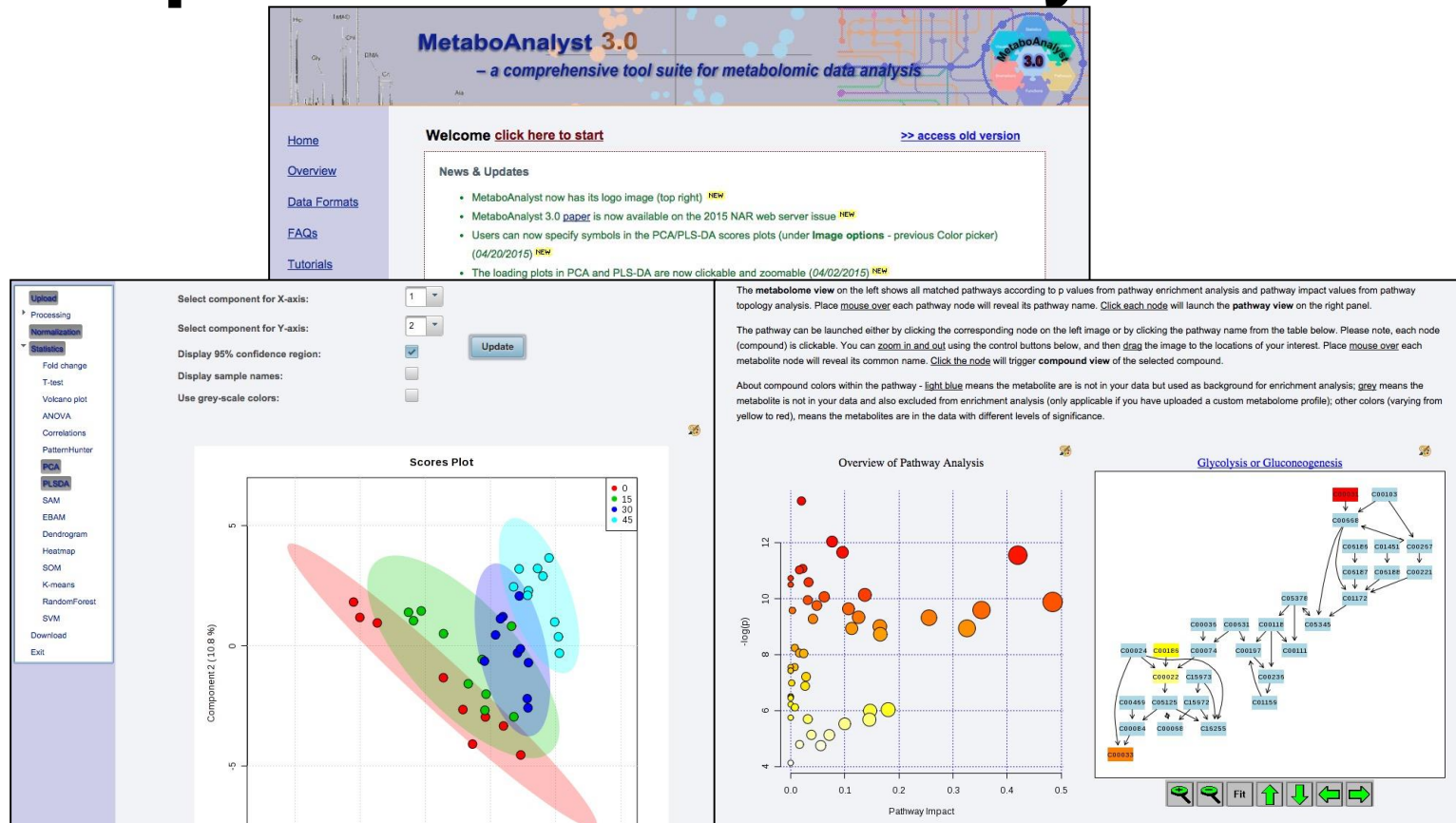


Dr. Rosa Vázquez-Fresno  
Postdoctoral Researcher  
Dr. David Wishart Lab  
University of Alberta



# MetaboAnalyst

<http://www.metaboanalyst.ca>



**A comprehensive web server designed to process & analyze -omics data**

# MetaboAnalyst Modules

## ➤ Statistical Analysis

This module offers various commonly used statistical and machine learning methods including t-tests, ANOVA, PCA, PLS-DA and Orthogonal 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/Two-factor Design

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.

## ➤ Power Analysis

This module uses pilot data to calculate the minimum number of samples required to detect a statistically significant difference between two populations with a given degree of confidence (called Power Analysis).

## ➤ Biomarker Analysis

This module performs various ROC curve based biomarker analyses for a single or multiple biomarkers. It also allows users to manually specify biomarker models as well as new sample prediction.

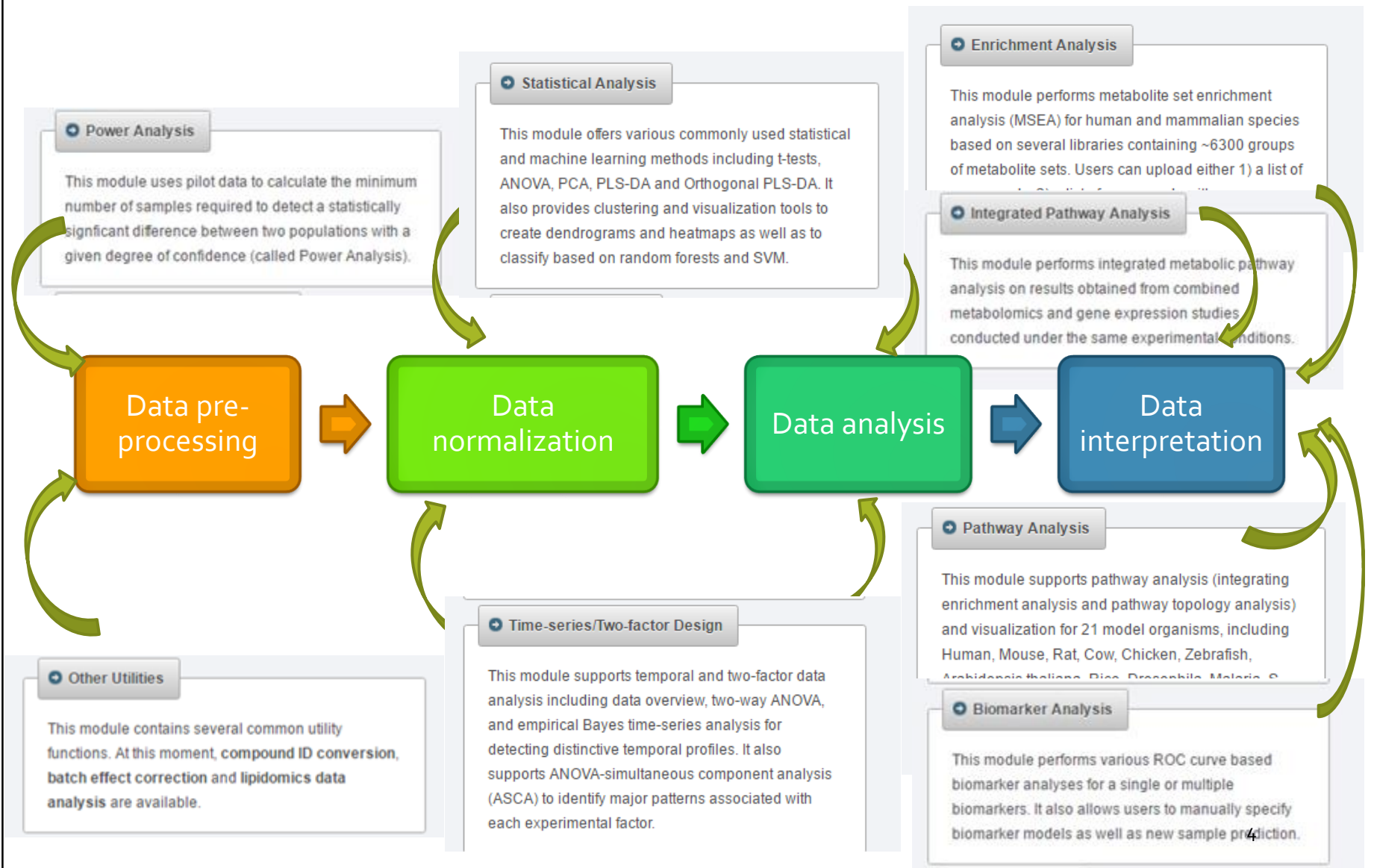
## ➤ Integrated Pathway Analysis

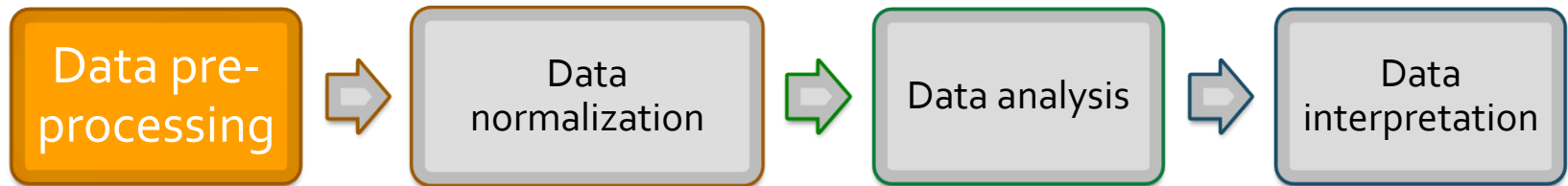
This module performs integrated metabolic pathway analysis on results obtained from combined metabolomics and gene expression studies conducted under the same experimental conditions.

## ➤ Other Utilities

This module contains several common utility functions. At this moment, **compound ID conversion**, **batch effect correction** and **lipidomics data analysis** are available.

# -Omics analysis



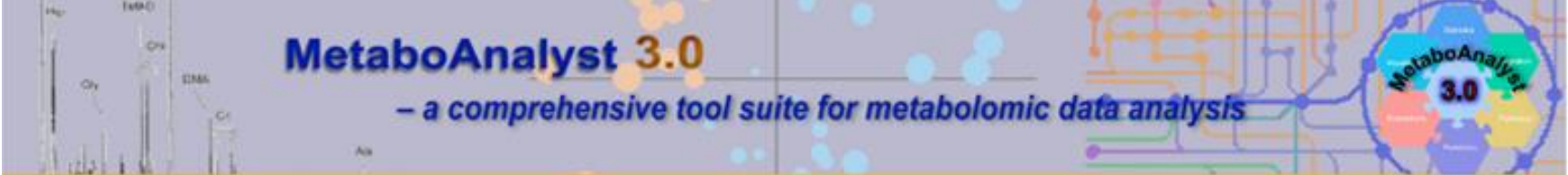


Purpose: to convert various raw data forms into data matrices suitable for statistical analysis

Supported data formats

- Concentration tables (Targeted Analysis)
- Peak lists (Untargeted)
- Spectral bins (Untargeted)
- Raw spectra (Untargeted)

# Data Formats



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

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

Welcome >> [click here to start](#) <<

### News & Updates

- Fixed the bug in feature table display in Biomarker Tester module (01/05/2017); **NEW**
- Updated the pathway result table to show all/matched compounds (11/25/2016); **NEW**
- Enhanced **Normalization** and **Data Editor** for better user experience (11/15/2016); **NEW**
- Added support for sparse PLS-DA (**sPLS-DA**) analysis (10/28/2016); **NEW**
- Added support for **quantile normalization** (08/29/2016);
- Improved name mapping functions for common metabolite names (08/18/2016);
- More than **1 million jobs** have been processed since 06/2015 (06/21/2016); **NEW**
- Updated Time Series module to support analysis of time-series only data (06/08/2016);
- Added support for Orthogonal PLS-DA (05/16/2016);
- Improved support for dealing with special characters and punctuations (05/11/2016);
- Minor feature updates and bug fixes based on user feedback (04/28/2016);
- Added support for batch effect correction for multiple data sets (**Other Utilities** module) (02/22/2016);
- Updated the web framework for better performance (02/18/2016);

[Read more](#)



# Example Datasets



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

Home  
Overview  
Data Formats  
FAQs  
Tutorials  
Troubleshooting  
Resources  
Update History  
User Stats  
About

  
**McGill**

  
**TMIC**

**Data Formats:**

**Example datasets for downloading, including :**

- Compound concentration data - cow, four groups ([download](#))
- Compound concentration data - human, two groups ([download](#))
- Binned NMR/MS spectra data ([download](#))
- Processed peak intensity table ([download](#))
- Time-series peak intensity data ([download](#))

**Zipped files (.zip) format datasets, including :**

- NMR peak lists (2 columns - chemical shift and intensity) ([download](#))
- MS peak lists (2 columns - mass and intensity) ([download](#))
- MS peak lists (3 columns - mass, retention time, and intensity) ([download](#))
- LC/GC - MS spectra (NetCDF, mzDATA, or mzXML) ([download](#))

*Note: please refer to detailed instructions and screenshots listed below.*

[General Introduction](#) [One-factor / Paired](#) [Time-series / Two-factor](#) [Peak lists / Spectra](#) [Biomarker data](#)

**Comma Separated Values (.csv) or Tab Delimited Text (.txt):**

These two formats are used for [concentration data](#), [peak intensity table](#), and [MS/NMR spectral bins](#). Samples can be in either rows or columns. Note,

1. Both sample or feature names must be unique and consist of a combination of common English letters, underscores and numbers for naming purpose. **Latin/Greek letters are not supported**
2. The class labels must follow immediately after the sample names; for two-factor and time series data, there must be two class labels corresponding to the two factors;
3. For time-series data, the time-point group must be named as **Time**. In addition, the samples collected from the same subjects at different time points should be consecutive (See the screenshots demo for "Two-factor / Time series")
4. Data values (concentrations, bins, peak intensities) should contain only numeric and positive values ([using empty or NA for missing values](#)).

# Data Formats

- **COMMA SEPARATED VALUES!! (.csv) or TAB DELIMITED TEXT (.txt)** → For quantitative (concentration tables) or qualitative (peak intensity or NMR/MS spectral bins).


Things to considere:

- Both samples and feature names **MUST** be **UNIQUE**. Can be combination fo letters and numbers separated by underscores [\_].
- The **class label must follow immediately after the sample name (for two-factors and time series data must be two class label columns)**
- **Metaboanalyst can also support .Zip files.**

Produced from either NMR, LC-MS, or GC-MS. In addition, GC/LC-MS spectra saved as open data format (NetCDF, mzDATA, mzXML) can also be processed using the XCMS packages



# Let's start!



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

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

Welcome >> [click here to start](#) <<



### News & Updates

- Fixed the bug in feature table display in Biomarker Tester module (01/05/2017); **NEW**
- Updated the pathway result table to show all/matched compounds (11/25/2016); **NEW**
- Enhanced **Normalization** and **Data Editor** for better user experience (11/15/2016); **NEW**
- Added support for sparse PLS-DA ( **sPLS-DA**) analysis (10/28/2016); **NEW**
- Added support for **quantile normalization** (08/29/2016);
- Improved name mapping functions for common metabolite names (08/18/2016);
- More than **1 million jobs** have been processed since 06/2015 (06/21/2016); **NEW**
- Updated Time Series module to support analysis of time-series only data (06/08/2016);
- Added support for Orthogonal PLS-DA (05/16/2016);
- Improved support for dealing with special characters and punctuations (05/11/2016);
- Minor feature updates and bug fixes based on user feedback (04/28/2016);
- Added support for batch effect correction for multiple data sets (**Other Utilities** module) (02/22/2016);
- Updated the web framework for better performance (02/18/2016);

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

# Select a Module :Statistical Analysis

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

**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.


➤ Power Analysis

This module uses pilot data to calculate the minimum number of samples required to detect a statistically significant difference between two populations with a given degree of confidence (called Power Analysis).

➤ Biomarker Analysis

This module performs various ROC curve based biomarker analyses for a single or multiple biomarkers. It also allows users to manually specify biomarker models as well as new sample prediction.

# Data Upload



**Upload**

- Processing
- Normalization
- Statistics
- Download
- Exit

## 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:**  cow\_diet.csv

**Zipped Files (.zip) :**

**Data Type:** ☒ NMR peak list ☐ MS peak list ☐ MS spectra

**Data File:**  No file chosen

**Pair File:**  No file chosen

# Data Integrity Check

**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 rows and features in columns

The uploaded file is in comma separated values (.csv) format.

The uploaded data file contains 39 (samples) by 47 (compounds) data matrix.

4 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

**Missing value estimation** **Skip**

# How to deal with missing values?

- Missing values should be presented either as **empty values or NA without quotes** in order to be accepted by MetaboAnalyst
- MetaboAnalyst offers a variety of methods to deal with missing values. By default, the missing values are treated **as the result of low signal intensity**. They will be replaced by half of the minimum positive values detected in your data. Users can also specify other methods, such as *replace by mean/median, Probabilistic PCA (PPCA), Bayesian PCA (BPCA) method, or Singular Value Decomposition (SVD) method* to impute the missing values ([Stacklies W. et al](#)).

# Data Integrity Check

**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 rows and features in columns

The uploaded file is in comma separated values (.csv) format.

The uploaded data file contains 39 (samples) by 47 (compounds) data matrix.

4 groups were detected in samples.

Samples are not paired.

All data values are numeric.

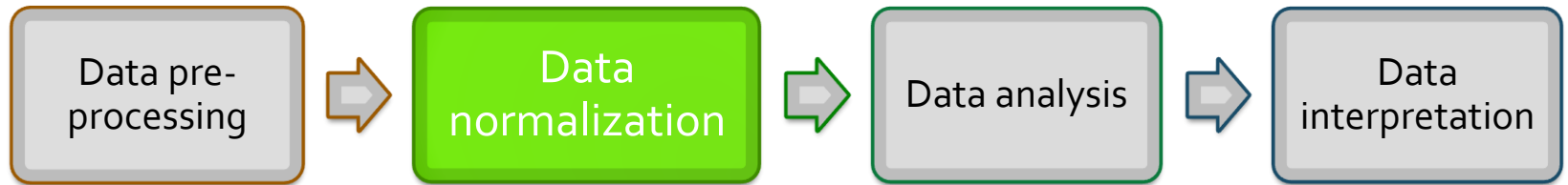
A total of 0 (0%) missing values were detected.

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

**Missing value estimation** **Skip**





# Data Normalization/Scaling

The normalization procedures are grouped into three categories.

The sample normalization allows general-purpose adjustment for differences among your sample

Data transformation and scaling are two different approaches to make individual features more comparable.

You can use one or combine them

The screenshot shows a web-based interface for data normalization and scaling, organized into three main sections:

- Sample normalization**
  - ☐ None
  - ☒ Sample-specific normalization (i.e. weight, volume) [Click here to specify](#)
  - ☐ Normalization by sum
  - ☐ Normalization by median
  - ☐ Normalization by a specific reference sample (with a dropdown menu showing '7')
  - ☐ Normalization by a pooled sample from group (with a dropdown menu showing 'C')
  - ☐ Normalization by reference feature (with a dropdown menu showing 'p-Hydroxyphenylacetic acid')
  - ☐ Quantile normalization
- Data transformation**
  - ☒ None
  - ☐ Log transformation (generalized logarithm transformation or glog)
  - ☐ Cube root transformation (take 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 standard deviation of each variable)
  - ☐ Range scaling (mean-centered and divided by the range of each variable)

# Data Normalization/Scaling

The screenshot shows a web-based interface for data normalization and scaling. It is divided into three main sections: Sample normalization, Data transformation, and Data scaling. The 'Sample normalization' section is highlighted with a red border. Annotations with arrows point to specific options: 'integrated area' points to 'Sample-specific normalization (i.e. weight, volume)', '=probabilistic quotient norm' points to 'Normalization by sum', and 'by a particular compound' points to the 'p-Hydroxyphenylacetic acid' dropdown in the 'Normalization by a specific reference sample' option.

**Sample normalization**

- ☐ None
- ☒ Sample-specific normalization (i.e. weight, volume) [Click here to specify](#)
- ☐ Normalization by sum
- ☐ Normalization by median
- ☐ Normalization by a specific reference sample (7) (C) (p-Hydroxyphenylacetic acid)
- ☐ Normalization by a pooled sample from group
- ☐ Normalization by reference feature
- ☐ Quantile normalization

**Data transformation**

- ☒ None
- ☐ Log transformation (generalized logarithm transformation or glog)
- ☐ Cube root transformation (take 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 standard deviation of each variable)
- ☐ Range scaling (mean-centered and divided by the range of each variable)

**To remove unwanted technical variation**

Account for different dilution effects of biofluids, drifts, instrument/injection

Aims to make each sample comparable to each other (i.e. urine samples with different dilution effects)

# Data Normalization/Scaling

## Sample normalization

- ☐ None
- ☒ Sample-specific normalization (i.e. weight, volume) [Click here to specify](#)
- ☐ Normalization by sum
- ☐ Normalization by median
- ☐ Normalization by a specific reference sample
- ☐ Normalization by a pooled sample from group
- ☐ Normalization by reference feature
- ☐ Quantile normalization

**Try to achieve a Normal distribution of your data**

## Data transformation

- ☒ None
- ☐ Log transformation (generalized logarithm transformation or glog)
- ☐ Cube root transformation (take cube root of data values)

Variation of log transform. Can deal with zeros or negative values. A strong transformation with a major effect on distribution shape

$$\text{glog}_2(x) = \log_2 \frac{x + \sqrt{x^2 + a^2}}{2}$$

## 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 standard deviation of each variable)
- ☐ Range scaling (mean-centered and divided by the range of each variable)

Fairly strong transformation. Weaker than the logarithm  
x to  $x^{(1/3)}$

# Data Normalization/Scaling

## Sample normalization

- ☐ None
- ☒ Sample-specific normalization (i.e. weight, volume) [Click here to specify](#)
- ☐ Normalization by sum
- ☐ Normalization by median
- ☐ Normalization by a specific reference sample
- ☐ Normalization by a pooled sample from group
- ☐ Normalization by reference feature
- ☐ Quantile normalization

## Data transformation

- ☒ None
- ☐ Log transformation (generalized logarithm transformation or glog)
- ☐ Cube root transformation (take 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 standard deviation of each variable)
- ☐ Range scaling (mean-centered and divided by the range of each variable)

**Transform your features in a same scale for suitable comparison of your variables**

This procedure is useful when variables are of very different orders of magnitude

# Scaling

Method	Formula	Goal	Advantages	Disadvantages
Autoscaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}$	Compare metabolites based on correlations	All metabolites become equally important	Inflation of the measurement errors
Range scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{(x_{i_{\max}} - x_{i_{\min}})}$	Compare metabolites relative to the biological response range	All metabolites become equally important. Scaling is related to biology	Inflation of the measurement errors and sensitive to outliers
Pareto scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{s_i}}$	Reduce the relative importance of large values, but keep data structure partially intact	Stays closer to the original measurement than autoscaling	Sensitive to large fold changes

# Data Normalization

## Normalization overview:

The normalization procedures are grouped into three categories. The sample normalization allows general-purpose adjustment for differences among your sample; data transformation and scaling are two different approaches to make individual features more comparable. You can use one or combine them to achieve better results.

