Data Analysis & Biomarker Discovery

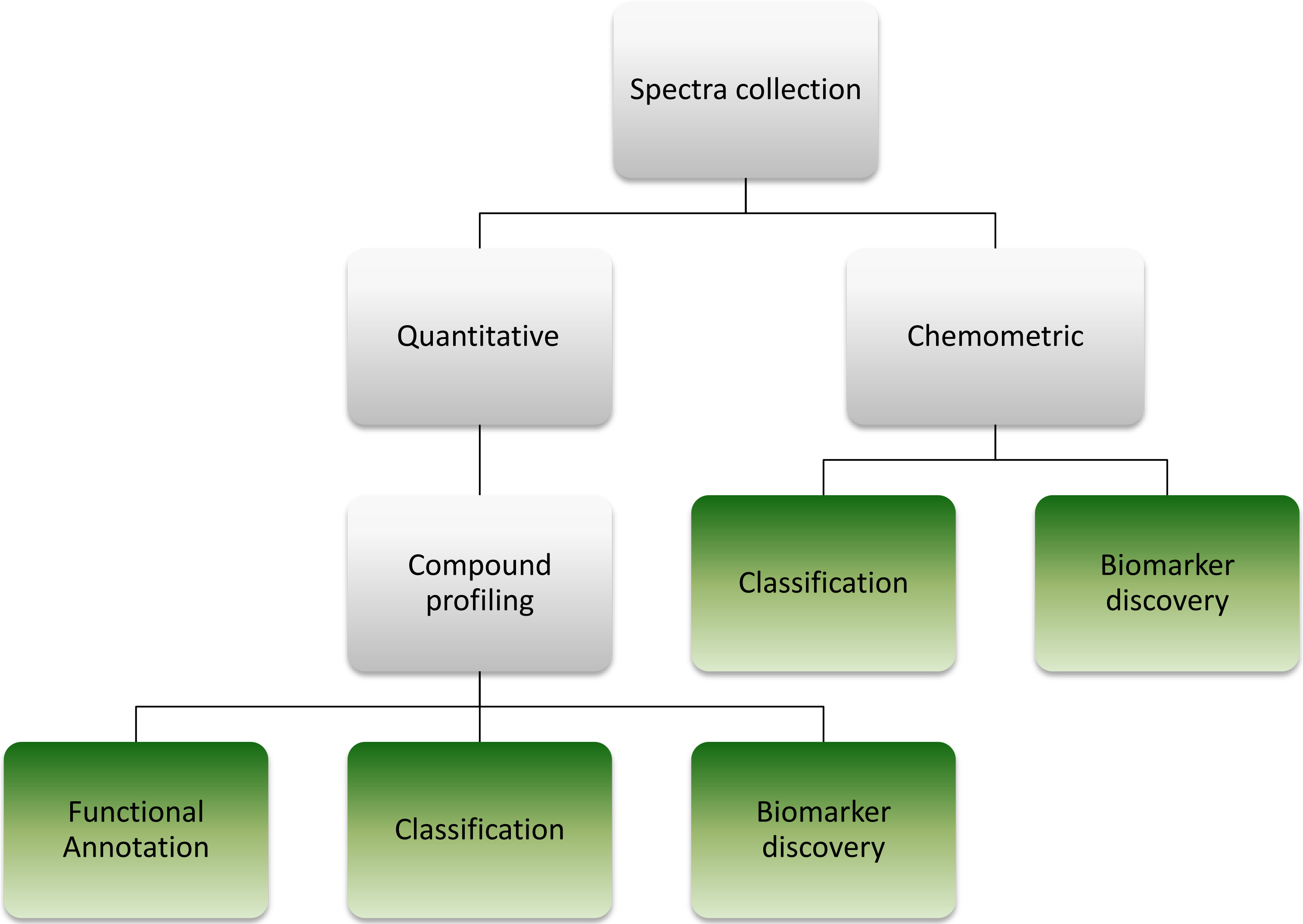
# MetaboAnalyst 2.0 & ROCCET

Jianguo Xia, PhD
University of Alberta, Canada

# Outline

- Introduction (updates) of two free web application
  - MetaboAnalyst 2.0
  - ROCCET
- Background & basic concepts
- Screenshot tutorials
- Live demo (if we have time)

# Metabolomic Data Analysis



Spectra collection

Quantitative

Chemometric

Compound profiling

Classification

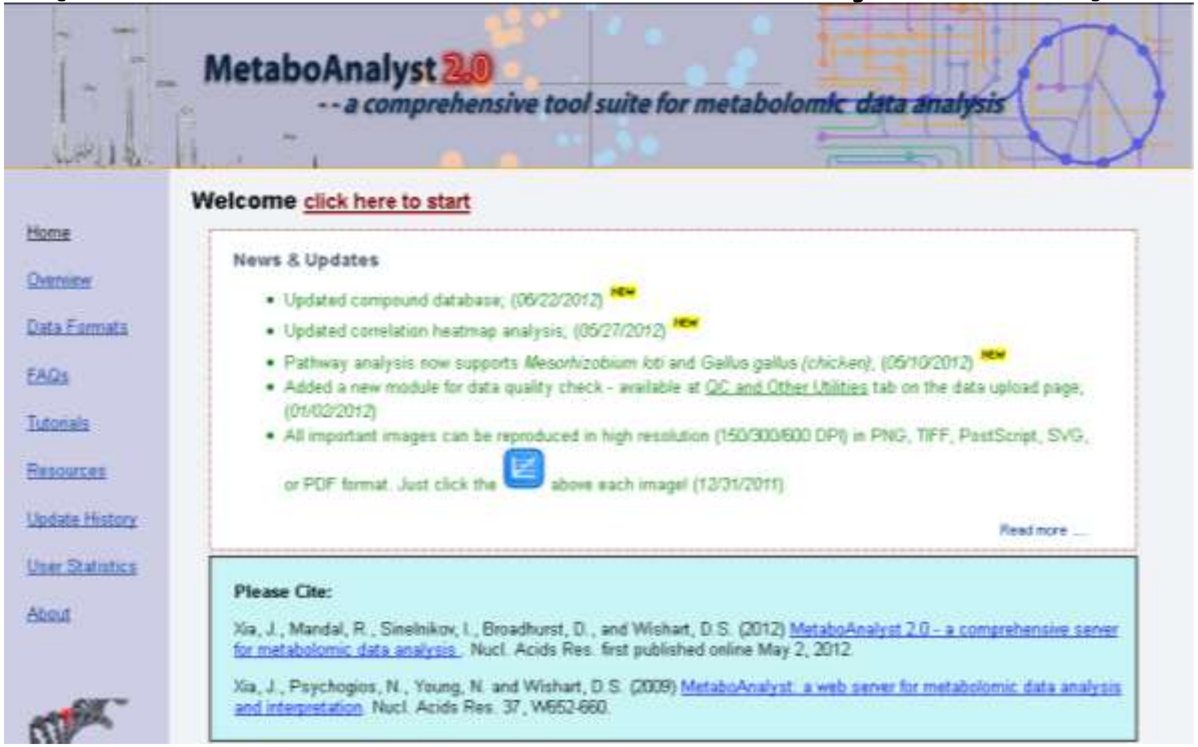
Biomarker discovery

Functional Annotation

Classification

Biomarker discovery

# MetaboAnalyst (www.metaboanalyst.ca)



**Welcome** click here to start

**News & Updates**

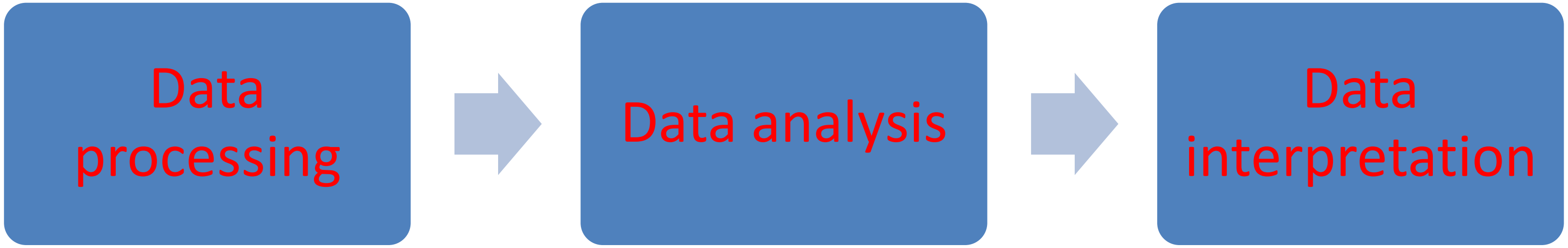
- Updated compound database; (06/22/2012) NEW
- Updated correlation heatmap analysis; (05/27/2012) NEW
- Pathway analysis now supports *Mesorhizobium loti* and Gallus gallus *(chicken)*; (05/10/2012) NEW
- Added a new module for data quality check - available at QC and Other Utilities tab on the data upload page; (01/02/2012)
- All important images can be reproduced in high resolution (150/300/600 DPI) in PNG, TIFF, PostScript, SVG, or PDF format. Just click the above each image! (12/31/2011)

Read more ...

**Please Cite:**

Xia, J., Mandal, R., Sinelnikov, I., Broadhurst, D., and Wishart, D.S. (2012) MetaboAnalyst 2.0 - a comprehensive server for metabolomic data analysis. Nucl. Acids Res. first published online May 2, 2012.

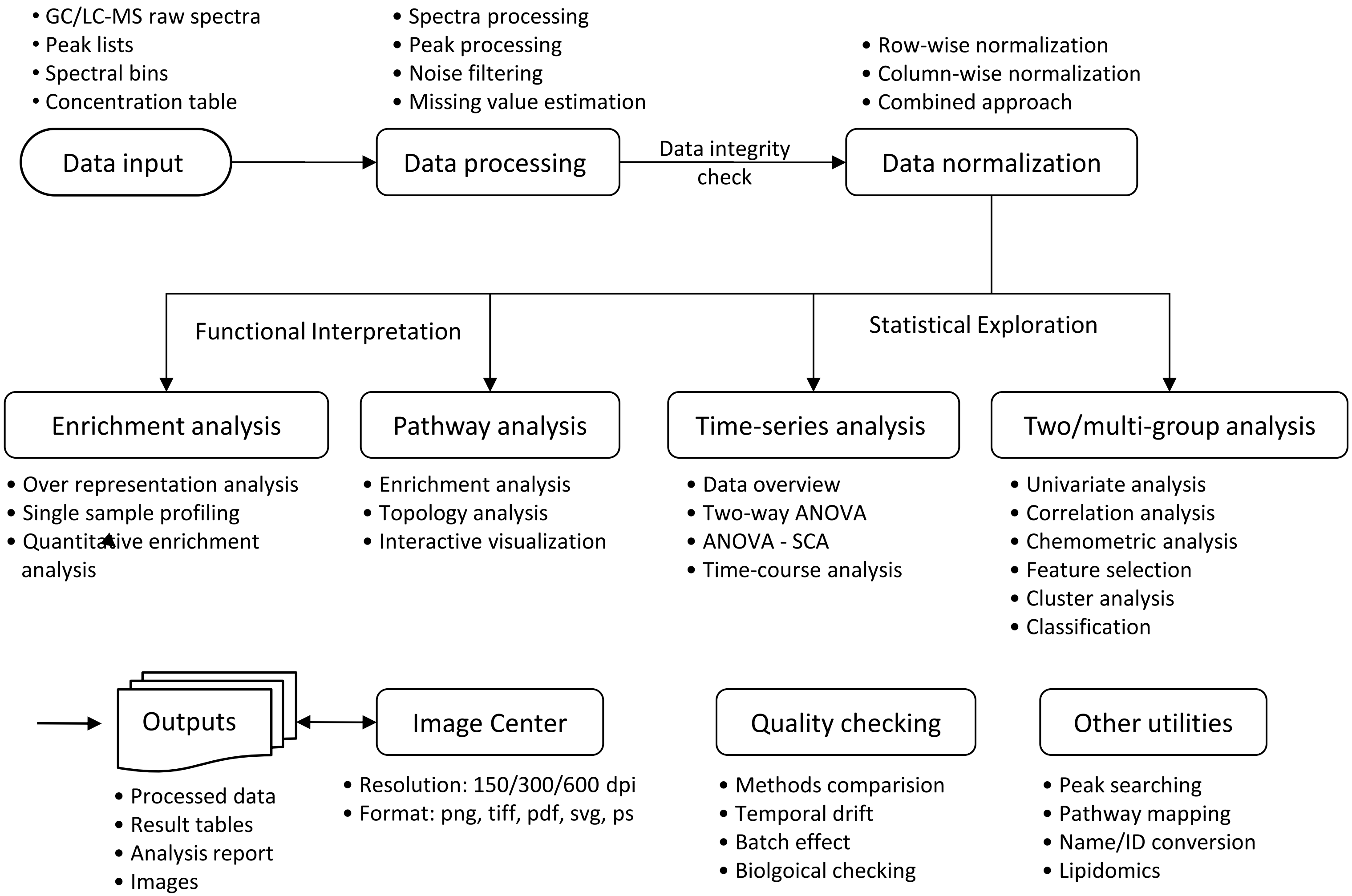
Xia, J., Psychogios, N., Young, N. and Wishart, D.S. (2009) MetaboAnalyst: a web server for metabolomic data analysis and interpretation. Nucl. Acids Res. 37, W652-660.



Data processing

Data analysis

Data interpretation



- GC/LC-MS raw spectra
- Peak lists
- Spectral bins
- Concentration table

- Spectra processing
- Peak processing
- Noise filtering
- Missing value estimation

- Row-wise normalization
- Column-wise normalization
- Combined approach

- Over representation analysis
- Single sample profiling
- Quantitative enrichment analysis

- Enrichment analysis
- Topology analysis
- Interactive visualization

- Data overview
- Two-way ANOVA
- ANOVA - SCA
- Time-course analysis

- Univariate analysis
- Correlation analysis
- Chemometric analysis
- Feature selection
- Cluster analysis
- Classification

- Processed data
- Result tables
- Analysis report
- Images

- Resolution: 150/300/600 dpi
- Format: png, tiff, pdf, svg, ps

- Methods comparision
- Temporal drift
- Batch effect
- Biolgoical checking

- Peak searching
- Pathway mapping
- Name/ID conversion
- Lipidomics

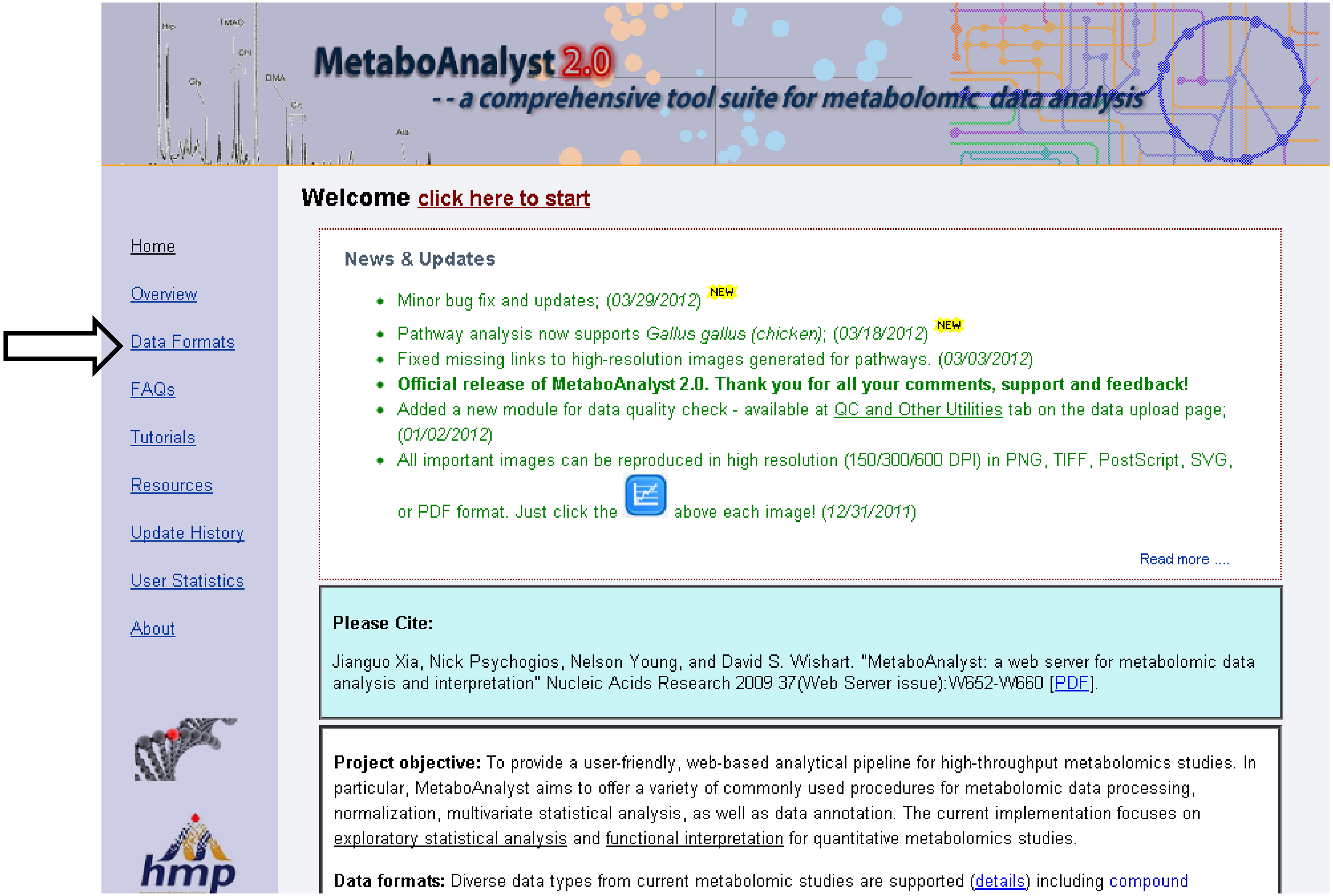
# MetaboAnalyst Overview

- Raw data processing
- Data reduction & statistical analysis
- Functional enrichment analysis
- Metabolic pathway analysis
- Quality control analysis

# Data processing overview

- Supported data formats
  - Concentration tables
  - Peak lists
  - Spectral bins
  - Raw spectra (* not recommended)

# Example Datasets



**Welcome** click here to start

**News & Updates**

- Minor bug fix and updates; *(03/29/2012)* NEW
- Pathway analysis now supports *Gallus gallus (chicken)*; *(03/18/2012)* NEW
- Fixed missing links to high-resolution images generated for pathways. *(03/03/2012)*
- **Official release of MetaboAnalyst 2.0. Thank you for all your comments, support and feedback!**
- Added a new module for data quality check - available at QC and Other Utilities tab on the data upload page; *(01/02/2012)*
- All important images can be reproduced in high resolution (150/300/600 DPI) in PNG, TIFF, PostScript, SVG, or PDF format. Just click the above each image! *(12/31/2011)*

Read more ....

**Please Cite:**

Jianguo Xia, Nick Psychogios, Nelson Young, and David S. Wishart. "MetaboAnalyst: a web server for metabolomic data analysis and interpretation" Nucleic Acids Research 2009 37(Web Server issue):W652-W660 [PDF].

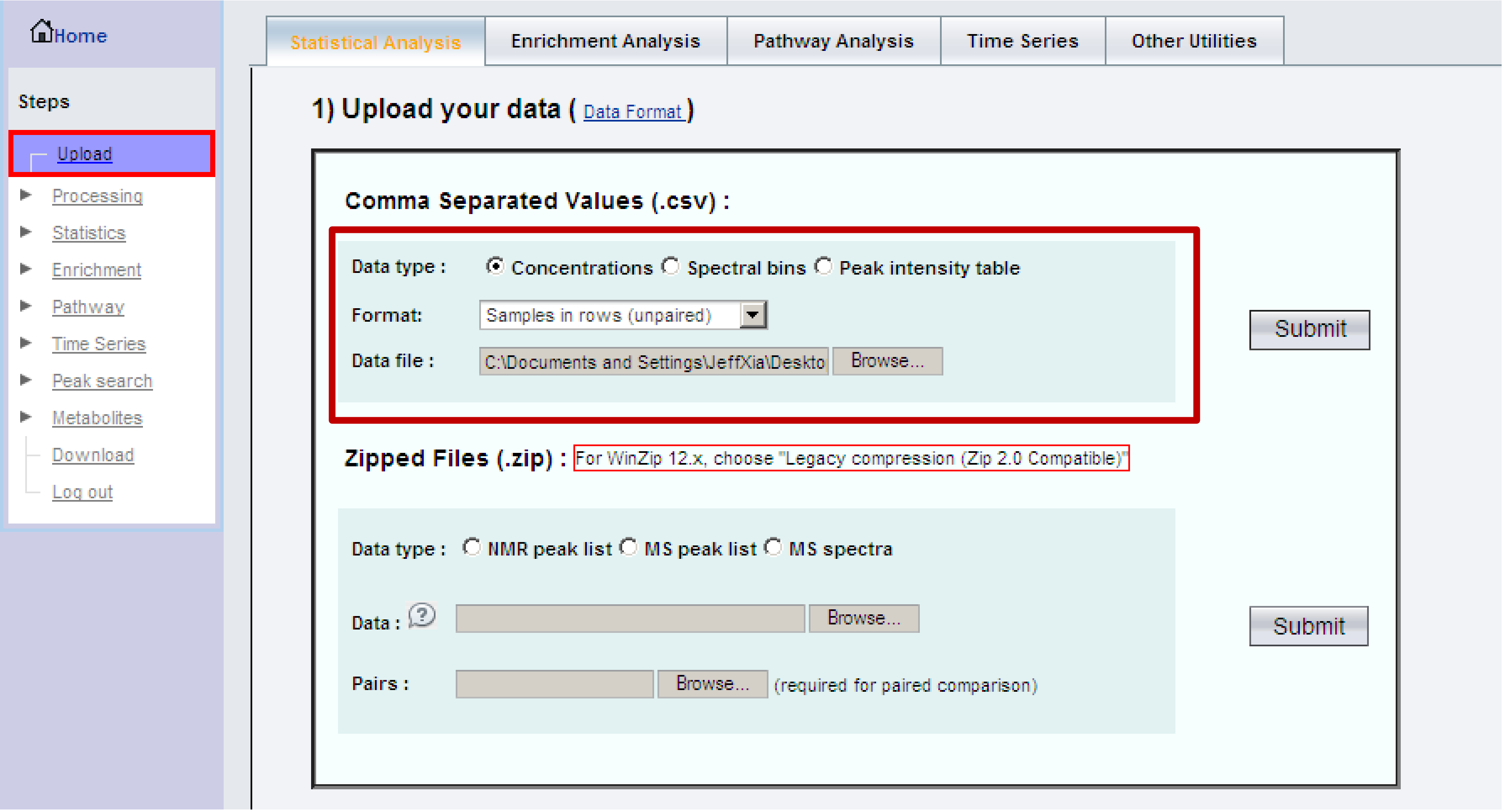
**Project objective:** To provide a user-friendly, web-based analytical pipeline for high-throughput metabolomics studies. In particular, MetaboAnalyst aims to offer a variety of commonly used procedures for metabolomic data processing, normalization, multivariate statistical analysis, as well as data annotation. The current implementation focuses on exploratory statistical analysis and functional interpretation for quantitative metabolomics studies.

**Data formats:** Diverse data types from current metabolomic studies are supported (details) including compound

# Data Processing

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

# Data Upload



1) Upload your data ( Data Format )

Comma Separated Values (.csv) :

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

Format: Samples in rows (unpaired)

Data file : C:\Documents and Settings\JeffXia\Deskto Browse...

Submit

Zipped Files (.zip) : For WinZip 12.x, choose "Legacy compression (Zip 2.0 Compatible)"

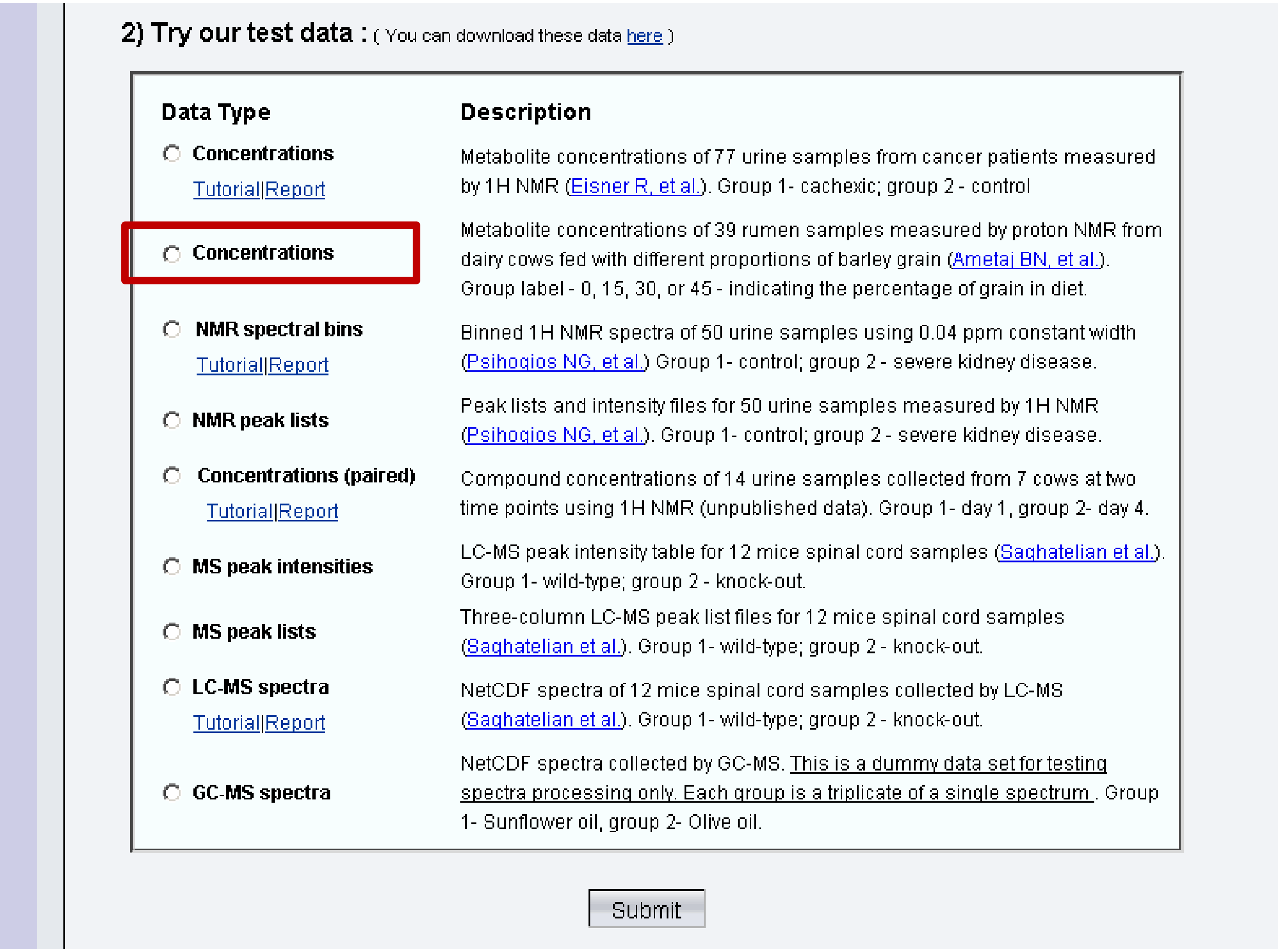
Data type : ○ NMR peak list ○ MS peak list ○ MS spectra

Data : Browse...