### Sample normalization

- ☒ None
- ☐ Sample-specific normalization (i.e. weight, volume) [Click here to specify](#)
- ☐ Normalization by sum
- ☐ Normalization by median
- ☐ Normalization by a specific reference sample
- ☐ Normalization by a pooled sample from group
- ☐ Normalization by reference feature
- ☐ Quantile normalization

### Data transformation

- ☒ None
- ☐ Log transformation (generalized logarithm transformation or glog)
- ☐ Cube root transformation (take 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 standard deviation of each variable)
- ☐ Range scaling (mean-centered and divided by the range of each variable)

Normalize

View Result

Proceed

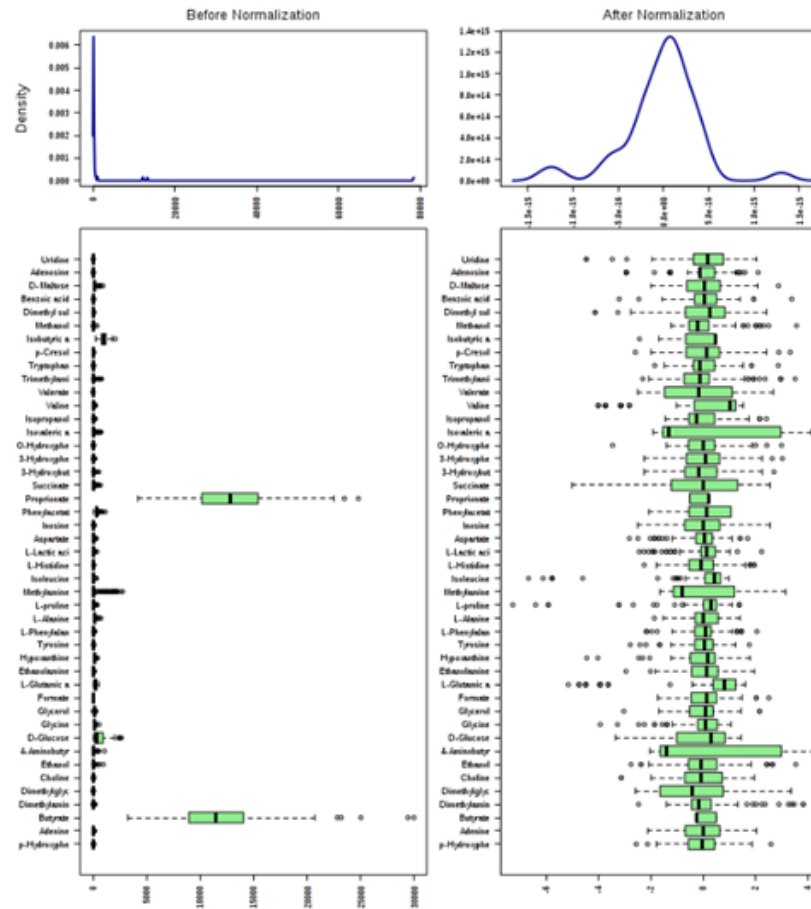
# Normalization Result



Feature View

Sample View

You can view the Sample normalization and the Features normalization



You cannot know a priori what the best normalization protocol will be. MetaboAnalyst allows you to interactively explore different normalization protocols and to visually inspect the degree of “normality” or Gaussian distribution



# Data Normalization/Scaling

## Sample normalization

- ☐ None
- ☒ Sample-specific normalization (i.e. weight, volume) [Click here to specify](#)
- ☐ Normalization by sum
- ☐ Normalization by median
- ☐ Normalization by a specific reference sample
- ☐ Normalization by a pooled sample from group
- ☐ Normalization by reference feature
- ☐ Quantile normalization

## Data transformation

- ☒ None
- ☐ (generalized logarithm transformation or glog)
- ☐ (take 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 standard deviation of each variable)
- ☐ Range scaling (mean-centered and divided by the range of each variable)

Choice depends on sample & circumstances

# Next Steps

After normalization has been completed it is a good idea to look at your data a little further to identify outliers or noise that could/should be removed



The image shows a magnifying glass focusing on a specific row of data in a table of normalized values. The table consists of four columns of decimal numbers. The magnifying glass is positioned over the second column, highlighting the value 0.000001 in the third row of the second section of the table.

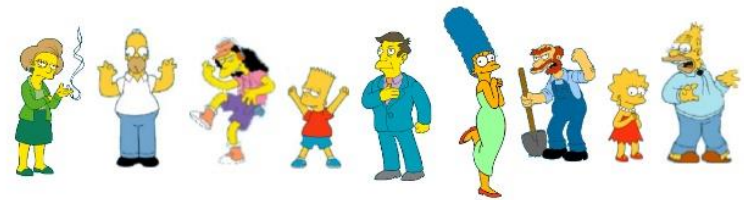
0.0000012	0.000002	0.0000013	0.000005
0.0001117	0.000728	0.0001113	0.000728
0.0000001	0.000001	0.0000021	0.000001
0.0009901	0.003349	0.0001901	0.003749
0.0000402	0.000265	0.0000402	0.000265
0.0000012	0.000002	0.0000012	0.000002
0.0001117	0.000728	0.0001117	0.000728
0.0000001	0.000001	0.0000001	0.000001
0.0009901	0.003349	0.0009901	0.003349
0.0000402	0.000265	0.0000402	0.000265
0.0000012	0.000002	0.0000012	0.000002
0.0001117	0.000728	0.0001117	0.000728
0.0000034	0.000001	0.0000001	0.000001
0.0009901	0.003349	0.0009901	0.003349
0.0000402	0.000265	0.0000402	0.000265
0.0004016	0.000002	0.000022	0.002235
0.0001117	0.000728	0.002203	0.005435
0.0004001	0.000001	0.000001	0.000000
0.0005901	0.006349	0.000020	0.000000
0.0000402	0.000865	0.003303	0.0037008

# Data QC, Outlier Removal & Data Reduction

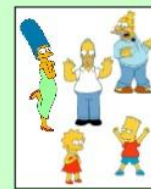
- Data filtering (remove solvent peaks, noise filtering, false positives, *outlier removal -- needs justification*)
- Dimensional reduction or feature selection to reduce number of features or factors to consider (PCA or PLS-DA)
- Clustering to find similarity



What is a natural grouping among these objects?



Clustering is subjective



Simpson's Family



School Employees



Females



Males

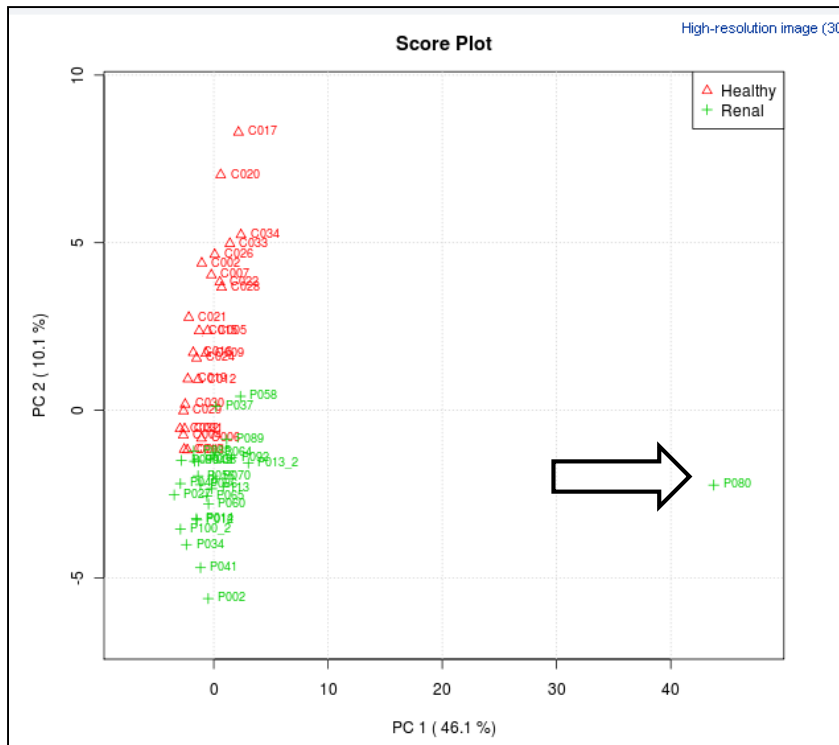
# Quality Control

- Dealing with outliers
  - Detected mainly by visual inspection
  - May be corrected by normalization
  - May be excluded
- Noise reduction
  - More of a concern for spectral bins/ peak lists
  - Usually improves downstream results

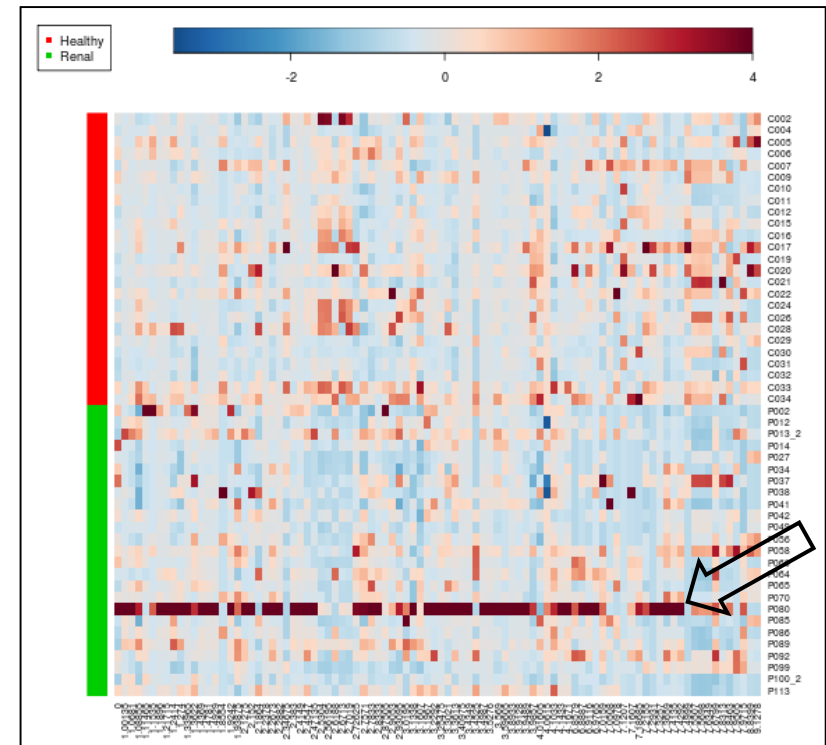


# Visual Inspection

- What does an outlier look like?



**Finding outliers via PCA**

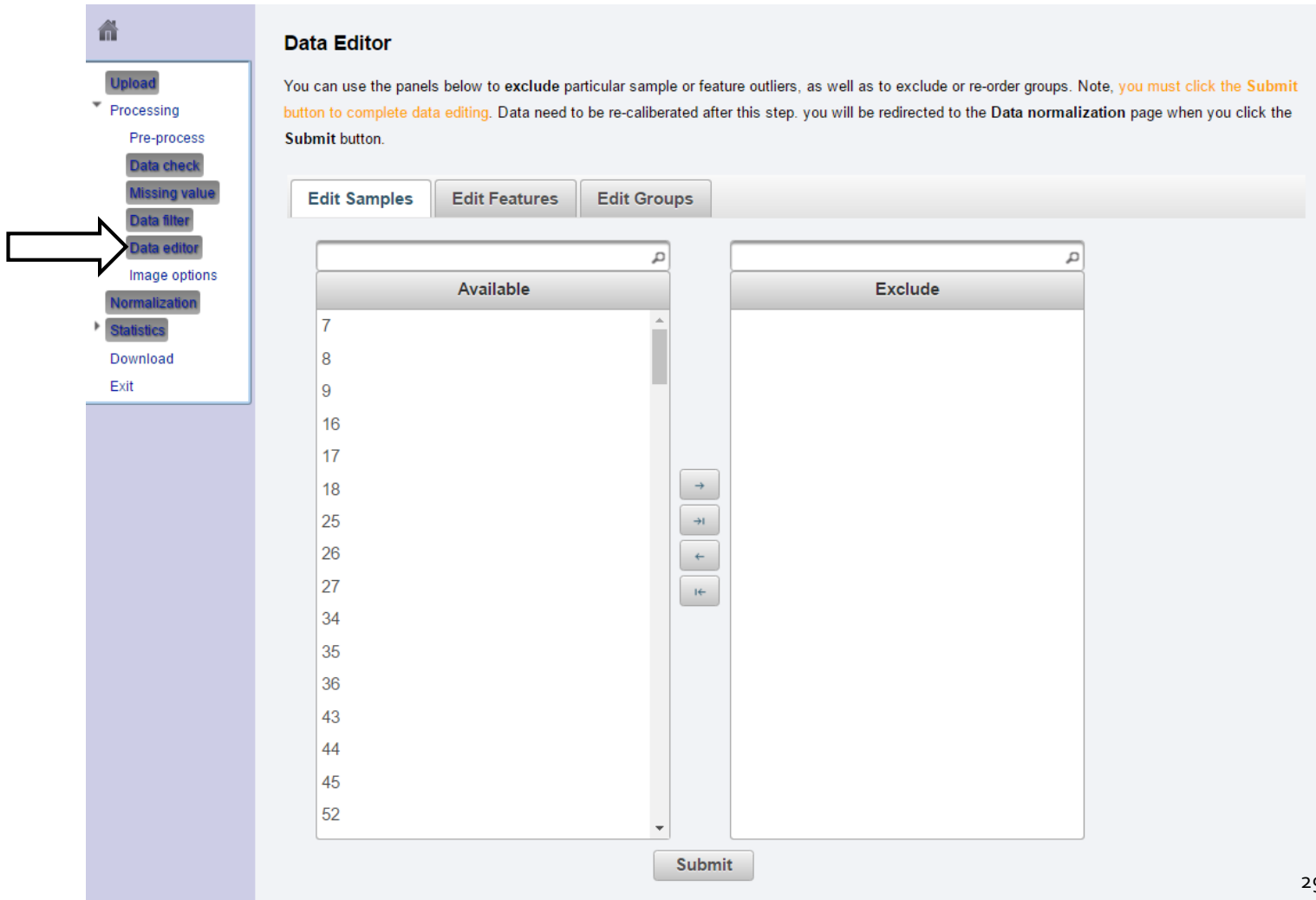


**Finding outliers via Heatmap**

# How to detect and deal with outlier?

- To deal with outliers, the first is to check if those samples / features are measured properly. In many cases, outliers are the result of operational errors during analytical process. If those values cannot be corrected, they should be removed from analysis, but ALWAYS justified.
- MetaboAnalyst provides **DataEditor** to enable easy removal of sample/feature outliers. Please note, you may need to re-normalize the data after outlier removal.

# Outlier Removal (Data Editor)



**Data Editor**

You can use the panels below to **exclude** particular sample or feature outliers, as well as to exclude or re-order groups. Note, **you must click the Submit button to complete data editing**. Data need to be re-calibrated after this step. you will be redirected to the **Data normalization** page when you click the **Submit** button.

**Edit Samples** **Edit Features** **Edit Groups**

**Available**

7  
8  
9  
16  
17  
18  
25  
26  
27  
34  
35  
36  
43  
44  
45  
52

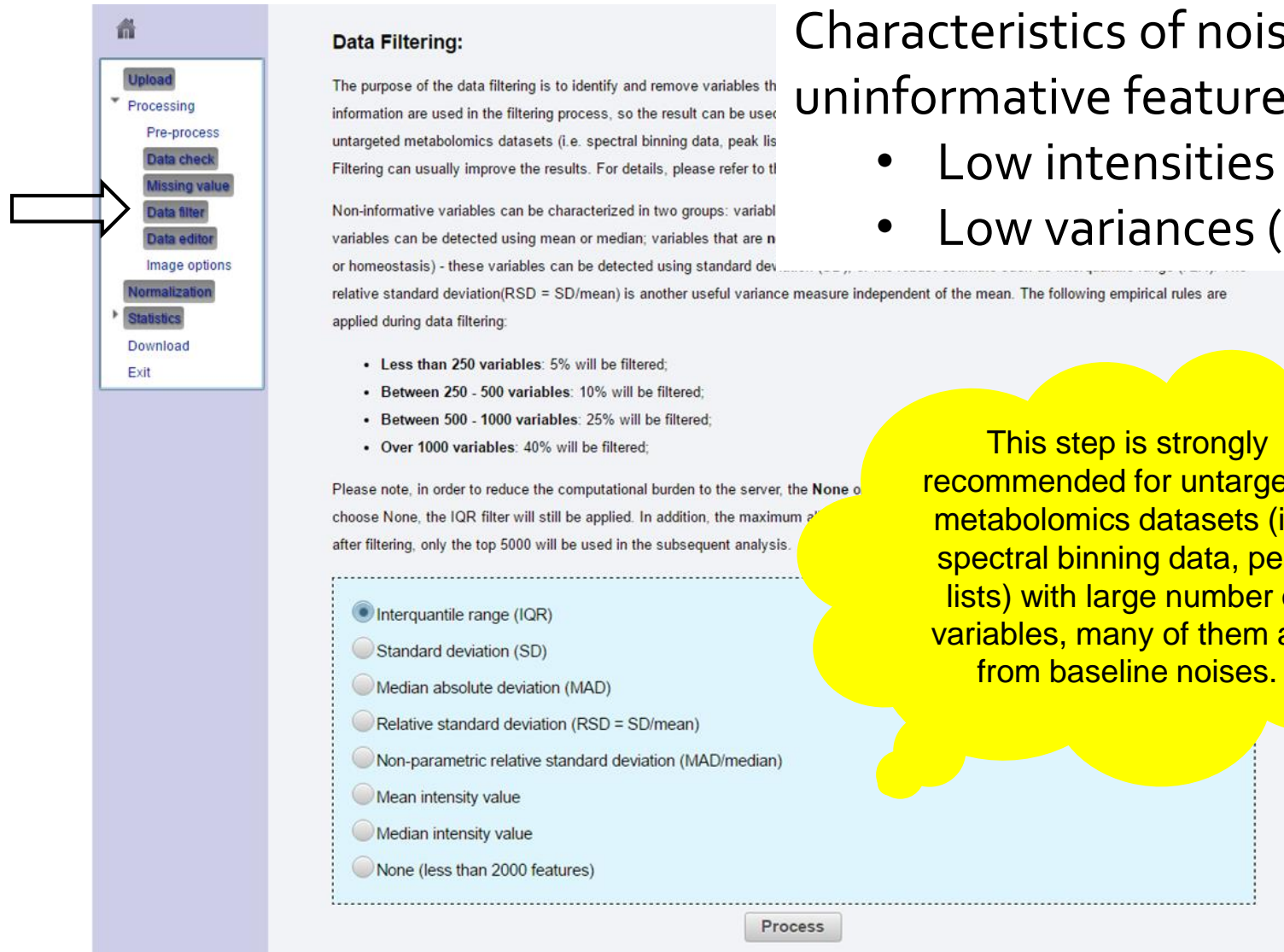
**Exclude**

**Submit**

The interface includes a sidebar on the left with a navigation menu. A large white arrow points to the 'Data editor' option, which is highlighted in blue. The main area contains a header 'Data Editor' with a descriptive paragraph. Below this are three tabs: 'Edit Samples', 'Edit Features', and 'Edit Groups'. The 'Edit Samples' tab is active, showing two panels: 'Available' and 'Exclude'. The 'Available' panel lists sample IDs (7, 8, 9, 16, 17, 18, 25, 26, 27, 34, 35, 36, 43, 44, 45, 52). The 'Exclude' panel is empty. Between the panels are four buttons: a right arrow, a right arrow with a vertical line, a left arrow, and a left arrow with a vertical line. At the bottom is a 'Submit' button.



# Noise Reduction (Data Filtering)



**Data Filtering:**

The purpose of the data filtering is to identify and remove variables that do not contain useful information are used in the filtering process, so the result can be used for untargeted metabolomics datasets (i.e. spectral binning data, peak lists). Filtering can usually improve the results. For details, please refer to the documentation.

Non-informative variables can be characterized in two groups: variable with low intensities and low variances. Variables with low intensities can be detected using mean or median; variables that are not in homeostasis - these variables can be detected using standard deviation. Relative standard deviation ( $RSD = SD/mean$ ) is another useful variance measure independent of the mean. The following empirical rules are applied during data filtering:

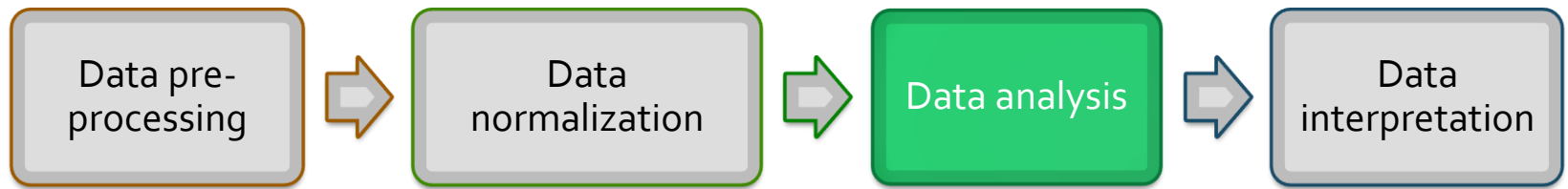
- **Less than 250 variables:** 5% will be filtered;
- **Between 250 - 500 variables:** 10% will be filtered;
- **Between 500 - 1000 variables:** 25% will be filtered;
- **Over 1000 variables:** 40% will be filtered;

Please note, in order to reduce the computational burden to the server, the **None** option should be chosen. If **None** is chosen, the IQR filter will still be applied. In addition, the maximum number of features 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**

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.



Data Reduction and  
statistical analysis

# Common Tasks

Data analysis

- To identify important features
- To detect interesting patterns
- To assess difference between the phenotypes
- To facilitate classification or prediction
- There are several statistical analyses that you can perform in Metaboanalyst. However, not all can be covered here- We will look at ANOVA, Multivariate Analysis (PCA, PLS-DA) and Clustering

## Select an analysis path to explore :

### Univariate Analysis

Fold Change Analysis T-tests Volcano plot

[One-way Analysis of Variance \(ANOVA\)](#)

[Correlation Analysis](#) [Pattern Searching](#)



### 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)

# ANOVA

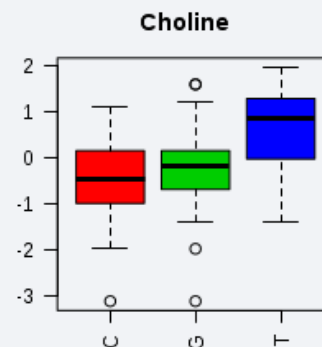
## One-way ANOVA & post-hoc Tests

You can choose to perform one-way ANOVA or its non-parametric version (Kruskal Wallis Test). Note, the post-hoc tests have only been implemented for parametric version.

Non-parametric ANOVA: ☐

Significance Level (alpha): raw p value <

Post-hoc analysis:

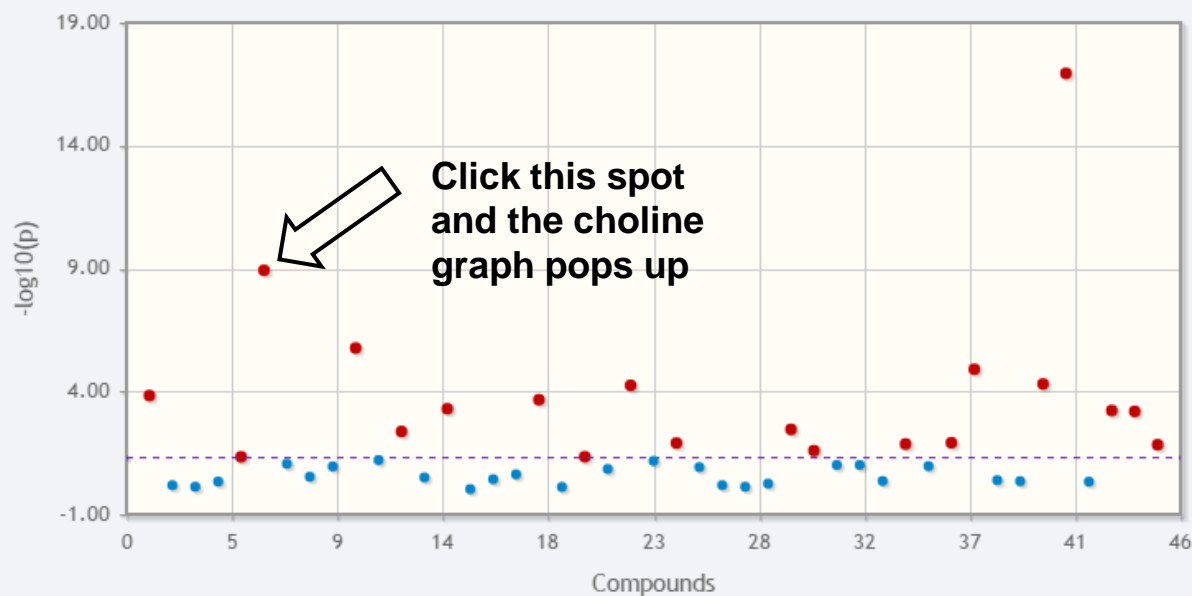


Click this  
to view  
the table



Click on a point to view, drag to zoom

[Reset](#)



# What's Next?

- Click and compare different compounds to see which ones are most different or most similar between the groups
- Click on the Correlation link (under the ANOVA link) to generate a heat map that displays the pairwise compound correlations and compound clusters

## Select an analysis path to explore :

### Univariate Analysis

Fold Change Analysis   T-tests   Volcano plot

[One-way Analysis of Variance \(ANOVA\)](#)

[Correlation Analysis](#)   [Pattern Searching](#)

### 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 \(SOM\)](#)

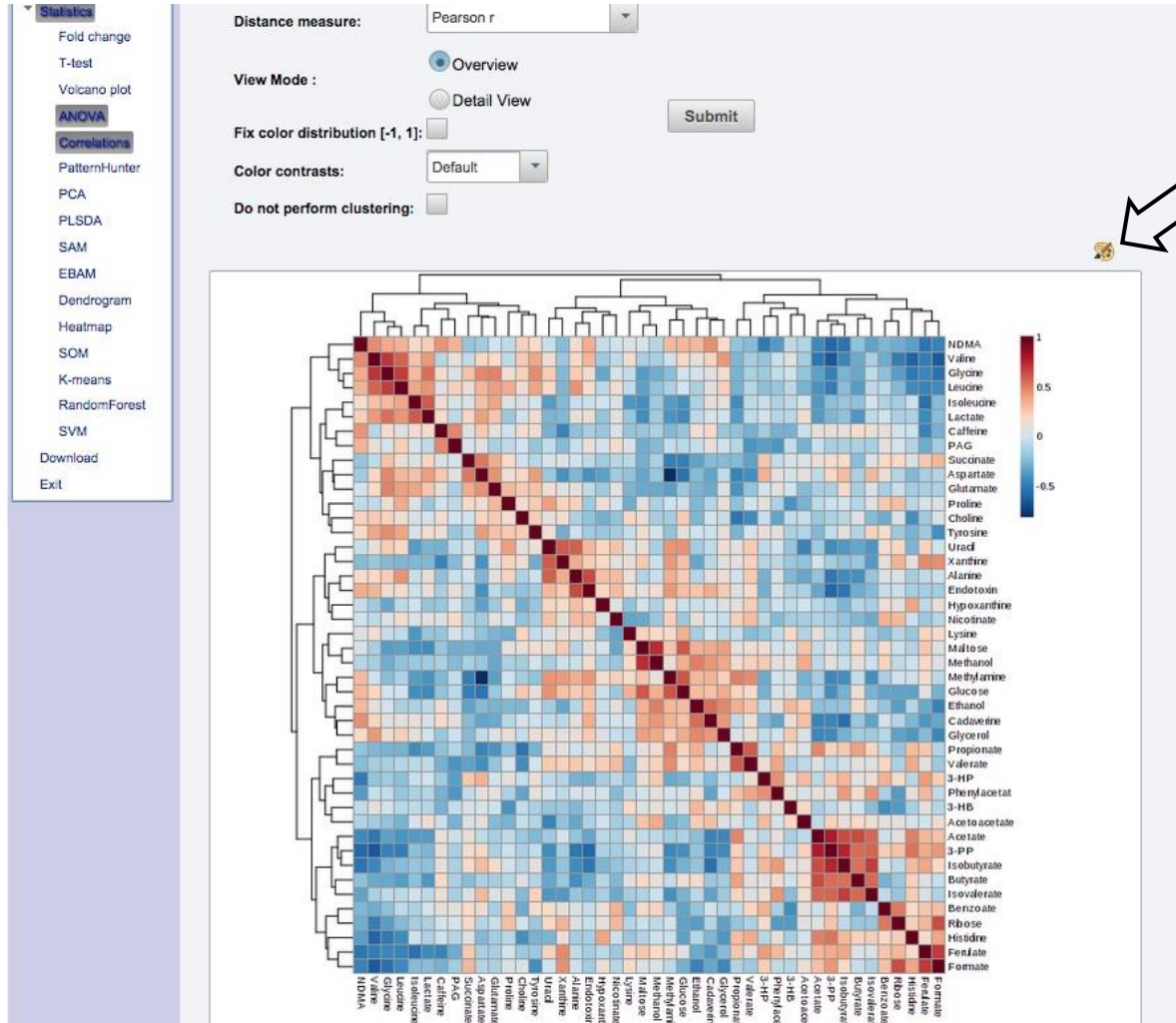
### Classification & Feature Selection

[Random Forest](#)

Support Vector Machine (SVM)



# Overall Correlation Pattern

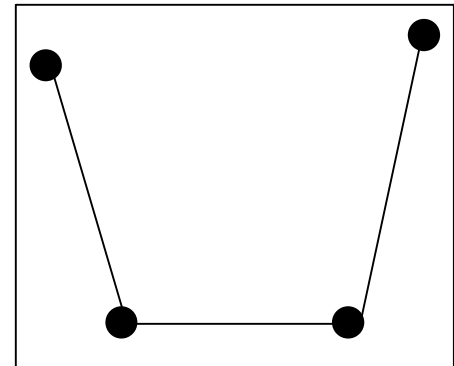
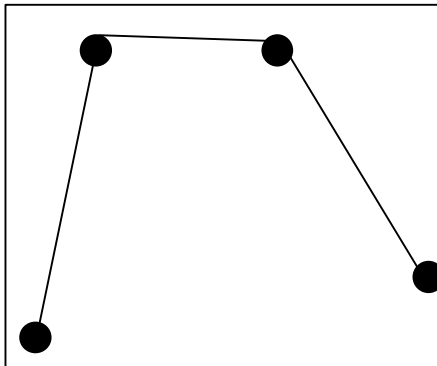
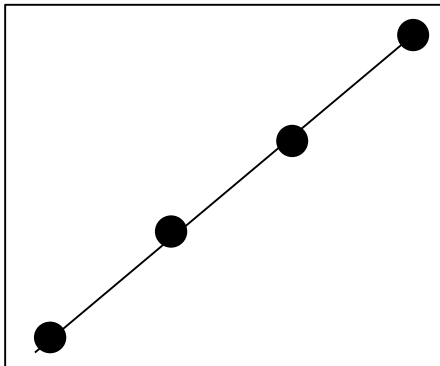


Click this to save  
a high res. image



# What's Next?

- When looking at >2 groups it is often useful to look for patterns or trends within particular metabolites
- Use Pattern Hunter to find these trends



## Select an analysis path to explore :

### Univariate Analysis

Fold Change Analysis   T-tests   Volcano plot

[One-way Analysis of Variance \(ANOVA\)](#)

[Correlation Analysis](#)   [Pattern Searching](#)



### 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)

# Pattern Searching

- Looking for compounds showing interesting patterns of change
- Essentially a method to look for linear trends or periodic trends in the data

Correlation analysis can be performed either against a given feature or against a given pattern. The pattern is specified as a series of numbers separated by "-". Each number corresponds to the expected expression pattern in the corresponding group. For example, a 1-2-3-4 pattern is used to search for features that increase linearly with time in a time-series data with four time points (or four groups). The order of the groups is given as the first item in the predefined patterns.

**Define a pattern using:**

☐ a feature of interest: 1,3-D

☒ a predefined profile: 1-2-3-4

☐ a custom profile:

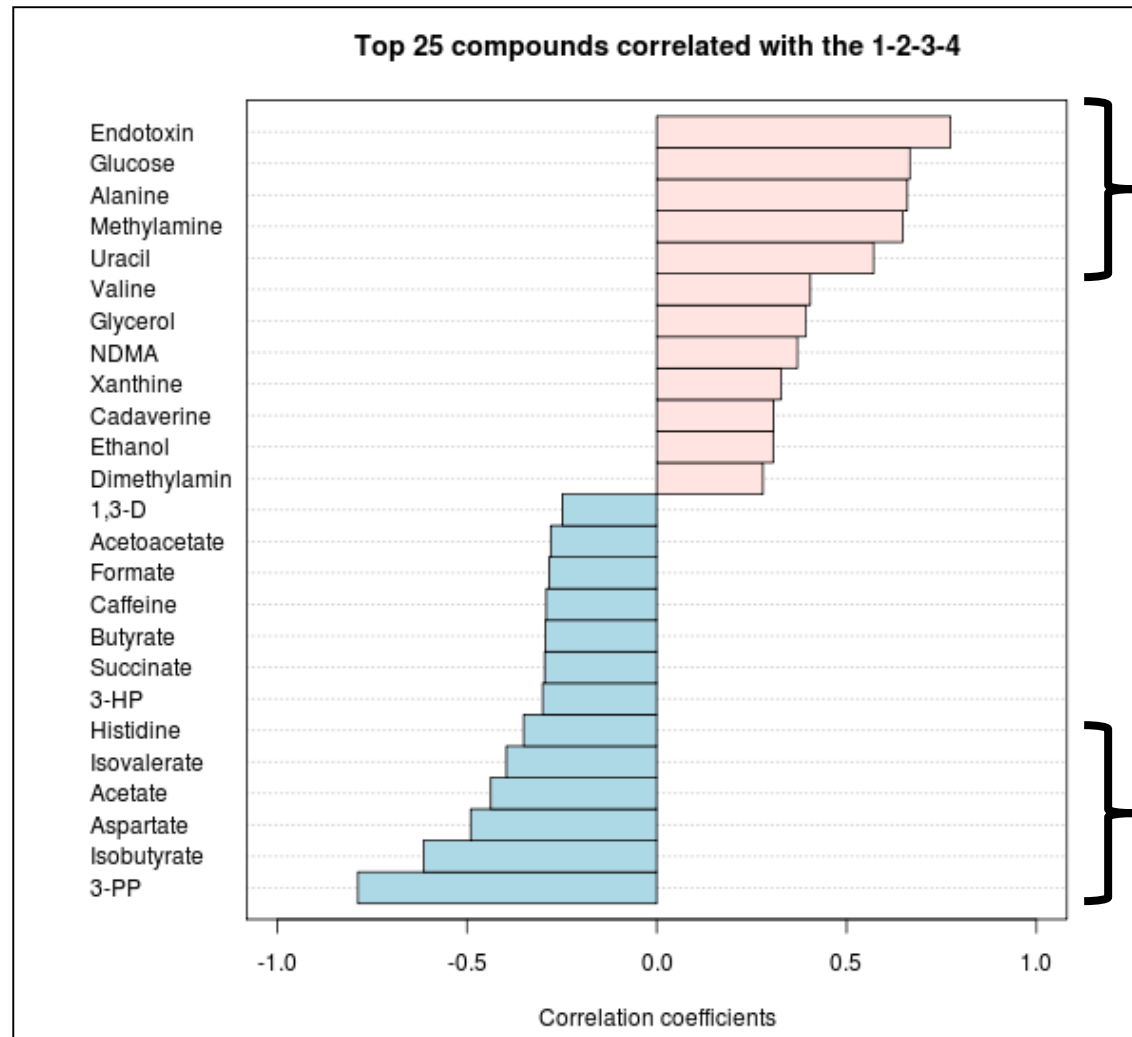
**Choose a distance measure:** Pearson r

**Submit**

**Navigation Menu:**

- Upload
- Processing
- Normalization
- Statistics
  - Fold change
  - T-test
  - Volcano plot
  - ANOVA
  - Correlations
  - PatternHunter**
  - PCA

# Pattern Matching (cont.)



**Strong linear  
+ correlation**

**Strong linear  
- correlation**

## Select an analysis path to explore :

### Univariate Analysis

Fold Change Analysis   T-tests   Volcano plot

[One-way Analysis of Variance \(ANOVA\)](#)

[Correlation Analysis](#)   [Pattern Searching](#)

### 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)

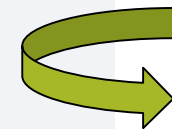
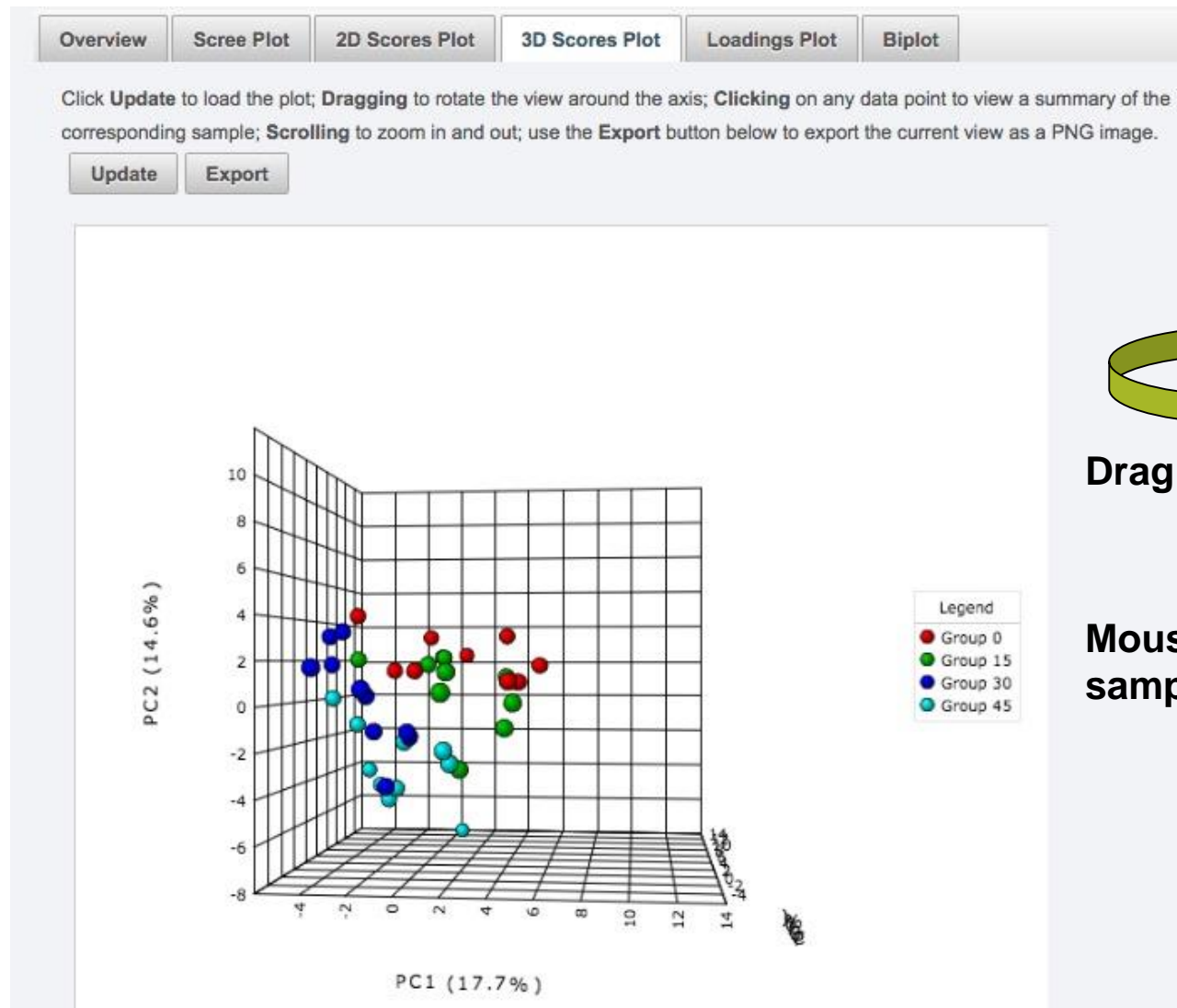
# PCA Scores Plot



Use PCA option to view the separation (if any) in the groups



# 3D Score Plot



**Drag to rotate**

**Mouse over to see  
sample names**

## Select an analysis path to explore :

### Univariate Analysis

Fold Change Analysis   T-tests   Volcano plot

[One-way Analysis of Variance \(ANOVA\)](#)

[Correlation Analysis](#)   [Pattern Searching](#)

### 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)

# Multivariate Analysis

- Use PLS-DA option to view the separation of the (labeled) groups
- PLS-DA “rotates” the PCA axes to maximize separation
- Look at the 2D PLS Scores Plot
- Look at the  $Q^2$  and  $R^2$  (Cross Validation) values
- Use the VIP plot to ID important metabolites

# PLS-DA Score Plot

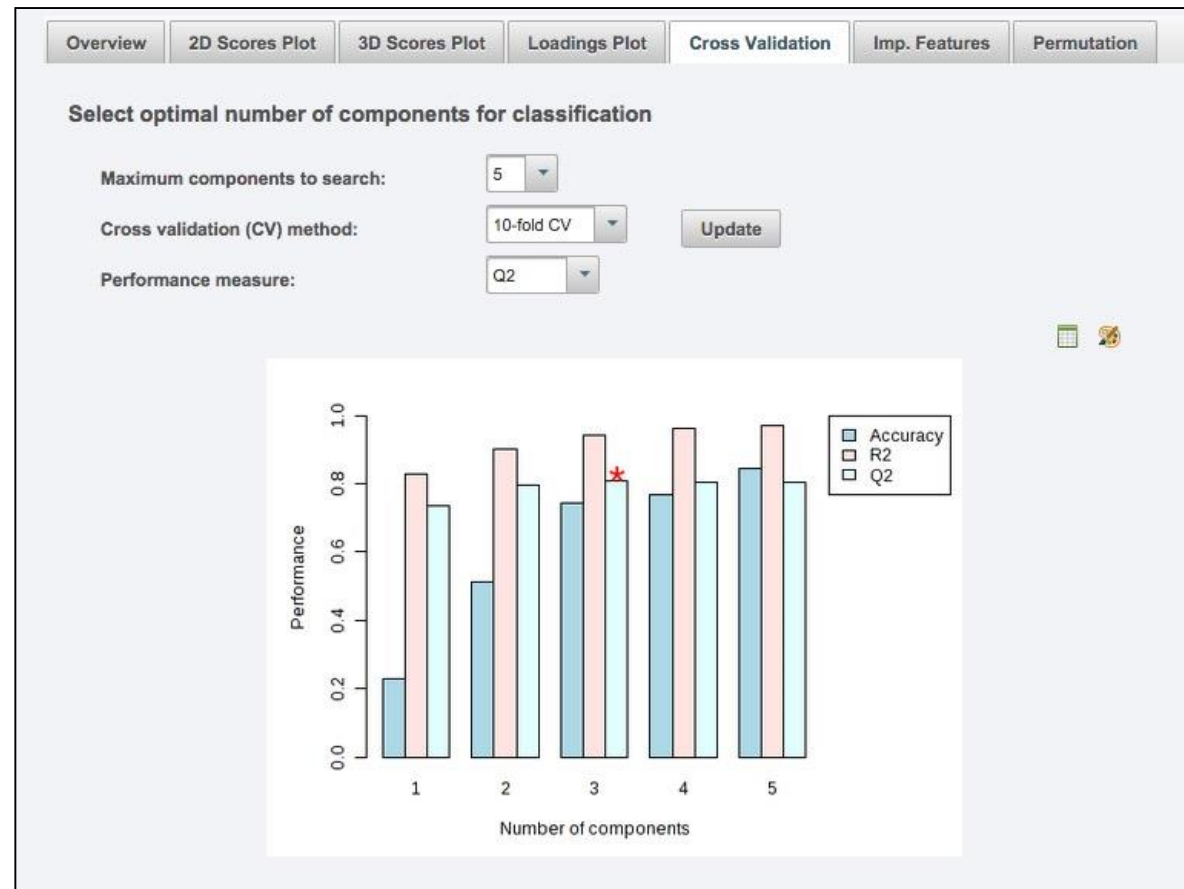


- Use PLS-DA option to view the separation of the (labeled) groups
- PLS-DA “rotates” the PCA axes to maximize separation
- Look at the 2D PLS Scores Plot
- Look at the  $Q^2$  and  $R^2$  (Cross Validation) values
- Use the VIP plot to ID important metabolites