Pairs : Browse... (required for paired comparison)

Submit

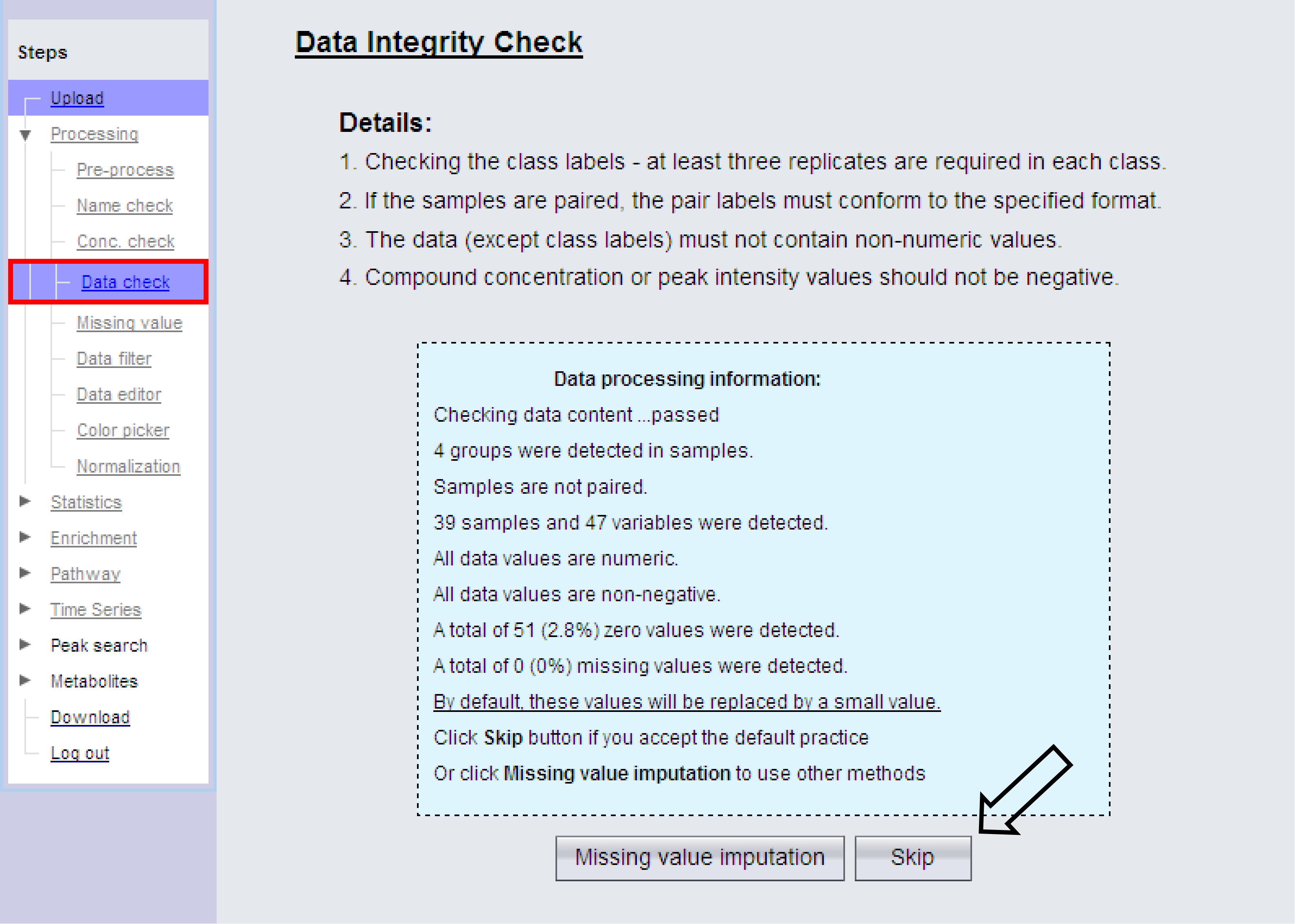
# Alternatively …



2) Try our test data : ( You can download these data here )

| Data Type | Description |
|---|---|
| ○ Concentrations<br>Tutorial\|Report | Metabolite concentrations of 77 urine samples from cancer patients measured by 1H NMR (Eisner R, et al.). Group 1- cachexic; group 2 - control |
| ○ Concentrations | Metabolite concentrations of 39 rumen samples measured by proton NMR from dairy cows fed with different proportions of barley grain (Ametaj BN, et al.). Group label - 0, 15, 30, or 45 - indicating the percentage of grain in diet. |
| ○ NMR spectral bins<br>Tutorial\|Report | Binned 1H NMR spectra of 50 urine samples using 0.04 ppm constant width (Psihogios NG, et al.) Group 1- control; group 2 - severe kidney disease. |
| ○ NMR peak lists | Peak lists and intensity files for 50 urine samples measured by 1H NMR (Psihogios NG, et al.). Group 1- control; group 2 - severe kidney disease. |
| ○ Concentrations (paired)<br>Tutorial\|Report | Compound concentrations of 14 urine samples collected from 7 cows at two time points using 1H NMR (unpublished data). Group 1- day 1, group 2- day 4. |
| ○ MS peak intensities | LC-MS peak intensity table for 12 mice spinal cord samples (Saghatelian et al.). Group 1- wild-type; group 2 - knock-out. |
| ○ MS peak lists | Three-column LC-MS peak list files for 12 mice spinal cord samples (Saghatelian et al.). Group 1- wild-type; group 2 - knock-out. |
| ○ LC-MS spectra<br>Tutorial\|Report | NetCDF spectra of 12 mice spinal cord samples collected by LC-MS (Saghatelian et al.). Group 1- wild-type; group 2 - knock-out. |
| ○ GC-MS spectra | NetCDF spectra collected by GC-MS. This is a dummy data set for testing spectra processing only. Each group is a triplicate of a single spectrum . Group 1- Sunflower oil, group 2- Olive oil. |

# Data Integrity Check



## Data Integrity Check

**Details:**

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. Compound concentration or peak intensity values should not be negative.

**Data processing information:**

Checking data content ...passed

4 groups were detected in samples.

Samples are not paired.

39 samples and 47 variables were detected.

All data values are numeric.

All data values are non-negative.

A total of 51 (2.8%) zero values were detected.

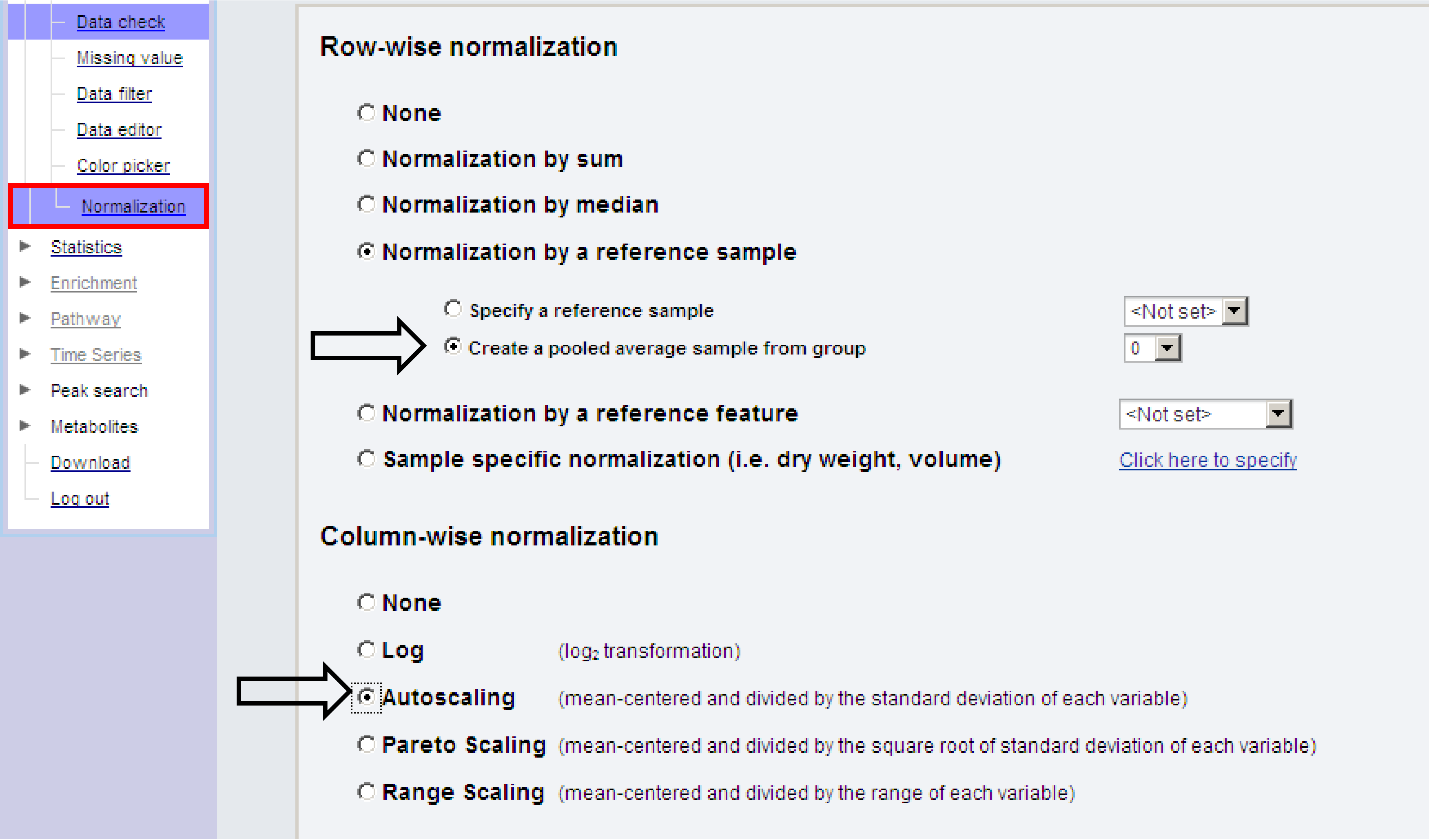
A total of 0 (0%) missing values were detected.

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

# Data Normalization



**Row-wise normalization**

○ None

○ Normalization by sum

○ Normalization by median

● Normalization by a reference sample

○ Specify a reference sample <Not set>

● Create a pooled average sample from group 0

○ Normalization by a reference feature <Not set>

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

**Column-wise normalization**

○ None

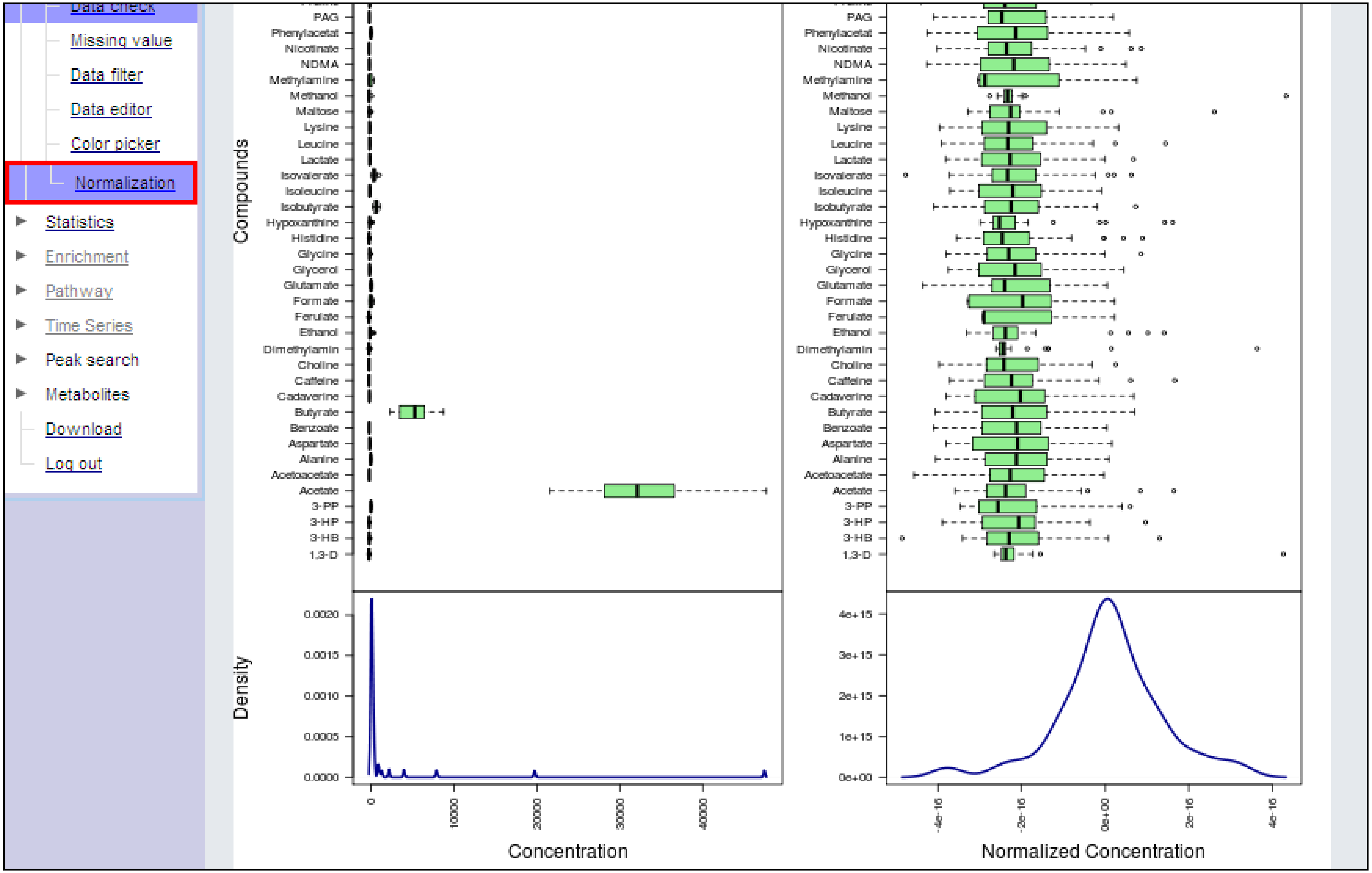
○ Log (log$_2$ transformation)

● Autoscaling (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)

# Normalization Result

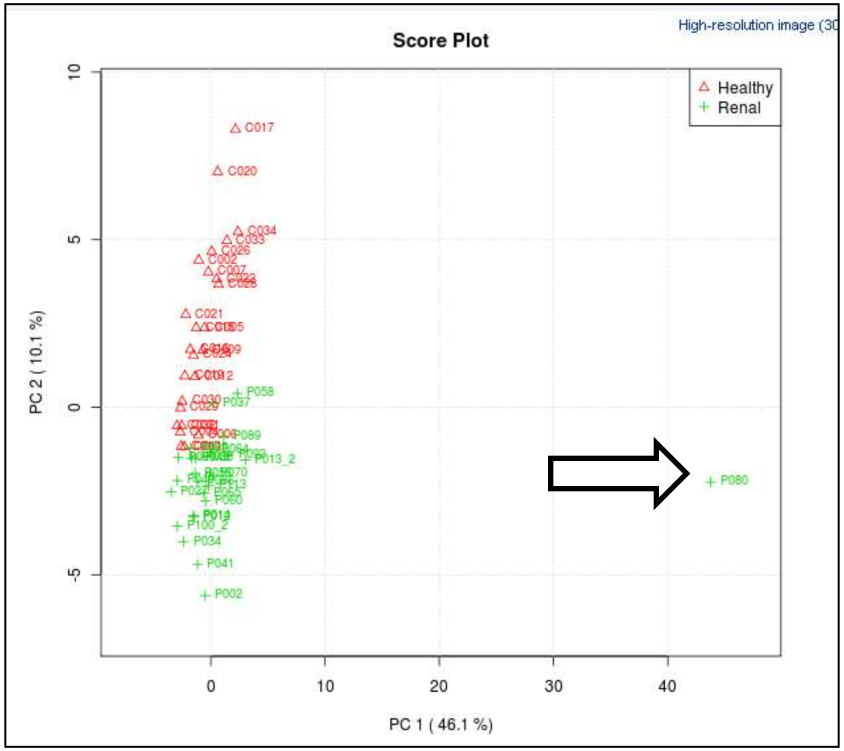


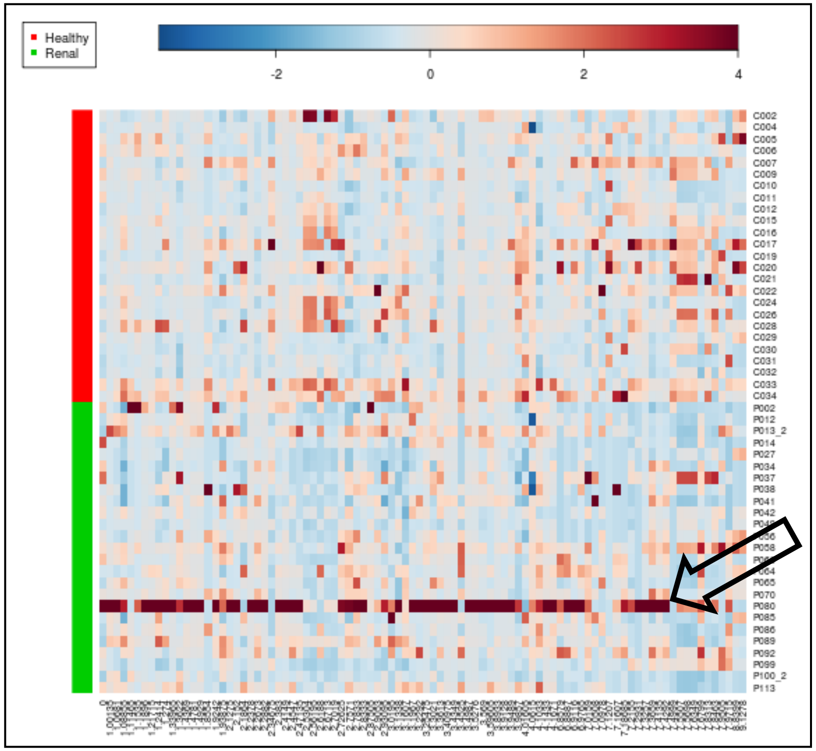
# Quality Control

- Dealing with outliers
  - Detected mainly by visual inspection
  - May be corrected by normalization
  - May be excluded
- Dealing with missing values
- Noise reduction

# Visual Inspection

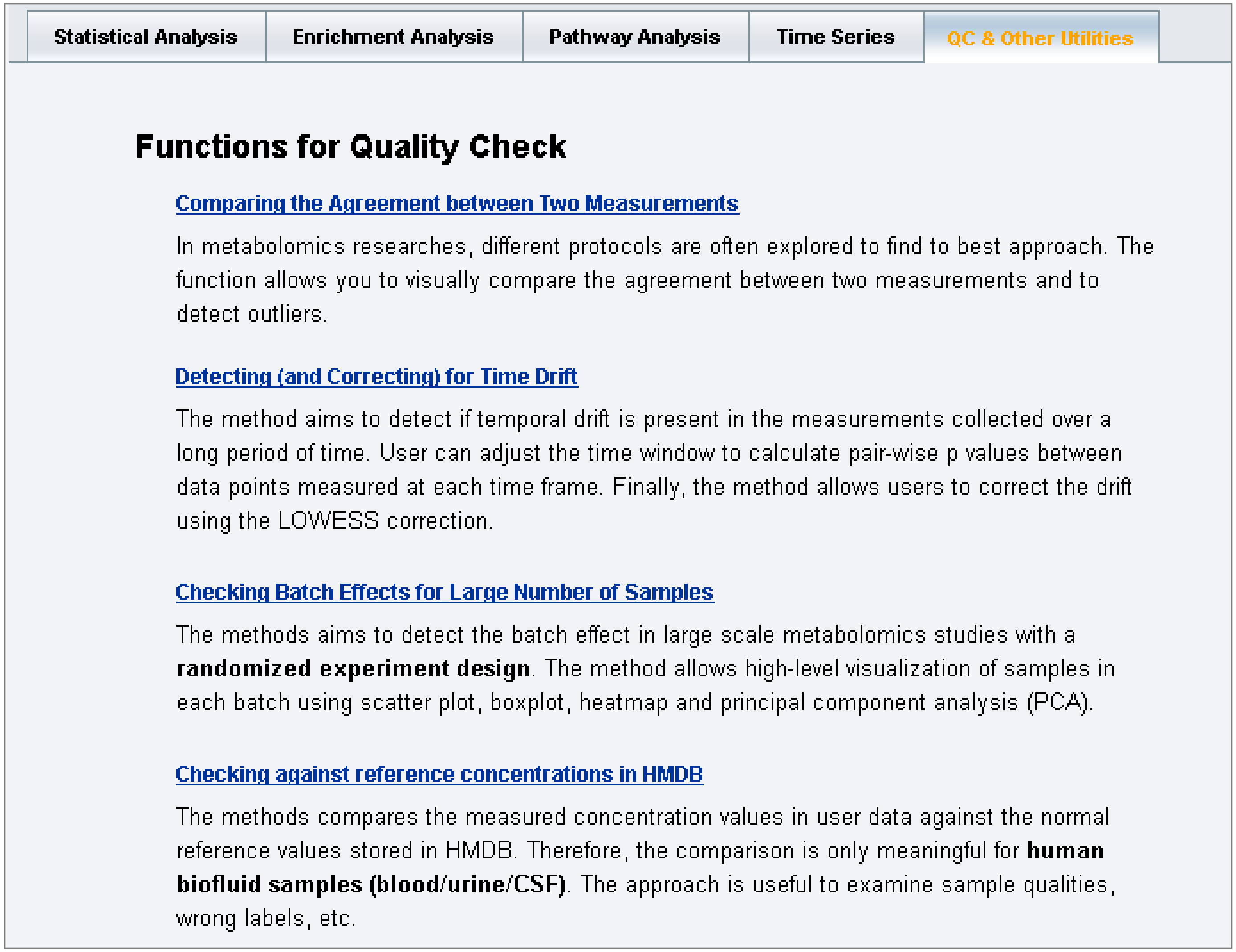
- What does an outlier look like?





Finding outliers via PCA

Finding outliers via Heatmap



## Functions for Quality Check

Comparing the Agreement between Two Measurements

In metabolomics researches, different protocols are often explored to find to best approach. The function allows you to visually compare the agreement between two measurements and to detect outliers.

Detecting (and Correcting) for Time Drift

The method aims to detect if temporal drift is present in the measurements collected over a long period of time. User can adjust the time window to calculate pair-wise p values between data points measured at each time frame. Finally, the method allows users to correct the drift using the LOWESS correction.

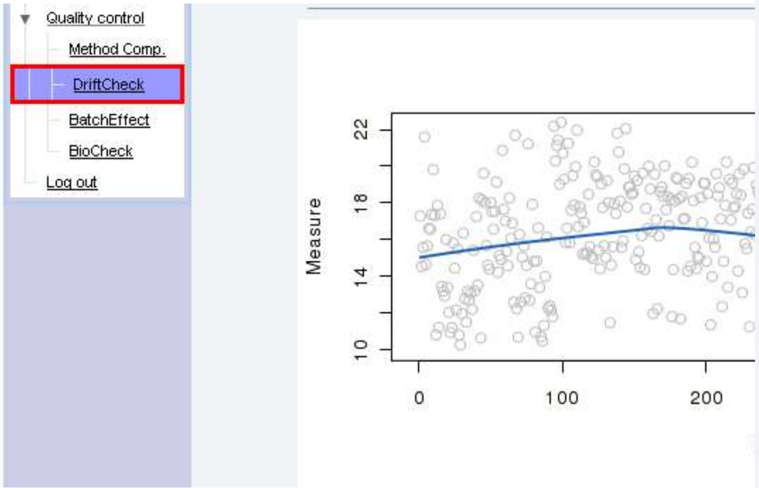
Checking Batch Effects for Large Number of Samples

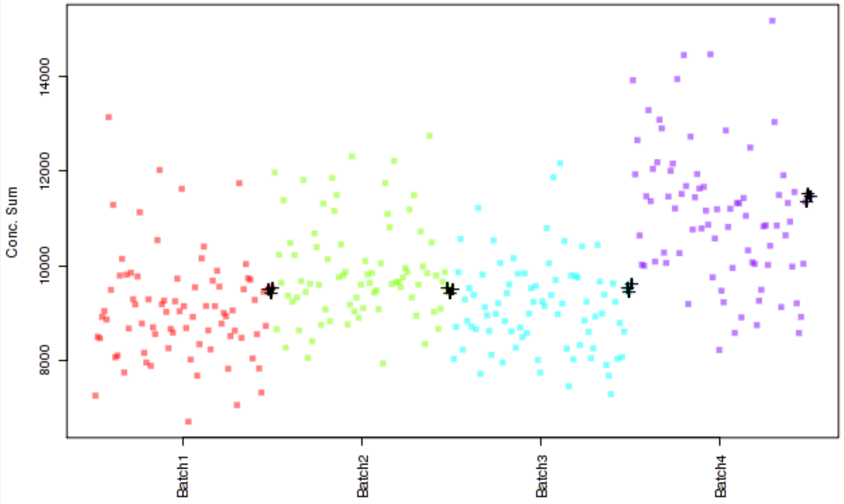
The methods aims to detect the batch effect in large scale metabolomics studies with a **randomized experiment design**. The method allows high-level visualization of samples in each batch using scatter plot, boxplot, heatmap and principal component analysis (PCA).

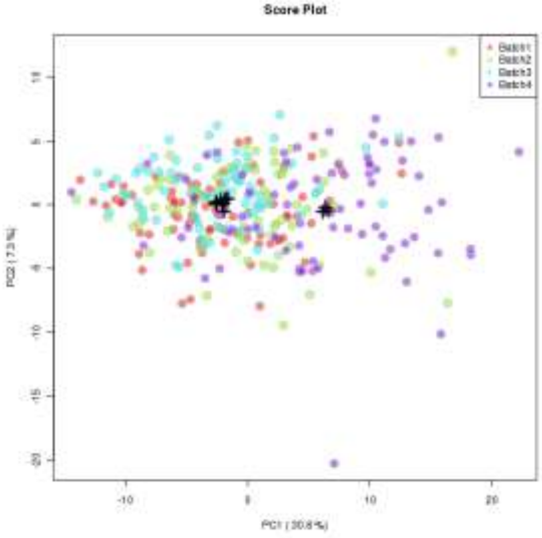
Checking against reference concentrations in HMDB

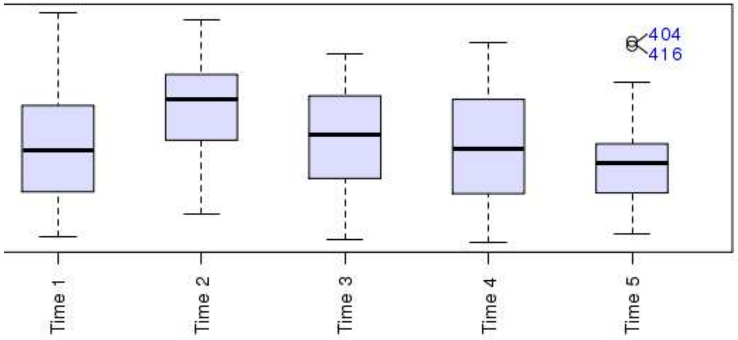
The methods compares the measured concentration values in user data against the normal reference values stored in HMDB. Therefore, the comparison is only meaningful for **human biofluid samples (blood/urine/CSF)**. The approach is useful to examine sample qualities, wrong labels, etc.

# Quality Check Module

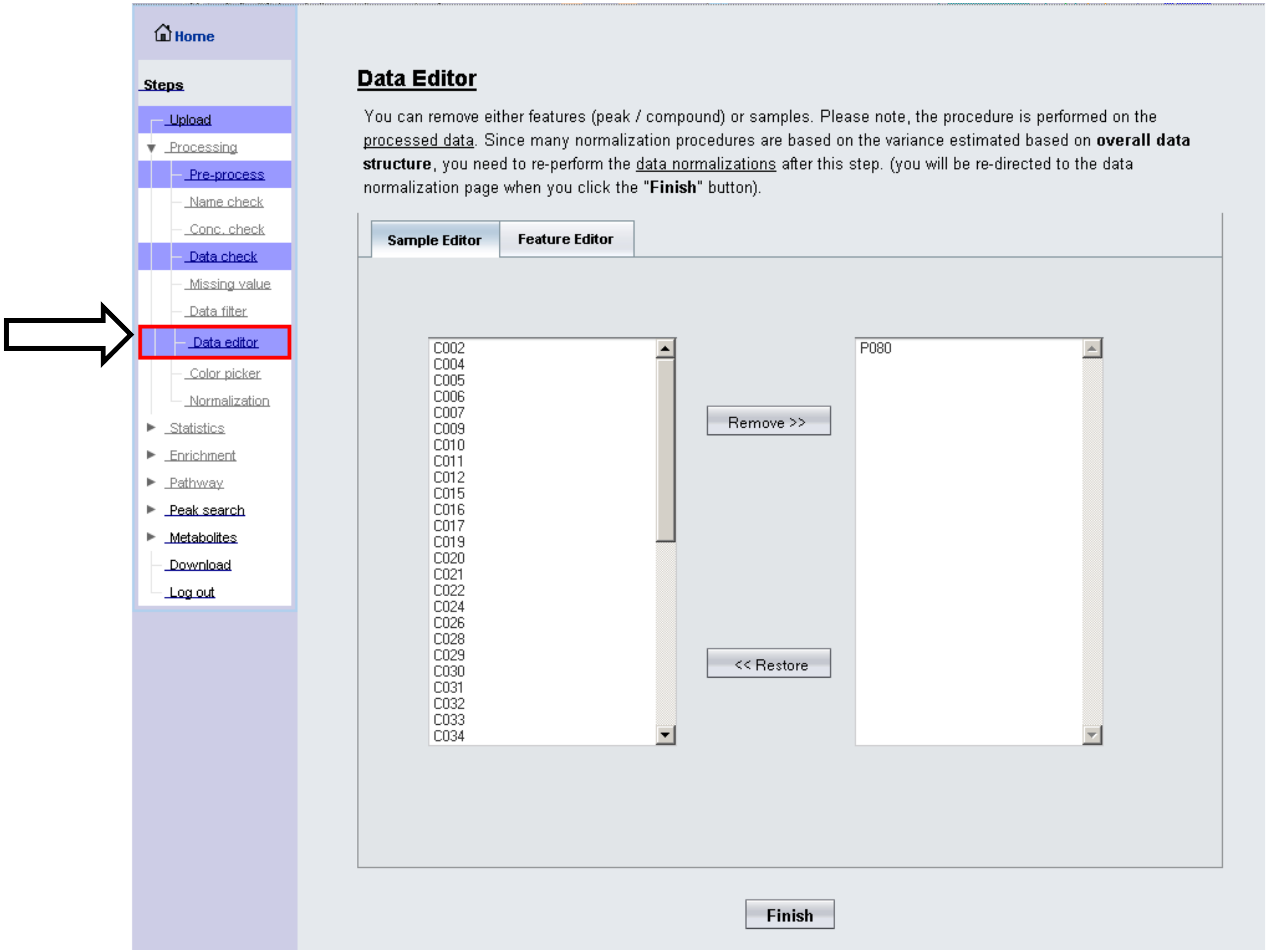








# Outlier Removal



## Data Editor

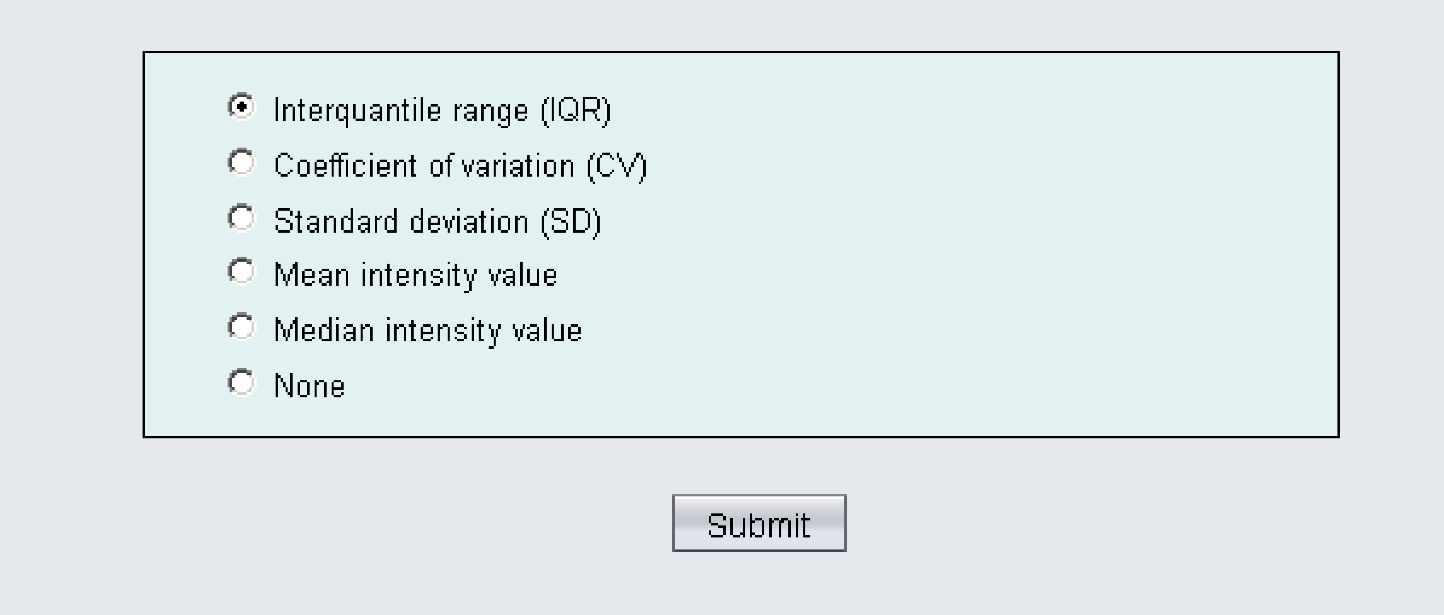
You can remove either features (peak / compound) or samples. Please note, the procedure is performed on the processed data. Since many normalization procedures are based on the variance estimated based on **overall data structure**, you need to re-perform the data normalizations after this step. (you will be re-directed to the data normalization page when you click the "**Finish**" button).