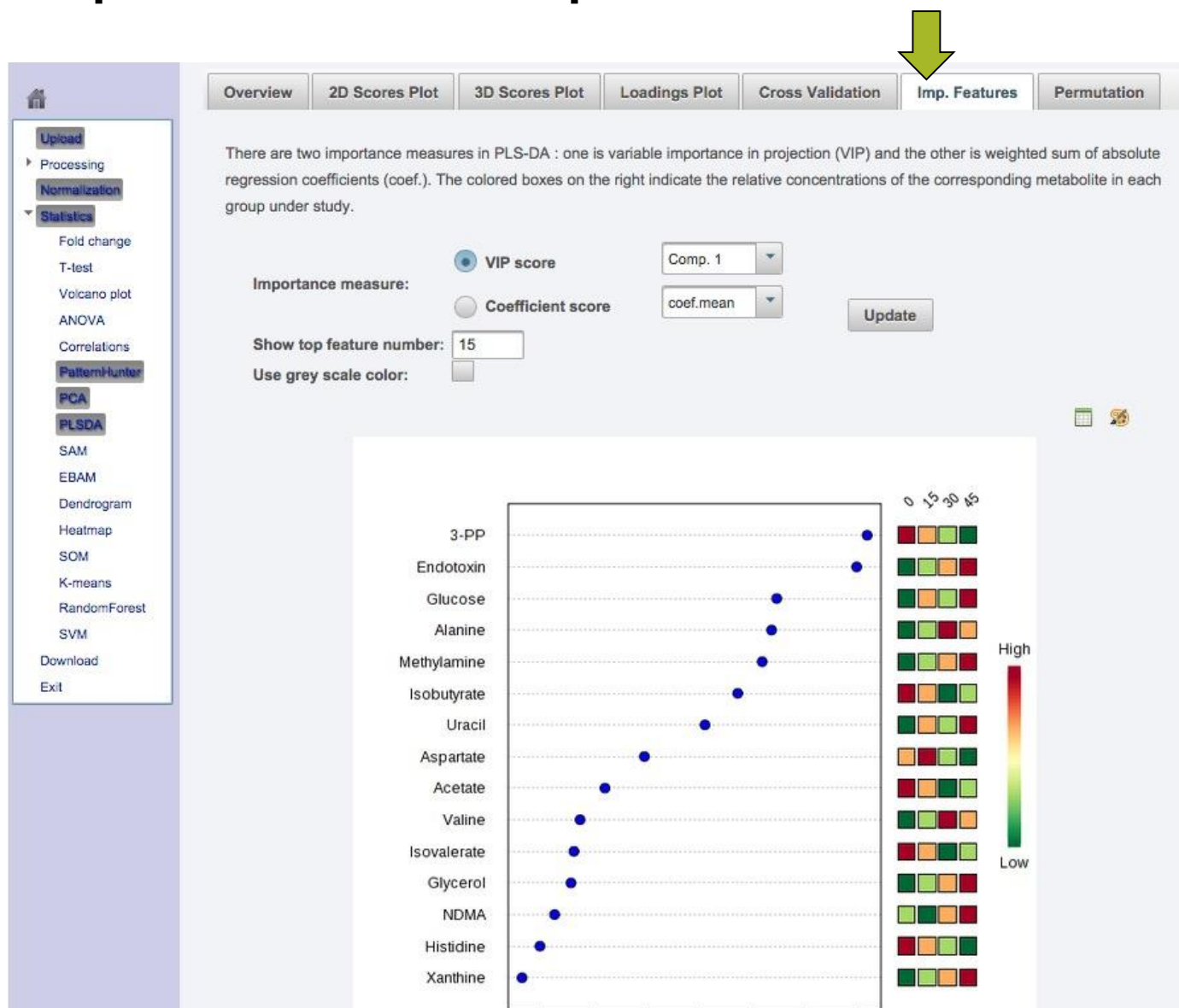


# Evaluation of PLS-DA Model

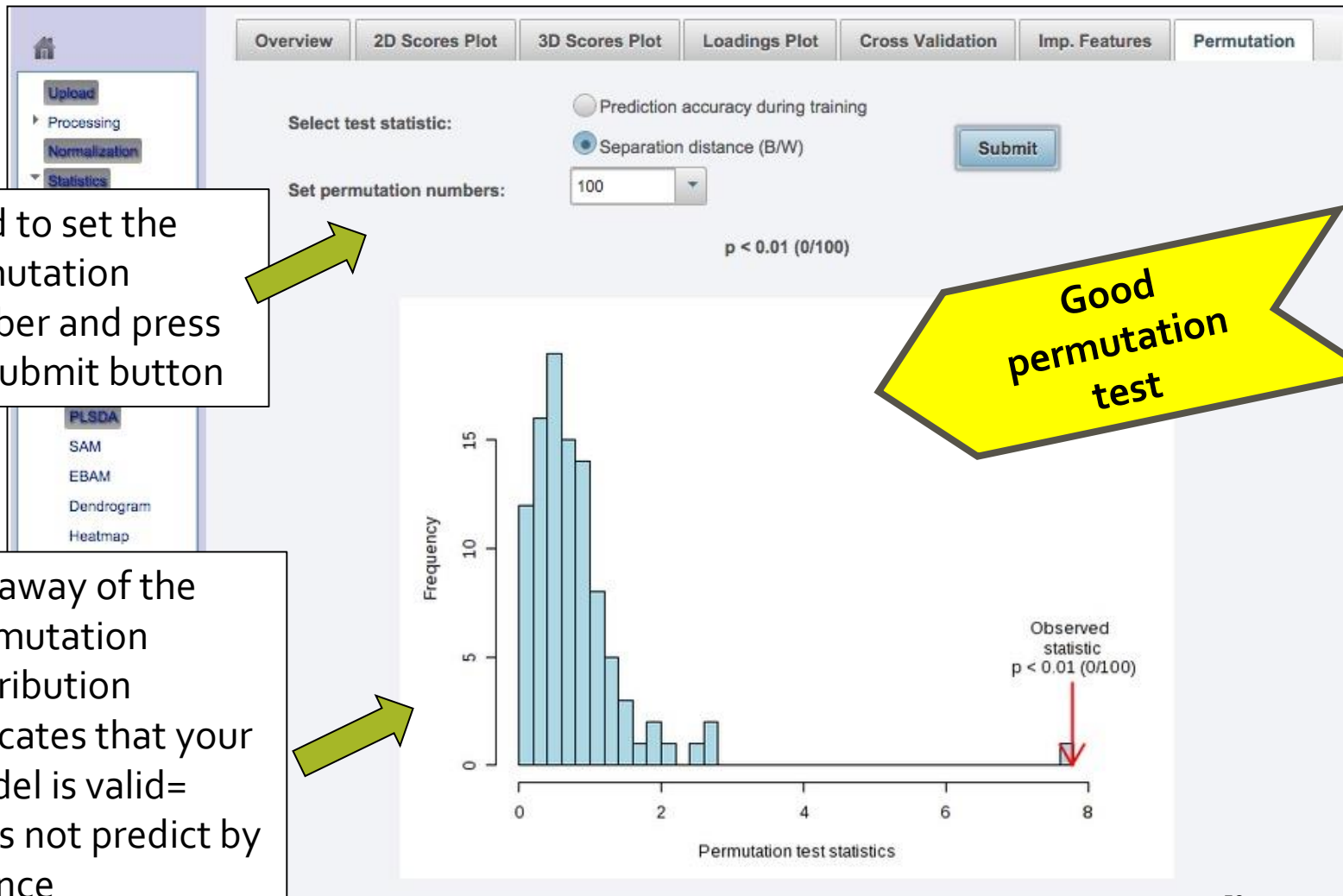
- PLS-DA Model evaluated by cross validation of  $Q^2$  and  $R^2$
- Using too many components can over-fit
- 3 component model seems to be a good compromise here
- Better  $R^2$  and  $Q^2$  as closer to 1



# Important Compounds in the Model



# Model Validation







Upload

Processing

Pre-process

Data check

Missing value

Data filter

Data editor

Image options

Normalization

Statistics

Download

Exit

## Select an analysis path to explore :

### 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 \(SOM\)](#)

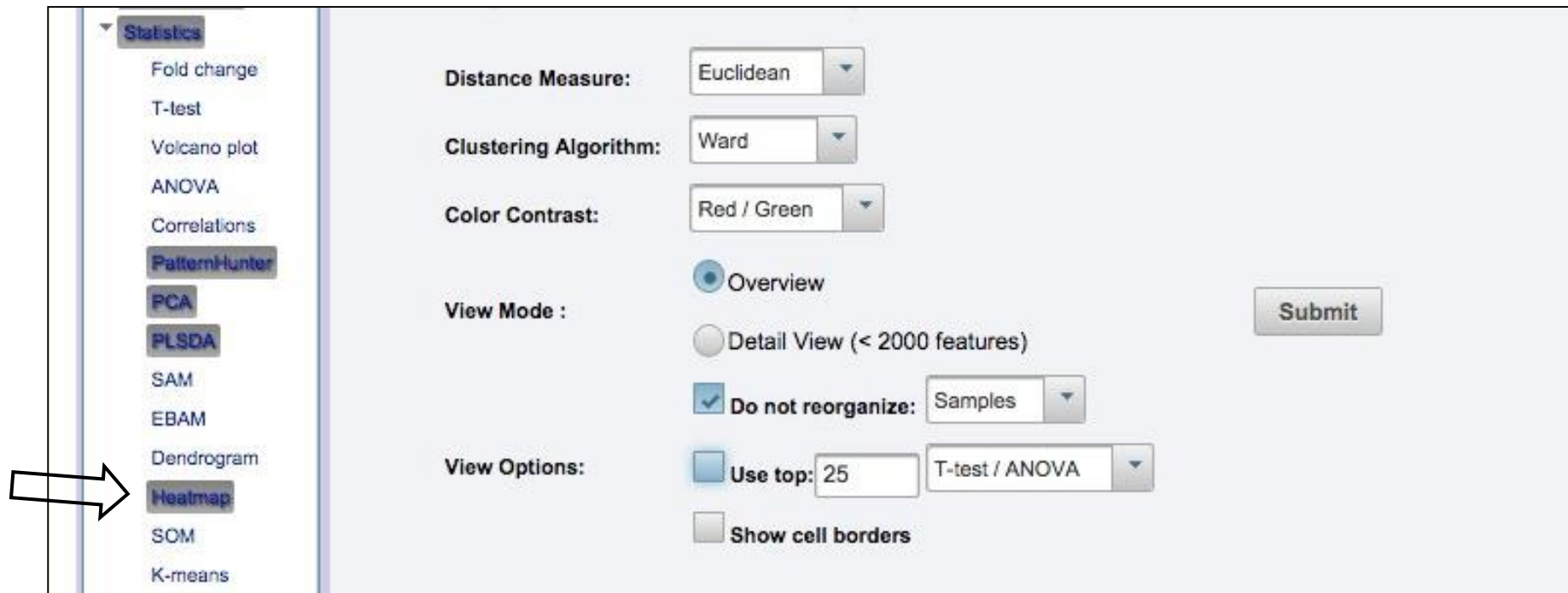
### Classification & Feature Selection

[Random Forest](#)

Support Vector Machine (SVM)



# Heatmap Visualization



**Statistics**

- Fold change
- T-test
- Volcano plot
- ANOVA
- Correlations
- PatternHunter**
- PCA**
- PLSDA**
- SAM
- EBAM
- Dendrogram
- Heatmap**
- SOM
- K-means

**Distance Measure:** Euclidean

**Clustering Algorithm:** Ward

**Color Contrast:** Red / Green

**View Mode :**

- ☒ Overview
- ☐ Detail View (< 2000 features)

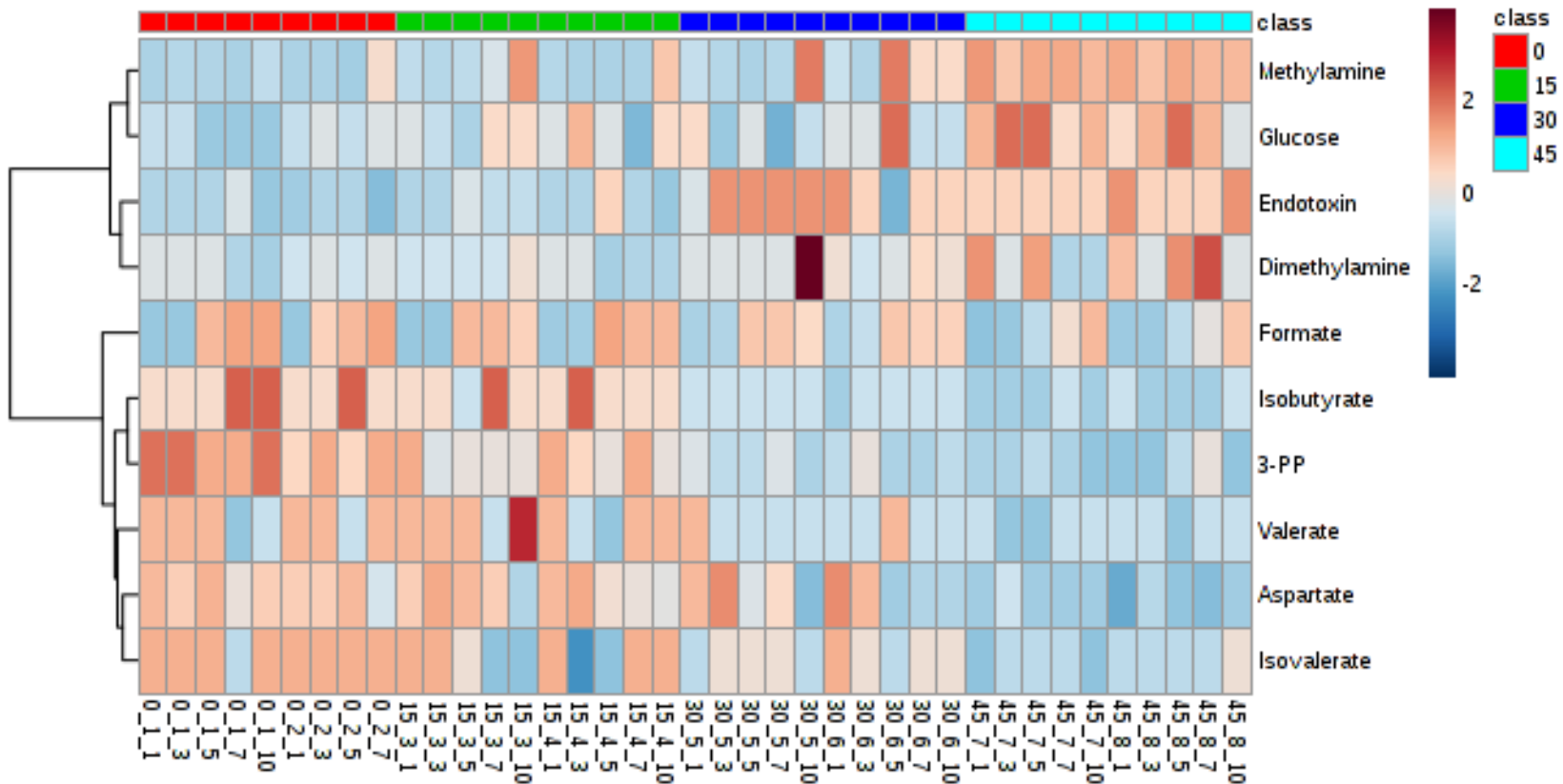
☒ **Do not reorganize:** Samples

**View Options:**

- ☒ **Use top:** 25 T-test / ANOVA
- ☐ **Show cell borders**

**Submit**


# Heatmap Visualization (cont.)



# What's Next?

- Most of the multivariate analysis is now done
- MetaboAnalyst has been keeping track of the plots or graphs you have generated
- Now its time to generate a printed report that summarizes what you've done and what you've found

# Download Results



Upload

Processing

Normalization

Statistics

Fold change

T-test

Volcano plot

ANOVA

Correlations

PatternHunter

PCA

PLSDA

SAM

EBAM

Dendrogram

Heatmap

SOM

K-means

RandomForest

SVM

Download

Exit

## Result Download

The "Download.zip" contains all the files in your home directory. These data will remain in the server for 72 hours before being deleted automatically.

<a href="#">Download.zip</a>	<a href="#">pca_loading_0 dpi72.png</a>
<a href="#">Analysis_Report.pdf</a>	<a href="#">data_processed.csv</a>
<a href="#">heatmap_3 dpi72.png</a>	<a href="#">plsda_coef.csv</a>
<a href="#">pca_loadings.csv</a>	<a href="#">pls_score2d_0 dpi72.png</a>
<a href="#">heatmap_0 dpi72.png</a>	<a href="#">pca_biplot_0 dpi72.png</a>
<a href="#">3-PP dpi72.png</a>	<a href="#">plsda_score.csv</a>
<a href="#">data_original.csv</a>	<a href="#">pca_score.csv</a>
<a href="#">pls_imp_0 dpi72.png</a>	<a href="#">heatmap_1 dpi72.png</a>
<a href="#">heatmap_4 dpi72.png</a>	<a href="#">pca_score2d_0 dpi72.png</a>
<a href="#">correlation_pattern.csv</a>	<a href="#">pls_perm_1 dpi72.png</a>
<a href="#">pls_cv_0 dpi72.png</a>	<a href="#">norm_0 dpi72.png</a>
<a href="#">Rhistory.R</a>	<a href="#">ptn_1 dpi72.png</a>
<a href="#">pca_pair_0 dpi72.png</a>	<a href="#">plsda_vip.csv</a>
<a href="#">pls_loading_0 dpi72.png</a>	<a href="#">heatmap_2 dpi72.png</a>
<a href="#">pca_scree_0 dpi72.png</a>	<a href="#">Isobutyrate dpi72.png</a>
<a href="#">data_normalized.csv</a>	<a href="#">heatmap_5 dpi72.png</a>
<a href="#">plsda_loadings.csv</a>	<a href="#">pls_pair_0 dpi72.png</a>

Logout

# Analysis Report

## 2.2 Correlation Analysis

Correlation analysis can be used to identify which features are correlated with a feature of interest. Correlation analysis can also be used to identify if certain features show particular patterns under different conditions. Users first need to define a pattern in the form of a series of hyphenated numbers. For example, in a time-series study with four time points, a pattern of 1-2-3-4 is used to search compounds with increasing the concentration as time changes; while a pattern of 3-2-1-3 can be used to search compounds that decrease at first, then bounce back to the original level.

Figure 3 shows the important features identified by correlation analysis. Table 3 shows the details of these features.

Table 3: Important features identified by Pattern search using correlation analysis

Compound	Correlation	t-test	p-value
1 Butyrate	-0.81252	12833	1.4467e-05
2 Isobutyrate	-0.89758	15784	8.9015e-06
3 3-PP	-0.87255	15835	0.0014063
4 Acetate	-0.84853	15209	0.0024011
5 3-HB	-0.41943	14024	0.007892
6 Isovalerate	-0.38861	13818	0.011096
7 Lysine	-0.34401	12291	0.15439
8 Methanol	-0.34257	12277	0.13678
9 Ferulate	-0.22920	12144	0.16028
10 Fumarate	-0.21966	12050	0.17906
11 Histidine	-0.2189	12023	0.18474
12 Propionate	-0.21015	11968	0.15912
13 Malic acid	-0.2003	11859	0.22148
14 Acetoacetate	-0.17772	11635	0.27907
15 Choline	-0.11856	11094	0.47111
16 Tyrosine	-0.10857	10933	0.51847
17 P.A.G	-0.079788	10668	0.62921
18 5-EP	-0.074018	10620	0.63035
19 Formate	-0.051547	10387	0.79923
20 Aspartate	-0.031681	10198	0.84674
21 Caffeine	0.011841	9763	0.94297
22 Ribose	0.038203	9495.1	0.81387
23 1,3-D	0.043188	9453.8	0.75419
24 Succinate	0.04904	9435	0.75542
25 Glucose	0.057544	9311.8	0.72757
26 Cadaverine	0.058543	9280.3	0.71352
27 Phenylacetate	0.053742	9250.2	0.69986
28 Hypoxanthine	0.10911	8802	0.50847
29 Ethanol	0.15234	8371.8	0.26471
30 NDMA	0.18492	8063	0.25975
31 Proline	0.18713	8031.2	0.26309
32 Glutamate	0.19334	7969.9	0.23820
33 Benzamide	0.21978	7758.8	0.17854
34 Valerate	0.23938	7515.1	0.14221
35 Glycerol	0.26901	7251.3	0.06969
36 Glycine	0.28054	7107.3	0.05393
37 Nicotinate	0.28512	7083	0.078511
38 Methylamine	0.28909	7024.2	0.07431
39 Isobutyrine	0.30358	6881	0.050303
40 Xanthine	0.32054	6561.3	0.048905
41 Dimethylamine	0.33268	6306.1	0.038306
42 Leucine	0.35142	6407.9	0.029204
43 Valine	0.3805	6116.7	0.016744
44 Lactate	0.42384	5692.5	0.0071709
45 Ureacil	0.45172	5417	0.0038928
46 Endotoxin	0.80141	4926.1	0.0011471
47 Alanine	0.62028	3751.5	2.9537e-05

## 2.3 Hierarchical Clustering

In (agglomerative) hierarchical cluster analysis, each sample begins as a separate cluster and the algorithm proceeds to combine them until all samples belong to one cluster. Two parameters need to be considered when performing hierarchical clustering. The first one is similarity measure - Euclidean distance, Pearson's correlation, Spearman's rank correlation. The other parameter is clustering algorithms, including average linkage (clustering uses the centroids of the observations), complete linkage (clustering uses the farthest pair of observations between the two groups), single linkage (clustering uses the closest pair of observations) and Ward's linkage (clustering to minimize the sum of squares of any two clusters). Heatmap is often presented as a visual aid in addition to the dendrogram.

Hierarchical clustering is performed with the `hclust` function in package `stat`. Figure 15 shows the clustering result in the form of a dendrogram. Figure 16 shows the clustering result in the form of a heatmap.

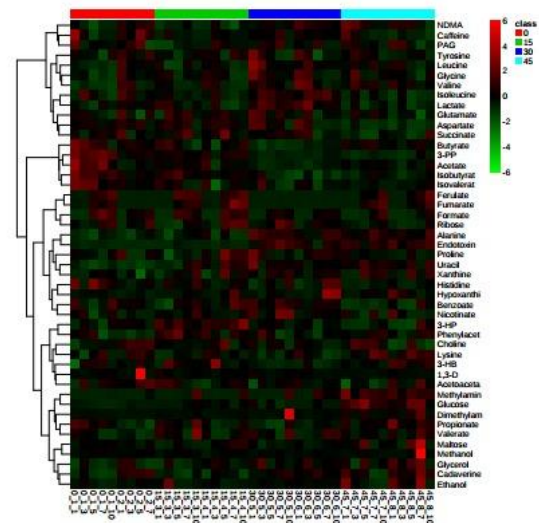
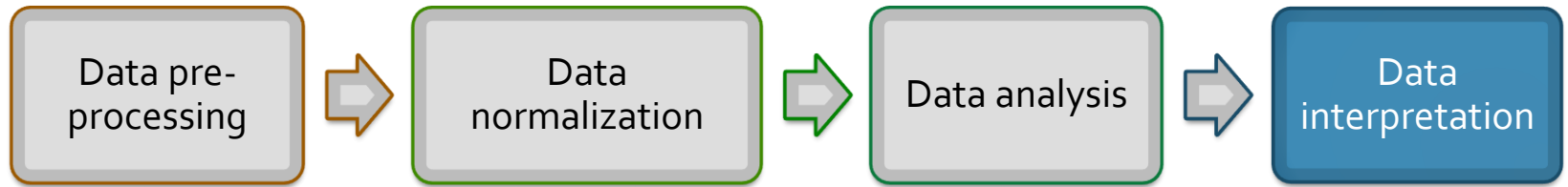


Figure 15: Clustering result shown as heatmap (distance measure using euclidean, and clustering algorithm using ward).





Metabolite enrichment  
analysis

Pathway analysis

Biomarker Analysis

# Select a Module (Enrichment Analysis)

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

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.

➤ Power Analysis

This module uses pilot data to calculate the minimum number of samples required to detect a statistically significant difference between two populations with a given degree of confidence (called Power Analysis).

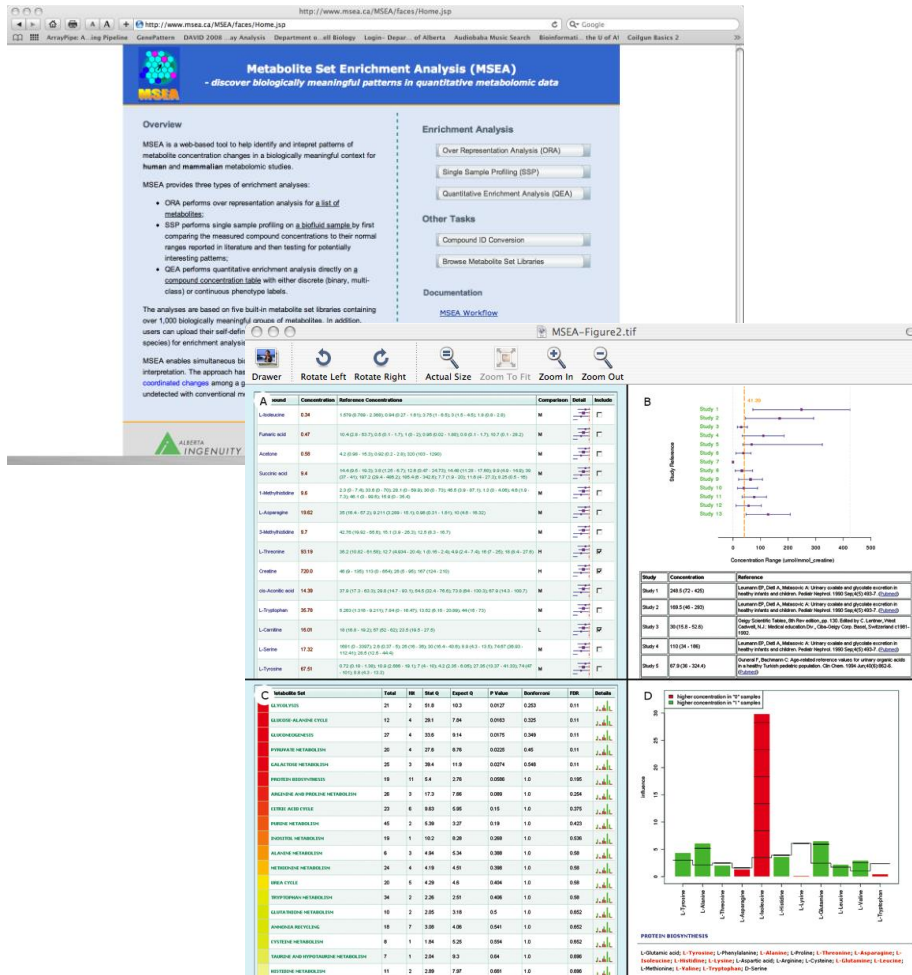
➤ Biomarker Analysis

This module performs various ROC curve based biomarker analyses for a single or multiple biomarkers. It also allows users to manually specify biomarker models as well as new sample prediction.



# Metabolite Set Enrichment Analysis (MSEA)

- Designed to handle lists of metabolites (with or without concentration data)
- Modeled after Gene Set Enrichment Analysis (GSEA)
- Supports over representation analysis (ORA), single sample profiling (SSP) and quantitative enrichment analysis (QEA)
- Contains a library of 6300 pre-defined metabolite sets including 85 pathway sets & 850 disease sets



# Enrichment Analysis

- Purpose: To test if there are **biologically meaningful** groups of metabolites that are significantly enriched in your data
- Biological meaningful in terms of:
  - Pathways
  - Disease
  - Localization
- Currently, MSEA only supports human metabolomic data

# Upload Compound List

Choose one of the following options to proceed

☒ A list of compound names (over representation analysis)

Please enter a one-column compound list:

Acetoacetic acid  
Beta-Alanine  
Creatine  
Dimethylglycine  
Fumaric acid  
Glycine  
Homocysteine  
L-Cysteine  
L-Isolucine  
L-Phenylalanine  
L-Serine  
L-Threonine  
L-Tyrosine  
L-Valine  
Phenylpyruvic acid  
Propionic acid  
Pyruvic acid  
Sarcosine

Input Type:

☒ Use example data (input type: compound names)

☐ A list of compounds with concentration values (single sample profiling)

☐ A concentration table (quantitative enrichment analysis)

**Normally GSEA would require a list of all known genes for the given platform. Here we just use the list of metabolites found in KEGG**

# Perform Compound Name Standardization

Upload

Processing

Name check

Conc. check

Data check

Missing value

Data filter

Data editor

Image options

Normalization

Enrichment

Download

Exit

## Compound Name/ID Standardization:

Please note:


- Query names in normal white indicate exact match - marked by "1" in the download file;
- Query names highlighted in red indicate **no match** - marked by "0" in the downloaded file;
- For compound name mapping, the no match query names will be highlighted in yellow indicate **no exact match found**. You should click the **View** link to perform **approximate search** and manually select the correct match if found;
- Greek alphabets are not recognized, they should be replaced by English names (i.e. alpha, beta)

Query	Hit	HMDB	PubChem	KEGG	Details
Acetoacetic acid	Acetoacetic acid	<a href="#">HMDB00060</a>	<a href="#">96</a>	<a href="#">C00164</a>	
Beta-Alanine	Beta-Alanine	<a href="#">HMDB00056</a>	<a href="#">239</a>	<a href="#">C00099</a>	
Creatine	Creatine	<a href="#">HMDB00064</a>	<a href="#">586</a>	<a href="#">C00300</a>	
Dimethylglycine	Dimethylglycine	<a href="#">HMDB00092</a>	<a href="#">673</a>	<a href="#">C01026</a>	
Fumaric acid	Fumaric acid	<a href="#">HMDB00134</a>	<a href="#">444972</a>	<a href="#">C00122</a>	
Glycine	Glycine	<a href="#">HMDB00123</a>	<a href="#">750</a>	<a href="#">C00037</a>	
Homocysteine	Homocysteine	<a href="#">HMDB00742</a>	<a href="#">778</a>	<a href="#">C05330</a>	
L-Cysteine	L-Cysteine	<a href="#">HMDB00574</a>	<a href="#">5862</a>	<a href="#">C00097</a>	
<span style="background-color: #ffffcc;">L-Isolucine</span>		-	-	-	<a href="#">View</a>
L-Phenylalanine	L-Phenylalanine	<a href="#">HMDB00159</a>	<a href="#">6140</a>	<a href="#">C00079</a>	
L-Serine	L-Serine	<a href="#">HMDB00187</a>	<a href="#">5951</a>	<a href="#">C00065</a>	
L-Threonine	L-Threonine	<a href="#">HMDB00167</a>	<a href="#">6288</a>	<a href="#">C00188</a>	
L-Tyrosine	L-Tyrosine	<a href="#">HMDB00158</a>	<a href="#">6057</a>	<a href="#">C00082</a>	
L-Valine	L-Valine	<a href="#">HMDB00883</a>	<a href="#">6287</a>	<a href="#">C00183</a>	
Phenylpyruvic acid	Phenylpyruvic acid	<a href="#">HMDB00205</a>	<a href="#">997</a>	<a href="#">C00166</a>	
Propionic acid	Propionic acid	<a href="#">HMDB00237</a>	<a href="#">1032</a>	<a href="#">C00163</a>	
Pyruvic acid	Pyruvic acid	<a href="#">HMDB00243</a>	<a href="#">1060</a>	<a href="#">C00022</a>	
Sarcosine	Sarcosine	<a href="#">HMDB00271</a>	<a href="#">1088</a>	<a href="#">C00213</a>	

You can download the result [here](#)

Submit

# Select a Metabolite Set Library



Upload

Processing

Normalization

Enrichment

Set parameter

View result

Download

Exit

## Set parameters for enrichment analysis:

Please select a metabolite set library

☒ **Pathway-associated metabolite sets**

This library contains 88 metabolite sets based on normal metabolic pathways.

☐ **Disease-associated metabolite sets (Blood)**

This library contains 416 metabolite sets reported in human blood.

☐ **Disease-associated metabolite sets (Urine)**

This library contains 346 metabolite sets reported in human urine.

☐ **Disease-associated metabolite sets (CSF)**

This library contains 124 metabolite sets reported in human cerebral spinal fluid (CSF).

☐ **SNP-associated metabolite sets**

This library contains 4,500 metabolite sets based on their associations with the detected single nucleotide polymorphisms (SNPs) loci.

☐ **Predicted metabolite sets**

This library contains 912 metabolic sets that are predicted to be changed in the case of dysfunctional enzymes using genome-scale network model of human metabolism.

☐ **Location-based metabolite sets**

This library contains 57 metabolite sets based on organ, tissue, and subcellular localizations.

☐ **Self-defined metabolite sets**

[Click here to upload your own customized metabolite set library](#)

☐ Only use metabolite sets containing at least

Please specify a reference metabolome

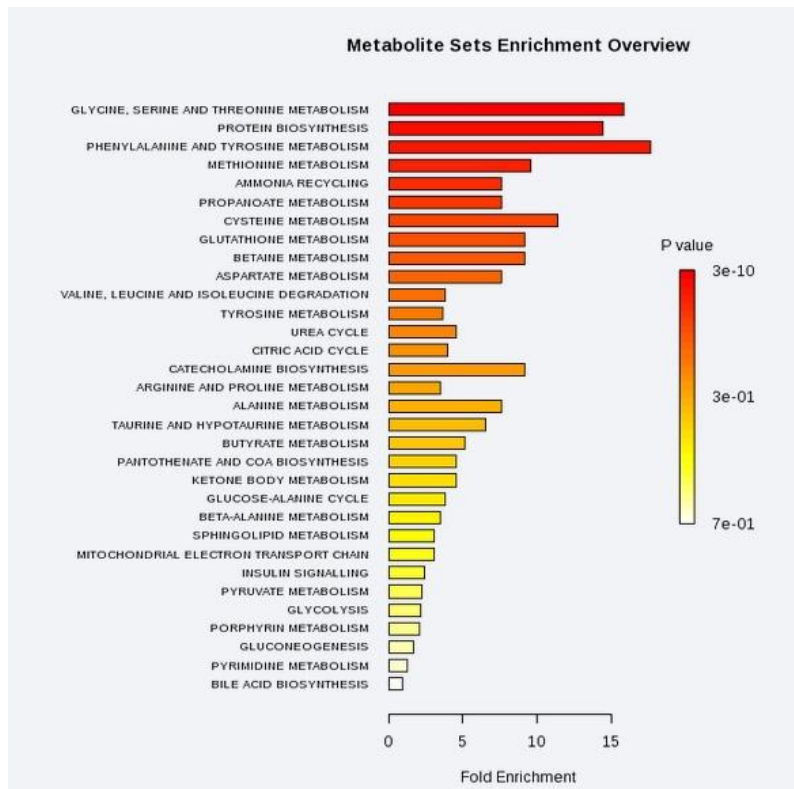
☒ **Use all the compounds in the selected metabolite set library**

☐ [Upload a reference metabolome based on your analytical platform](#)

Submit



# Result



Click on details  
to see more

Metabolite Set	Total	Hits	Expect	P value	Holm P	FDR	Details
GLYCINE, SERINE AND THREONINE METABOLISM	26	9	0.567	2.74E-10	2.19E-8	2.19E-8	<a href="#">View</a>
PROTEIN BIOSYNTHESIS	19	6	0.415	9.93E-7	7.85E-5	3.97E-5	<a href="#">View</a>
PHENYLALANINE AND TYROSINE METABOLISM	13	5	0.284	3.15E-6	2.46E-4	8.4E-5	<a href="#">View</a>
METHIONINE METABOLISM	24	5	0.524	8.98E-5	0.00691	0.0018	<a href="#">View</a>
AMMONIA RECYCLING	18	3	0.393	0.00581	0.441	0.0774	<a href="#">View</a>
PROPANOATE METABOLISM	18	3	0.393	0.00581	0.441	0.0774	<a href="#">View</a>
CYSTEINE METABOLISM	8	2	0.175	0.0117	0.863	0.133	<a href="#">View</a>
GLUTATHIONE METABOLISM	10	2	0.218	0.0183	1.0	0.162	<a href="#">View</a>
BETAINE METABOLISM	10	2	0.218	0.0183	1.0	0.162	<a href="#">View</a>
ASPARTATE METABOLISM	12	2	0.262	0.0261	1.0	0.209	<a href="#">View</a>
VALINE, LEUCINE AND ISOLEUCINE DEGRADATION	36	3	0.785	0.0397	1.0	0.288	<a href="#">View</a>
TYROSINE METABOLISM	38	3	0.829	0.0456	1.0	0.304	<a href="#">View</a>
UREA CYCLE	20	2	0.436	0.0677	1.0	0.417	<a href="#">View</a>
CITRIC ACID CYCLE	23	2	0.502	0.0868	1.0	0.496	<a href="#">View</a>
CATECHOLAMINE BIOSYNTHESIS	5	1	0.109	0.105	1.0	0.536	<a href="#">View</a>
ARGININE AND PROLINE METABOLISM	26	2	0.567	0.107	1.0	0.536	<a href="#">View</a>
ALANINE METABOLISM	6	1	0.131	0.124	1.0	0.585	<a href="#">View</a>
TAURINE AND HYPOTAURINE METABOLISM	7	1	0.153	0.144	1.0	0.638	<a href="#">View</a>
BUTYRATE METABOLISM	9	1	0.196	0.181	1.0	0.758	<a href="#">View</a>
PANTOTHENATE AND COA BIOSYNTHESIS	10	1	0.218	0.199	1.0	0.758	<a href="#">View</a>
KETONE BODY METABOLISM	10	1	0.218	0.199	1.0	0.758	<a href="#">View</a>
GLUCOSE-ALANINE CYCLE	12	1	0.262	0.234	1.0	0.851	<a href="#">View</a>
BETA-ALANINE METABOLISM	13	1	0.284	0.251	1.0	0.873	<a href="#">View</a>
SPHINGOLIPID METABOLISM	15	1	0.327	0.284	1.0	0.908	<a href="#">View</a>
MITOCHONDRIAL ELECTRON TRANSPORT CHAIN	15	1	0.327	0.284	1.0	0.908	<a href="#">View</a>
INSULIN SIGNALLING	19	1	0.415	0.345	1.0	1.0	<a href="#">View</a>
PYRUVATE METABOLISM	20	1	0.436	0.36	1.0	1.0	<a href="#">View</a>
GLYCOLYSIS	21	1	0.458	0.374	1.0	1.0	<a href="#">View</a>
PORPHYRIN METABOLISM	22	1	0.48	0.388	1.0	1.0	<a href="#">View</a>
GLUCONEOGENESIS	27	1	0.589	0.454	1.0	1.0	<a href="#">View</a>
PYRIMIDINE METABOLISM	36	1	0.785	0.556	1.0	1.0	<a href="#">View</a>
BILE ACID BIOSYNTHESIS	49	1	1.07	0.672	1.0	1.0	<a href="#">View</a>

Submit

# The Matched Metabolite Set

Metabolite Set	Total	Hits	Expect	P value	Holm P	FDR	Details
GLYCINE, SERINE AND THREONINE METABOLISM	26	9	0.567	2.74E-10	2.19E-8	2.19E-8	<a href="#">View</a>
PROTEIN BIOSYNTHESIS	19	6	0.415	9.93E-7	7.85E-5	3.97E-5	<a href="#">View</a>
PHENYLALANINE AND TYROSINE METABOLISM	13	5	0.284	3.15E-6	2.46E-4	8.4E-5	<a href="#">View</a>
METHIONINE METABOLISM	8	1	0.131	0.124	1.0	0.585	<a href="#">View</a>
AMMONIA RECYCLING	4	1	0.153	0.144	1.0	0.638	<a href="#">View</a>
PROPANOATE METABOLISM	4	1	0.196	0.181	1.0	0.758	<a href="#">View</a>
CYSTEINE METABOLISM	4	1	0.196	0.181	1.0	0.758	<a href="#">View</a>
GLUTATHIONE METABOLISM	4	1	0.196	0.181	1.0	0.758	<a href="#">View</a>
BETAINES METABOLISM	4	1	0.196	0.181	1.0	0.758	<a href="#">View</a>
ASPARTATE METABOLISM	4	1	0.196	0.181	1.0	0.758	<a href="#">View</a>
VALINE, LEUCINE AND ISOLEUCINE METABOLISM	4	1	0.196	0.181	1.0	0.758	<a href="#">View</a>
TYROSINE METABOLISM	4	1	0.196	0.181	1.0	0.758	<a href="#">View</a>
UREA CYCLE	4	1	0.196	0.181	1.0	0.758	<a href="#">View</a>
CITRIC ACID CYCLE	4	1	0.196	0.181	1.0	0.758	<a href="#">View</a>
CATECHOLAMINE BIOSYNTHESIS	4	1	0.196	0.181	1.0	0.758	<a href="#">View</a>
ARGININE AND PROLINE METABOLISM	4	1	0.196	0.181	1.0	0.758	<a href="#">View</a>
ALANINE METABOLISM	6	1	0.131	0.124	1.0	0.585	<a href="#">View</a>
TAURINE AND HYPOTAURINE METABOLISM	7	1	0.153	0.144	1.0	0.638	<a href="#">View</a>
BUTYRATE METABOLISM	9	1	0.196	0.181	1.0	0.758	<a href="#">View</a>

Current metabolite set:

Set Name	Metabolites	References
PHENYLALANINE AND TYROSINE METABOLISM	Ammonia; <b>Acetoacetic acid</b> ; Homogentisic acid; <b>Fumaric acid</b> ; <b>L-Tyrosine</b> ; <b>L-Phenylalanine</b> ; <b>Phenylpyruvic acid</b> ; 4-Hydroxyphenylpyruvic acid; 4-Fumarylacetoacetic acid; Oxygen; Maleylacetoacetic acid; Water; Hydrogen peroxide	<a href="#">SMPDB</a>

OK

Click on **SMPDB** to see more information

# Metabolism of the compound of interest in SMPDB

The screenshot displays the SMPDB website interface. At the top, there is a navigation bar with the SMPDB logo, a search bar, and links for Browse, Search, About, Downloads, and Contact Us. Below the navigation bar, a large, detailed metabolic pathway diagram is shown, centered around a liver icon. The diagram illustrates the metabolic pathways for Phenylalanine and Tyrosine, including their conversion to various products like L-DOPA, dopamine, norepinephrine, and epinephrine. The pathway is color-coded and includes chemical structures for key molecules. To the right of the diagram, a sidebar titled 'Pathway Description' provides information about the pathway, including its title, organism, and a detailed description of the metabolic process.

**Pathway Description**

**Phenylalanine and Tyrosine Metabolism**

*Homo sapiens*

**Metabolic Pathway**

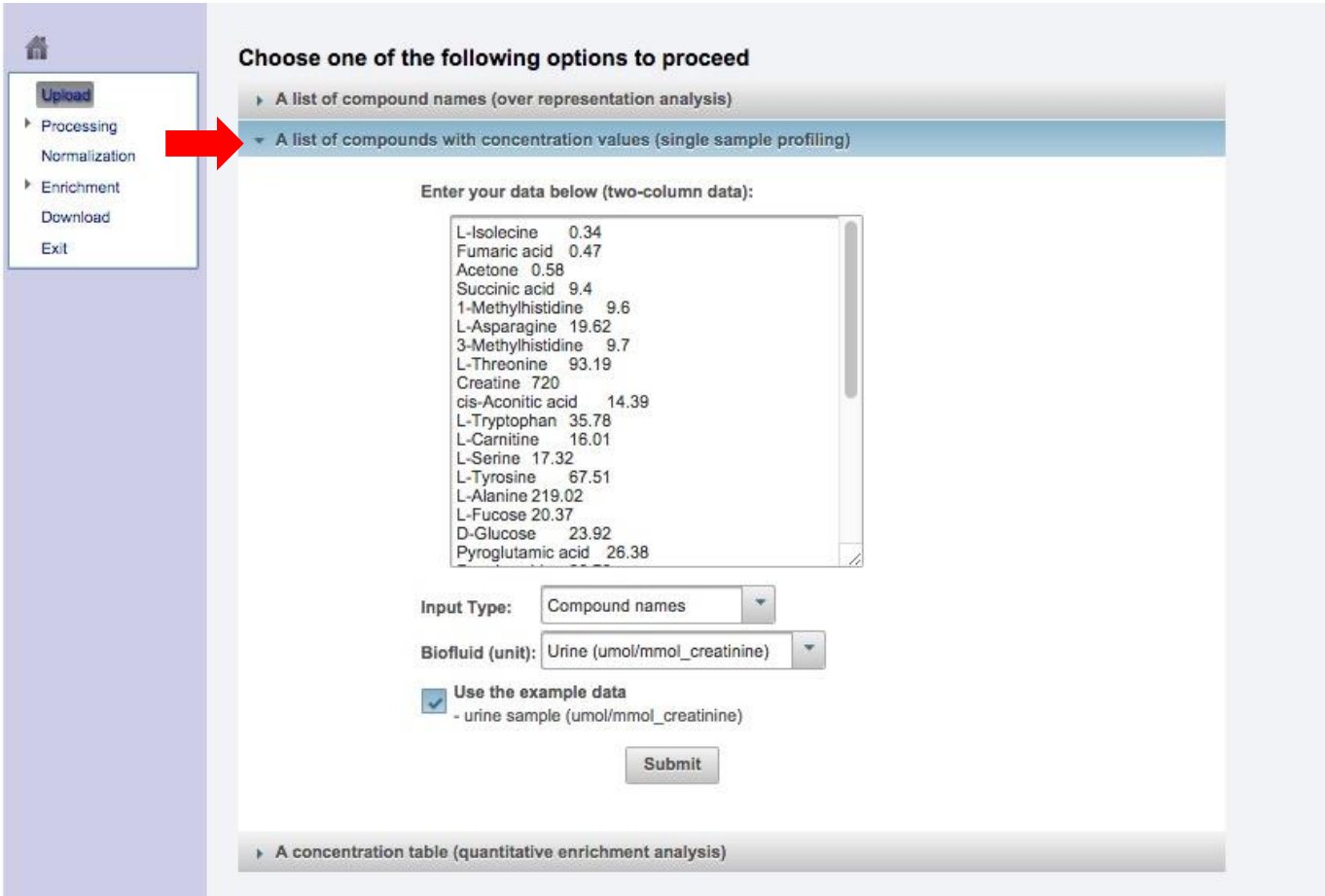
In man, phenylalanine is an essential amino acid which must be supplied in the dietary proteins. Once in the body, phenylalanine may follow any of three paths. It may be (1) incorporated into cellular proteins, (2) converted to phenylpyruvic acid, or (3) converted to tyrosine. Tyrosine is found in many high protein food products such as soy products, chicken, turkey, fish, peanuts, almonds, avocados, bananas, milk, cheese, yogurt, cottage cheese, lima beans, pumpkin seeds, and sesame seeds. Tyrosine can be converted into L-DOPA, which is further converted into dopamine, norepinephrine (noradrenaline), and epinephrine (adrenaline). Depicted in this pathway is the conversion of phenylalanine to phenylpyruvate (via amino acid oxidase or tyrosine amino transferase acting on phenylalanine), the incorporation of phenylalanine and/or tyrosine into polypeptides (via tyrosyl tRNA synthetase and phenylalanyl tRNA synthetase) and the conversion of phenylalanine to tyrosine via phenylalanine hydroxylase. Deficiencies in this enzyme are responsible for the commonest form of phenylketonuria (PKU) in humans. This reaction functions both as the first step in tyrosine/phenylalanine catabolism by which the