C002
C004
C005
C006
C007
C009
C010
C011
C012
C015
C016
C017
C019
C020
C021
C022
C024
C026
C028
C029
C030
C031
C032
C033
C034

P080

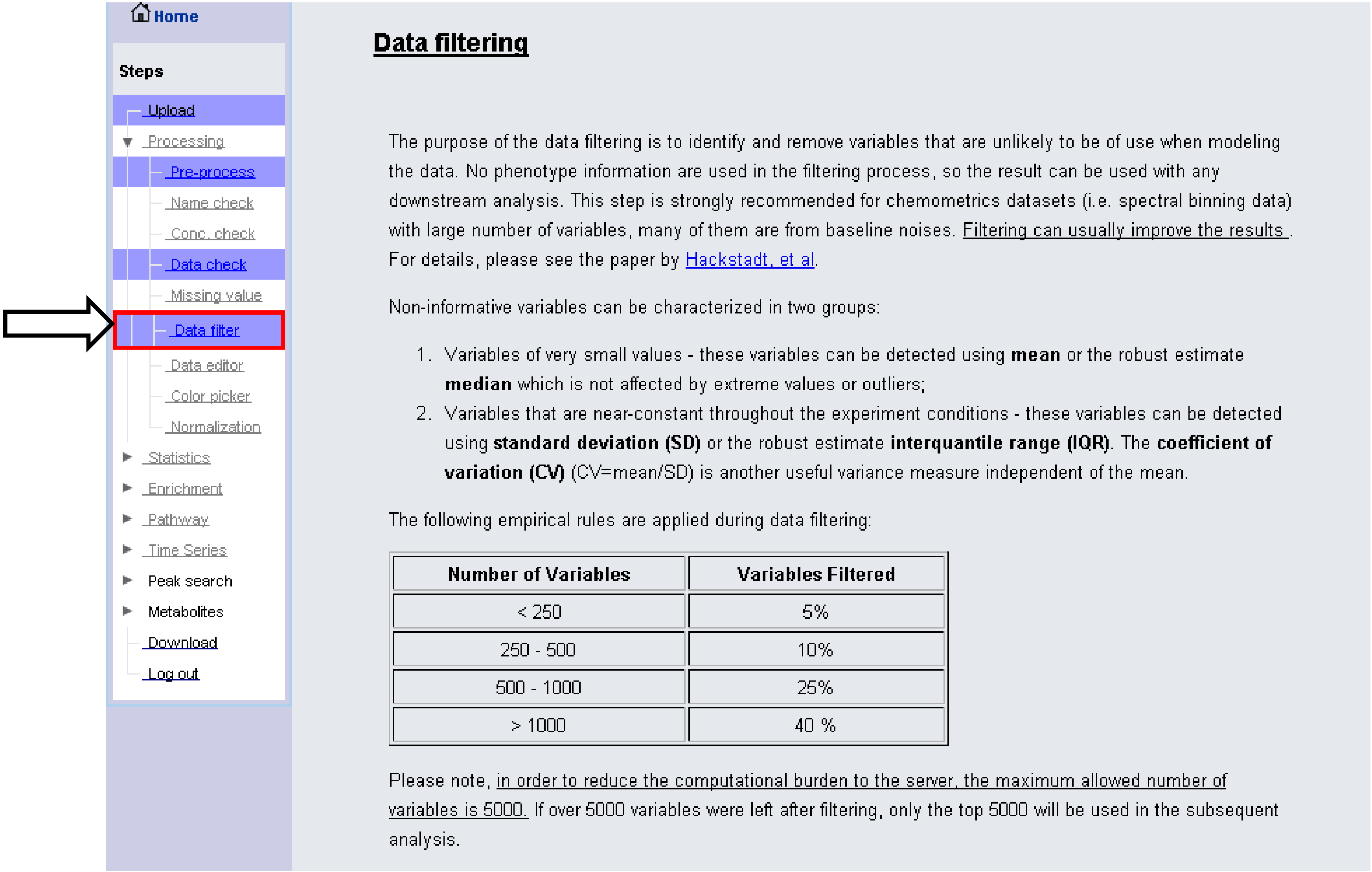
# Data Filtering

- Characteristics of noise & uninformative features
  - Low intensities
  - Low variances (default)



● Interquantile range (IQR)
○ Coefficient of variation (CV)
○ Standard deviation (SD)
○ Mean intensity value
○ Median intensity value
○ None

# Noise Reduction



## Data filtering

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

Non-informative variables can be characterized in two groups:

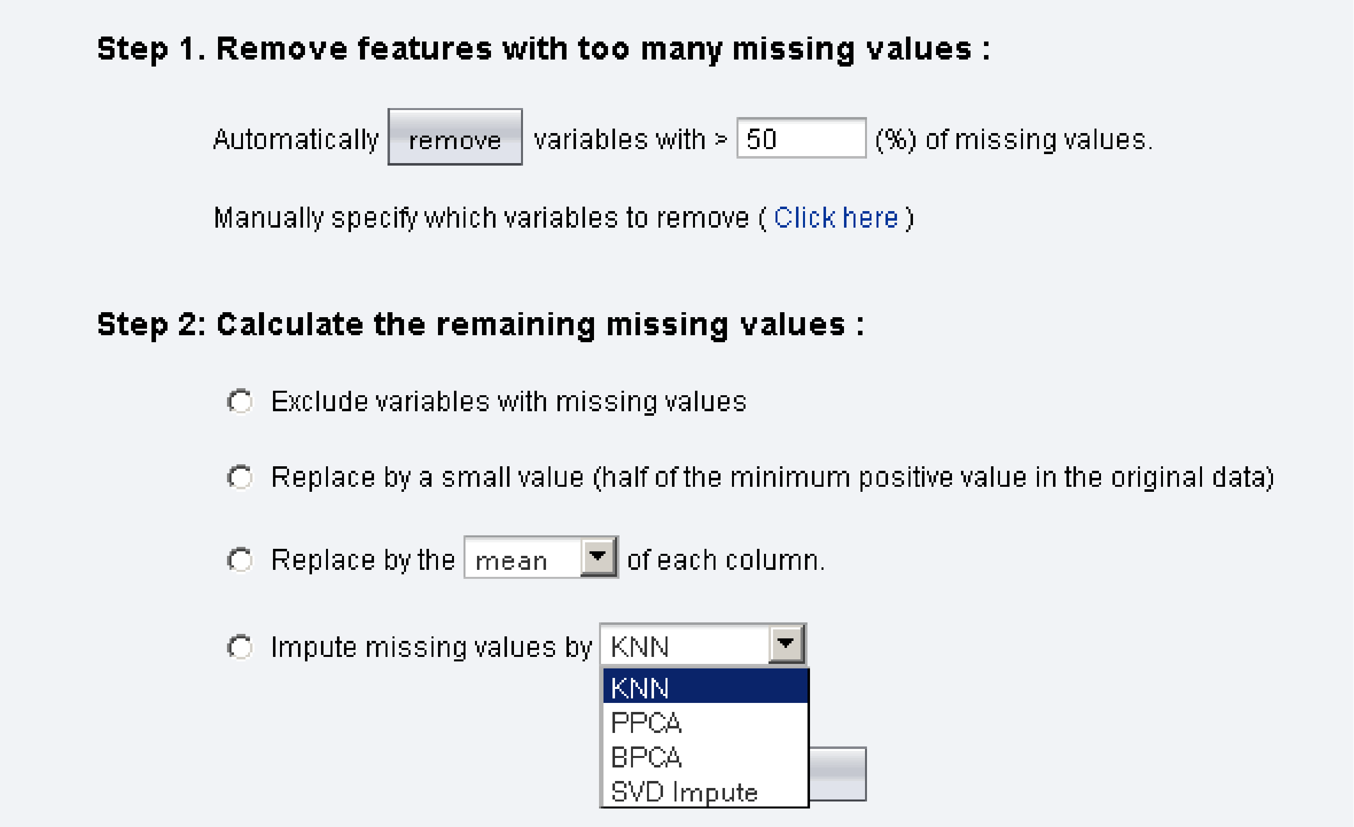
1. Variables of very small values - these variables can be detected using **mean** or the robust estimate **median** which is not affected by extreme values or outliers;
2. Variables that are near-constant throughout the experiment conditions - these variables can be detected using **standard deviation (SD)** or the robust estimate **interquantile range (IQR)**. The **coefficient of variation (CV)** (CV=mean/SD) is another useful variance measure independent of the mean.

The following empirical rules are applied during data filtering:

| Number of Variables | Variables Filtered |
|---|---|
| < 250 | 5% |
| 250 - 500 | 10% |
| 500 - 1000 | 25% |
| > 1000 | 40 % |

Please note, in order to reduce the computational burden to the server, the maximum allowed number of variables is 5000. If over 5000 variables were left after filtering, only the top 5000 will be used in the subsequent analysis.

# Missing values



**Step 1. Remove features with too many missing values :**

Automatically remove variables with > 50 (%) of missing values.

Manually specify which variables to remove ( Click here )

**Step 2: Calculate the remaining missing values :**

○ Exclude variables with missing values

○ Replace by a small value (half of the minimum positive value in the original data)

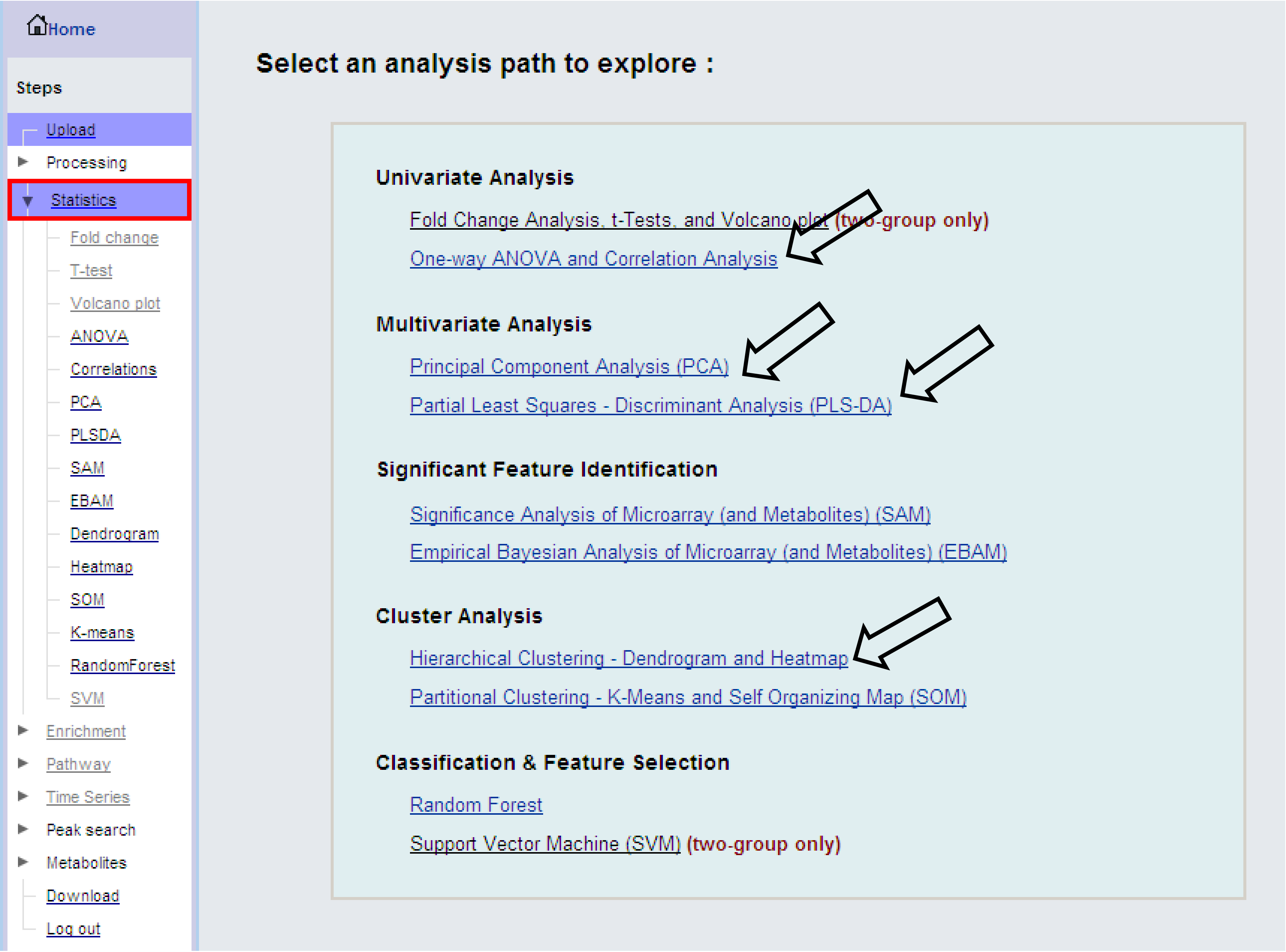
○ Replace by the mean of each column.

○ Impute missing values by KNN

# Dimension Reduction & Statistical Analysis

# Common tasks

- To identify important features;
- To detect interesting patterns;
- To assess difference between the phenotypes
- To facilitate classification / prediction



Home

Steps

- Upload
- Processing
- Statistics
  - Fold change
  - T-test
  - Volcano plot
  - ANOVA
  - Correlations
  - PCA
  - PLSDA
  - SAM
  - EBAM
  - Dendrogram
  - Heatmap
  - SOM
  - K-means
  - RandomForest
  - SVM
- Enrichment
- Pathway
- Time Series
- Peak search
- Metabolites
- Download
- Log out

Select an analysis path to explore :

**Univariate Analysis**

Fold Change Analysis, t-Tests, and Volcano plot (two-group only)

One-way ANOVA and Correlation Analysis

**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 and Heatmap

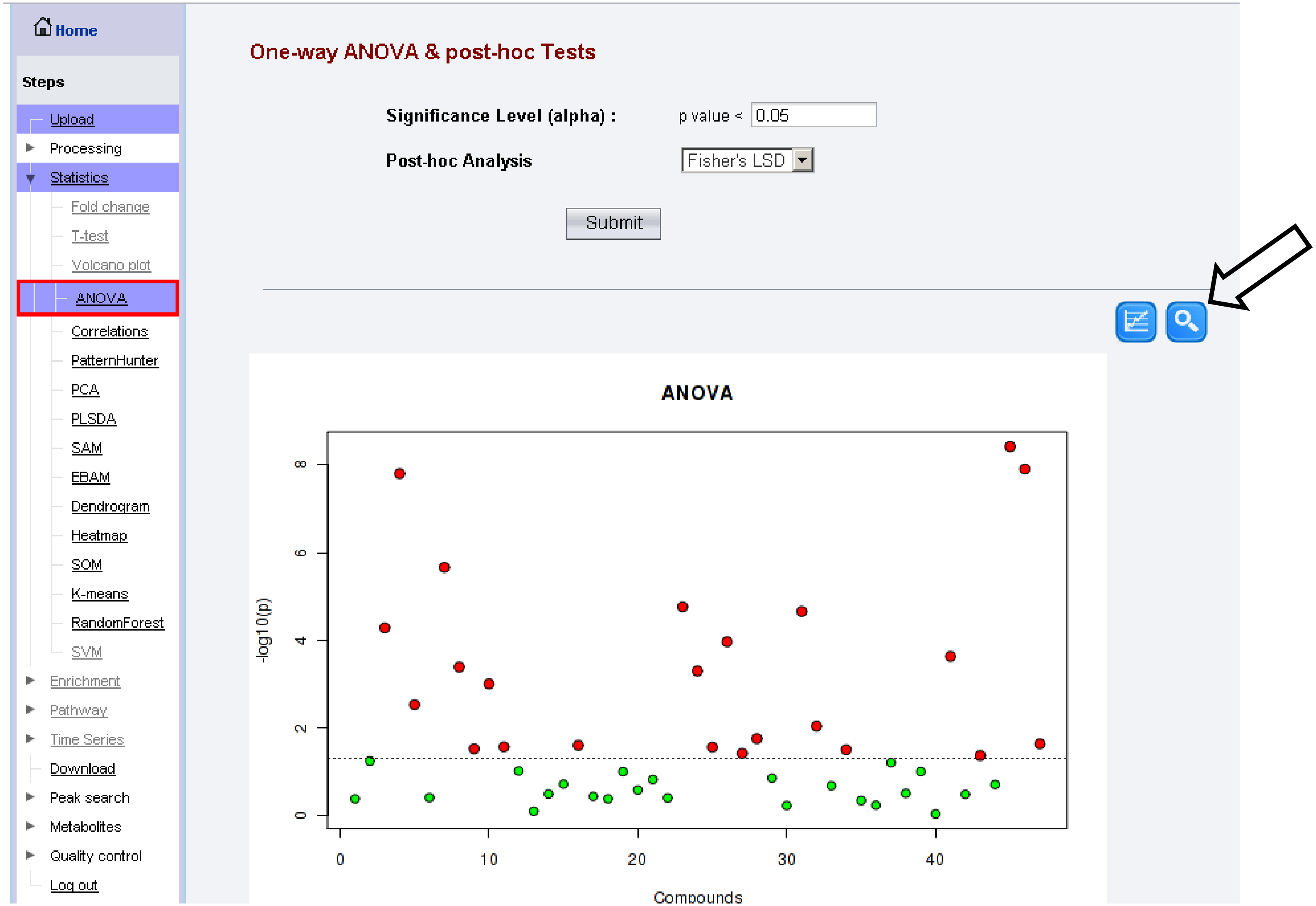
Partitional Clustering - K-Means and Self Organizing Map (SOM)

**Classification & Feature Selection**

Random Forest

Support Vector Machine (SVM) (two-group only)

# ANOVA

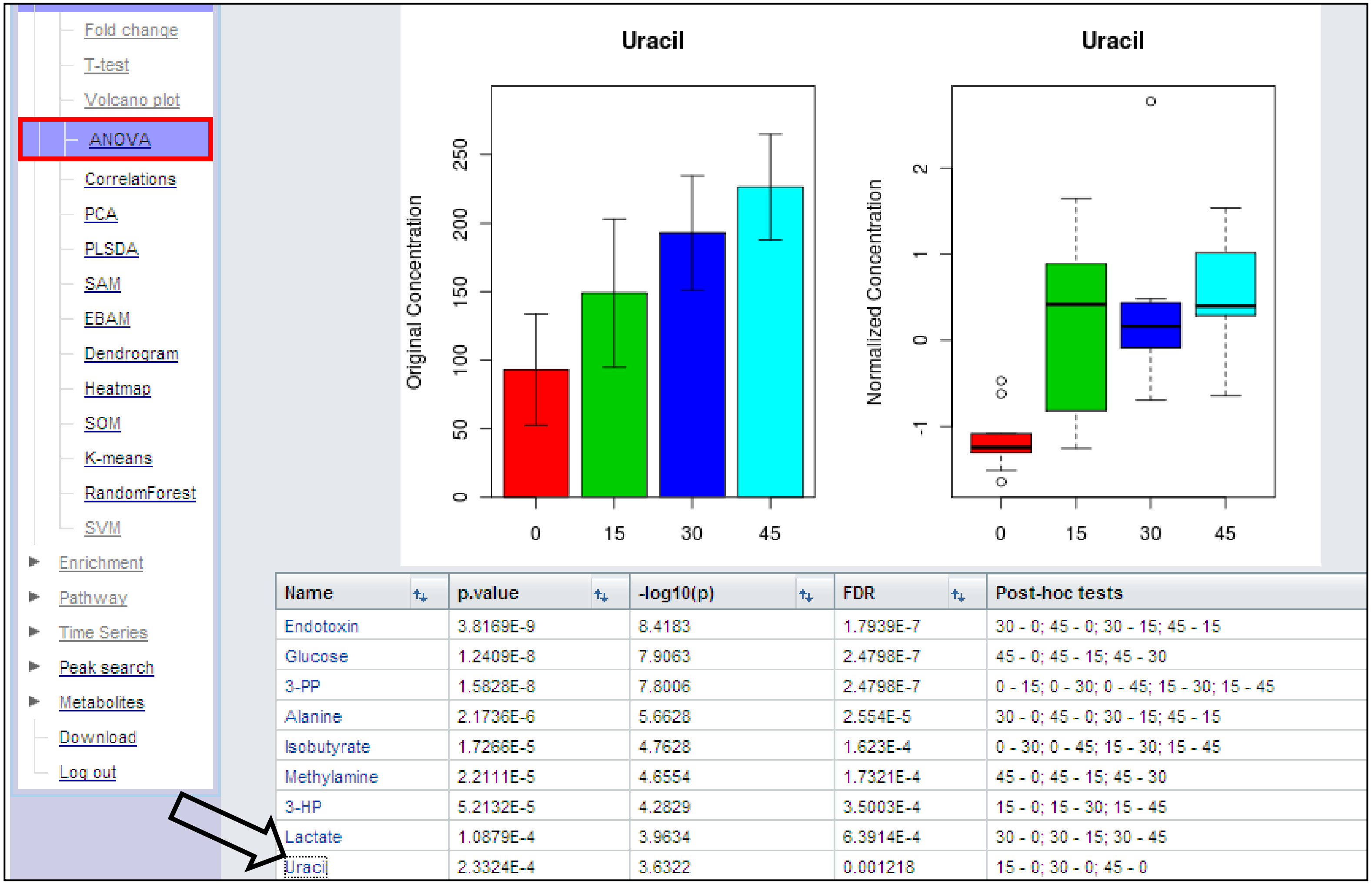


One-way ANOVA & post-hoc Tests

Significance Level (alpha) : p value < 0.05

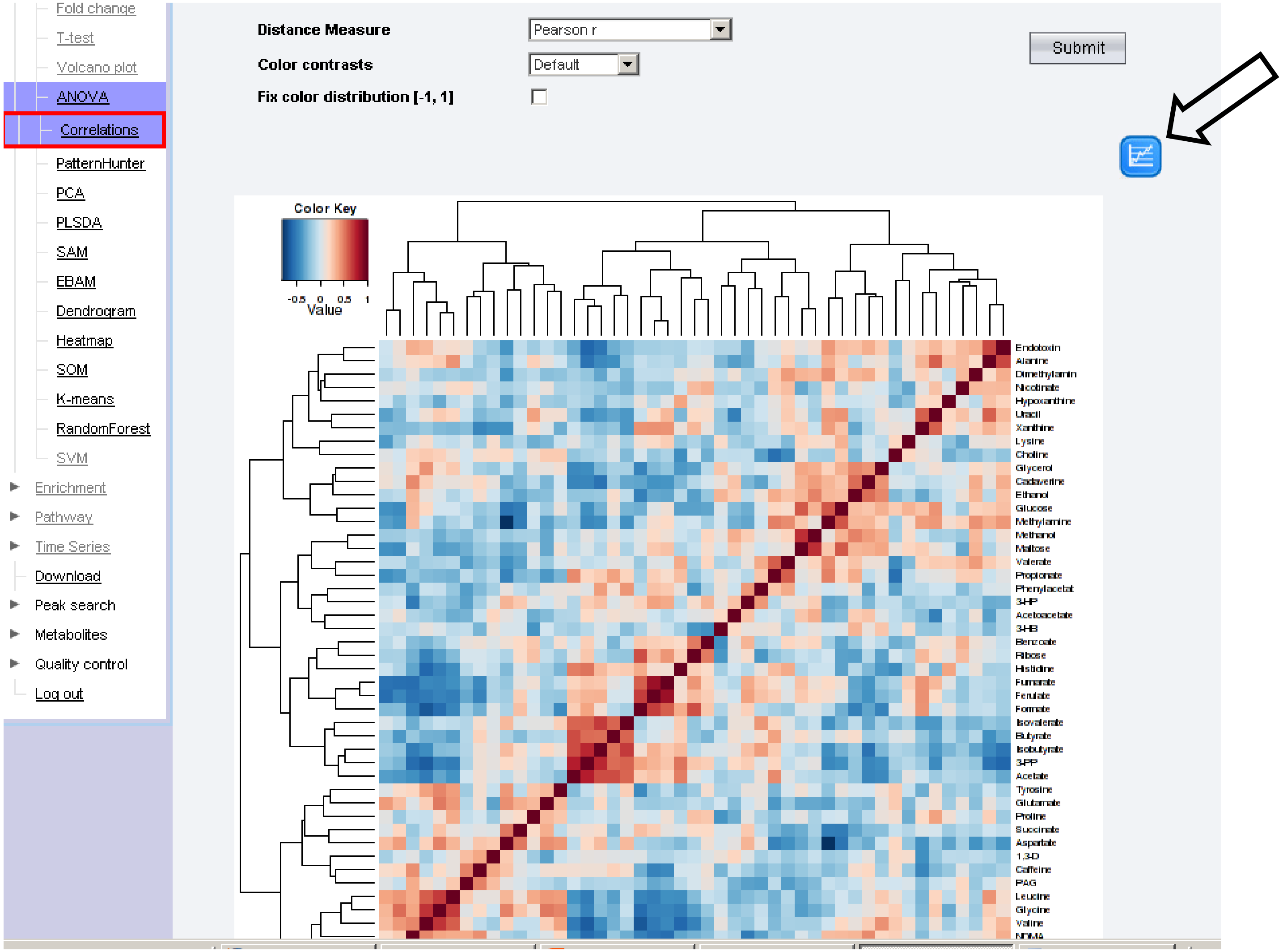
Post-hoc Analysis: Fisher's LSD

# View Individual Compounds



| Name | p.value | -log10(p) | FDR | Post-hoc tests |
|---|---|---|---|---|
| Endotoxin | 3.8169E-9 | 8.4183 | 1.7939E-7 | 30 - 0; 45 - 0; 30 - 15; 45 - 15 |
| Glucose | 1.2409E-8 | 7.9063 | 2.4798E-7 | 45 - 0; 45 - 15; 45 - 30 |
| 3-PP | 1.5828E-8 | 7.8006 | 2.4798E-7 | 0 - 15; 0 - 30; 0 - 45; 15 - 30; 15 - 45 |
| Alanine | 2.1736E-6 | 5.6628 | 2.554E-5 | 30 - 0; 45 - 0; 30 - 15; 45 - 15 |
| Isobutyrate | 1.7266E-5 | 4.7628 | 1.623E-4 | 0 - 30; 0 - 45; 15 - 30; 15 - 45 |
| Methylamine | 2.2111E-5 | 4.6554 | 1.7321E-4 | 45 - 0; 45 - 15; 45 - 30 |
| 3-HP | 5.2132E-5 | 4.2829 | 3.5003E-4 | 15 - 0; 15 - 30; 15 - 45 |
| Lactate | 1.0879E-4 | 3.9634 | 6.3914E-4 | 30 - 0; 30 - 15; 30 - 45 |
| Uracil | 2.3324E-4 | 3.6322 | 0.001218 | 15 - 0; 30 - 0; 45 - 0 |

# Overall correlation pattern

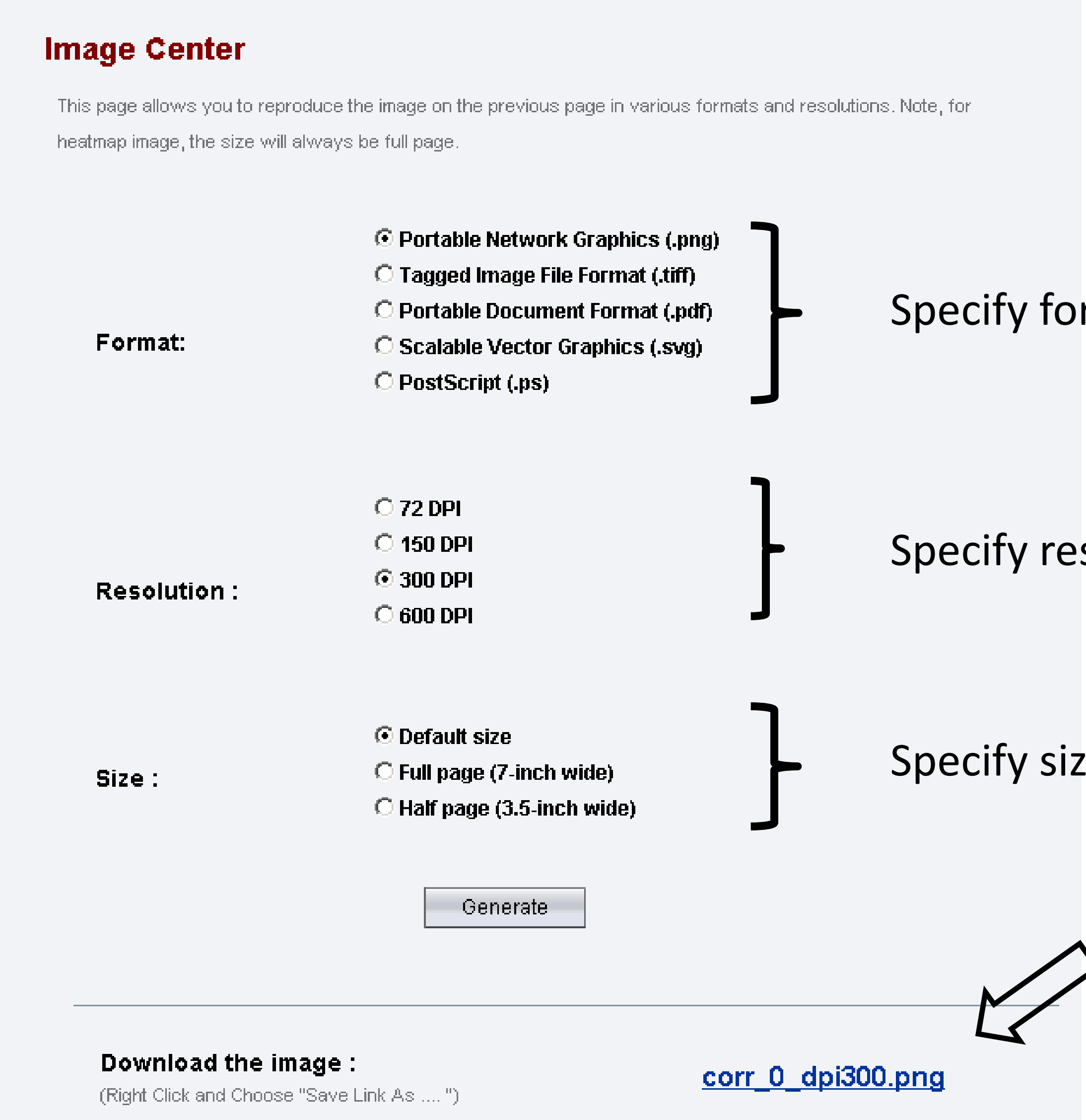


**Distance Measure** Pearson r

**Color contrasts** Default

**Fix color distribution [-1, 1]** □

# High resolution image



## Image Center

This page allows you to reproduce the image on the previous page in various formats and resolutions. Note, for heatmap image, the size will always be full page.