# Single Sample Profiling (SSP)

Basically used by doctors to analyze a patient

Aim: compare to normal references



**Choose one of the following options to proceed**

- ▶ A list of compound names (over representation analysis)
- ▼ A list of compounds with concentration values (single sample profiling)

Enter your data below (two-column data):

L-Isoleucine	0.34
Fumaric acid	0.47
Acetone	0.58
Succinic acid	9.4
1-Methylhistidine	9.6
L-Asparagine	19.62
3-Methylhistidine	9.7
L-Threonine	93.19
Creatine	720
cis-Aconitic acid	14.39
L-Tryptophan	35.78
L-Carnitine	16.01
L-Serine	17.32
L-Tyrosine	67.51
L-Alanine	219.02
L-Fucose	20.37
D-Glucose	23.92
Pyroglutamic acid	26.38

Input Type:

Biofluid (unit):

☒ Use the example data  
- urine sample (umol/mmol\_creatinine)

▶ A concentration table (quantitative enrichment analysis)

# Concentration Comparison

Processing

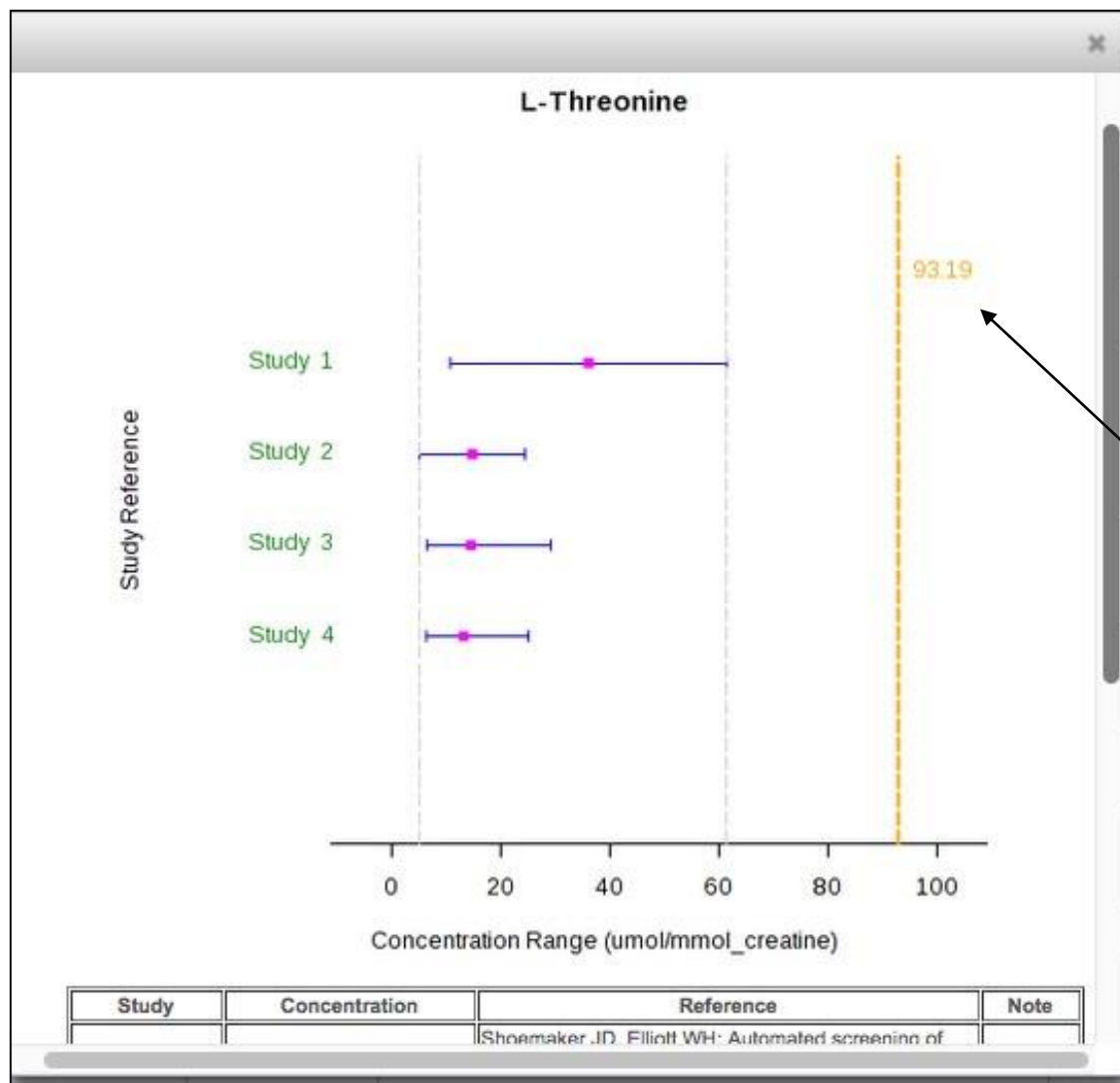
Enrichment

## Comparison with Reference Concentration

Note: *reference concentrations* are in the form of **mean(min - max)** format. In cases where the ranges were not reported in the original literature, the min and max were calculated using the 95% confidence intervals. In the *Comparison* column, **H, M, L** means **higher, medium (within range), lower** compared to the reference concentrations. Click the **Image Icon** link to see a graphical summary for the comparisons.


Compound	Concentration	Reference concentrations	Comparison	Detail	Include
<a href="#">L-Isoleucine</a>	0.34	3.75 (1 - 6.5); 3.55 (1.7 - 5.4); 0.02125 (0.0086 - 0.0339); 1.3 (0.5 - 2.7); 1.3 (0.4 - 2.6)	M	<a href="#">View</a>	<input type="checkbox"/>
<a href="#">Fumaric acid</a>	0.47	0.95 (0.02 - 1.88); 0.4 (0.2 - 0.8); 10.4 (2.8 - 53.7); 0.5 (0.1 - 1.7); 10.7 (0.1 - 28.2); 0.1 (0.1 - 1.7); 0.25 (0.1 - 0.4); 0.7 (0.2 - 1.7)	M	<a href="#">View</a>	<input type="checkbox"/>
<a href="#">Acetone</a>	0.58	2.24 (0 - 6.37); 3.9 (0.8 - 17.6)	M	<a href="#">View</a>	<input type="checkbox"/>
<a href="#">Succinic acid</a>	9.4	12.6 (0.47 - 24.73); 7.5 (0.5 - 16); 7.7 (1.9 - 20); 197.2 (29.4 - 486.2); 185.4 (6 - 342.6); 11.6 (4 - 27.3); 14.48 (11.28 - 17.68); 8.25 (0.5 - 16); 5.6 (1.8 - 9.4); 9.9 (4.9 - 14.9); 14.4 (9.5 - 19.3); 6.2 (2.5 - 13.5); 4.7 (1.1 - 14.5); 6 (0.3 - 33.3)	M	<a href="#">View</a>	<input type="checkbox"/>
<a href="#">1-Methylhistidine</a>	9.6	4.6 (1.9 - 7.3); 2.3 (0 - 7.4); 46.1 (0 - 99.6); 15.9 (0 - 35.4); 28.1 (0 - 59.9); 1.3 (0 - 4.06); 45.5 (3.9 - 87.1); 33.6 (0 - 70); 15.9 (0 - 35.4); 30 (0 - 73); 0.00285 (0.0019 - 0.0038); 8.3 (2.4 - 28.4)	M	<a href="#">View</a>	<input type="checkbox"/>
<a href="#">L-Asparagine</a>	19.62	0.96 (0.31 - 1.61); 10.52 (6.67 - 14.37); 10 (4.6 - 16.32); 10.595 (4.66 - 16.53); 8.8 (4.6 - 17.7); 9.5 (3 - 26); 10.1 (4.6 - 17.8)	M	<a href="#">View</a>	<input type="checkbox"/>
<a href="#">3-Methylhistidine</a>	9.7	42.76 (19.92 - 65.6); 12.5 (8.3 - 16.7); 0.0149 (0.0012 - 0.0286); 16.5 (2.8 - 59.8)	M	<a href="#">View</a>	<input type="checkbox"/>
<a href="#">L-Threonine</a>	93.19	36.2 (10.82 - 61.58); 14.88 (5.17 - 24.59); 14.6 (6.6 - 29.3); 13.3 (6.4 - 25.2)	H	<a href="#">View</a>	<input checked="" type="checkbox"/>
<a href="#">Creatine</a>	720	113 (0 - 654); 113 (0 - 654); 46 (3 - 448)	H	<a href="#">View</a>	<input checked="" type="checkbox"/>
<a href="#">cis-Aconitic acid</a>	14.39	13 (2.7 - 44); 67.9 (14.3 - 100.7); 73.8 (64 - 130.3); 37.9 (17.3 - 63.3); 29.8 (14.7 - 93.1); 54.5 (32.4 - 76.6); 10.3 (5.2 - 16.3); 20.9 (3.8 - 95.3)	M	<a href="#">View</a>	<input type="checkbox"/>
<a href="#">L-Tryptophan</a>	35.78	13.52 (6.15 - 20.89); 5.6 (9.3 - 2.1); 6.3 (3.4 - 11.1)	H	<a href="#">View</a>	<input checked="" type="checkbox"/>
<a href="#">L-Carnitine</a>	16.01	4.5 (0.62 - 15.2); 5 (0.7 - 16.4)	M	<a href="#">View</a>	<input type="checkbox"/>

# Concentration Comparison (cont.)



Your value.  
It's out of  
the normal  
range

# Quantitative Enrichment Analysis (QEA)

  
**Upload**  
Processing  
Normalization  
Enrichment  
Download  
Exit

## Choose one of the following options to proceed

- A list of compound names (over representation analysis)
- A list of compounds with concentration values (single sample profiling)
- A concentration table (quantitative enrichment analysis)**

### Upload your concentration data (.csv or .txt)

**Group Label:** ☒ Discrete (Classification) ☐ Continuous (Regression)

**ID Type:**

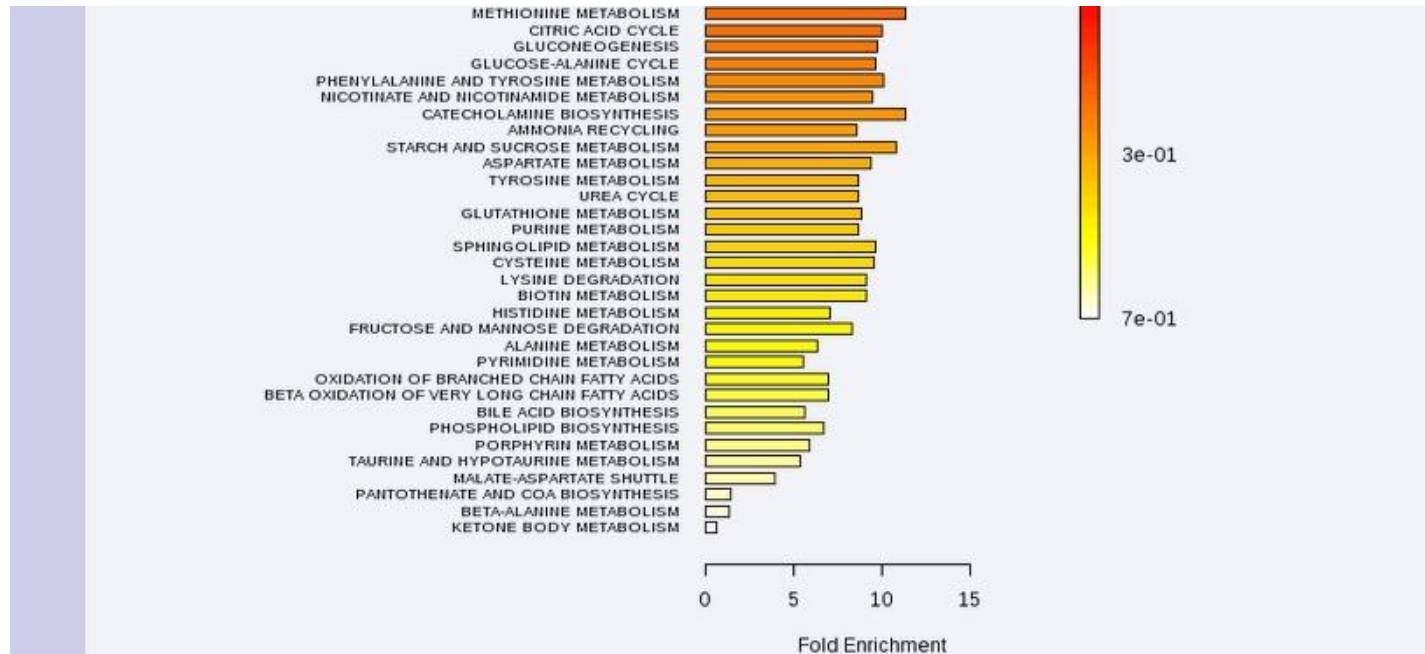
**Data File:**  No file chosen

---

### Try our test data:

Data	ID Type	Group Label	Description
<input checked="" type="radio"/> <a href="#">Data 1</a>	Common name	Discrete	Urinary metabolite concentrations from 77 cancer patients measured by 1H NMR. Phenotype: <b>N</b> - cachexic; <b>Y</b> - control
<input type="radio"/> <a href="#">Data 2</a>	PubChem CID	Continuous	Urinary metabolite concentrations from 97 cancer patients measured by 1H NMR. Phenotype: <b>muscle gain</b> (percentage within 100 days, negative values indicate muscle loss)

# Result

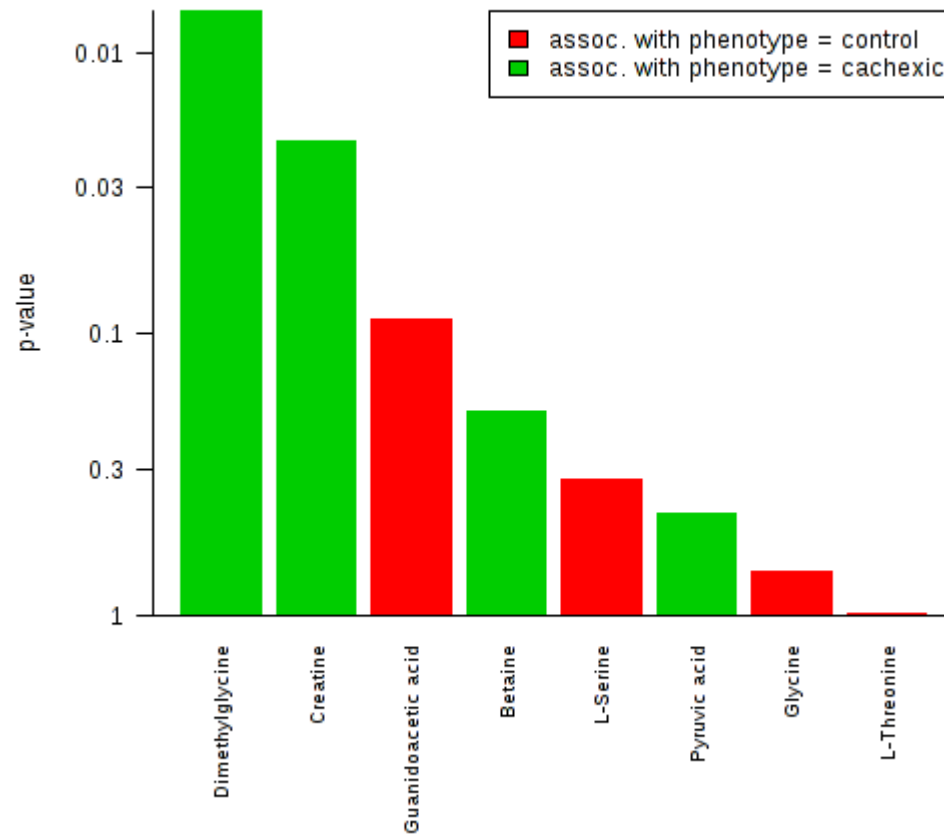


	Metabolite Set	Total	Hits	Statistic	Expected	P value	Holm P	FDR	Details
	GALACTOSE METABOLISM	25	3	18.866	1.3158	1.4154E-6	6.5107E-5	6.5107E-5	<a href="#">View</a>
	TRYPTOPHAN METABOLISM	34	1	25.111	1.3158	3.4524E-6	1.5536E-4	7.9406E-5	<a href="#">View</a>
	VALINE, LEUCINE AND ISOLEUCINE DEGRADATION	36	2	21.24	1.3158	1.569E-5	6.9038E-4	1.559E-4	<a href="#">View</a>
	GLYCOLYSIS	21	2	17.511	1.3158	2.0894E-5	8.9845E-4	1.559E-4	<a href="#">View</a>
	INSULIN SIGNALLING	19	2	17.511	1.3158	2.0894E-5	8.9845E-4	1.559E-4	<a href="#">View</a>
	PYRUVATE METABOLISM	20	3	15.116	1.3158	2.11E-5	8.9845E-4	1.559E-4	<a href="#">View</a>
	BETAINE METABOLISM	10	2	19.344	1.3158	2.5834E-5	0.0010334	1.559E-4	<a href="#">View</a>
	MITOCHONDRIAL ELECTRON TRANSPORT CHAIN	15	2	17.669	1.3158	2.9636E-5	0.0011558	1.559E-4	<a href="#">View</a>
	PROPANOATE METABOLISM	18	1	20.811	1.3158	3.0502E-5	0.0011591	1.559E-4	<a href="#">View</a>
	GLYCOLIC ACID METABOLISM	10	1	18.866	1.3158	1.4154E-6	6.5107E-5	6.5107E-5	<a href="#">View</a>

Click on details to see more





# The Matched Metabolite Set



# Pathway Analysis Module

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

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.

➤ Power Analysis

This module uses pilot data to calculate the minimum number of samples required to detect a statistically significant difference between two populations with a given degree of confidence (called Power Analysis).

➤ Biomarker Analysis

This module performs various ROC curve based biomarker analyses for a single or multiple biomarkers. It also allows users to manually specify biomarker models as well as new sample prediction.



# Pathway Analysis

- Purpose: to extend and enhance metabolite set enrichment analysis for pathways by
  - Considering pathway structures
  - Supporting pathway visualization
- Currently supports analysis for 21 diverse (model) organisms such as humans, mouse, drosophila, arabidopsis, *E. coli*, yeast, etc. (KEGG pathways only)



# Data Upload

**Upload**

- Processing
- Normalization
- Pathway
- Download
- Exit

**Please enter a one-column compound list:**

Input Type: -- Please specify

☐ Use our example data

Submit

---

**Or upload a concentration table (.csv or .txt):**

Group Label: ☐ Discrete (Classification) ☐ Continuous (Regression)

ID Type: -- Please specify

Data File:  No file chosen

☒ Use the example data

Data	Description
<a href="#">Dataset</a>	Urinary metabolite concentrations from 77 cancer patients measured by 1H NMR. Phenotype: N - cachexic; Y - control

Submit

# Perform Data Normalization

## Data Normalization:

The normalization procedures are grouped into three categories. The sample normalization allows general-purpose adjustment for differences among samples; data transformation and scaling are two different approaches to make features more comparable. You can use one or combine them to achieve better results.