Format:

● Portable Network Graphics (.png)
○ Tagged Image File Format (.tiff)
○ Portable Document Format (.pdf)
○ Scalable Vector Graphics (.svg)
○ PostScript (.ps)

Specify format

Resolution :

○ 72 DPI
○ 150 DPI
● 300 DPI
○ 600 DPI

Specify resolution

Size :

● Default size
○ Full page (7-inch wide)
○ Half page (3.5-inch wide)

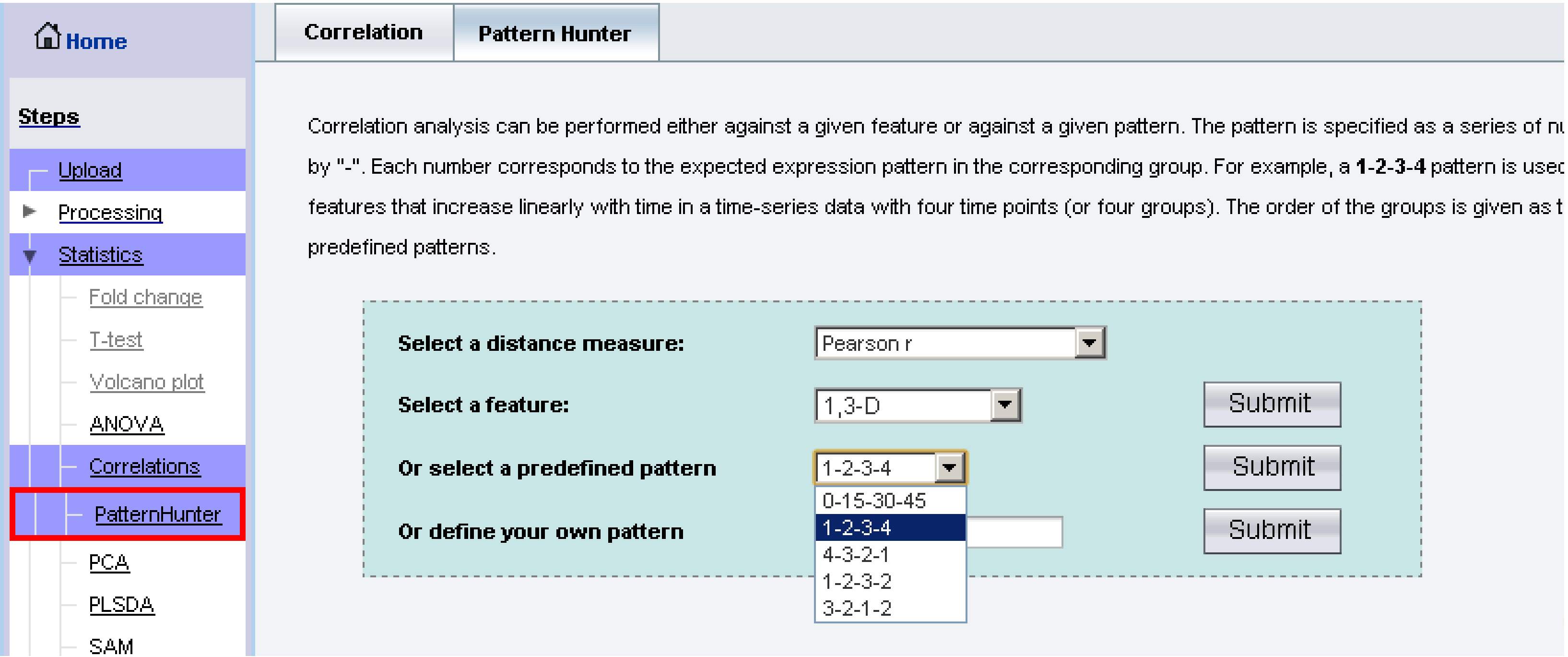
Specify size

**Download the image :**
(Right Click and Choose "Save Link As .... ")

corr_0_dpi300.png

# Template Matching

- Looking for compounds showing interesting patterns of change
- Essentially a method to look for linear trends or periodic trends in the data
- Best for data that has 3 or more groups



Correlation analysis can be performed either against a given feature or against a given pattern. The pattern is specified as a series of nu by "-". Each number corresponds to the expected expression pattern in the corresponding group. For example, a **1-2-3-4** pattern is usec 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 t predefined patterns.

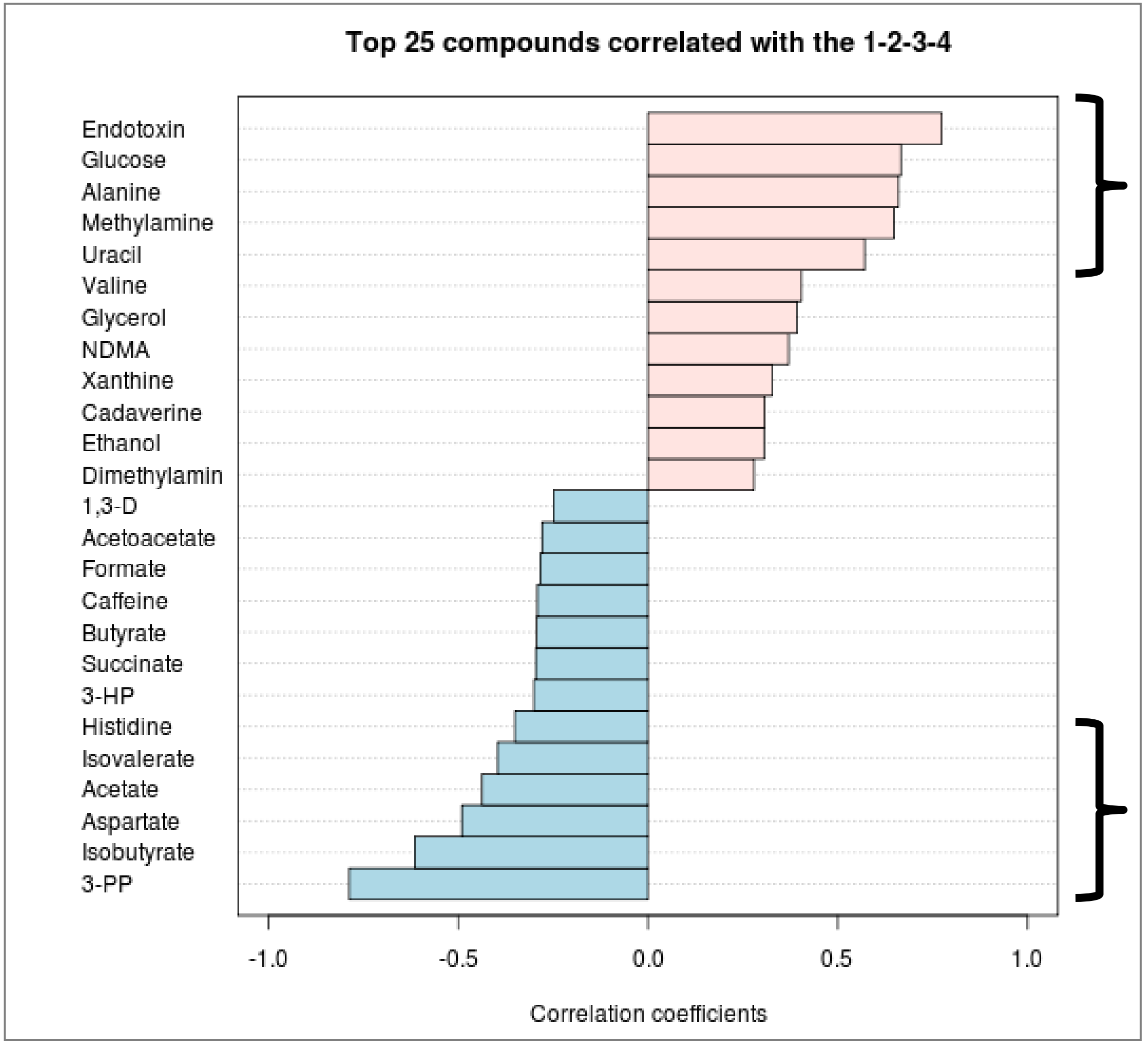
**Select a distance measure:** Pearson r

**Select a feature:** 1,3-D

**Or select a predefined pattern** 1-2-3-4

**Or define your own pattern**

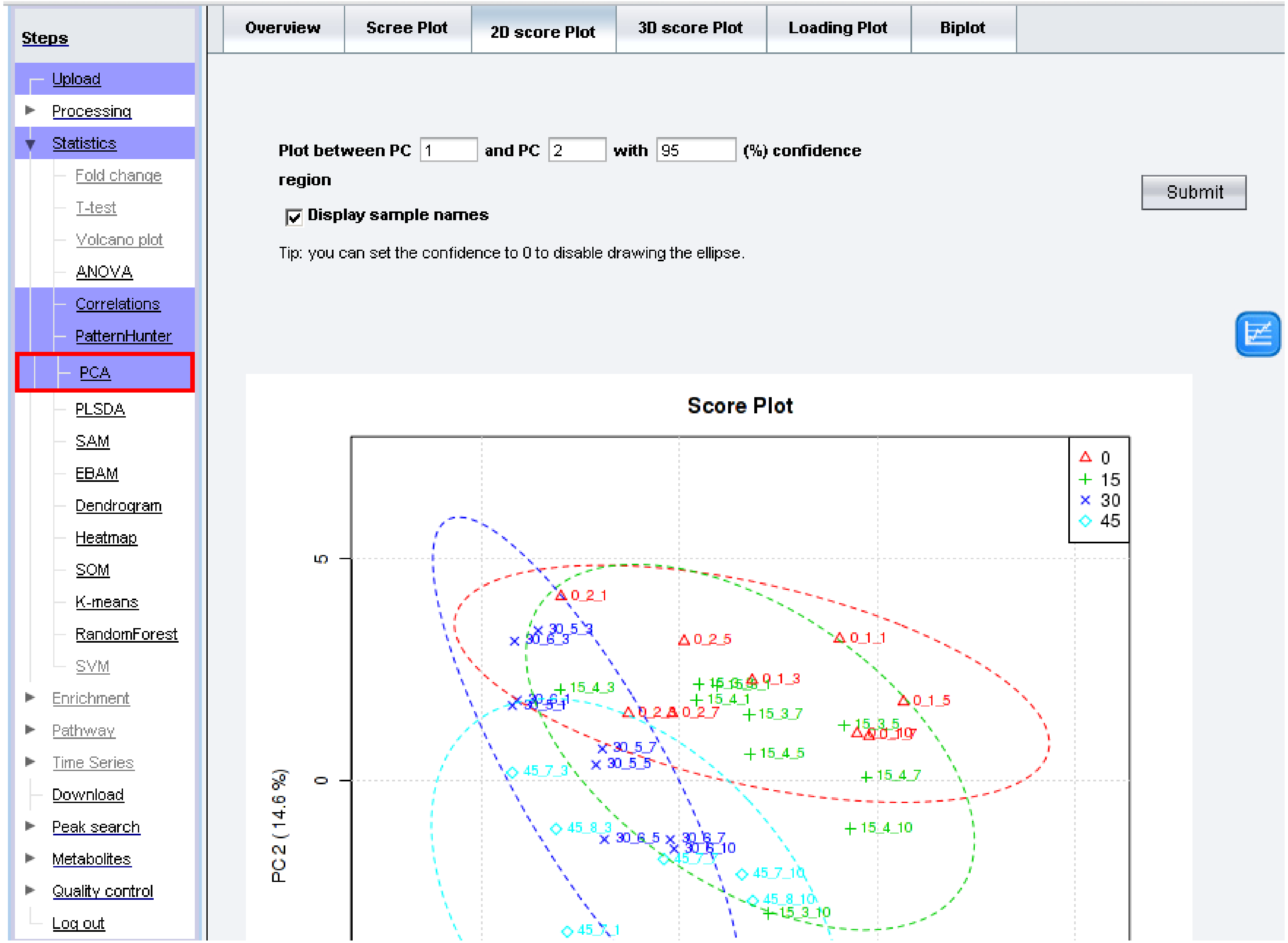
# Template Matching (cont.)



Strong linear + correlation to grain %

Strong linear - correlation to grain %

# PCA Scores Plot

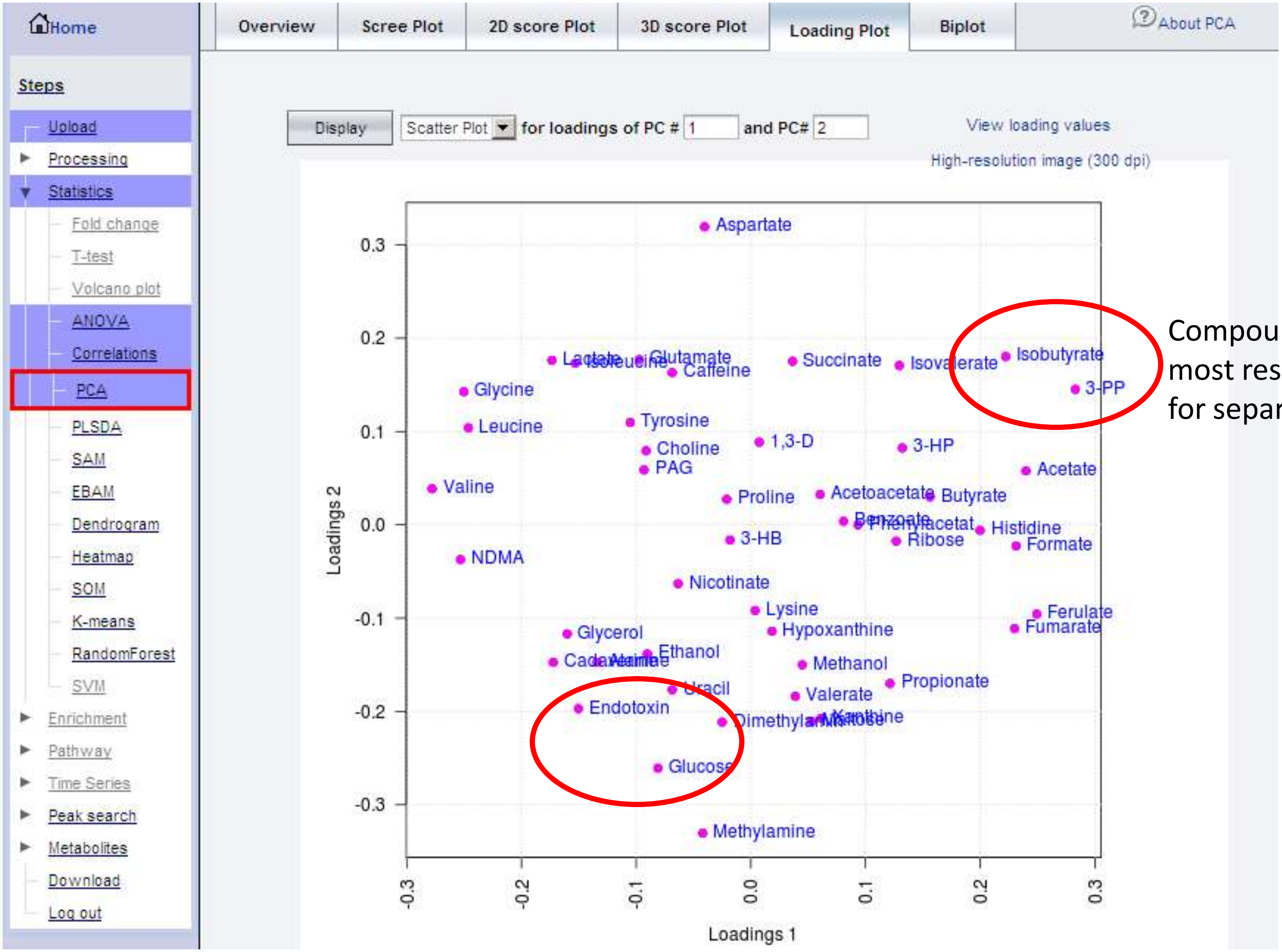


Plot between PC 1 and PC 2 with 95 (%) confidence region

☑ Display sample names

Tip: you can set the confidence to 0 to disable drawing the ellipse.

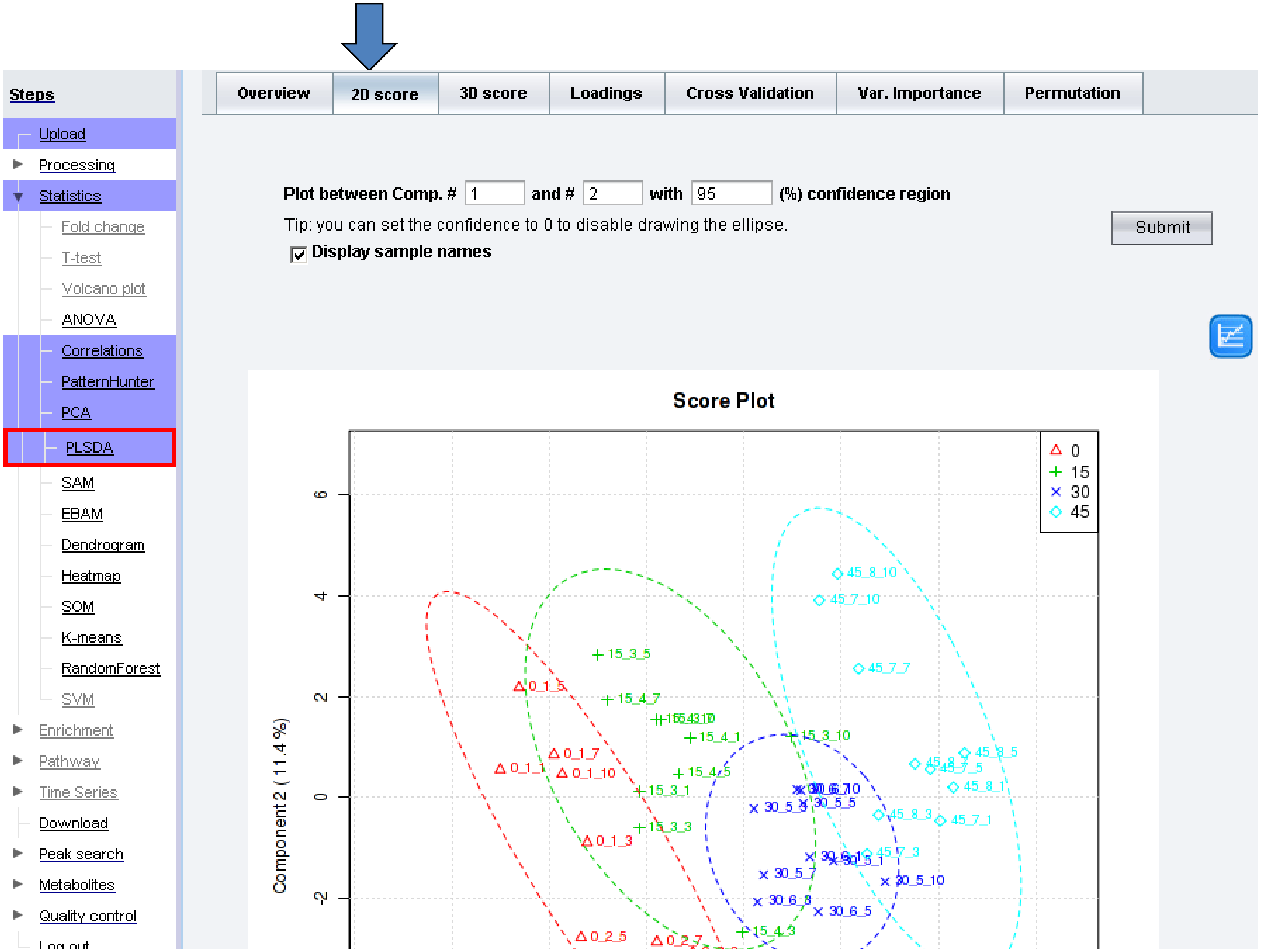
# PCA Loading Plot



Display Scatter Plot for loadings of PC # 1 and PC# 2

Compounds most responsible for separation

# PLS-DA Score Plot



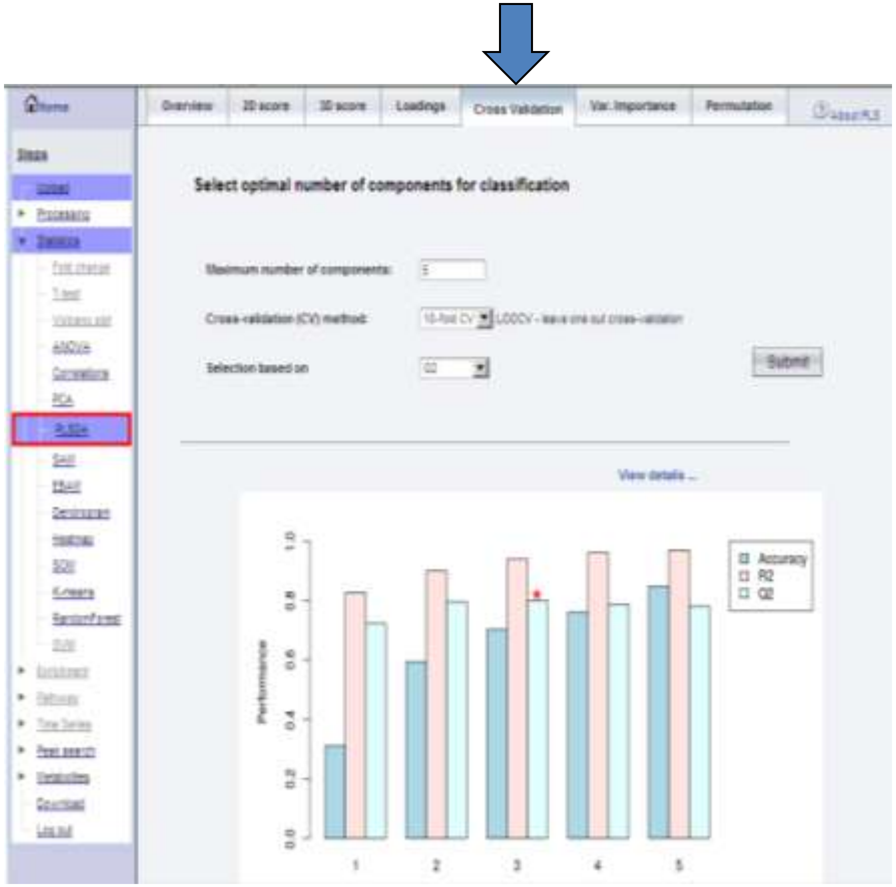
Plot between Comp. # 1 and # 2 with 95 (%) confidence region

Tip: you can set the confidence to 0 to disable drawing the ellipse.