### 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 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)

Submit

# Select Pathway Libraries

  
**Upload**  
Processing  
**Normalization**  
Pathway  
**Set parameter**  
View result  
Download  
Exit

Please select a pathway library:

Mammals	<input checked="" type="radio"/> Homo sapiens (human) [80] <input type="radio"/> Mus musculus (mouse) [82] <input type="radio"/> Rattus norvegicus (rat) [81] <input type="radio"/> Bos taurus (cow) [81]
Birds	<input type="radio"/> Gallus gallus (chicken) [78]
Fish	<input type="radio"/> Danio rerio (zebrafish) [81]
Insects	<input type="radio"/> Drosophila melanogaster (fruit fly) [79]
Nematodes	<input type="radio"/> Caenorhabditis elegans (nematode) [78]
Fungi	<input type="radio"/> Saccharomyces cerevisiae (yeast) [65]
Plants	<input type="radio"/> Oryza sativa japonica (Japanese rice) [83] <input type="radio"/> Arabidopsis thaliana (thale cress) [87]
Parasites	<input type="radio"/> Schistosoma mansoni [69] <input type="radio"/> Plasmodium falciparum 3D7 (Malaria) [47] <input type="radio"/> Trypanosoma brucei [54]
Prokaryotes	<input type="radio"/> Escherichia coli K-12 MG1655 [87] <input type="radio"/> Bacillus subtilis [80] <input type="radio"/> Pseudomonas putida KT2440 [89] <input type="radio"/> Staphylococcus aureus N315 (MRSA/VSSA) [73] <input type="radio"/> Thermotoga maritima [57] <input type="radio"/> Synechococcus elongatus PCC7942 [75] <input type="radio"/> Mesorhizobium loti [86]

# Perform Network Topology Analysis

Please specify a reference metabolome:

- ☒ Use all compounds in the selected pathways
- ☐ [Upload a reference metabolome based on your technical platform](#)

Identifies which metabolic pathways have compounds (from the input lists) that are over-represented and have significant perturbations to their concentrations

Specify pathway analysis algorithms:

Pathway Enrichment Analysis

- ☒ Global Test (Goeman et al., 2004)
- ☐ Global Ancova (Hummel et al., 2008)

Pathway Topology Analysis

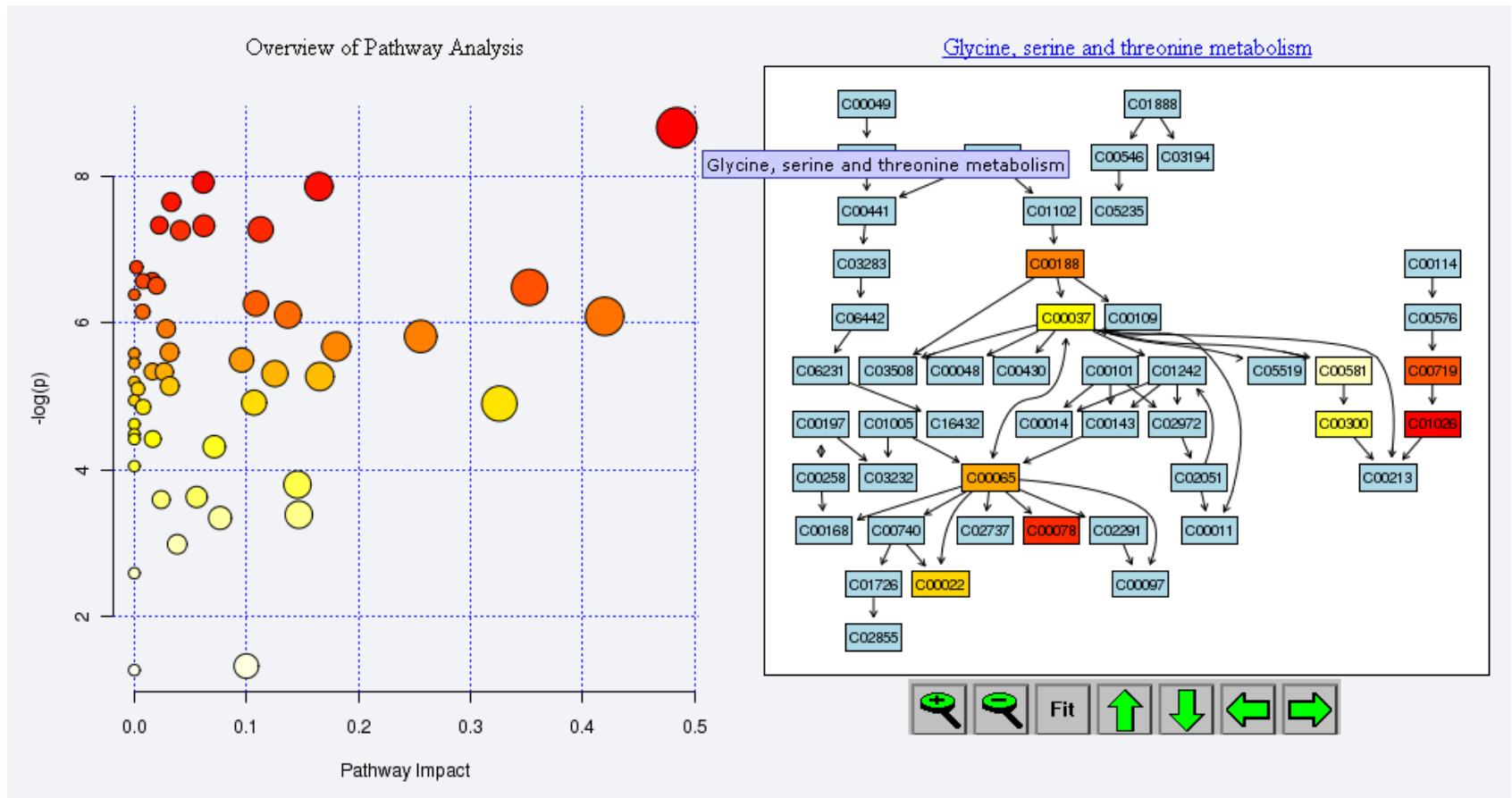
- ☒ Relative-betweenness Centrality
- ☐ Out-degree Centrality

Submit

Topological Analysis measures the centrality of a metabolite in a metabolic network or a metabolic pathway.

MetPA's pathway topological analysis is based on the centrality measures of a metabolite in a given metabolic network. Centrality is a local quantitative measure of the position of a node relative to the other nodes, and is often used to estimate a node's relative importance or role in network organization. Since metabolic networks are directed graphs, MetPA uses relative betweenness centrality and out degree centrality measures to calculate compound importance.

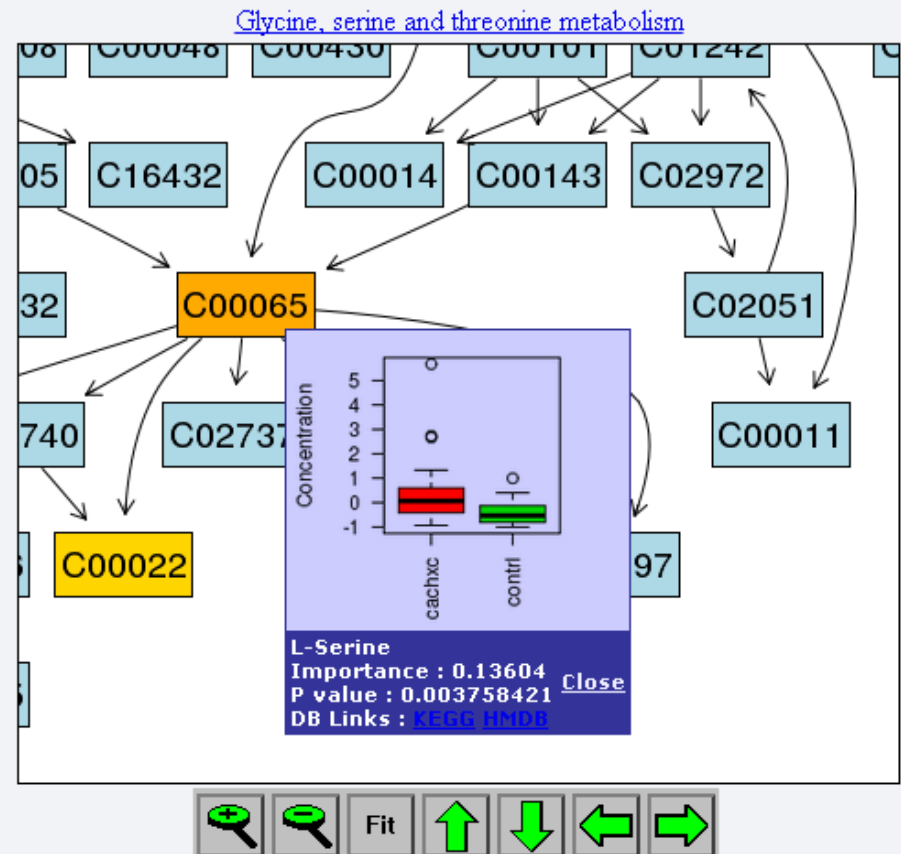
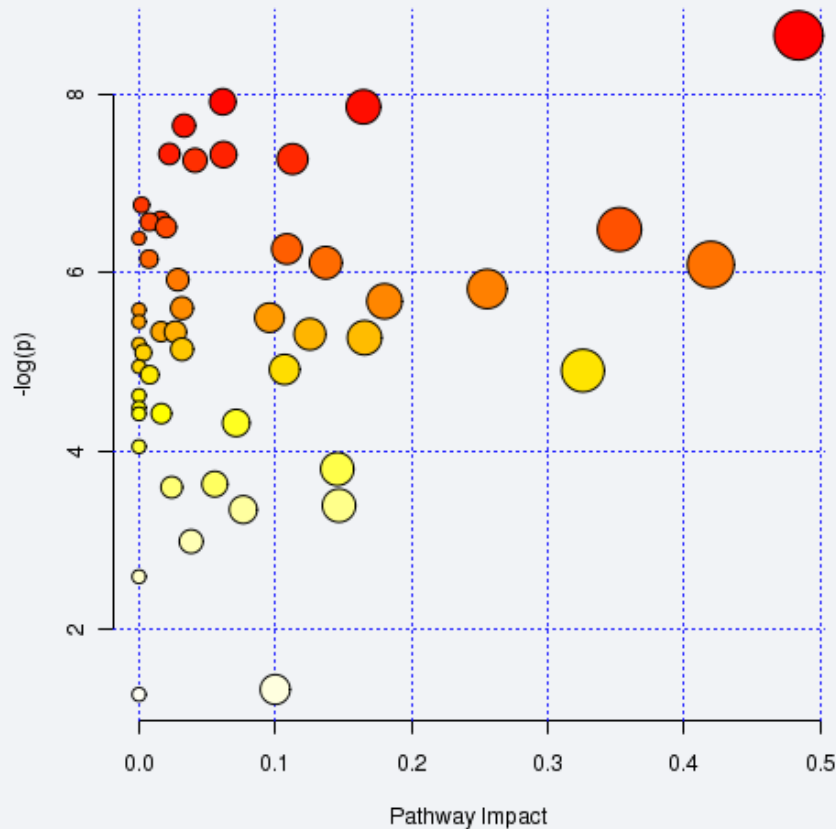
# Pathway Visualization



The pathway impact is calculated as the sum of the importance measures of the matched metabolites normalized by the sum of the importance measures of all metabolites in each pathway.

# Pathway Visualization (cont.)

Overview of Pathway Analysis





# Result



Pathway Name	Total	Hits	p	-log(p)	Holm p	FDR	Impact	Details
<a href="#">Valine, leucine and isoleucine degradation</a>	40	2	1.1954E-4	9.0319	0.0059769	0.0031356	0.02232	<a href="#">KEGG</a> <a href="#">SMP</a>
<a href="#">Valine, leucine and isoleucine biosynthesis</a>	27	4	1.2542E-4	8.9838	0.0061458	0.0031356	0.04823	<a href="#">KEGG</a> <a href="#">SMP</a>
<a href="#">Glycine, serine and threonine metabolism</a>	48	8	2.4586E-4	8.3107	0.011801	0.0040977	0.48394	<a href="#">KEGG</a> <a href="#">SMP</a>
<a href="#">Methane metabolism</a>	34	6	3.8485E-4	7.8626	0.018088	0.0043833	0.16466	<a href="#">KEGG</a>
<a href="#">Sulfur metabolism</a>	18	2	4.755E-4	7.6512	0.021873	0.0043833	0.03307	<a href="#">KEGG</a> <a href="#">SMP</a>
<a href="#">Arginine and proline metabolism</a>	77	6	6.578E-4	7.3266	0.029601	0.0043833	0.06203	<a href="#">KEGG</a> <a href="#">SMP</a>
<a href="#">Aminoacyl-tRNA biosynthesis</a>	75	10	6.6275E-4	7.3191	0.029601	0.0043833	0.11268	<a href="#">KEGG</a>
<a href="#">Nicotinate and nicotinamide metabolism</a>	44	5	7.0133E-4	7.2625	0.030157	0.0043833	0.04113	<a href="#">KEGG</a> <a href="#">SMP</a>
<a href="#">Glutathione metabolism</a>	38	2	0.0011587	6.7605	0.048664	0.0063514	0.0019	<a href="#">KEGG</a> <a href="#">SMP</a>
<a href="#">Propanoate metabolism</a>	35	4	0.0013934	6.576	0.057129	0.0063514	0.01603	<a href="#">KEGG</a> <a href="#">SMP</a>
<a href="#">Galactose metabolism</a>	41	3	0.001486	6.5116	0.059441	0.0063514	0.01992	<a href="#">KEGG</a> <a href="#">SMP</a>
<a href="#">Taurine and hypotaurine metabolism</a>	20	3	0.0015243	6.4862	0.059449	0.0063514	0.35252	<a href="#">KEGG</a> <a href="#">SMP</a>
<a href="#">Cyanoamino acid metabolism</a>	16	4	0.0016826	6.3874	0.06394	0.0064716	0.0	<a href="#">KEGG</a>
<a href="#">Nitrogen metabolism</a>	39	7	0.0021434	6.1454	0.079305	0.0070701	0.00763	<a href="#">KEGG</a> <a href="#">SMP</a>
<a href="#">Inositol phosphate metabolism</a>	39	1	0.002215	6.1125	0.079741	0.0070701	0.13703	<a href="#">KEGG</a> <a href="#">SMP</a>
<a href="#">Pyruvate metabolism</a>	32	4	0.0022624	6.0913	0.079741	0.0070701	0.41957	<a href="#">KEGG</a> <a href="#">SMP</a>
<a href="#">Cysteine and methionine metabolism</a>	56	2	0.0026796	5.9221	0.091106	0.0078811	0.02846	<a href="#">KEGG</a> <a href="#">SMP</a> <a href="#">SMP</a>
<a href="#">Alanine, aspartate and glutamate metabolism</a>	24	6	0.0029727	5.8183	0.0981	0.0082576	0.25546	<a href="#">KEGG</a> <a href="#">SMP</a> <a href="#">SMP</a> <a href="#">SMP</a>
<a href="#">Pantothenate and CoA biosynthesis</a>	27	4	0.0034143	5.6798	0.10926	0.0089486	0.18014	<a href="#">KEGG</a> <a href="#">SMP</a>
<a href="#">Phenylalanine metabolism</a>	45	6	0.0036884	5.6026	0.11434	0.0089486	0.0315	<a href="#">KEGG</a> <a href="#">SMP</a>

Submit



# Select a Module (Biomarker Analysis)

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

**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.

**Power Analysis**

This module uses pilot data to calculate the minimum number of samples required to detect a statistically significant difference between two populations with a given degree of confidence (called Power Analysis).

**Biomarker Analysis**

This module performs various ROC curve based biomarker analyses for a single or multiple biomarkers. It also allows users to manually specify biomarker models as well as new sample prediction.

# Biomarker Analysis

- Purpose is to find biomarkers using ROC (receiver operator characteristic) curves with high sensitivity and specificity
- Maximize AUC under ROC curve while minimizing the number of metabolites used in the biomarker panel
- 3 different modules
  - univariate – single marker at a time
  - multivariate – many combinations of biomarkers
  - manual – user choice

# Select Test Data Set 1

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

Home

Upload

- Processing
- Normalization
- ROC Analysis
- Download
- Exit

**Upload your data table (.csv or .txt):**

**Data Type:** ☒ Concentrations ☐ Spectral bins ☐ Peak intensity table

**Format:**

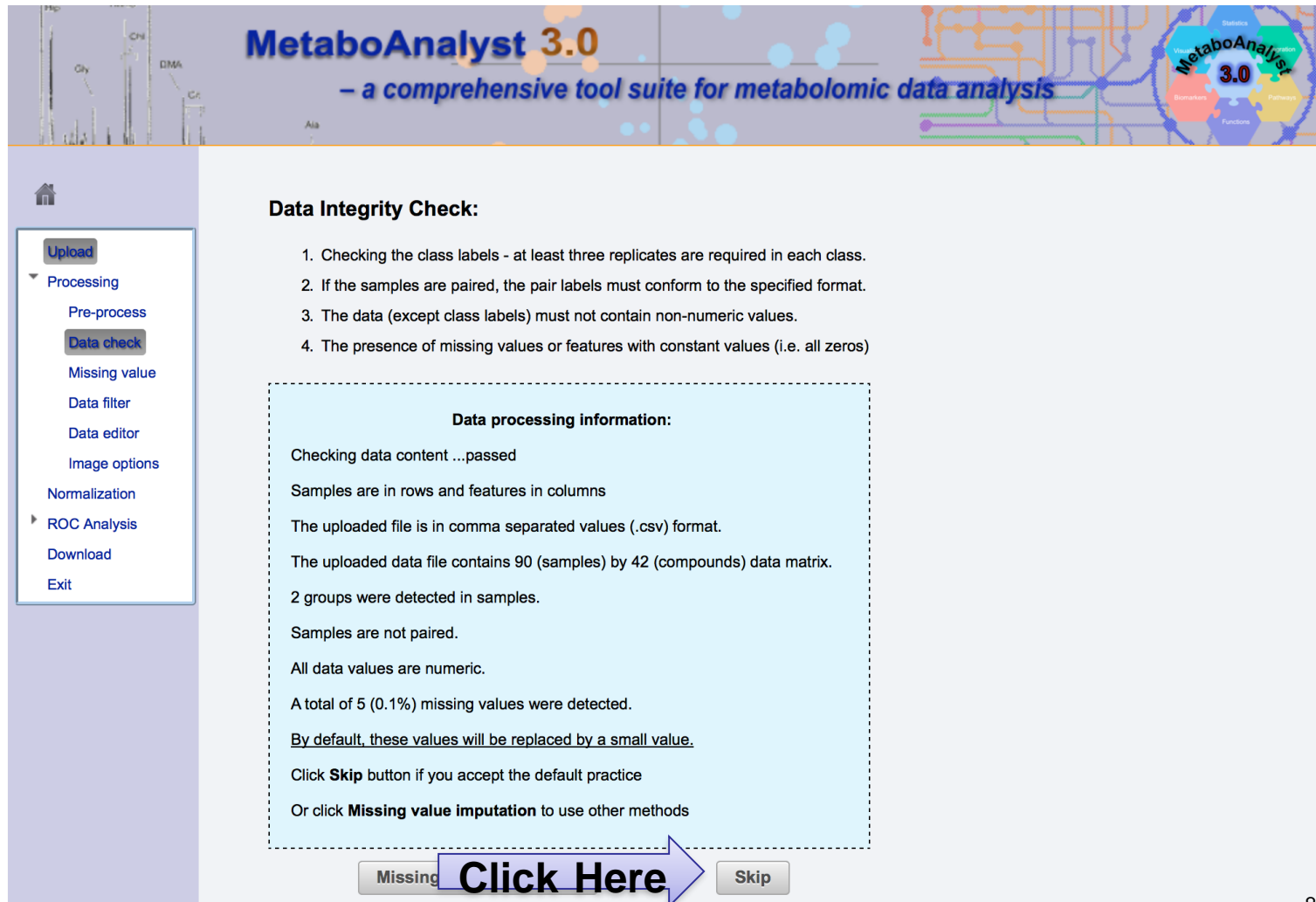
**Data File:**  no file selected

☐ Use the example data

Data	Description
<input checked="" type="radio"/> <a href="#">Dataset1</a>	Metabolite concentrations of 90 human plasma samples measured by <sup>1</sup> H NMR. Phenotype labels: 0 - Controls; 1 - Patients.
<input type="radio"/> <a href="#">Dataset2</a>	Metabolite concentrations of 77 human plasma samples. Among them, the phenotypes of 12 samples are empty/unknown. Their class can be predicted using the <b>Tester</b> module.

**Click Here**

# Perform Data Integrity Check



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

**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 rows and features in columns

The uploaded file is in comma separated values (.csv) format.

The uploaded data file contains 90 (samples) by 42 (compounds) data matrix.

2 groups were detected in samples.

Samples are not paired.

All data values are numeric.

A total of 5 (0.1%) 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

Missing **Click Here** Skip

# Perform Normalization

Upload

Processing

Pre-process

Data check

Missing value

Data filter

Data editor

Image options

Normalization

ROC Analysis

Download

Exit

**Data Normalization:**

The normalization procedures are grouped into three categories. The sample normalization allows general-purpose adjustment for differences among samples; data transformation and scaling are two different approaches to make features more comparable. You can use one or combine them to achieve better results.

**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

C01

☐ Create a pooled average sample from group

0

☐ Normalization by reference feature

2-Hydroxybutyrate

**Data transformation**

☐ None

☒ Log transformation (generalized logarithm transformation or glog)

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

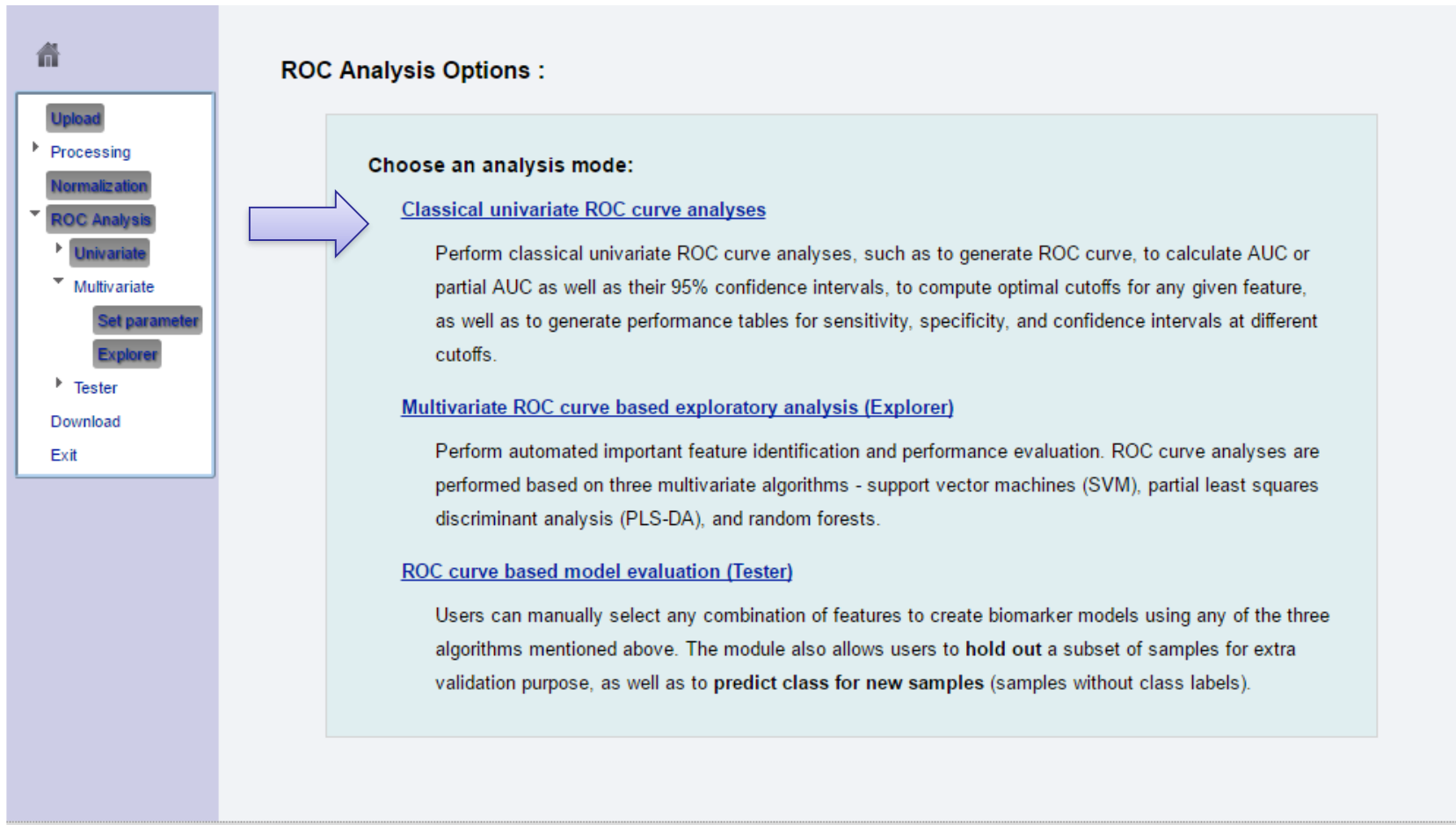
**Data scaling**

☒ None

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

86

# Select Multivariate Option



The screenshot shows a software interface with a sidebar on the left and a main content area on the right. The sidebar contains a home icon, an 'Upload' button, and a list of options: 'Processing', 'Normalization', 'ROC Analysis' (which is expanded to show 'Univariate' and 'Multivariate'), 'Tester', 'Download', and 'Exit'. The 'Multivariate' option is highlighted with a blue arrow pointing to the main content area. The main content area is titled 'ROC Analysis Options :' and contains three sections: 'Choose an analysis mode:', 'Classical univariate ROC curve analyses', 'Multivariate ROC curve based exploratory analysis (Explorer)', and 'ROC curve based model evaluation (Tester)'. Each section has a brief description of its functionality.

**ROC Analysis Options :**

**Choose an analysis mode:**

[Classical univariate ROC curve analyses](#)

Perform classical univariate ROC curve analyses, such as to generate ROC curve, to calculate AUC or partial AUC as well as their 95% confidence intervals, to compute optimal cutoffs for any given feature, as well as to generate performance tables for sensitivity, specificity, and confidence intervals at different cutoffs.

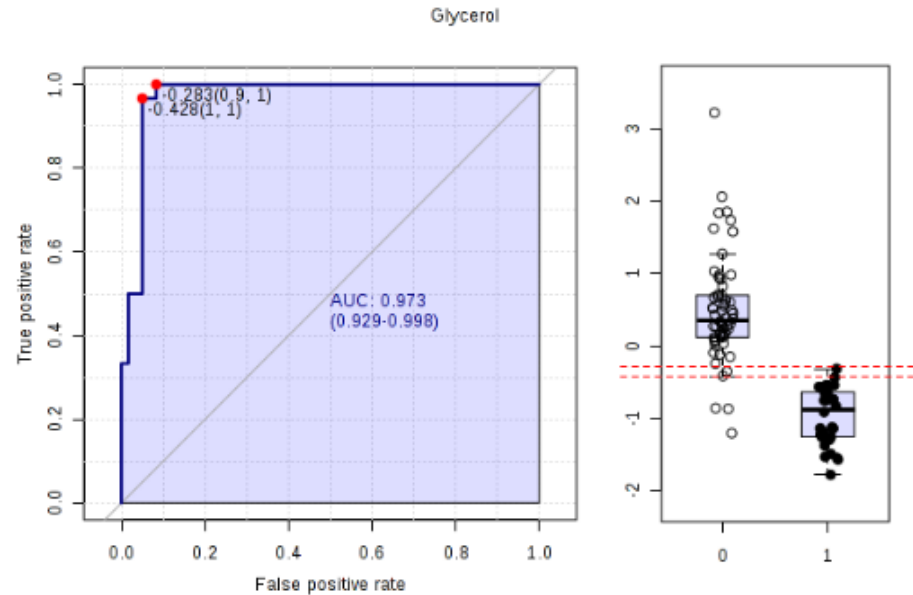
[Multivariate ROC curve based exploratory analysis \(Explorer\)](#)

Perform automated important feature identification and performance evaluation. ROC curve analyses are performed based on three multivariate algorithms - support vector machines (SVM), partial least squares discriminant analysis (PLS-DA), and random forests.

[ROC curve based model evaluation \(Tester\)](#)

Users can manually select any combination of features to create biomarker models using any of the three algorithms mentioned above. The module also allows users to **hold out** a subset of samples for extra validation purpose, as well as to **predict class for new samples** (samples without class labels).

# ROC curve analysis





In **Details** you  
get the cut-off  
point,  
Sensitivity and  
Specificity

Name ↕	AUC ↕	T-tests ↕	Log2 FC ↕	ROC Curve	Details
Glycerol	0.97111	1.6955E-16	1.3795	<a href="#">View</a>	→
Acetate	0.83278	2.9672E-6	1.0714	<a href="#">View</a>	→
Trimethylamine	0.77944	1.9543E-6	-0.65174	<a href="#">View</a>	→
Pyruvate	0.75111	1.143E-4	-0.45701	<a href="#">View</a>	→
Choline	0.74861	0.0029303	1.2869	<a href="#">View</a>	→
Propylene glycol	0.72583	7.8583E-5	-0.86155	<a href="#">View</a>	→
Alanine	0.705	2.6036E-4	0.44779	<a href="#">View</a>	→
Arginine	0.69639	1.6421E-4	0.44837	<a href="#">View</a>	→
Isoleucine	0.69167	1.2424E-4	0.71957	<a href="#">View</a>	→



# Select a Module (Power Analysis)

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

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.

➤ Power Analysis

This module uses pilot data to calculate the minimum number of samples required to detect a statistically significant difference between two populations with a given degree of confidence (called Power Analysis).

➤ 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.

➤ Biomarker Analysis

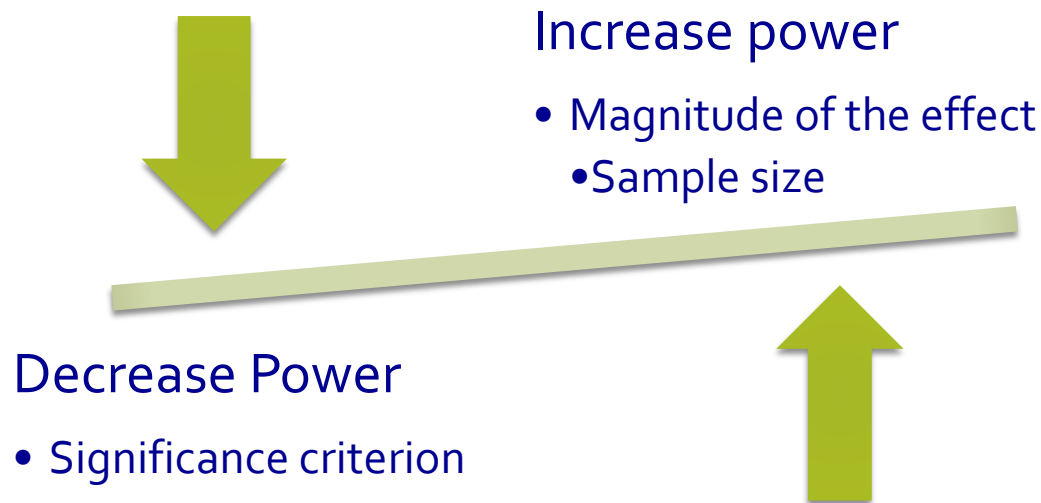
This module performs various ROC curve based biomarker analyses for a single or multiple biomarkers. It also allows users to manually specify biomarker models as well as new sample prediction.

# Statistical Power

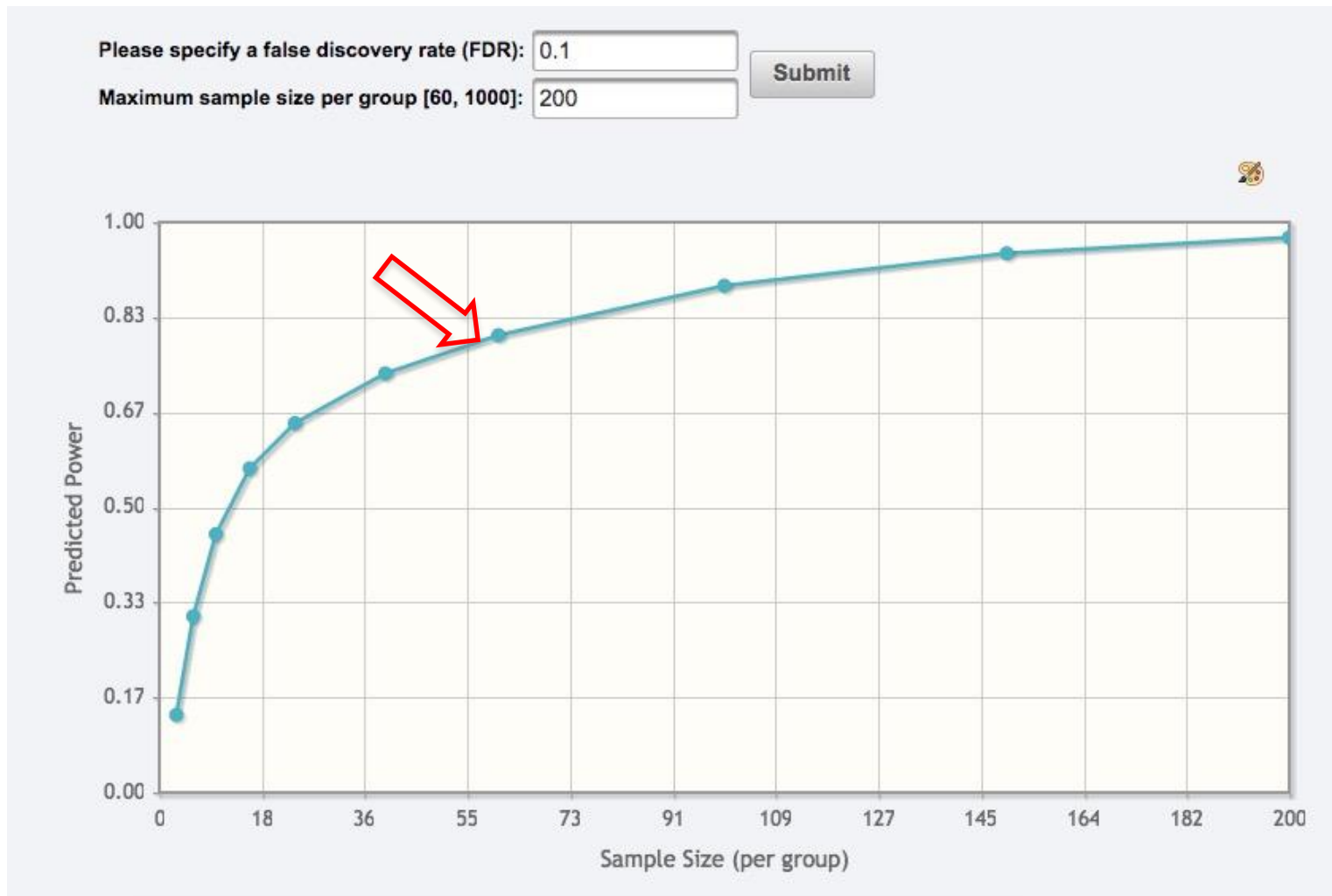
- Statistical power is the ability of a test to detect an effect, if the effect actually exists
  - A power of 0.8 in a clinical trial means that the study has a 80% chance of ending up with a statistically significant treatment effect if there really was an important difference between treatments.
- To answer research questions:
  - How powerful is my study?
  - How many samples do I need to have for what I want to get from the study?

# Statistical Power (cont.)

- The statistical power of a test depends:
  1. Sample size,
  2. Significance criterion (alpha)
  3. Magnitude of the effect



# Power vs. Sample size curve

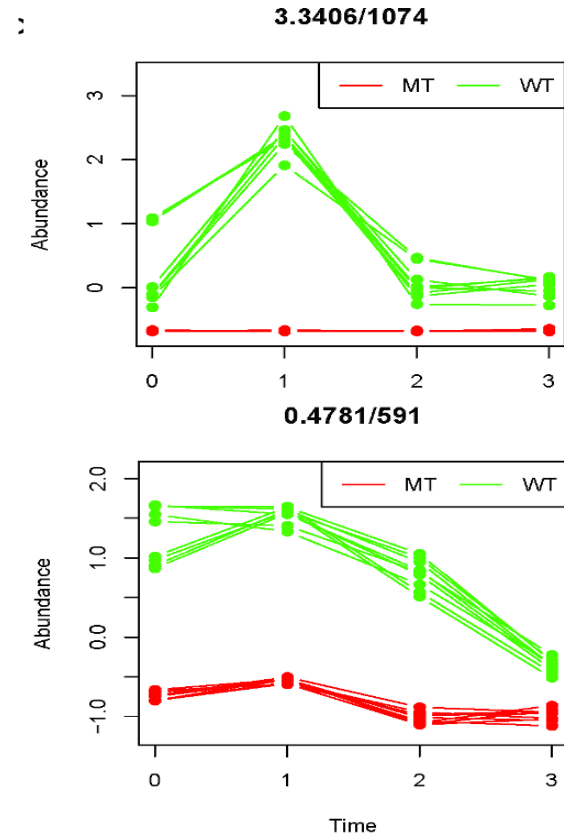
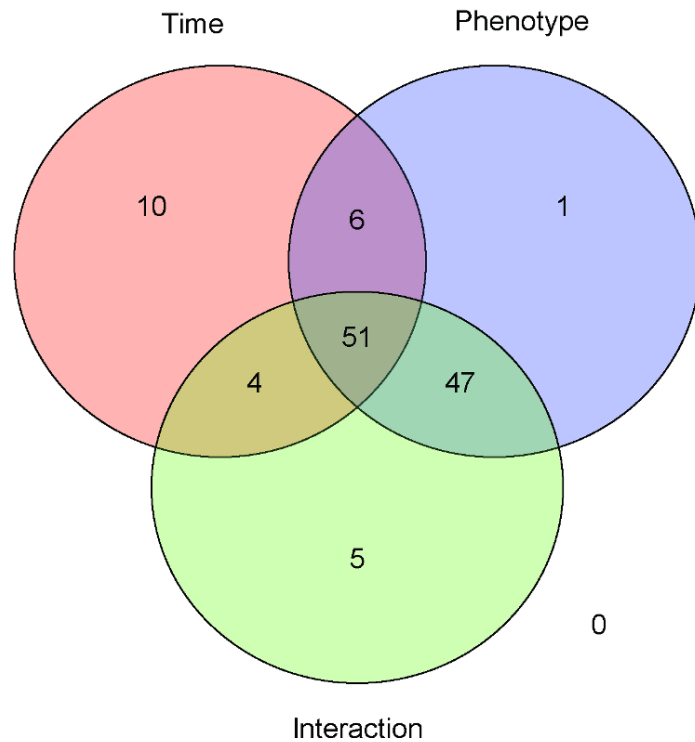


**At least 60 samples/group will needed to get a power of 0.8** <sup>92</sup>

# Not Everything Was Covered

- Clustering Methods (K-means, SOM)
- Classification Methods (SVM, Random Forests)
- SAM and EDAM – (used for identification of differentially expressed genes in microarray experiments)
- Time-series data analysis & Two factor data analysis
- Integrative pathway analysis (gene and metabolite)
- Batch effect correction - each batch contains roughly the same numbers of class labels (i.e. control vs. disease); It can not adjust batch effect if the control and disease are in different batches. Quality control samples should be named as QC. MetaboAnalyst will detect and align all the tables
- Lipidomics tool - Calculate the upper limit and most probable concentration from lipidomics data

# Time Series Analysis in MetaboAnalyst



# Integrative Pathway Analysis

**Gene List**

Gene list with optional fold changes

ID

Type:

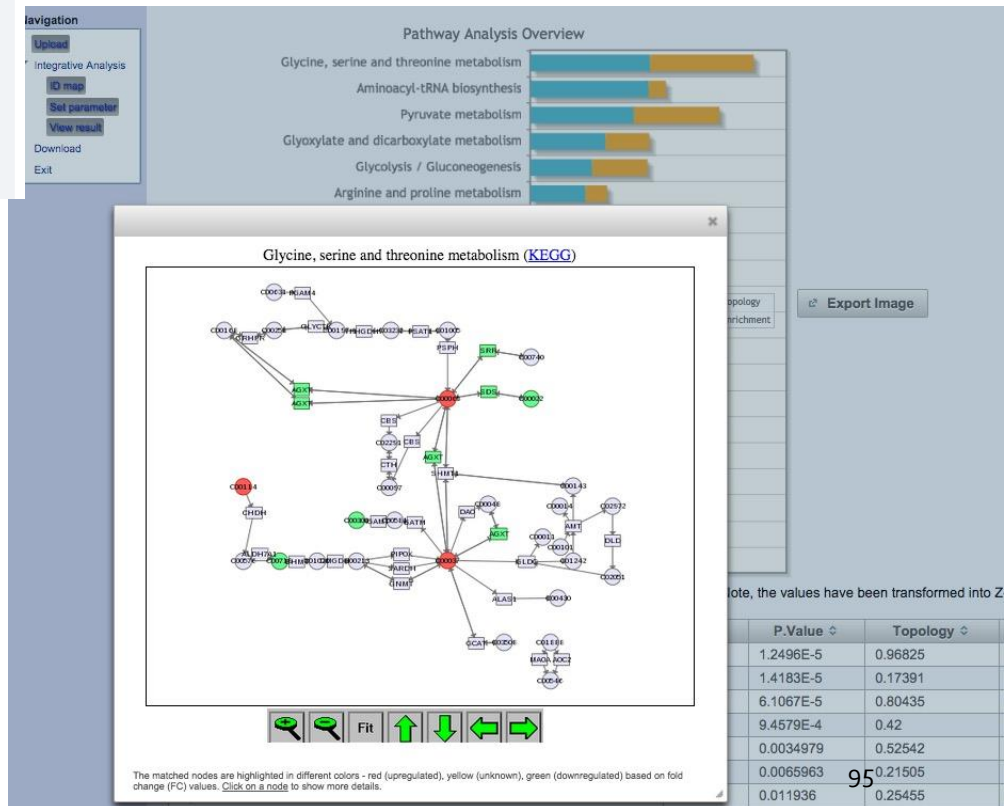
**Metabolite List**

Compound list with optional fold changes

ID Type:

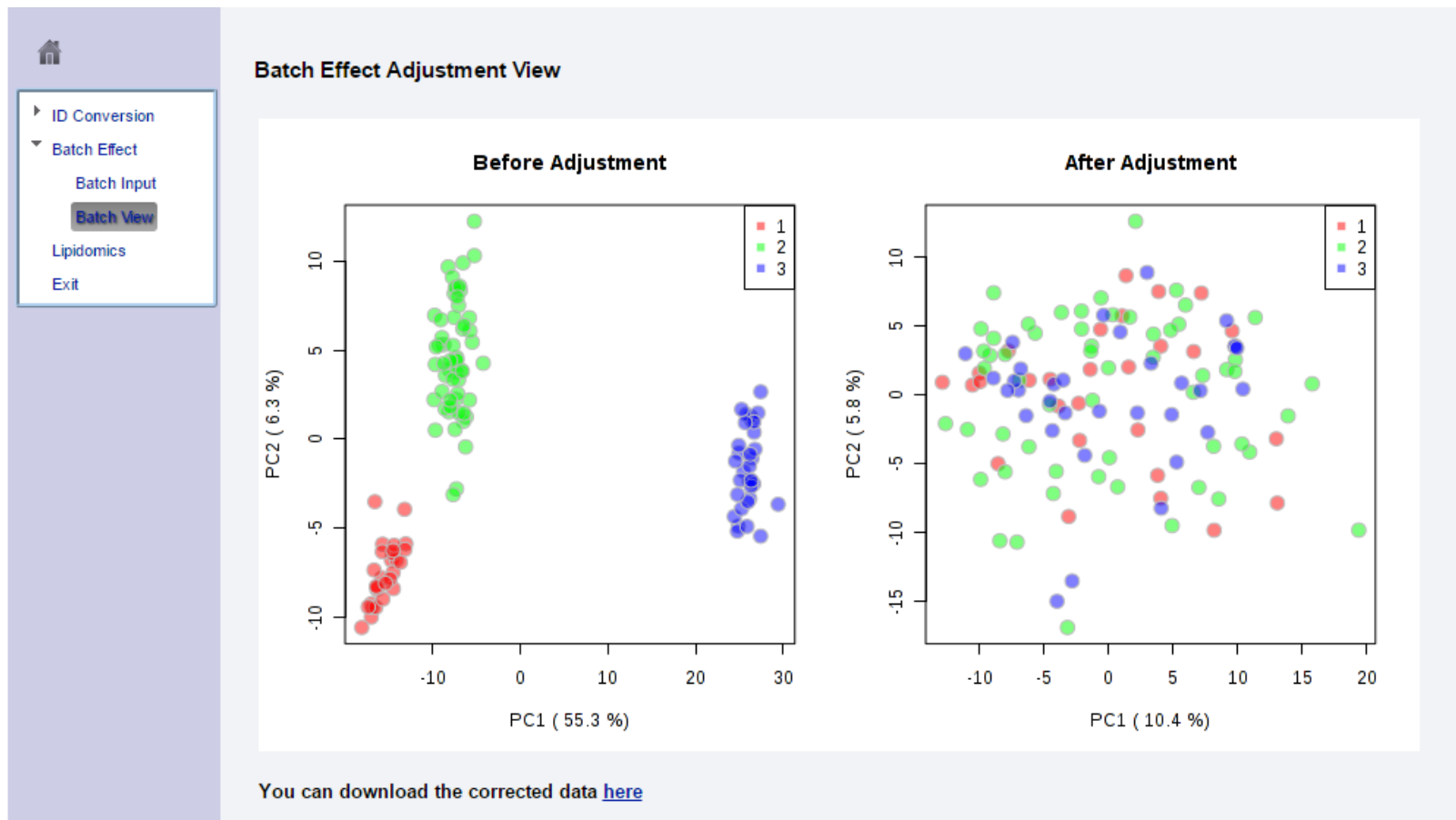
Specify organism:

☐ Use our example data





# Batch Adjustment



# Lipidomics

Calculate the upper limit and most probable concentration from lipomics data:

Upload your lipid concentration file (see below for instructions):

No file chosen

☒ Isomers merged as [iso #] ☐ Isomers listed individually

## File Format

- The file must be in comma separated format (.csv);
- The first column is the sample name;
- The second column is lipid class names. Currently, only the following lipid classes are supported:
  - DG: Diacylglycerol
  - PC: Phosphatidylcholine
  - PE: Phosphatidylethanolamine
  - TG: Triacylglycerol
- The first row are free fatty acid names;
- No missing values are allowed (please replace by 0);

A screenshot of sample data is shown below:

Sample ID	Lipid Class	14:0	15:0	16:0	18:0	20:0	22:0	24:0	14:1(9Z)	16:1(9Z)
S-FB	DG	2.21	0.83	15.75	8.3	0.21	0.2	0.22	0.59	1.65
P-2007-07-06	DG	5.48	1.54	16.74	9.19	0.38	0.55	0.49	0.61	0.97
P-2007-07-09	DG	4.26	1.12	16.45	9.89	0.45	0.64	0.47	0.36	1.35
S-FB	PC	18.87	11.31	1290.74	538.46	1.92	0.41	0.47	1.21	29.96
P-2007-07-06	PC	12.57	10.14	860.77	432.63	1.32	0.33	0.95	0.27	11.15
P-2007-07-09	PC	19.24	10.27	1355.83	585.11	2.69	0.48	0.65	0.75	34.09
S-FB	PE	2.03	0.67	37.35	81.39	0.44	0.44	3.29	0.41	2.07
P-2007-07-06	PE	6.79	3.34	42.03	55.62	1.14	0.83	1.47	0.92	0.91
P-2007-07-09	PE	4.65	1.58	77.05	102.06	0.62	0.69	0.47	0.35	3.41
S-FB	TG	35.71	6.26	319.44	71.63	1.52	0.56	0.81	5.05	40.52
P-2007-07-06	TG	56.4	8.53	311.11	80.71	2.32	0.82	1.54	5.09	28.51
P-2007-07-09	TG	65.76	9.43	479.55	107.13	2.41	1.18	1.77	8.73	69.54

The complete sample data can be downloaded [here](#)