☑ Display sample names

# Evaluation of PLS-DA Model

- PLS-DA Model evaluated by cross validation of $Q^2$ and $R^2$
- More components to model improves quality of fit, but try to minimize this value
- 3 Component model seems to be a good compromise here
- Good $R^2/Q^2$ (>0.7)



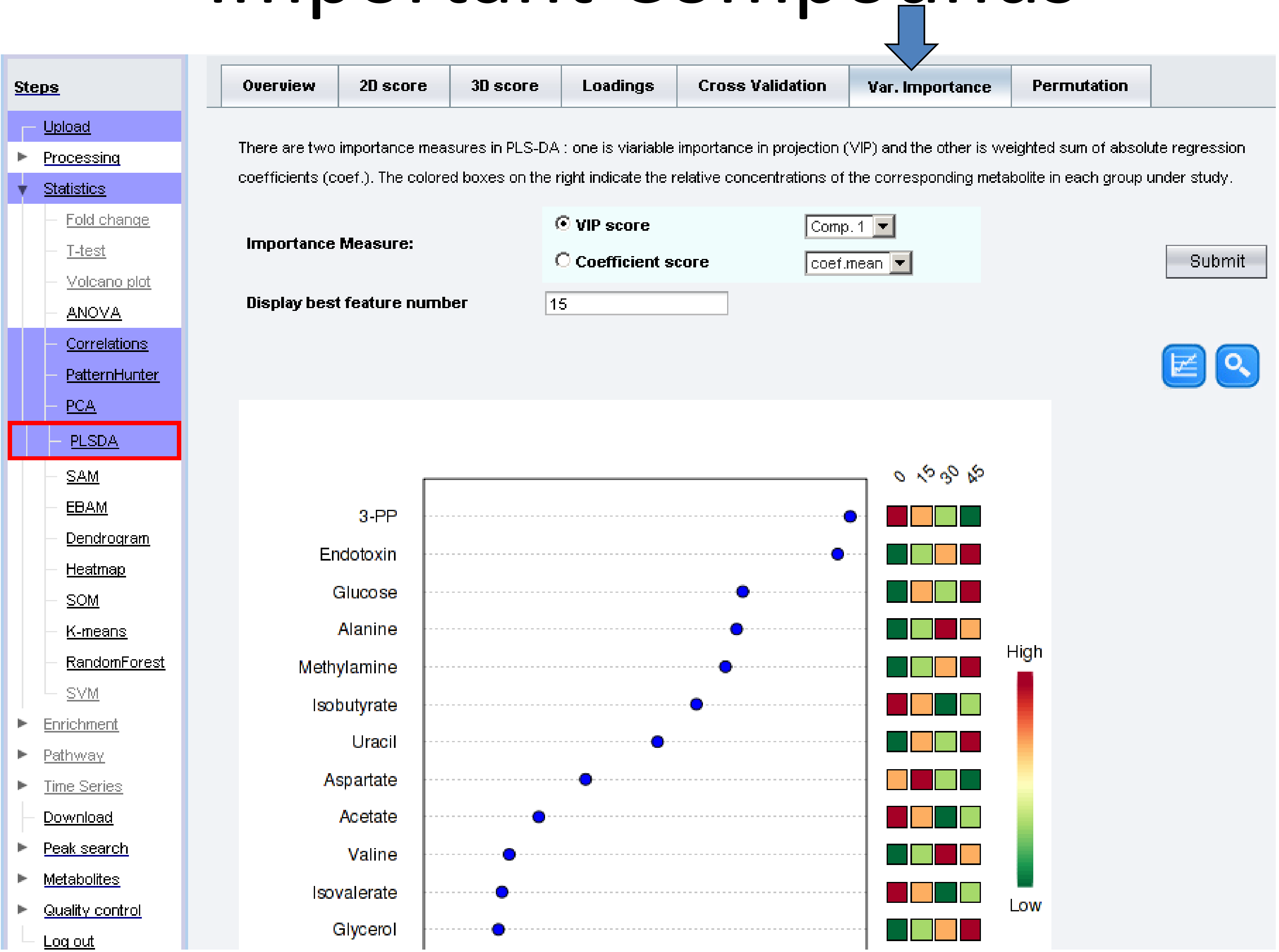
Select optimal number of components for classification

Maximum number of components: 5

Cross-validation (CV) method: 10-fold CV LOOCV - leave one out cross-validation

Selection based on Q2

# Important Compounds

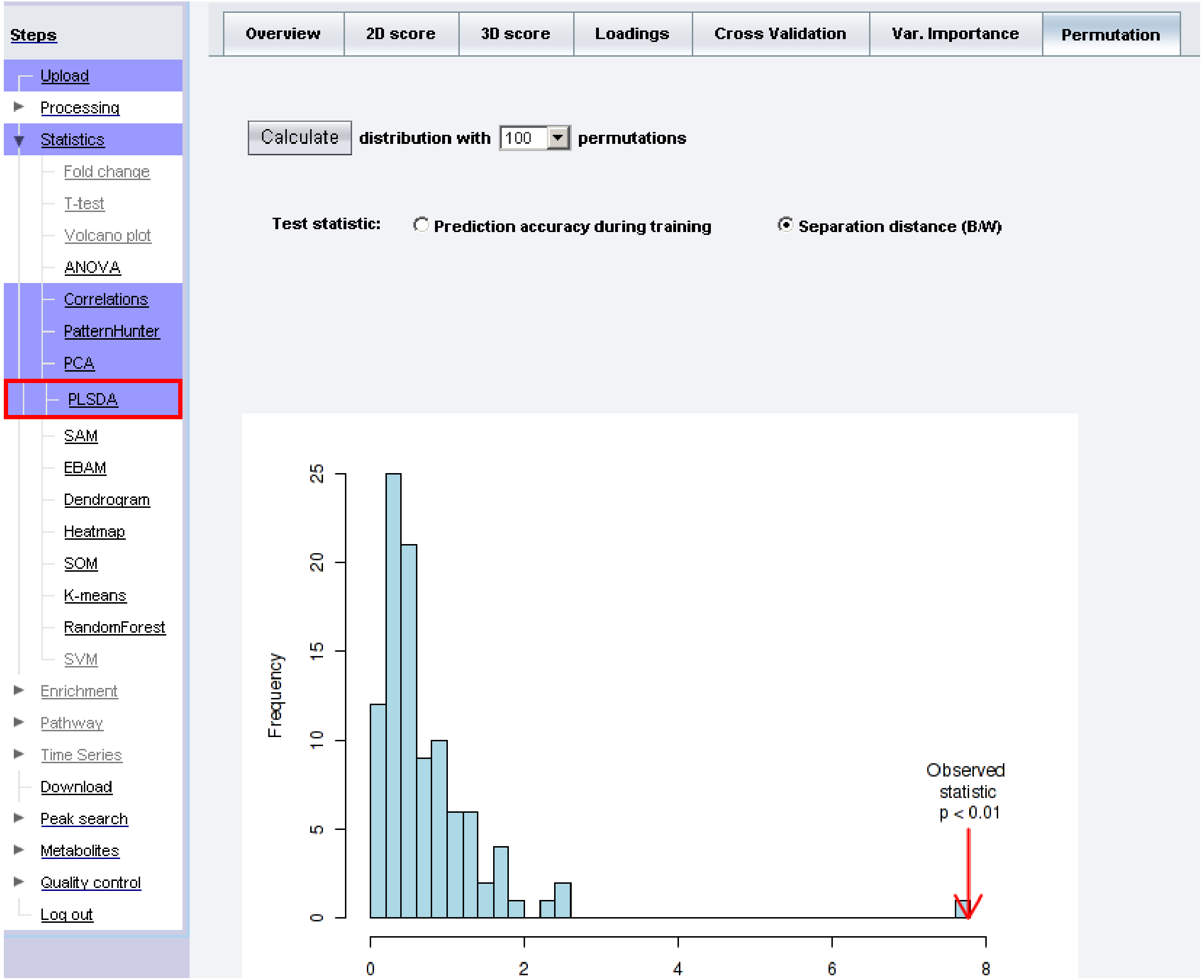


There are two importance measures in PLS-DA : one is viariable importance in projection (VIP) and the other is weighted sum of absolute regression coefficients (coef.). The colored boxes on the right indicate the relative concentrations of the corresponding metabolite in each group under study.

Importance Measure: ● VIP score — Comp. 1; ○ Coefficient score — coef.mean

Display best feature number: 15

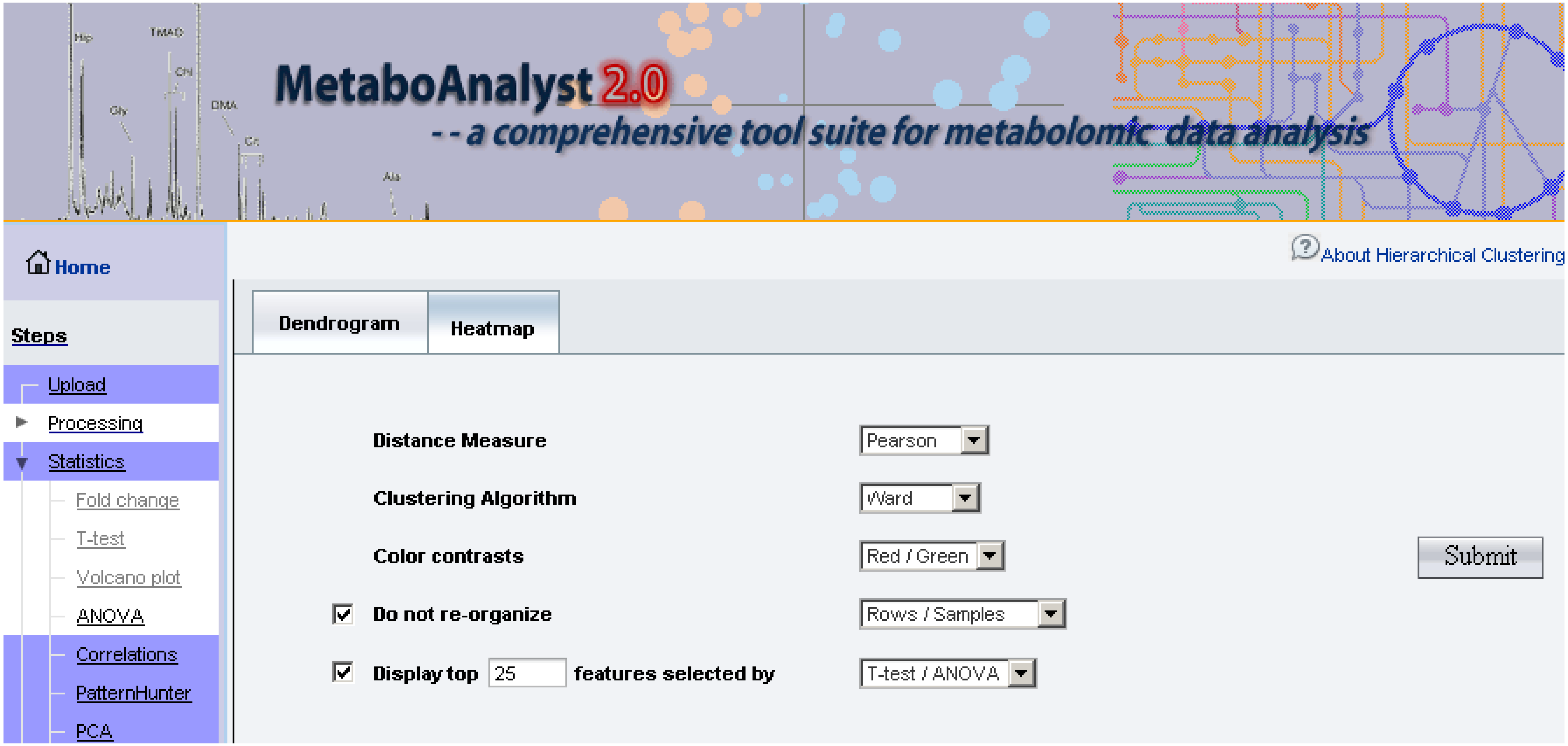
# Model Validation



Calculate distribution with 100 permutations

Test statistic: ○ Prediction accuracy during training ● Separation distance (B/W)

# Heatmap Visualization



Distance Measure: Pearson

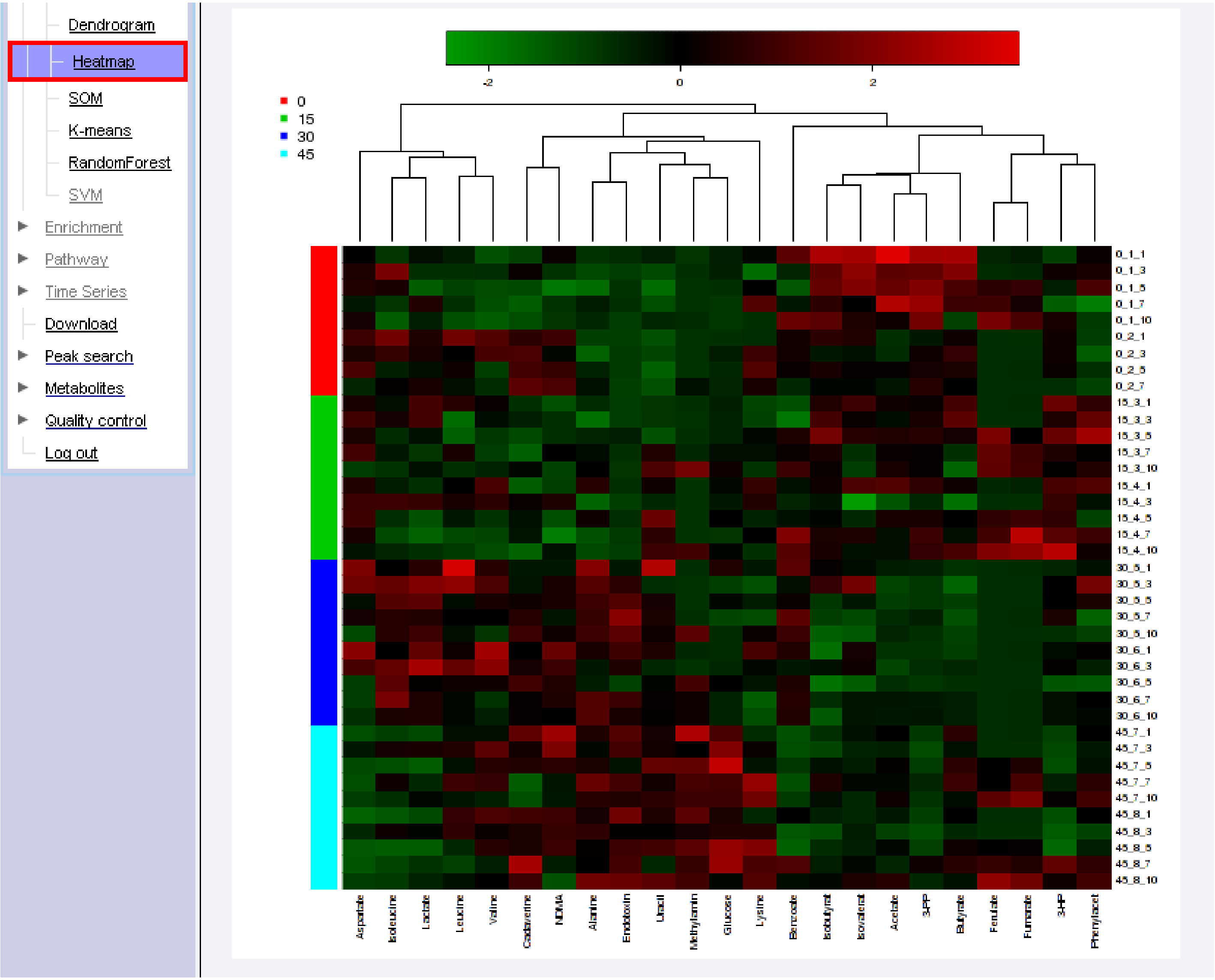
Clustering Algorithm: Ward

Color contrasts: Red / Green

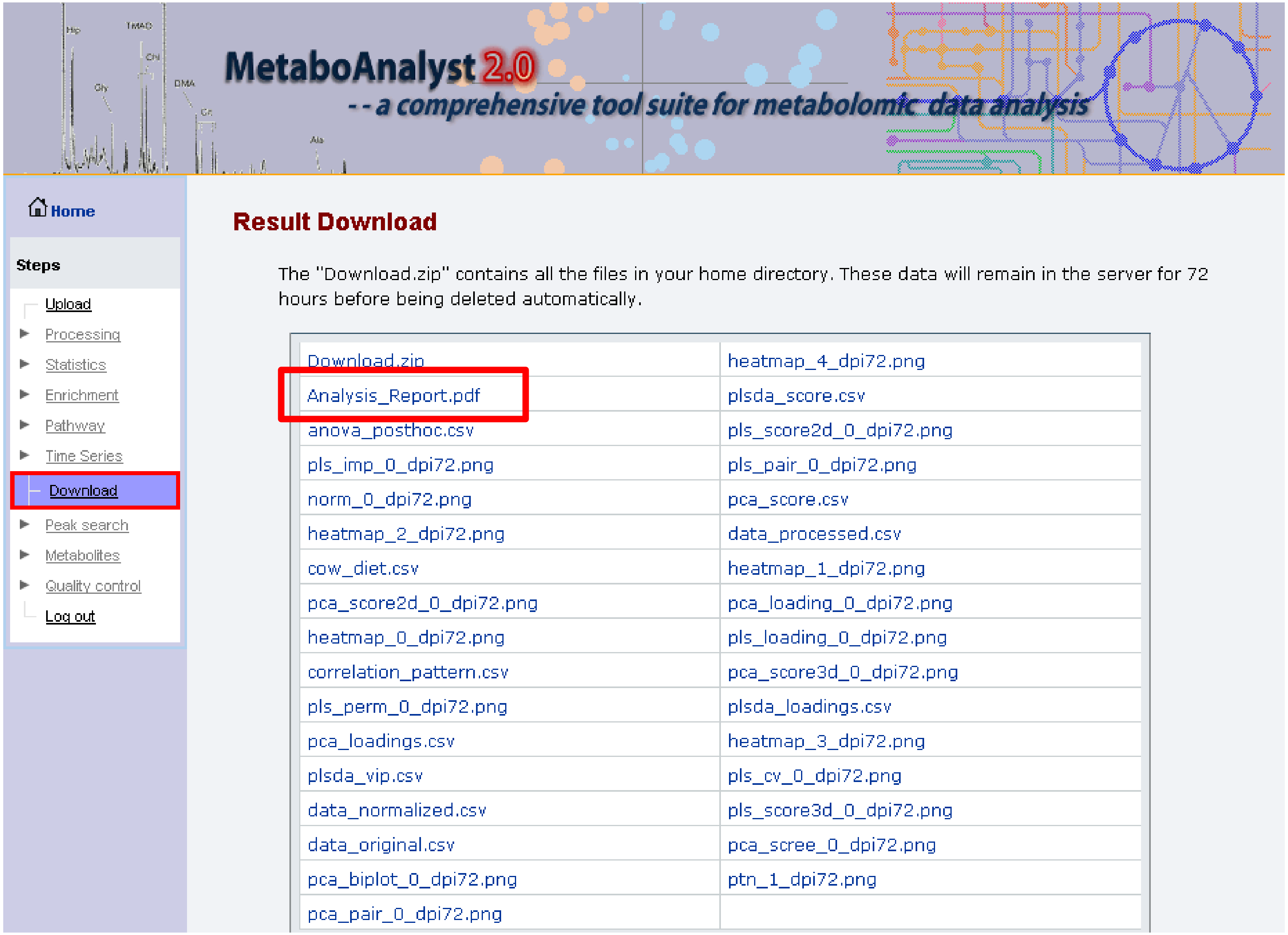
☑ Do not re-organize: Rows / Samples

☑ Display top 25 features selected by: T-test / ANOVA

# Heatmap Visualization (cont.)



# Download Results

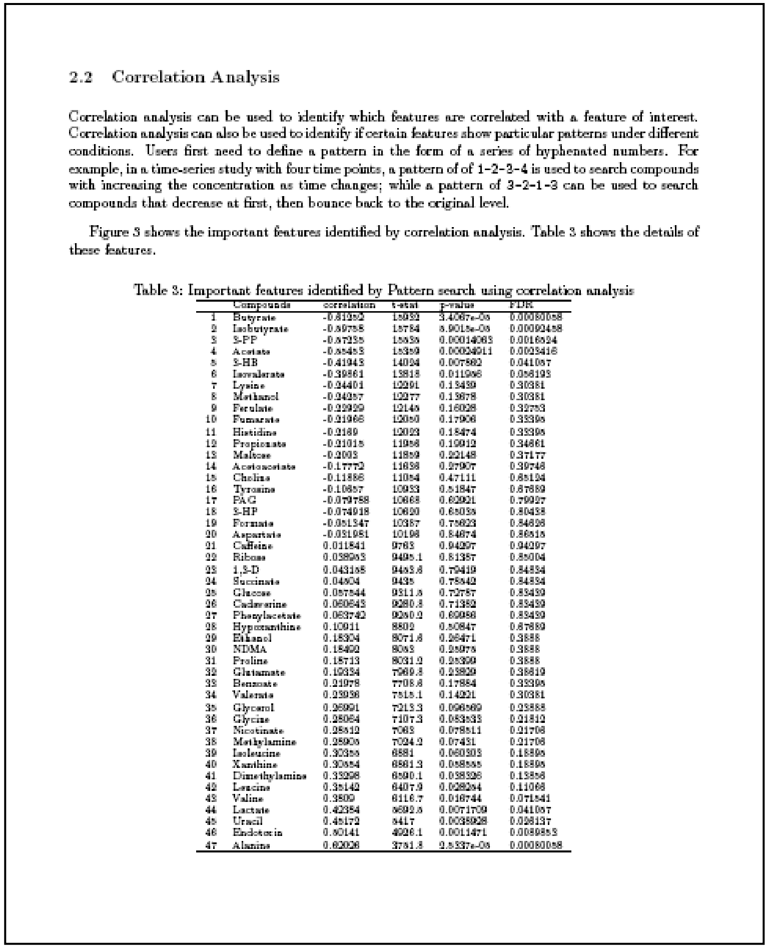


## Result Download

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.

| | |
|---|---|
| Download.zip | heatmap_4_dpi72.png |
| Analysis_Report.pdf | plsda_score.csv |
| anova_posthoc.csv | pls_score2d_0_dpi72.png |
| pls_imp_0_dpi72.png | pls_pair_0_dpi72.png |
| norm_0_dpi72.png | pca_score.csv |
| heatmap_2_dpi72.png | data_processed.csv |
| cow_diet.csv | heatmap_1_dpi72.png |
| pca_score2d_0_dpi72.png | pca_loading_0_dpi72.png |
| heatmap_0_dpi72.png | pls_loading_0_dpi72.png |
| correlation_pattern.csv | pca_score3d_0_dpi72.png |
| pls_perm_0_dpi72.png | plsda_loadings.csv |
| pca_loadings.csv | heatmap_3_dpi72.png |
| plsda_vip.csv | pls_cv_0_dpi72.png |
| data_normalized.csv | pls_score3d_0_dpi72.png |
| data_original.csv | pca_scree_0_dpi72.png |
| pca_biplot_0_dpi72.png | ptn_1_dpi72.png |
| pca_pair_0_dpi72.png | |

# Analysis Report



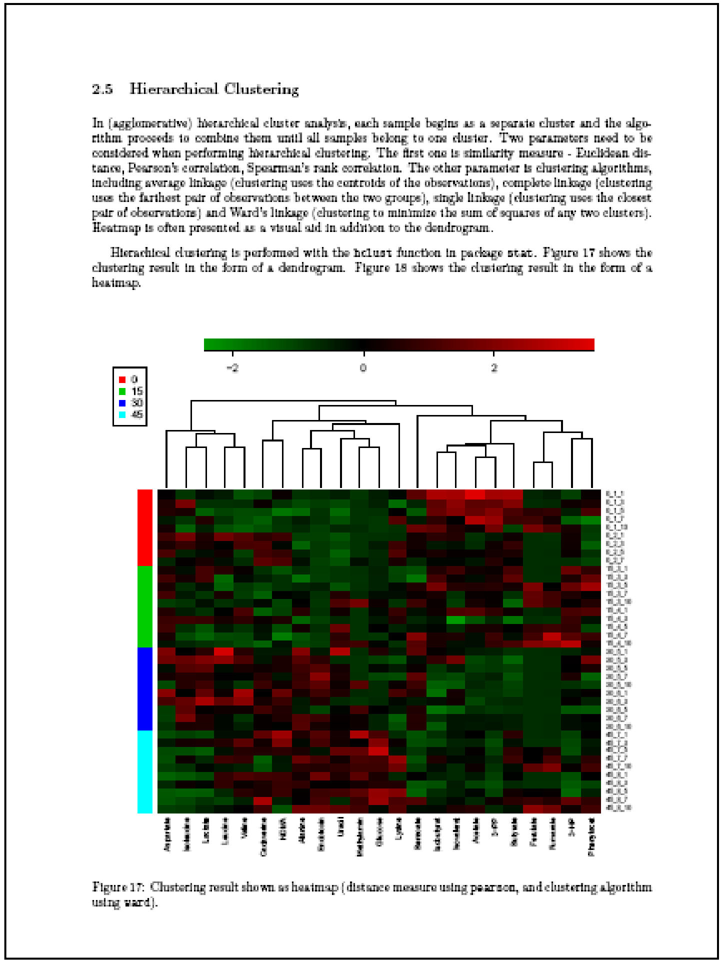
### 2.2 Correlation Analysis

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 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

| | Compounds | correlation | t-stat | p-value | FDR |
|---|---|---|---|---|---|
| 1 | Butyrate | -0.61282 | 15932 | 3.4087e-05 | 0.00080058 |
| 2 | Isobutyrate | -0.59758 | 15784 | 5.9015e-05 | 0.00092458 |
| 3 | 3-PP | -0.57235 | 15535 | 0.00014063 | 0.0016524 |
| 4 | Acetate | -0.55453 | 15359 | 0.00024911 | 0.0023416 |
| 5 | 3-HB | -0.41943 | 14094 | 0.007582 | 0.041087 |
| 6 | Isovalerate | -0.39861 | 13815 | 0.011966 | 0.056193 |
| 7 | Lysine | -0.24401 | 12291 | 0.13459 | 0.30381 |
| 8 | Methanol | -0.24257 | 12277 | 0.13878 | 0.30381 |
| 9 | Ferulate | -0.22929 | 12148 | 0.16028 | 0.32753 |
| 10 | Fumarate | -0.21968 | 12050 | 0.17906 | 0.33396 |
| 11 | Histidine | -0.2169 | 12023 | 0.18474 | 0.33396 |
| 12 | Propionate | -0.21018 | 11958 | 0.19912 | 0.34661 |
| 13 | Maltose | -0.2003 | 11859 | 0.22148 | 0.37177 |
| 14 | Acetoacetate | -0.17772 | 11638 | 0.27907 | 0.39748 |
| 15 | Choline | -0.11886 | 11084 | 0.47111 | 0.65124 |
| 16 | Tyrosine | -0.10857 | 10953 | 0.51847 | 0.67889 |
| 17 | PAG | -0.079758 | 10668 | 0.62921 | 0.79927 |
| 18 | 3-HP | -0.074918 | 10620 | 0.65038 | 0.80438 |
| 19 | Formate | -0.051347 | 10387 | 0.75623 | 0.84626 |
| 20 | Aspartate | -0.031951 | 10198 | 0.84874 | 0.86818 |
| 21 | Caffeine | 0.011541 | 9763 | 0.94297 | 0.94297 |
| 22 | Ribose | 0.038963 | 9495.1 | 0.81357 | 0.85004 |
| 23 | 1,3-D | 0.043185 | 9453.6 | 0.79419 | 0.84834 |
| 24 | Succinate | 0.04804 | 9435 | 0.75842 | 0.84834 |
| 25 | Glucose | 0.057544 | 9311.5 | 0.72787 | 0.83439 |
| 26 | Cadaverine | 0.060643 | 9280.3 | 0.71352 | 0.83439 |
| 27 | Phenylacetate | 0.063742 | 9260.2 | 0.69958 | 0.83439 |
| 28 | Hypoxanthine | 0.10911 | 8802 | 0.50847 | 0.67889 |
| 29 | Ethanol | 0.18304 | 8071.6 | 0.26471 | 0.3858 |
| 30 | NDMA | 0.18492 | 8063 | 0.25978 | 0.3858 |
| 31 | Proline | 0.18713 | 8031.2 | 0.25399 | 0.3858 |
| 32 | Glutamate | 0.19334 | 7969.5 | 0.23829 | 0.36819 |
| 33 | Benzoate | 0.21978 | 7708.6 | 0.17854 | 0.33396 |
| 34 | Valerate | 0.23938 | 7518.1 | 0.14221 | 0.30381 |
| 35 | Glycerol | 0.26991 | 7213.3 | 0.096569 | 0.23585 |
| 36 | Glycine | 0.28064 | 7107.3 | 0.083533 | 0.21812 |
| 37 | Nicotinate | 0.28512 | 7063 | 0.078511 | 0.21708 |
| 38 | Methylamine | 0.28906 | 7024.2 | 0.07431 | 0.21708 |
| 39 | Isoleucine | 0.30358 | 6881 | 0.060303 | 0.18595 |
| 40 | Xanthine | 0.30654 | 6851.3 | 0.058555 | 0.18595 |
| 41 | Dimethylamine | 0.33298 | 6590.1 | 0.038326 | 0.13556 |
| 42 | Leucine | 0.35142 | 6407.9 | 0.028284 | 0.11068 |
| 43 | Valine | 0.3509 | 6116.7 | 0.016744 | 0.071541 |
| 44 | Lactate | 0.42354 | 5692.5 | 0.0071709 | 0.041087 |
| 45 | Uracil | 0.45172 | 5417 | 0.0035928 | 0.028137 |
| 46 | Endotoxin | 0.50141 | 4926.1 | 0.0011471 | 0.0089853 |
| 47 | Alanine | 0.62026 | 3781.5 | 2.5337e-05 | 0.00080058 |



### 2.5 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.

Hierachical clustering is performed with the hclust function in package stat. Figure 17 shows the clustering result in the form of a dendrogram. Figure 18 shows the clustering result in the form of a heatmap.

Figure 17: Clustering result shown as heatmap (distance measure using pearson, and clustering algorithm using ward).

# Metabolite Set Enrichment Analysis (MSEA)

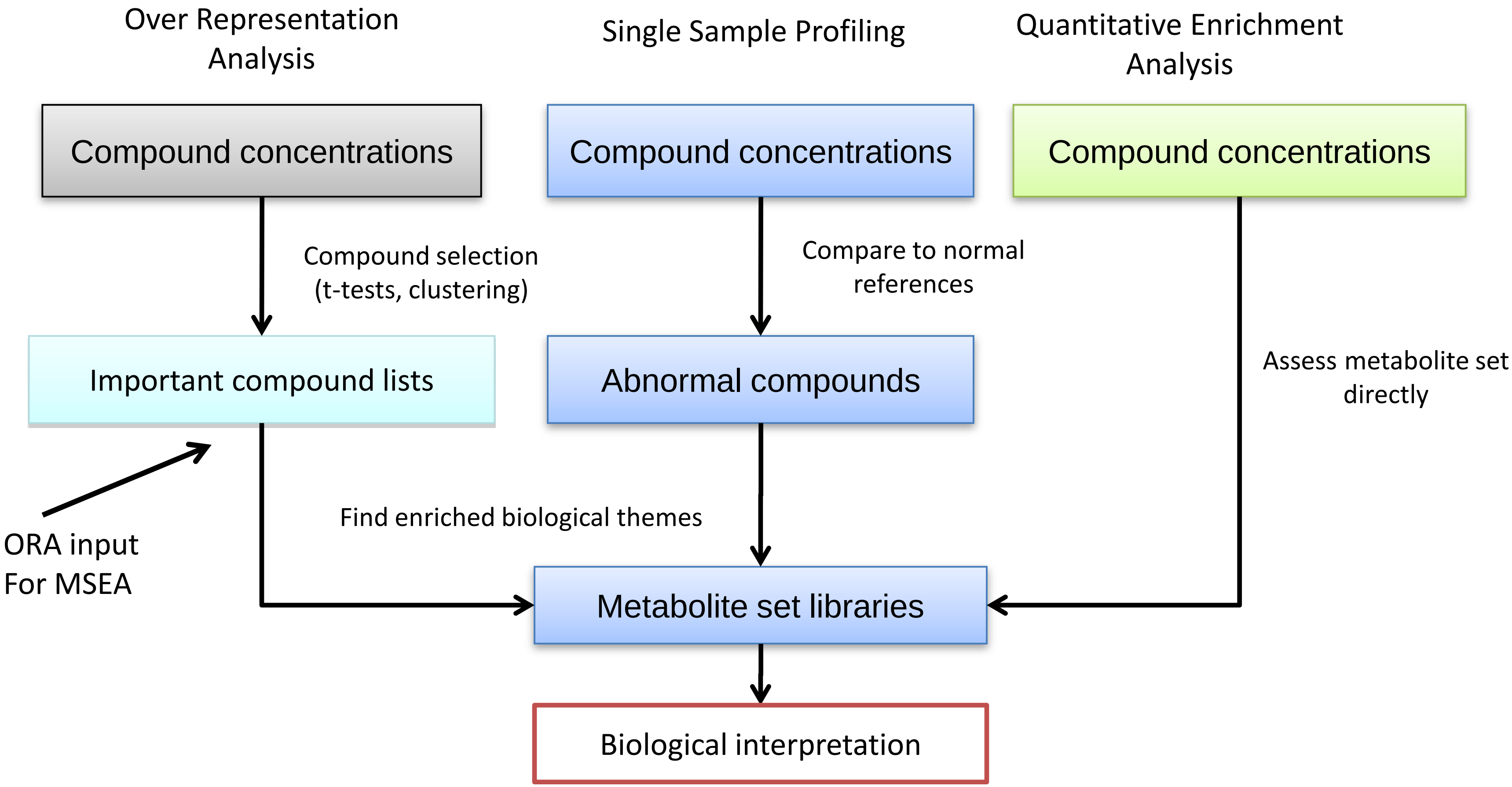
# Enrichment Analysis

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

# MSEA

- Accepts 3 kinds of input files
- 1) list of metabolite names only (ORA)
- 2) list of metabolite names + concentration data from a single sample (SSP)
- 3) a concentration table with a list of metabolite names + concentrations for multiple samples/patients (QEA)

# The MSEA approach



Over Representation Analysis

Single Sample Profiling

Quantitative Enrichment Analysis

Compound concentrations

Compound concentrations

Compound concentrations

Compound selection (t-tests, clustering)

Compare to normal references

Important compound lists

Abnormal compounds

Assess metabolite set directly

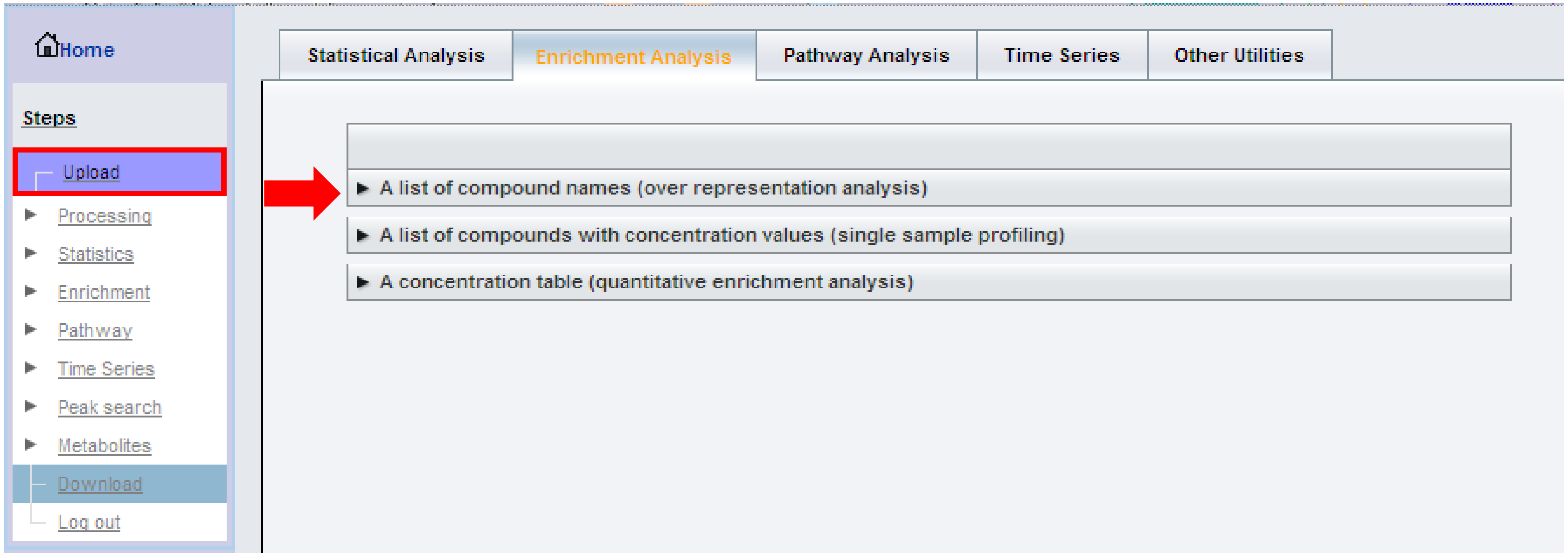
ORA input
For MSEA

Find enriched biological themes

Metabolite set libraries

Biological interpretation

# Start with a compound List

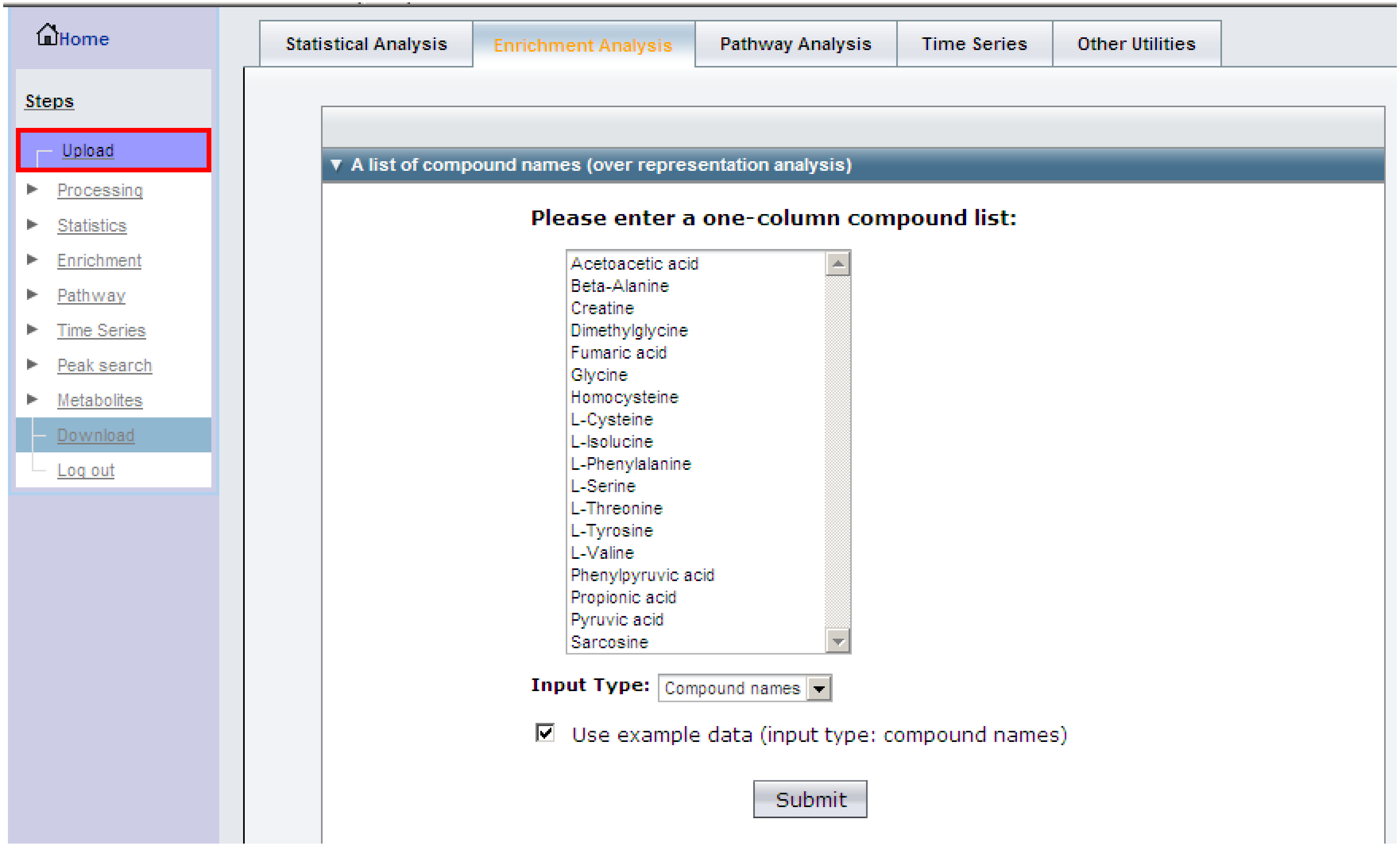


▶ A list of compound names (over representation analysis)

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

▶ A concentration table (quantitative enrichment analysis)

# Upload Compound List



▼ A list of compound names (over representation analysis)

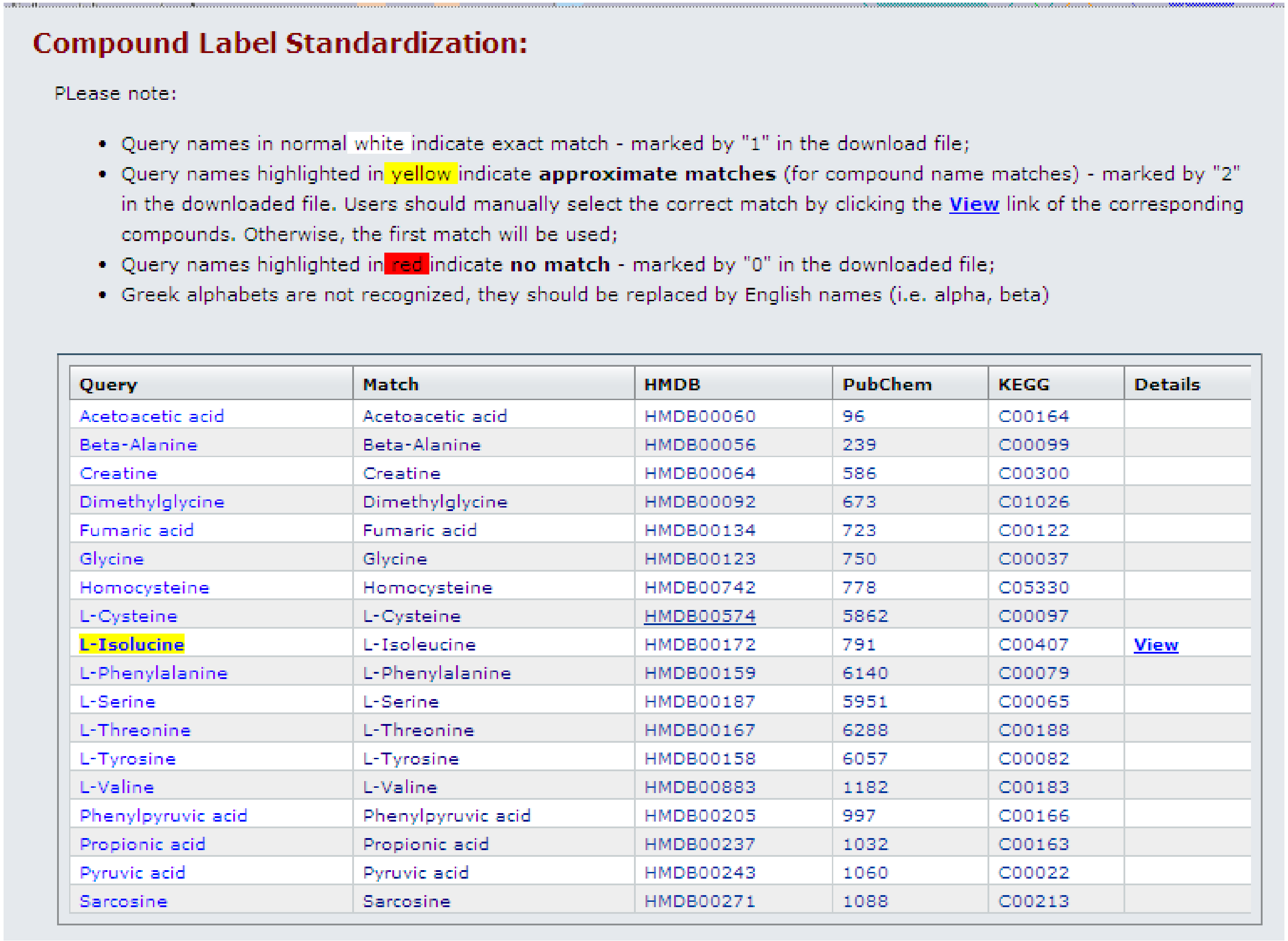
**Please enter a one-column compound list:**

```
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:** Compound names

☑ Use example data (input type: compound names)

# Compound Name Standardization



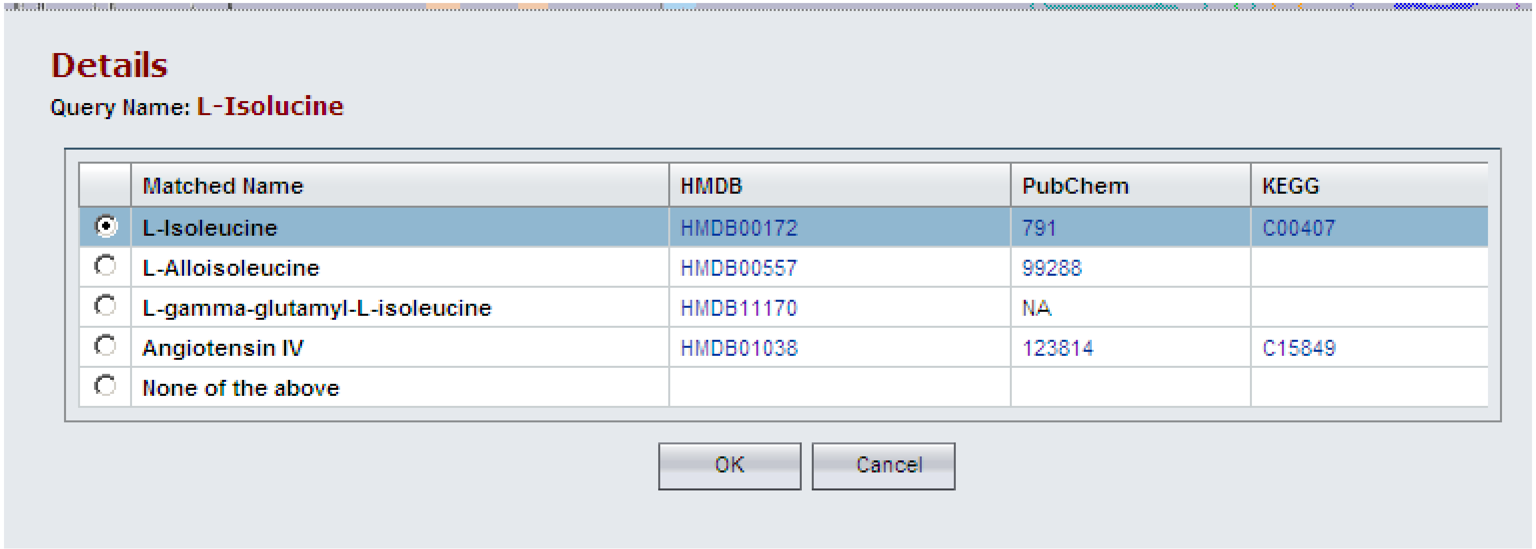
## Compound Label Standardization:

PLease note:

- Query names in normal white indicate exact match - marked by "1" in the download file;
- Query names highlighted in yellow indicate **approximate matches** (for compound name matches) - marked by "2" in the downloaded file. Users should manually select the correct match by clicking the **View** link of the corresponding compounds. Otherwise, the first match will be used;
- Query names highlighted in red indicate **no match** - marked by "0" in the downloaded file;
- Greek alphabets are not recognized, they should be replaced by English names (i.e. alpha, beta)

| Query | Match | HMDB | PubChem | KEGG | Details |
|---|---|---|---|---|---|
| Acetoacetic acid | Acetoacetic acid | HMDB00060 | 96 | C00164 | |
| Beta-Alanine | Beta-Alanine | HMDB00056 | 239 | C00099 | |
| Creatine | Creatine | HMDB00064 | 586 | C00300 | |
| Dimethylglycine | Dimethylglycine | HMDB00092 | 673 | C01026 | |
| Fumaric acid | Fumaric acid | HMDB00134 | 723 | C00122 | |
| Glycine | Glycine | HMDB00123 | 750 | C00037 | |
| Homocysteine | Homocysteine | HMDB00742 | 778 | C05330 | |
| L-Cysteine | L-Cysteine | HMDB00574 | 5862 | C00097 | |
| **L-Isolucine** | L-Isoleucine | HMDB00172 | 791 | C00407 | **View** |
| L-Phenylalanine | L-Phenylalanine | HMDB00159 | 6140 | C00079 | |
| L-Serine | L-Serine | HMDB00187 | 5951 | C00065 | |
| L-Threonine | L-Threonine | HMDB00167 | 6288 | C00188 | |
| L-Tyrosine | L-Tyrosine | HMDB00158 | 6057 | C00082 | |
| L-Valine | L-Valine | HMDB00883 | 1182 | C00183 | |
| Phenylpyruvic acid | Phenylpyruvic acid | HMDB00205 | 997 | C00166 | |
| Propionic acid | Propionic acid | HMDB00237 | 1032 | C00163 | |
| Pyruvic acid | Pyruvic acid | HMDB00243 | 1060 | C00022 | |
| Sarcosine | Sarcosine | HMDB00271 | 1088 | C00213 | |

# Name Standardization (cont.)

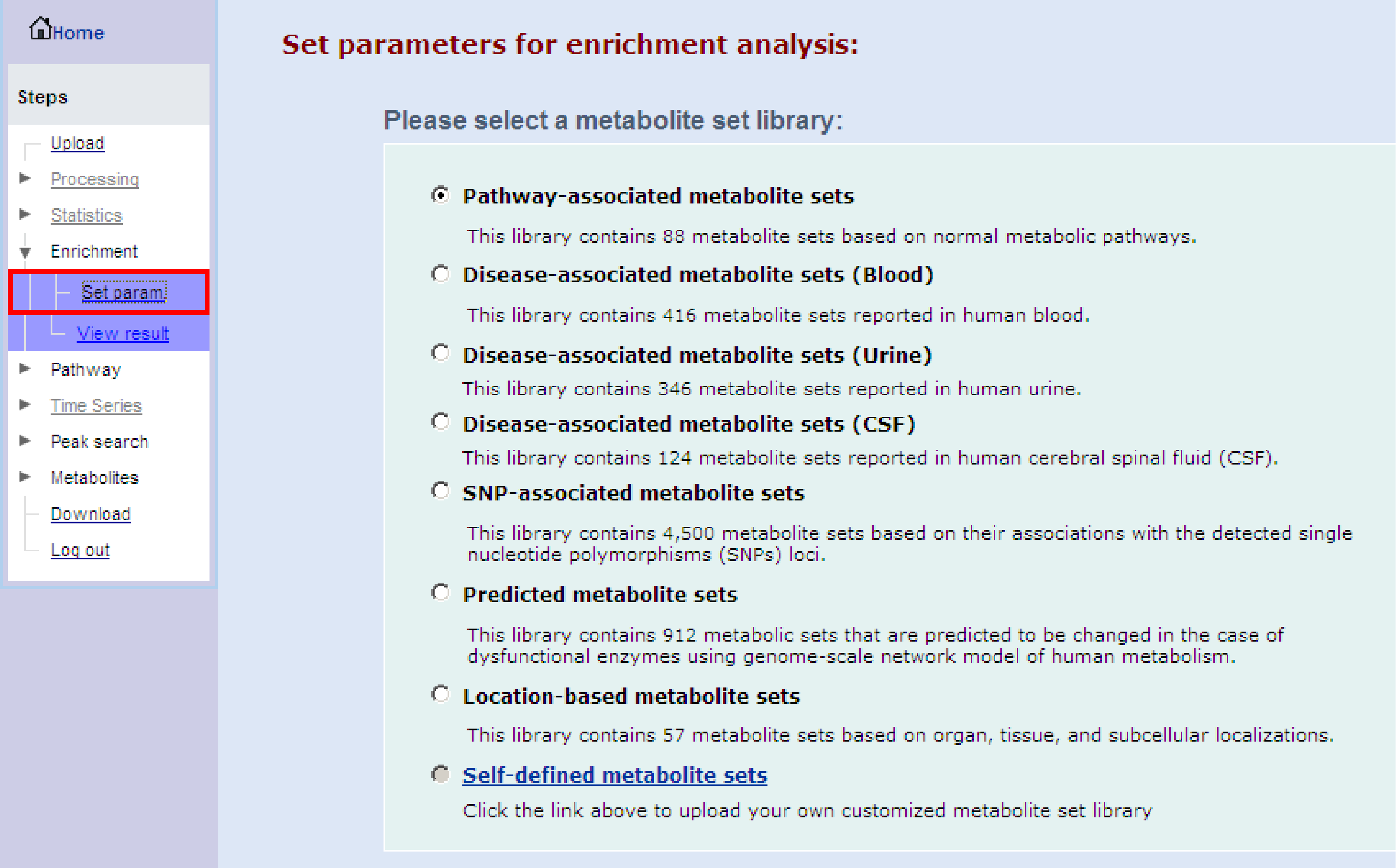


## Details

Query Name: L-Isolucine

|   | Matched Name | HMDB | PubChem | KEGG |
|---|---|---|---|---|
| ● | L-Isoleucine | HMDB00172 | 791 | C00407 |
| ○ | L-Alloisoleucine | HMDB00557 | 99288 |  |
| ○ | L-gamma-glutamyl-L-isoleucine | HMDB11170 | NA |  |
| ○ | Angiotensin IV | HMDB01038 | 123814 | C15849 |
| ○ | None of the above |  |  |  |

# Select a Metabolite Set Library



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.

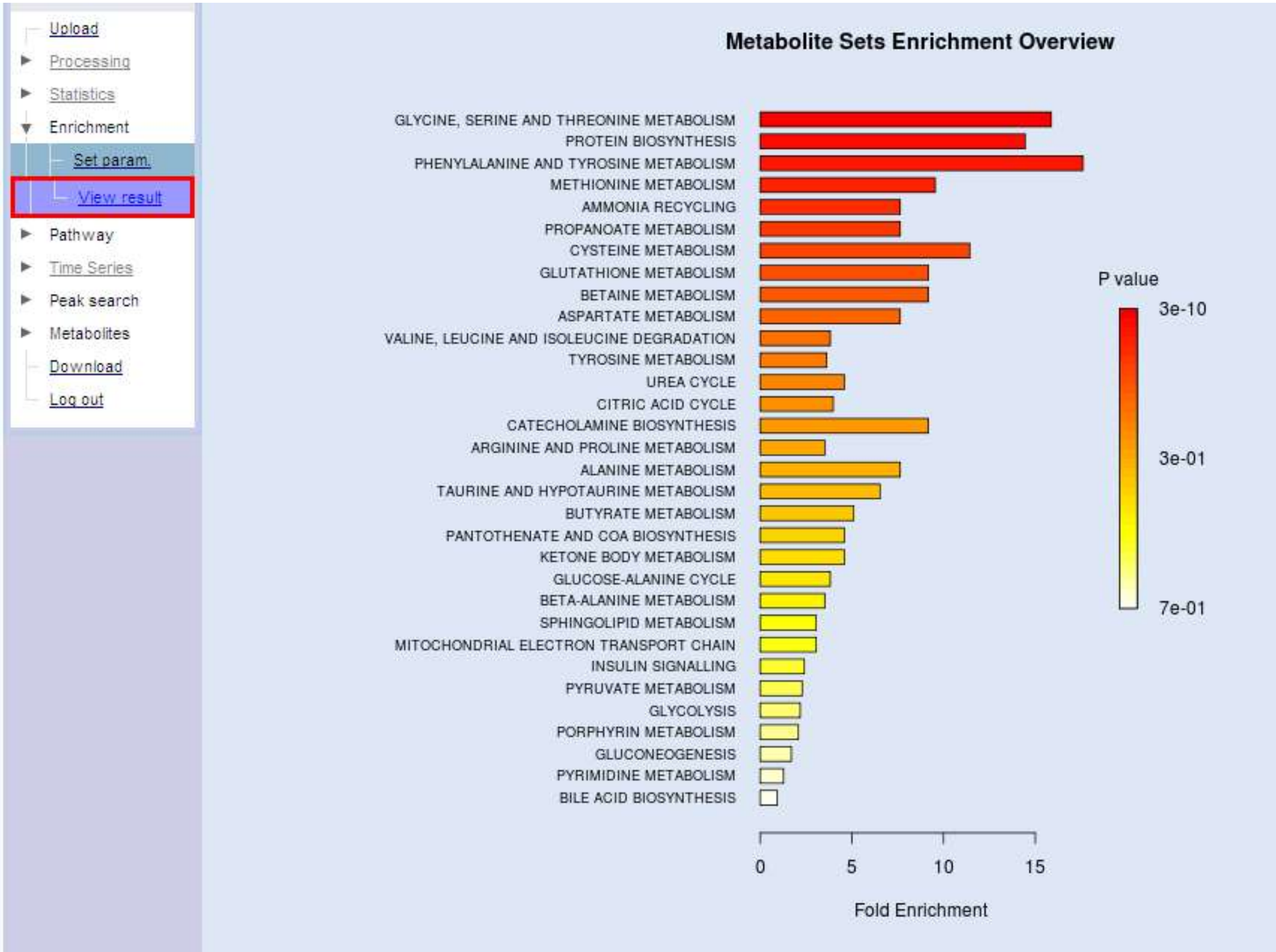
○ **Location-based metabolite sets**

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

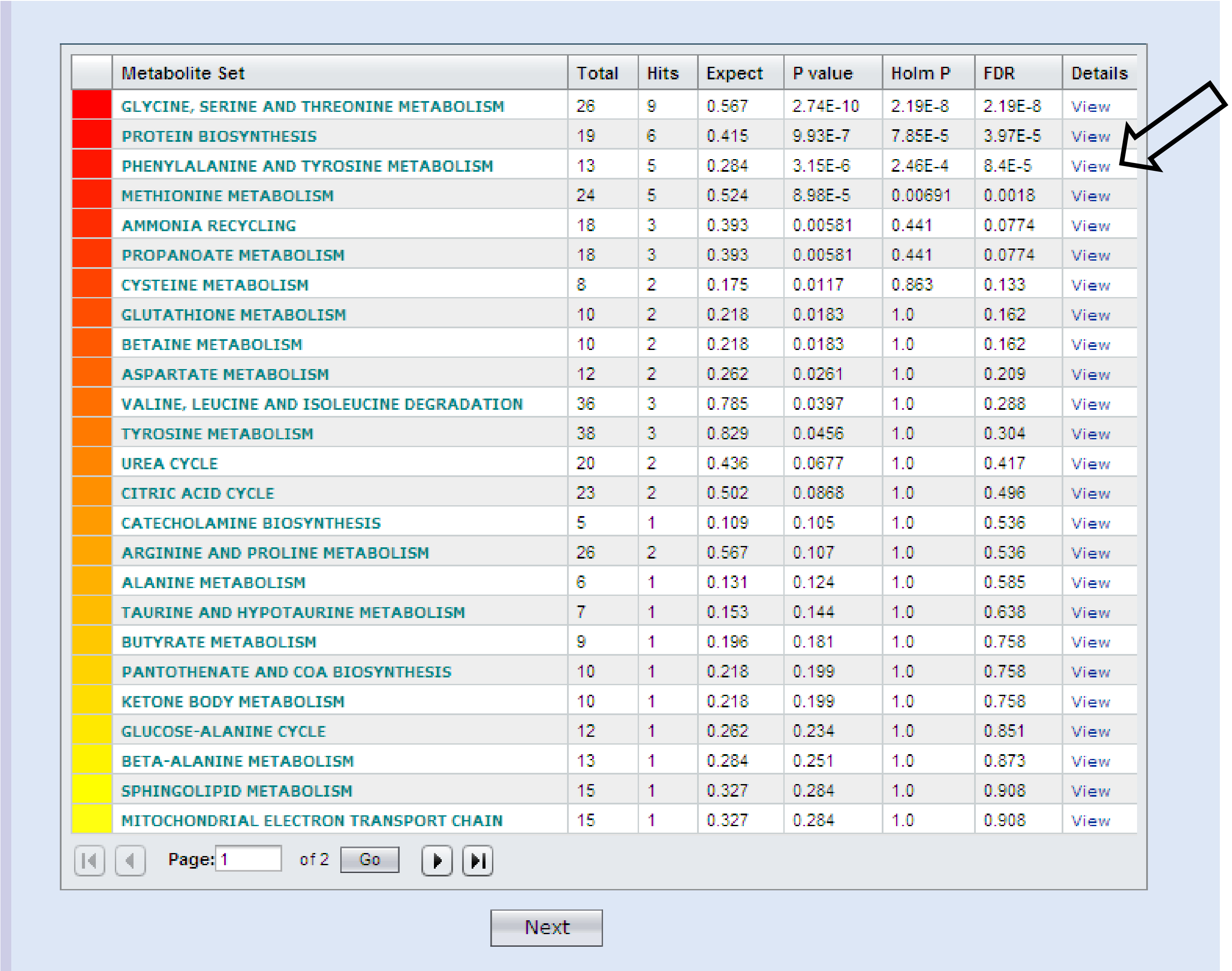
○ **Self-defined metabolite sets**

Click the link above to upload your own customized metabolite set library

# Result

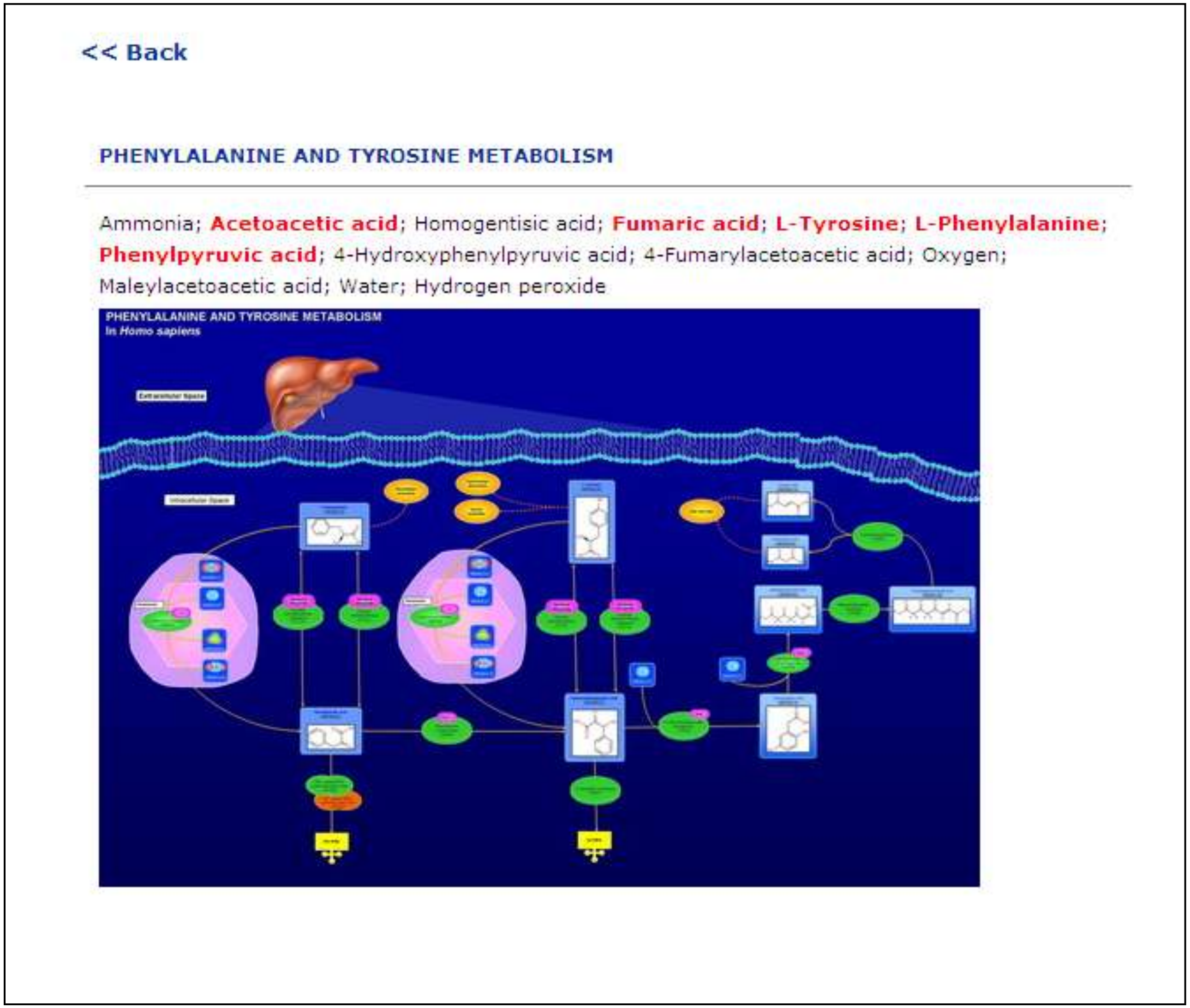

Upload
Processing
Statistics
Enrichment
Set param.
View result
Pathway
Time Series
Peak search
Metabolites
Download
Log out
Metabolite Sets Enrichment Overview
GLYCINE, SERINE AND THREONINE METABOLISM
PROTEIN BIOSYNTHESIS
PHENYLALANINE AND TYROSINE METABOLISM
METHIONINE METABOLISM
AMMONIA RECYCLING
PROPANOATE METABOLISM
CYSTEINE METABOLISM
GLUTATHIONE METABOLISM
BETAINE METABOLISM
ASPARTATE METABOLISM
VALINE, LEUCINE AND ISOLEUCINE DEGRADATION
TYROSINE METABOLISM
UREA CYCLE
CITRIC ACID CYCLE
CATECHOLAMINE BIOSYNTHESIS
ARGININE AND PROLINE METABOLISM
ALANINE METABOLISM
TAURINE AND HYPOTAURINE METABOLISM
BUTYRATE METABOLISM
PANTOTHENATE AND COA BIOSYNTHESIS
KETONE BODY METABOLISM
GLUCOSE-ALANINE CYCLE
BETA-ALANINE METABOLISM
SPHINGOLIPID METABOLISM
MITOCHONDRIAL ELECTRON TRANSPORT CHAIN
INSULIN SIGNALLING
PYRUVATE METABOLISM
GLYCOLYSIS
PORPHYRIN METABOLISM
GLUCONEOGENESIS
PYRIMIDINE METABOLISM
BILE ACID BIOSYNTHESIS
0
5
10
15
Fold Enrichment
P value
3e-10
3e-01
7e-01

# Result (cont.)



|  | 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 | View |
| 🟥 | PROTEIN BIOSYNTHESIS | 19 | 6 | 0.415 | 9.93E-7 | 7.85E-5 | 3.97E-5 | View |
| 🟥 | PHENYLALANINE AND TYROSINE METABOLISM | 13 | 5 | 0.284 | 3.15E-6 | 2.46E-4 | 8.4E-5 | View |
| 🟥 | METHIONINE METABOLISM | 24 | 5 | 0.524 | 8.98E-5 | 0.00691 | 0.0018 | View |
| 🟥 | AMMONIA RECYCLING | 18 | 3 | 0.393 | 0.00581 | 0.441 | 0.0774 | View |
| 🟥 | PROPANOATE METABOLISM | 18 | 3 | 0.393 | 0.00581 | 0.441 | 0.0774 | View |
| 🟥 | CYSTEINE METABOLISM | 8 | 2 | 0.175 | 0.0117 | 0.863 | 0.133 | View |
| 🟥 | GLUTATHIONE METABOLISM | 10 | 2 | 0.218 | 0.0183 | 1.0 | 0.162 | View |
| 🟧 | BETAINE METABOLISM | 10 | 2 | 0.218 | 0.0183 | 1.0 | 0.162 | View |
| 🟧 | ASPARTATE METABOLISM | 12 | 2 | 0.262 | 0.0261 | 1.0 | 0.209 | View |
| 🟧 | VALINE, LEUCINE AND ISOLEUCINE DEGRADATION | 36 | 3 | 0.785 | 0.0397 | 1.0 | 0.288 | View |
| 🟧 | TYROSINE METABOLISM | 38 | 3 | 0.829 | 0.0456 | 1.0 | 0.304 | View |
| 🟧 | UREA CYCLE | 20 | 2 | 0.436 | 0.0677 | 1.0 | 0.417 | View |
| 🟧 | CITRIC ACID CYCLE | 23 | 2 | 0.502 | 0.0868 | 1.0 | 0.496 | View |
| 🟧 | CATECHOLAMINE BIOSYNTHESIS | 5 | 1 | 0.109 | 0.105 | 1.0 | 0.536 | View |
| 🟧 | ARGININE AND PROLINE METABOLISM | 26 | 2 | 0.567 | 0.107 | 1.0 | 0.536 | View |
| 🟨 | ALANINE METABOLISM | 6 | 1 | 0.131 | 0.124 | 1.0 | 0.585 | View |
| 🟨 | TAURINE AND HYPOTAURINE METABOLISM | 7 | 1 | 0.153 | 0.144 | 1.0 | 0.638 | View |
| 🟨 | BUTYRATE METABOLISM | 9 | 1 | 0.196 | 0.181 | 1.0 | 0.758 | View |
| 🟨 | PANTOTHENATE AND COA BIOSYNTHESIS | 10 | 1 | 0.218 | 0.199 | 1.0 | 0.758 | View |
| 🟨 | KETONE BODY METABOLISM | 10 | 1 | 0.218 | 0.199 | 1.0 | 0.758 | View |
| 🟨 | GLUCOSE-ALANINE CYCLE | 12 | 1 | 0.262 | 0.234 | 1.0 | 0.851 | View |
| 🟨 | BETA-ALANINE METABOLISM | 13 | 1 | 0.284 | 0.251 | 1.0 | 0.873 | View |
| 🟨 | SPHINGOLIPID METABOLISM | 15 | 1 | 0.327 | 0.284 | 1.0 | 0.908 | View |
| 🟨 | MITOCHONDRIAL ELECTRON TRANSPORT CHAIN | 15 | 1 | 0.327 | 0.284 | 1.0 | 0.908 | View |

# The Matched Metabolite Set

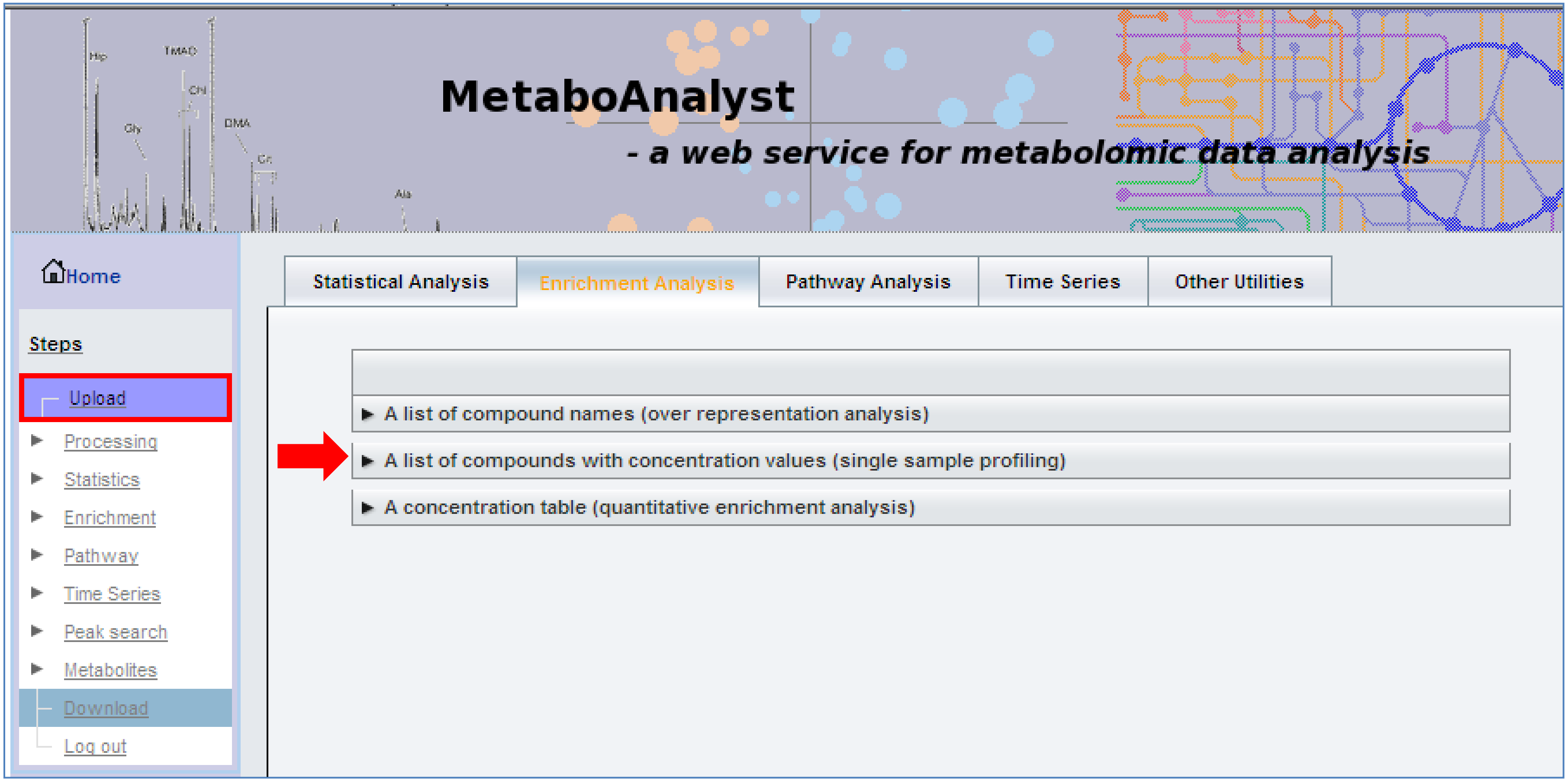


<< Back

PHENYLALANINE AND TYROSINE METABOLISM

Ammonia; **Acetoacetic acid**; Homogentisic acid; **Fumaric acid**; **L-Tyrosine**; **L-Phenylalanine**; **Phenylpyruvic acid**; 4-Hydroxyphenylpyruvic acid; 4-Fumarylacetoacetic acid; Oxygen; Maleylacetoacetic acid; Water; Hydrogen peroxide

# Single Sample Profiling

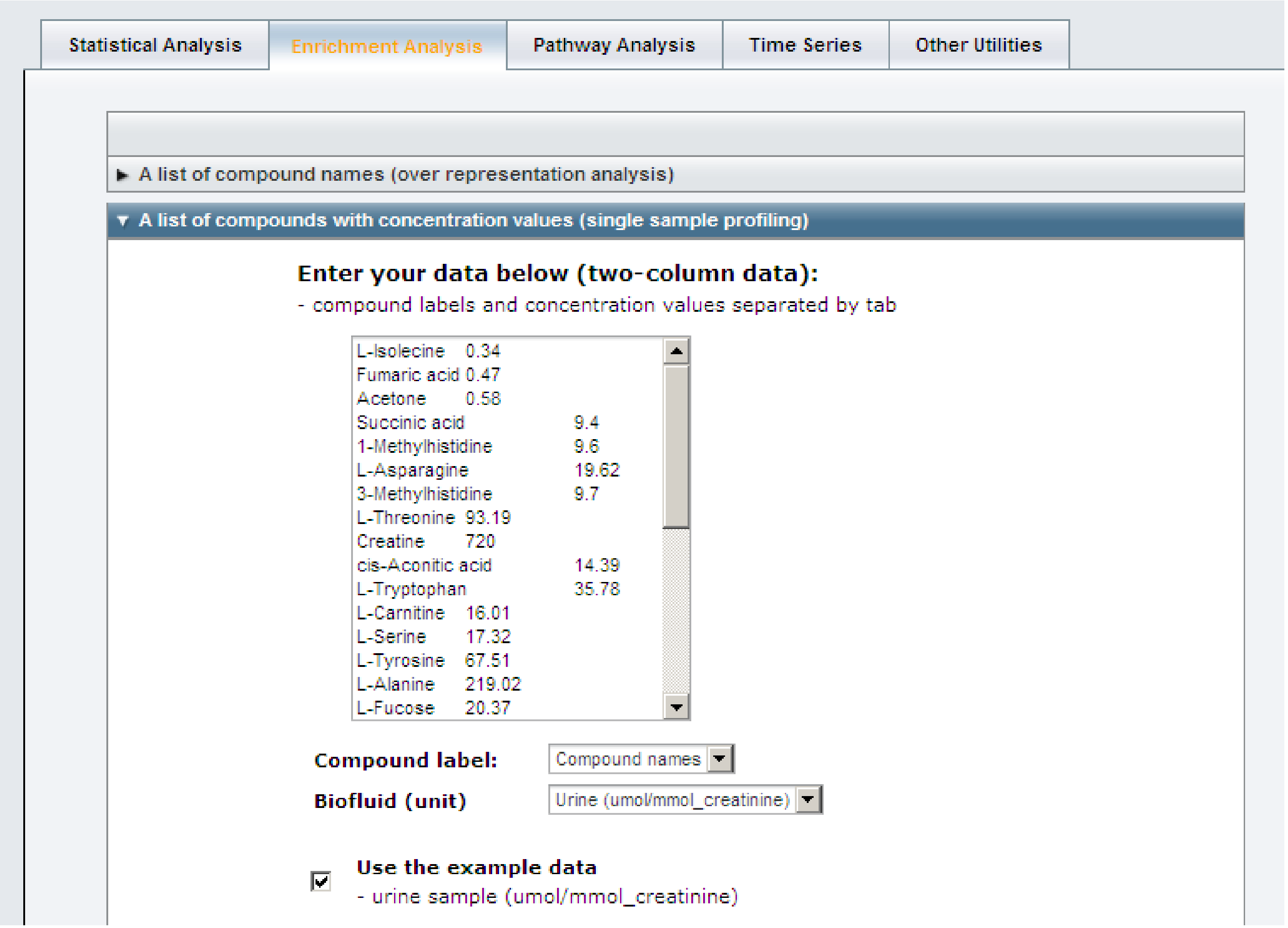


▶ A list of compound names (over representation analysis)

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

▶ A concentration table (quantitative enrichment analysis)

# Single Sample Profiling (cont.)



▶ 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):**

- compound labels and concentration values separated by tab

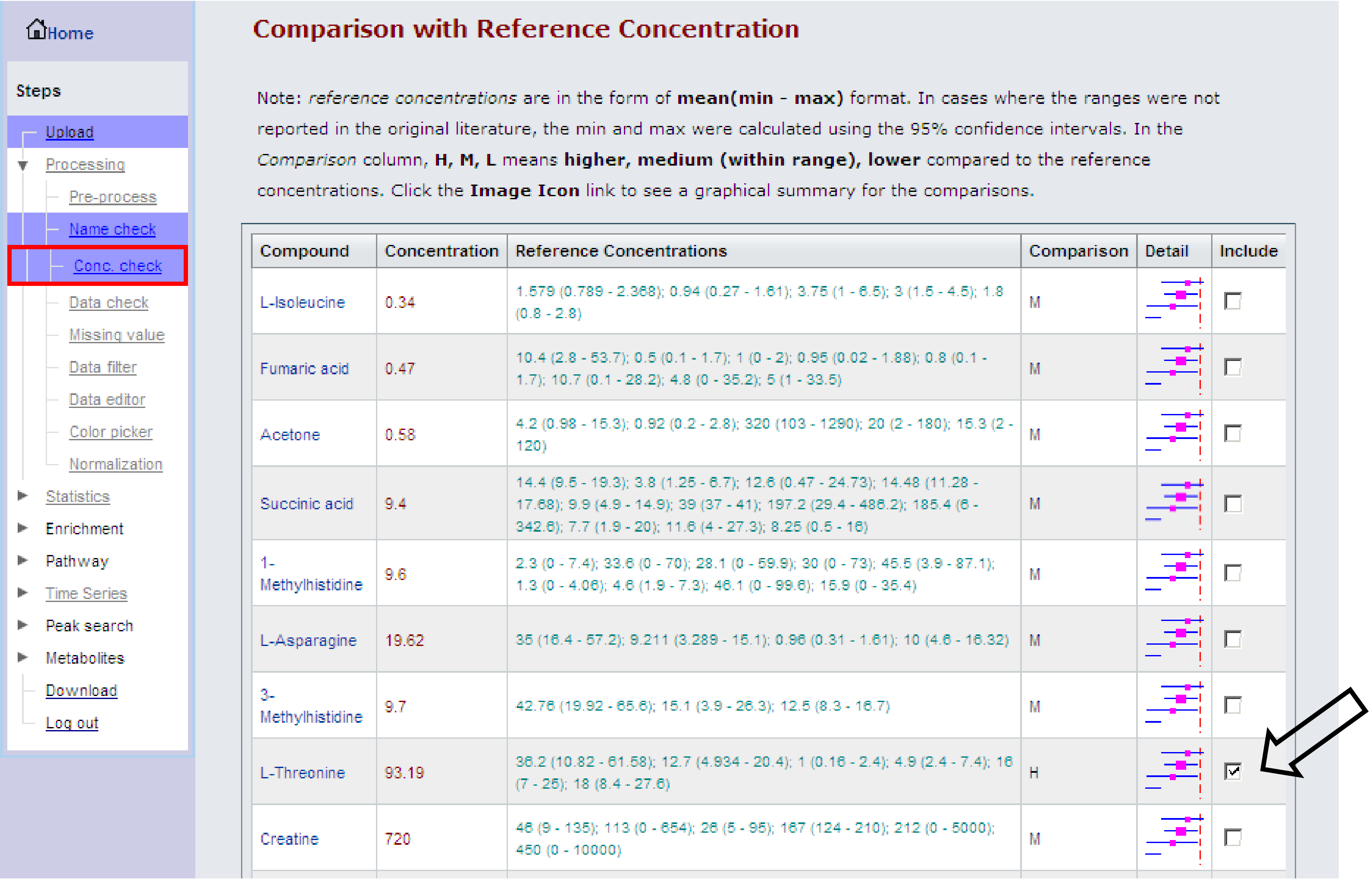
| | |
|---|---|
| L-Isolecine | 0.34 |
| Fumaric acid | 0.47 |
| Acetone | 0.58 |
| Succinic acid | 9.4 |
| 1-Methylhistidine | 9.6 |
| L-Asparagine | 19.62 |
| 3-Methylhistidine | 9.7 |
| L-Threonine | 93.19 |
| Creatine | 720 |
| cis-Aconitic acid | 14.39 |
| L-Tryptophan | 35.78 |
| L-Carnitine |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

**Compound label:** Compound names

**Biofluid (unit)** Urine (umol/mmol_creatinine)

☑ **Use the example data**
- urine sample (umol/mmol_creatinine)

# Concentration Comparison

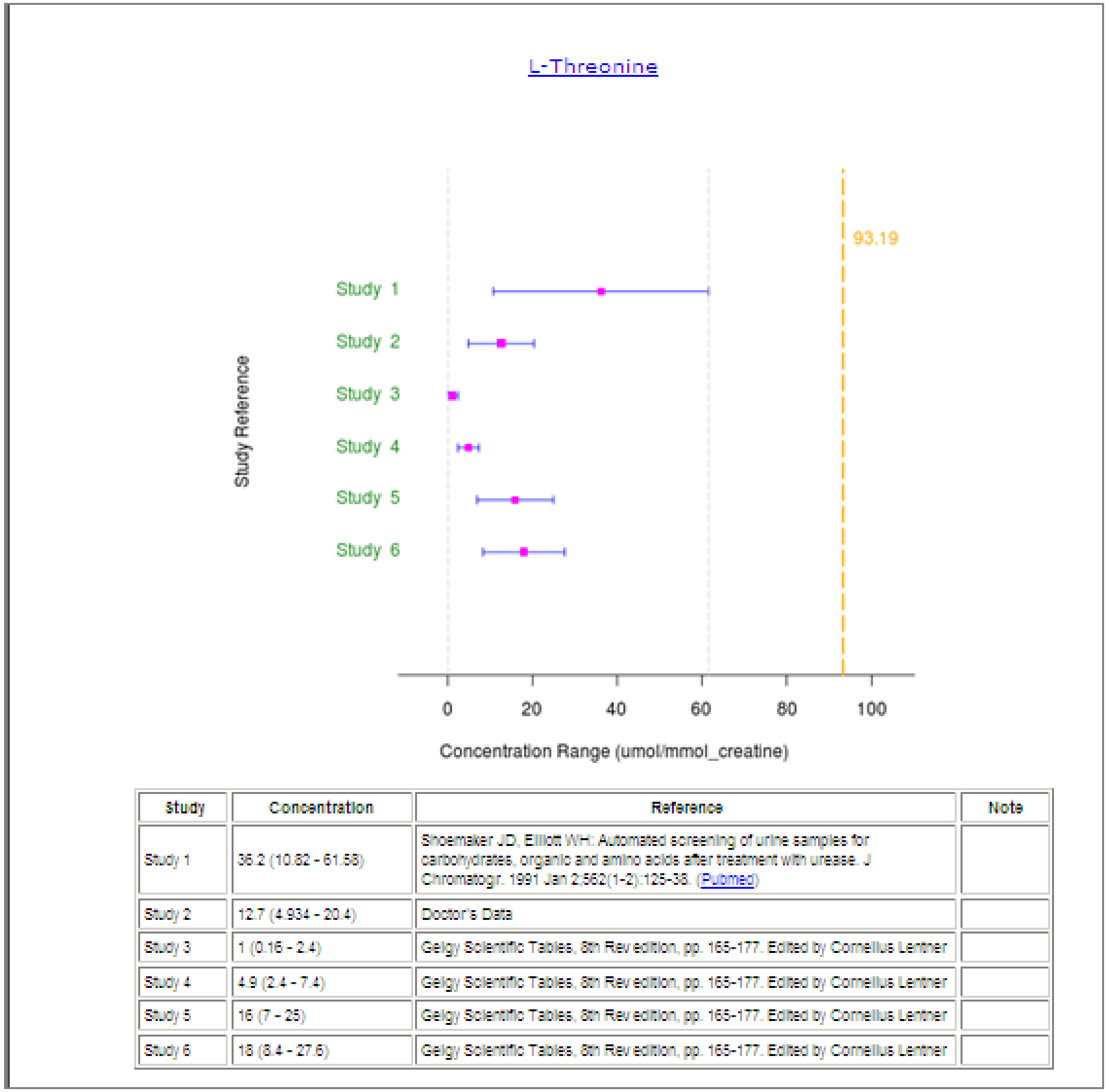


## Comparison with Reference Concentration

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 |
|---|---|---|---|---|---|
| L-Isoleucine | 0.34 | 1.579 (0.789 - 2.368); 0.94 (0.27 - 1.61); 3.75 (1 - 6.5); 3 (1.5 - 4.5); 1.8 (0.8 - 2.8) | M | | ☐ |
| Fumaric acid | 0.47 | 10.4 (2.8 - 53.7); 0.5 (0.1 - 1.7); 1 (0 - 2); 0.95 (0.02 - 1.88); 0.8 (0.1 - 1.7); 10.7 (0.1 - 28.2); 4.8 (0 - 35.2); 5 (1 - 33.5) | M | | ☐ |
| Acetone | 0.58 | 4.2 (0.98 - 15.3); 0.92 (0.2 - 2.8); 320 (103 - 1290); 20 (2 - 180); 15.3 (2 - 120) | M | | ☐ |
| Succinic acid | 9.4 | 14.4 (9.5 - 19.3); 3.8 (1.25 - 6.7); 12.6 (0.47 - 24.73); 14.48 (11.28 - 17.68); 9.9 (4.9 - 14.9); 39 (37 - 41); 197.2 (29.4 - 486.2); 185.4 (6 - 342.6); 7.7 (1.9 - 20); 11.6 (4 - 27.3); 8.25 (0.5 - 16) | M | | ☐ |
| 1-Methylhistidine | 9.6 | 2.3 (0 - 7.4); 33.6 (0 - 70); 28.1 (0 - 59.9); 30 (0 - 73); 45.5 (3.9 - 87.1); 1.3 (0 - 4.06); 4.6 (1.9 - 7.3); 46.1 (0 - 99.6); 15.9 (0 - 35.4) | M | | ☐ |
| L-Asparagine | 19.62 | 35 (16.4 - 57.2); 9.211 (3.289 - 15.1); 0.96 (0.31 - 1.61); 10 (4.6 - 16.32) | M | | ☐ |
| 3-Methylhistidine | 9.7 | 42.76 (19.92 - 65.6); 15.1 (3.9 - 26.3); 12.5 (8.3 - 16.7) | M | | ☐ |
| L-Threonine | 93.19 | 36.2 (10.82 - 61.58); 12.7 (4.934 - 20.4); 1 (0.16 - 2.4); 4.9 (2.4 - 7.4); 16 (7 - 25); 18 (8.4 - 27.6) | H | | ☑ |
| Creatine | 720 | 46 (9 - 135); 113 (0 - 654); 26 (5 - 95); 167 (124 - 210); 212 (0 - 5000); 450 (0 - 10000) | M | | ☐ |

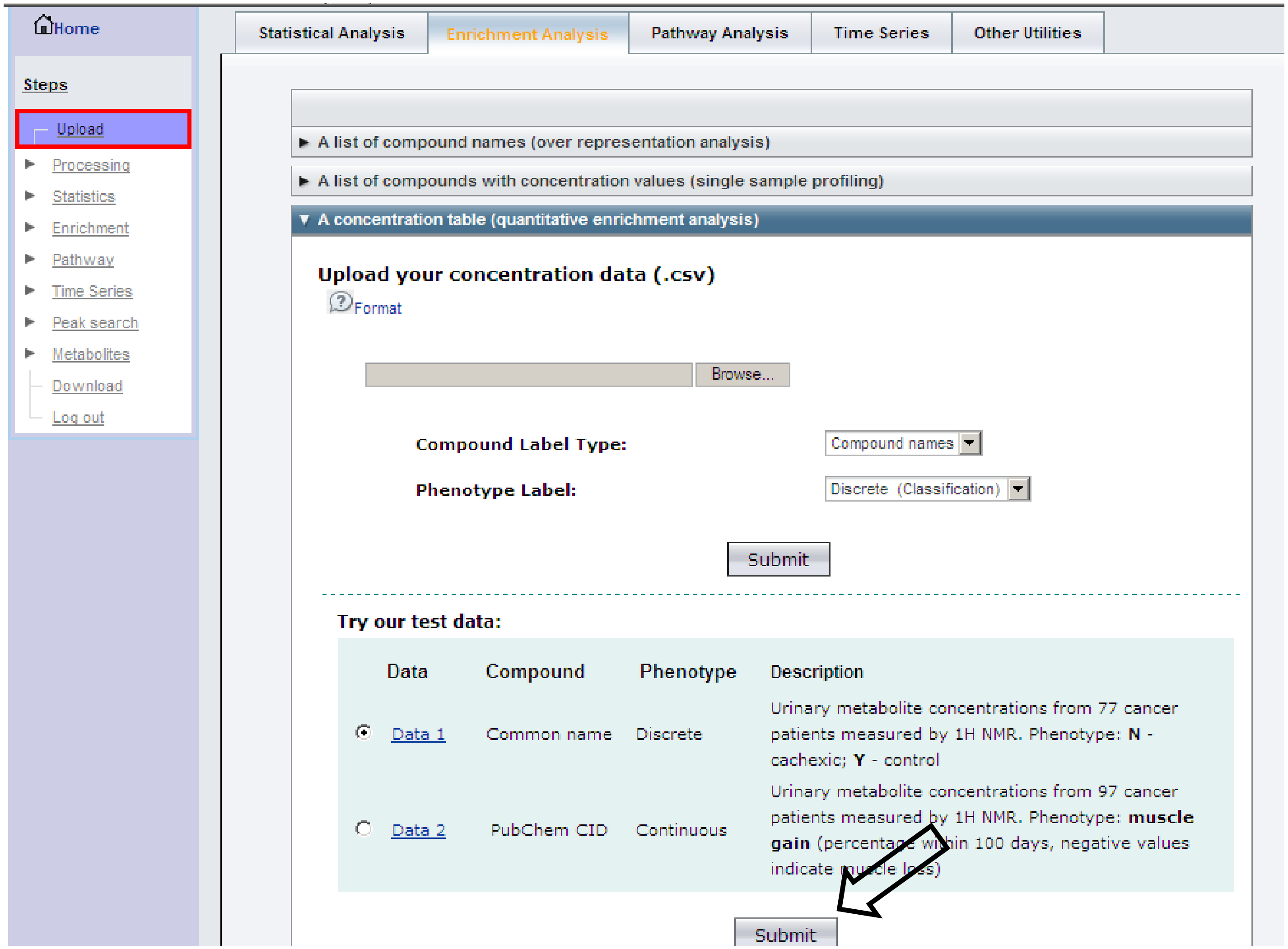
# Concentration Comparison (cont.)



L-Threonine

| Study | Concentration | Reference | Note |
|---|---|---|---|
| Study 1 | 36.2 (10.82 - 61.58) | Shoemaker JD, Elliott WH: Automated screening of urine samples for carbohydrates, organic and amino acids after treatment with urease. J Chromatogr. 1991 Jan 2;562(1-2):125-38. (Pubmed) | |
| Study 2 | 12.7 (4.934 - 20.4) | Doctor's Data | |
| Study 3 | 1 (0.16 - 2.4) | Geigy Scientific Tables, 8th Rev edition, pp. 165-177. Edited by Cornelius Lentner | |
| Study 4 | 4.9 (2.4 - 7.4) | Geigy Scientific Tables, 8th Rev edition, pp. 165-177. Edited by Cornelius Lentner | |
| Study 5 | 16 (7 - 25) | Geigy Scientific Tables, 8th Rev edition, pp. 165-177. Edited by Cornelius Lentner | |
| Study 6 | 18 (8.4 - 27.6) | Geigy Scientific Tables, 8th Rev edition, pp. 165-177. Edited by Cornelius Lentner | |

# Quantitative Enrichment Analysis



Upload your concentration data (.csv)

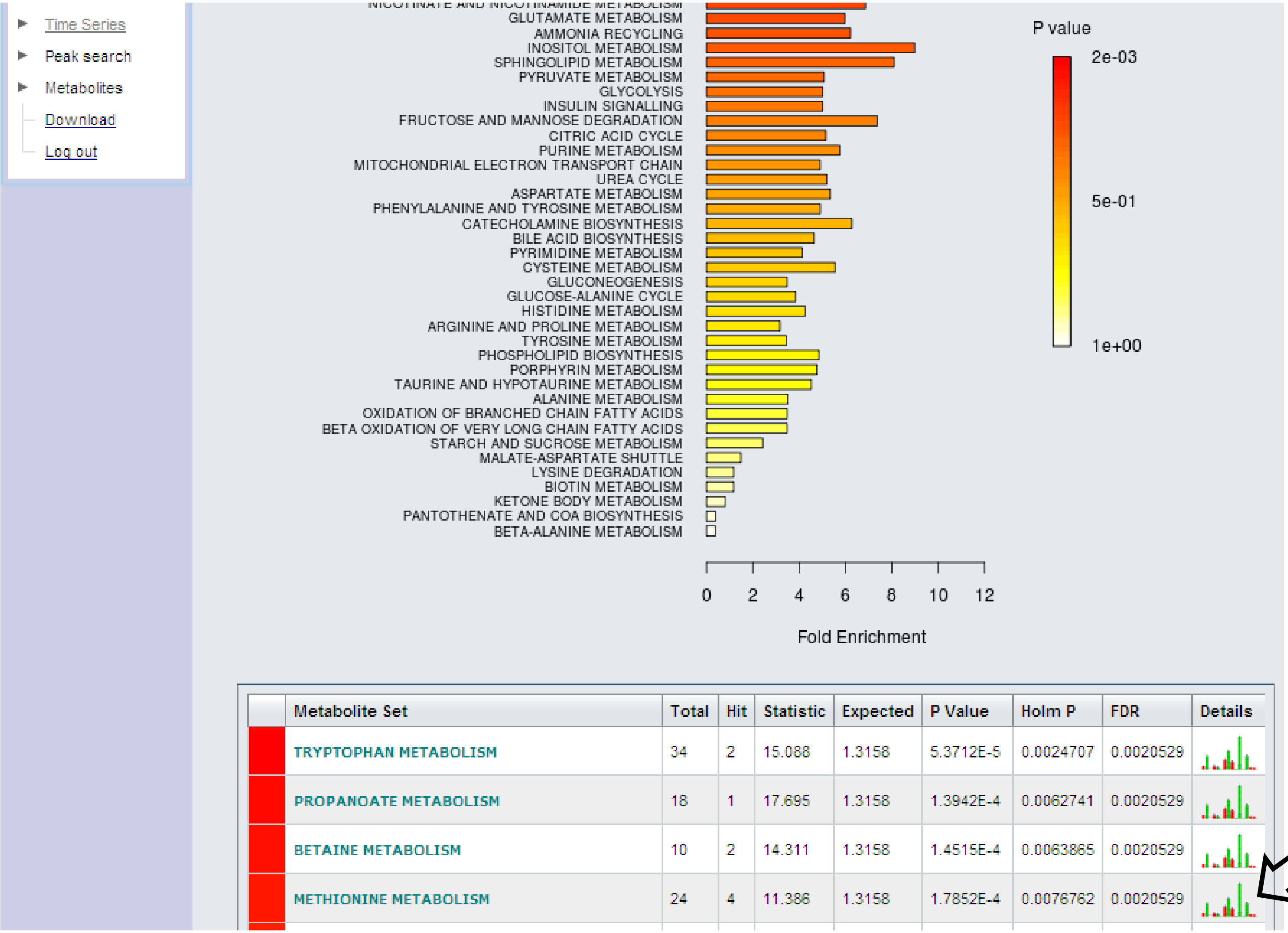
**Compound Label Type:** Compound names

**Phenotype Label:** Discrete (Classification)

Try our test data:

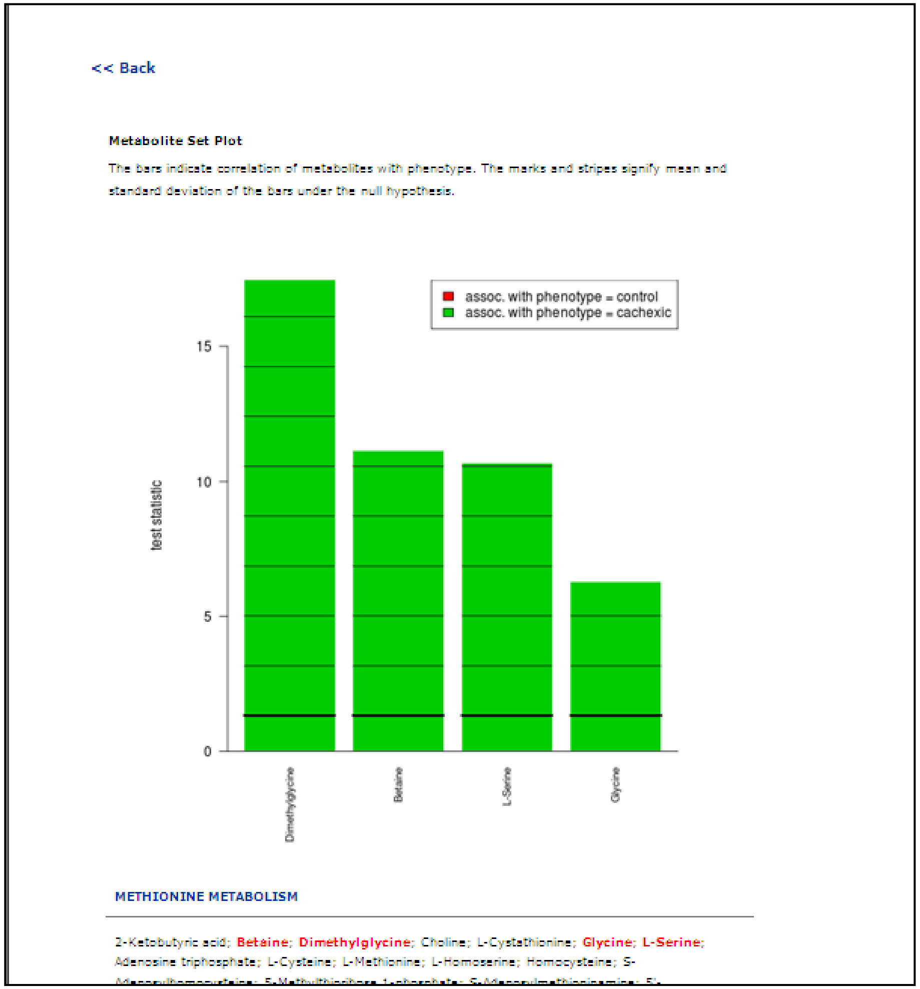
| | Data | Compound | Phenotype | Description |
|---|---|---|---|---|
| ● | Data 1 | Common name | Discrete | Urinary metabolite concentrations from 77 cancer patients measured by 1H NMR. Phenotype: **N** - cachexic; **Y** - control |
| ○ | Data 2 | PubChem CID | Continuous | Urinary metabolite concentrations from 97 cancer patients measured by 1H NMR. Phenotype: **muscle gain** (percentage within 100 days, negative values indicate muscle loss) |

# Result



| | Metabolite Set | Total | Hit | Statistic | Expected | P Value | Holm P | FDR | Details |
|---|---|---|---|---|---|---|---|---|---|
| 🟥 | TRYPTOPHAN METABOLISM | 34 | 2 | 15.088 | 1.3158 | 5.3712E-5 | 0.0024707 | 0.0020529 | |
| 🟥 | PROPANOATE METABOLISM | 18 | 1 | 17.695 | 1.3158 | 1.3942E-4 | 0.0062741 | 0.0020529 | |
| 🟥 | BETAINE METABOLISM | 10 | 2 | 14.311 | 1.3158 | 1.4515E-4 | 0.0063865 | 0.0020529 | |
| 🟥 | METHIONINE METABOLISM | 24 | 4 | 11.386 | 1.3158 | 1.7852E-4 | 0.0076762 | 0.0020529 | |

# The Matched Metabolite Set



<< Back

**Metabolite Set Plot**

The bars indicate correlation of metabolites with phenotype. The marks and stripes signify mean and standard deviation of the bars under the null hypothesis.

🟥 assoc. with phenotype = control
🟩 assoc. with phenotype = cachexic
test statistic
15
10
5
0
Dimethylglycine
Betaine
L-Serine
Glycine

**METHIONINE METABOLISM**

2-Ketobutyric acid; **Betaine**; **Dimethylglycine**; Choline; L-Cystathionine; **Glycine**; **L-Serine**; Adenosine triphosphate; L-Cysteine; L-Methionine; L-Homoserine; Homocysteine; S-Adenosylhomocysteine; 5-Methylthioribose 1-phosphate; S-Adenosylmethioninamine; 5'-

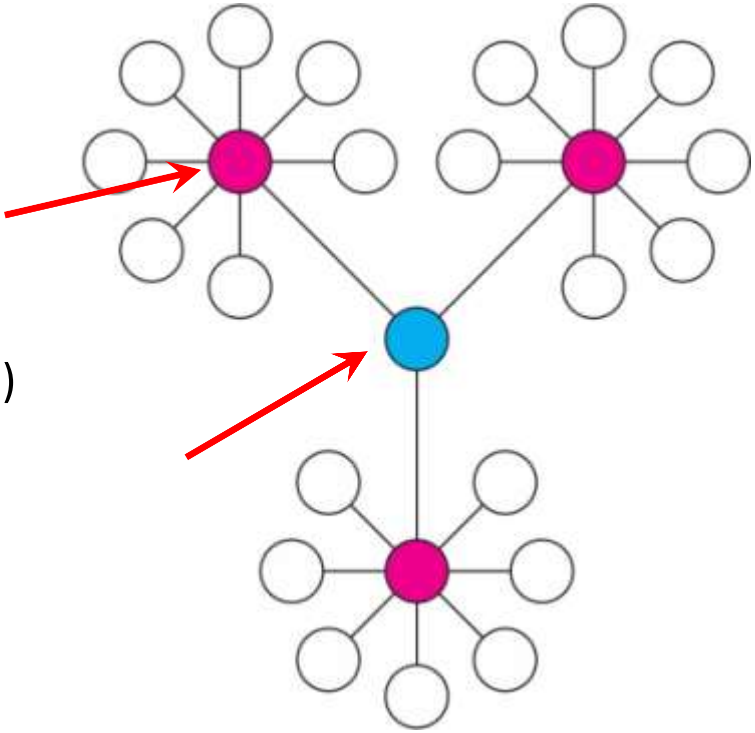
# Metabolic Pathway Analysis

# Pathway Analysis

- Purpose: to extend and enhance metabolite set enrichment analysis for pathways by
  - Considering the **structures of pathway**
  - Dynamic pathway visualization
- Currently supports ~1500 pathways covering 17 organisms (based on KEGG)

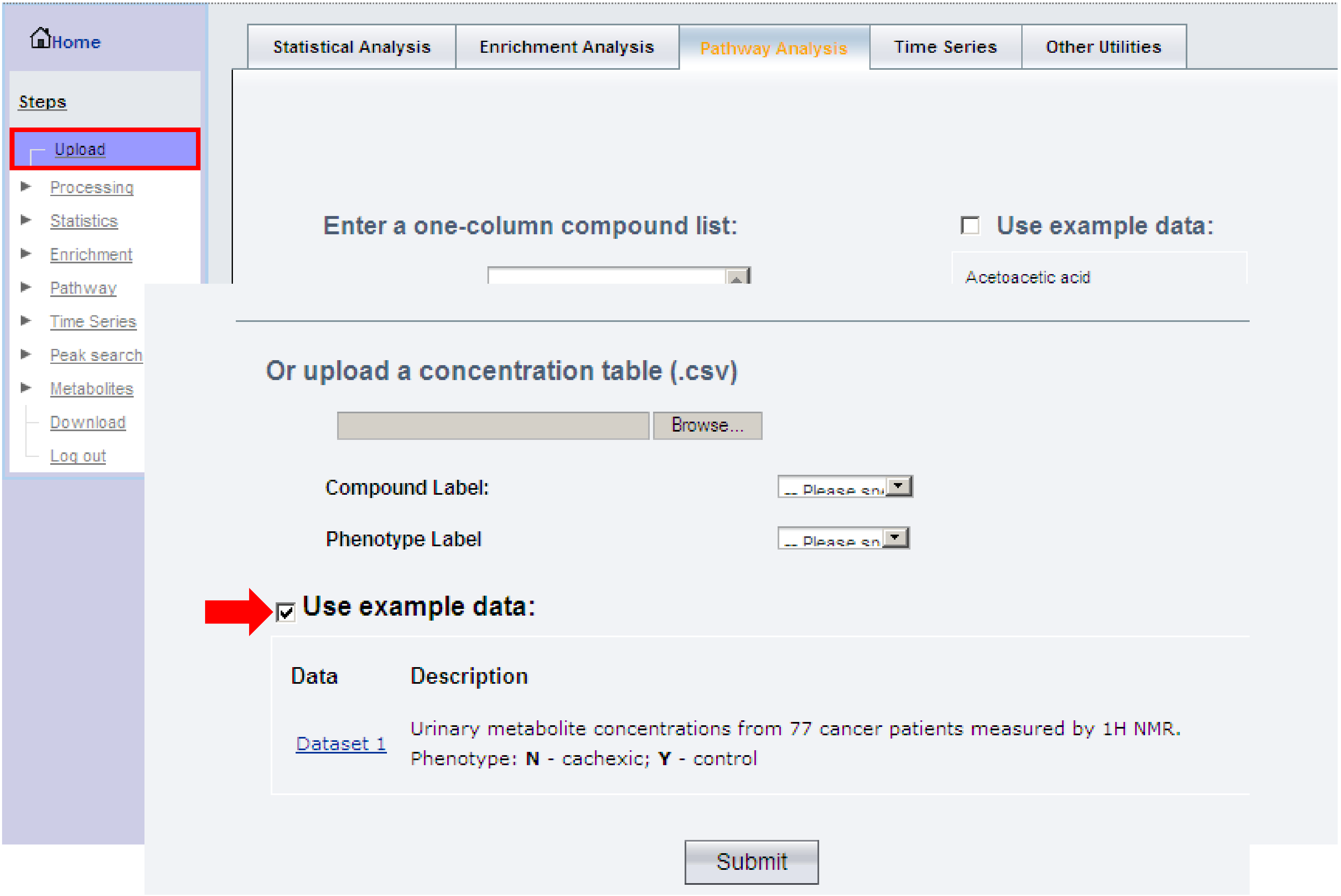
# Network Topology

- Which positions are important?
  - Hubs
    - Nodes that are highly connected (red ones)
  - Bottlenecks
    - Nodes on many shortest paths between other nodes (blue ones)
- Graph theory
  - Degree centrality
  - Betweenness centrality



Junker *et al. BMC Bioinformatics* 2006

# Data Upload



Enter a one-column compound list:

□ Use example data:

Acetoacetic acid

Or upload a concentration table (.csv)

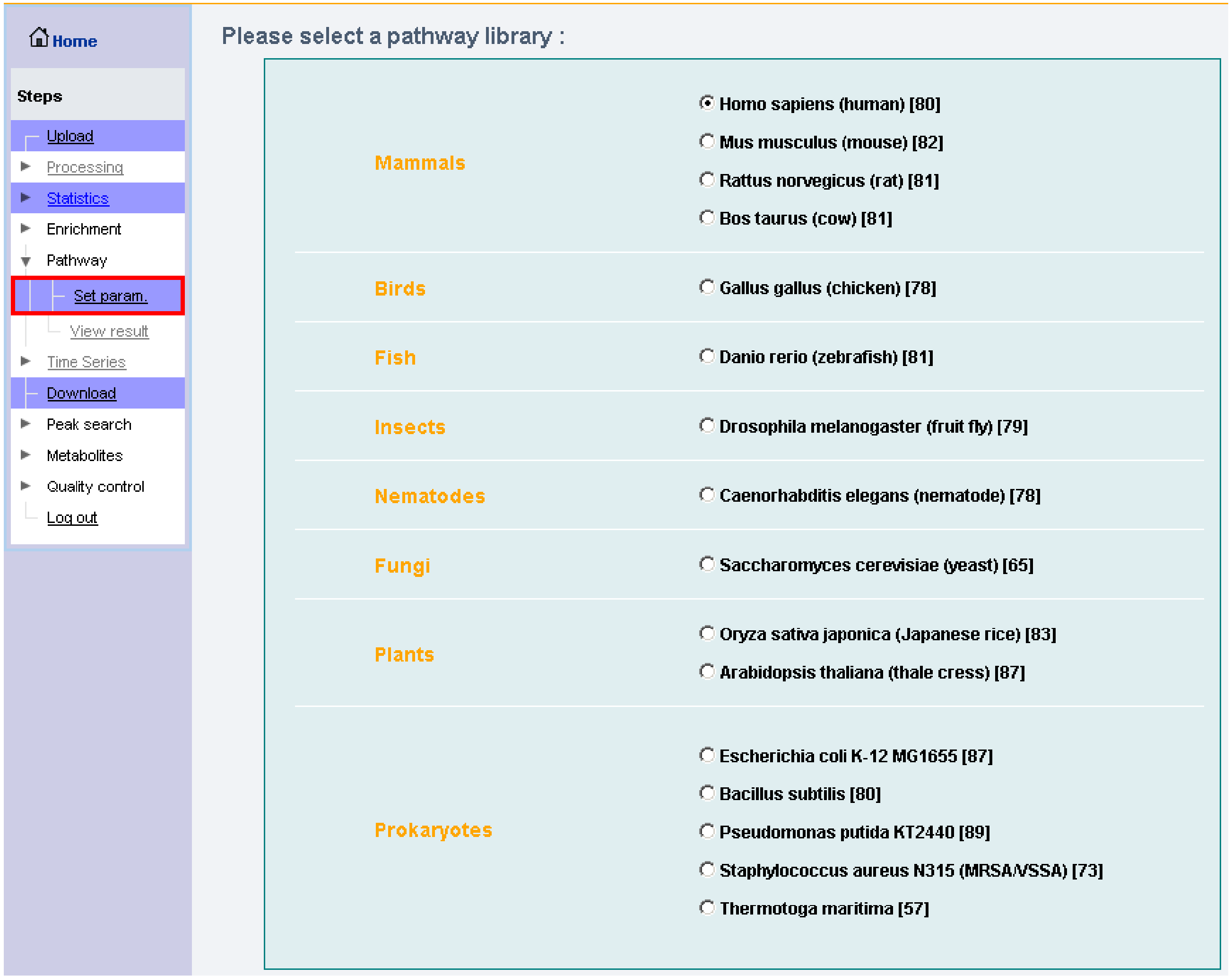
Compound Label: -- Please sp

Phenotype Label -- Please sp

☑ Use example data:

| Data | Description |
|---|---|
| Dataset 1 | Urinary metabolite concentrations from 77 cancer patients measured by 1H NMR.<br>Phenotype: **N** - cachexic; **Y** - control |

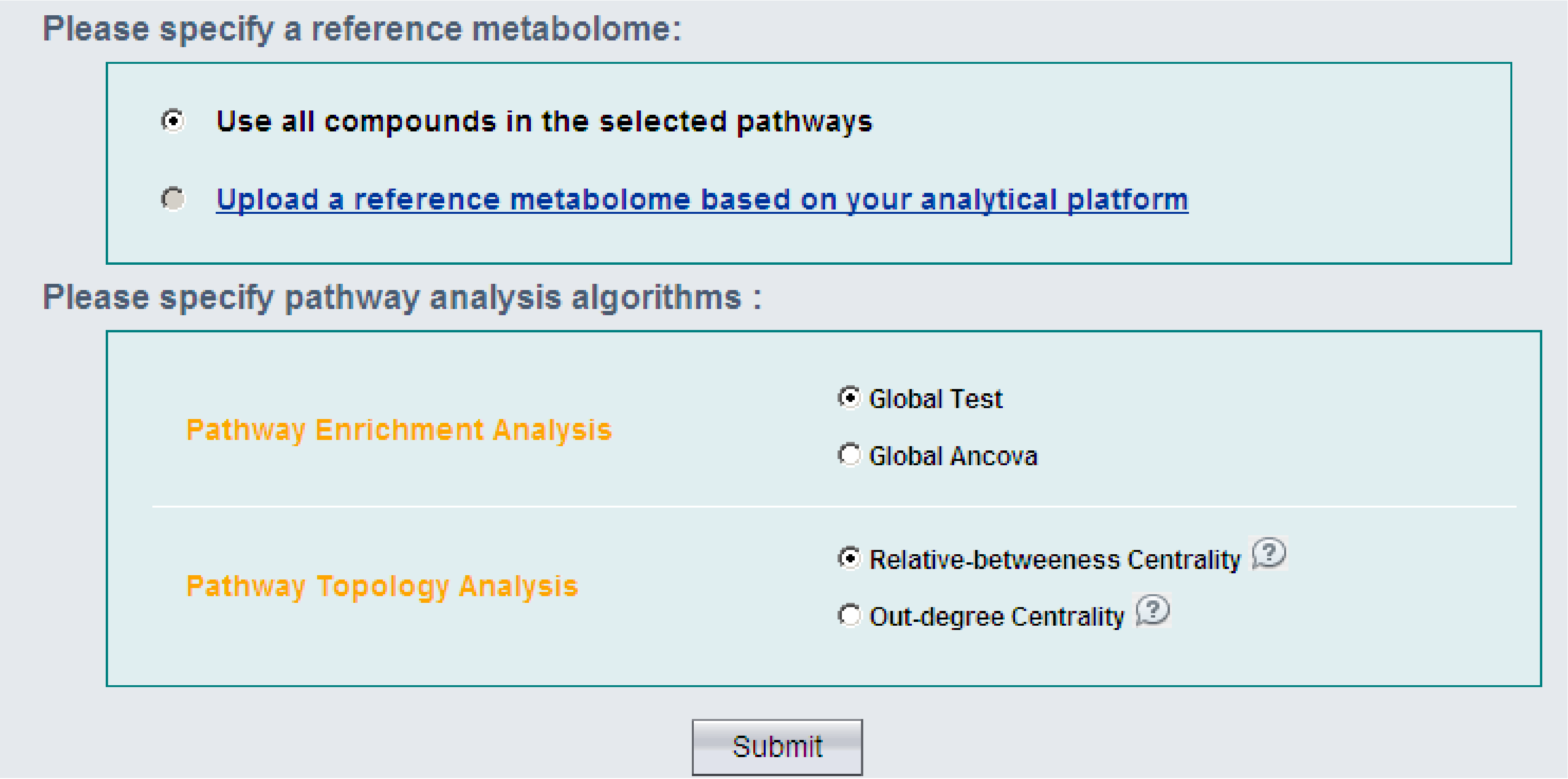
# Pathway Libraries



Please select a pathway library :

| | |
|---|---|
| Mammals | ● Homo sapiens (human) [80]<br>○ Mus musculus (mouse) [82]<br>○ Rattus norvegicus (rat) [81]<br>○ Bos taurus (cow) [81] |
| Birds | ○ Gallus gallus (chicken) [78] |
| Fish | ○ Danio rerio (zebrafish) [81] |
| Insects | ○ Drosophila melanogaster (fruit fly) [79] |
| Nematodes | ○ Caenorhabditis elegans (nematode) [78] |
| Fungi | ○ Saccharomyces cerevisiae (yeast) [65] |
| Plants | ○ Oryza sativa japonica (Japanese rice) [83]<br>○ Arabidopsis thaliana (thale cress) [87] |
| Prokaryotes | ○ Escherichia coli K-12 MG1655 [87]<br>○ Bacillus subtilis [80]<br>○ Pseudomonas putida KT2440 [89]<br>○ Staphylococcus aureus N315 (MRSA/VSSA) [73]<br>○ Thermotoga maritima [57] |

# Network Topology Analysis



Please specify a reference metabolome:

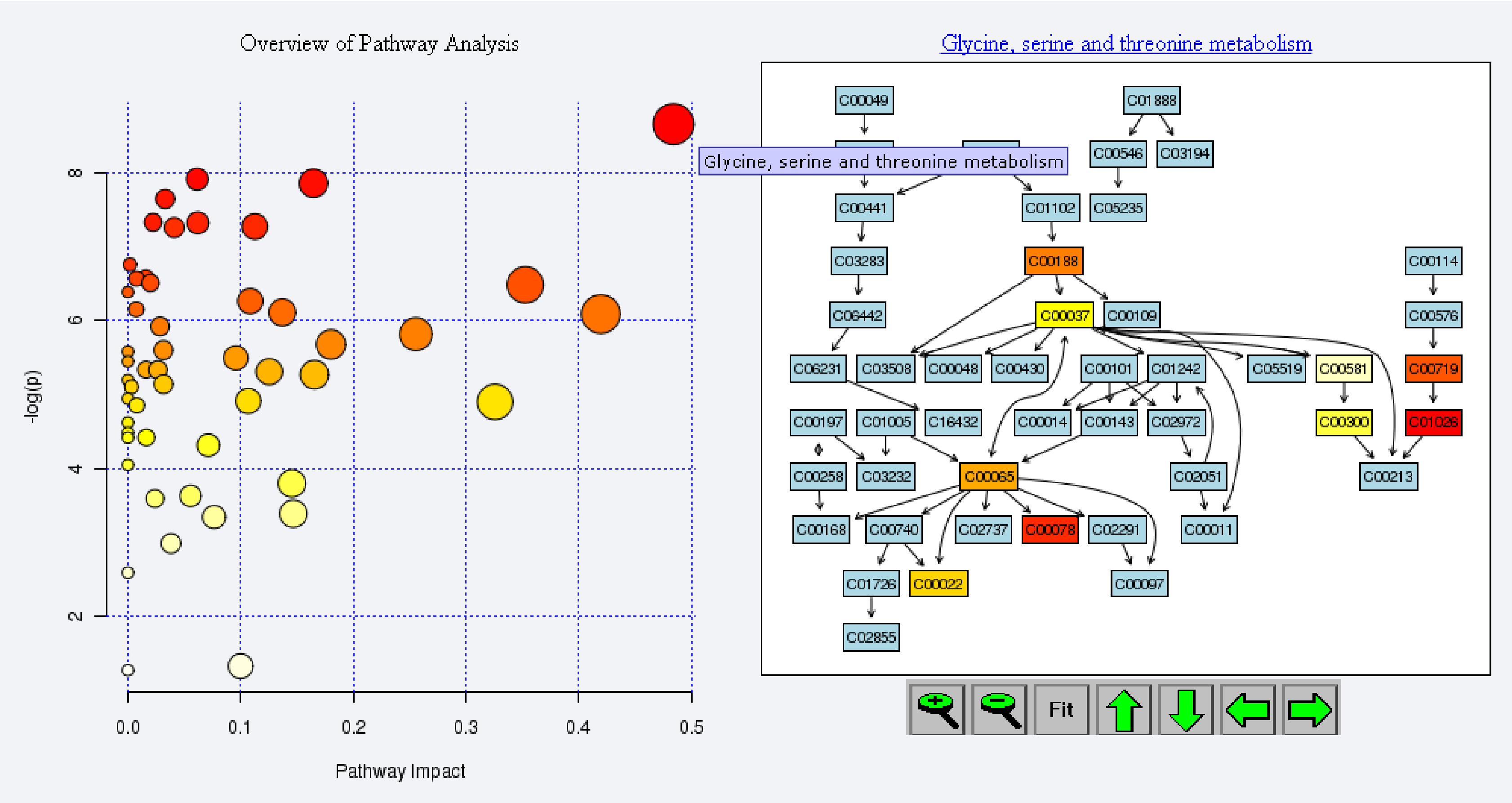
● Use all compounds in the selected pathways

○ Upload a reference metabolome based on your analytical platform

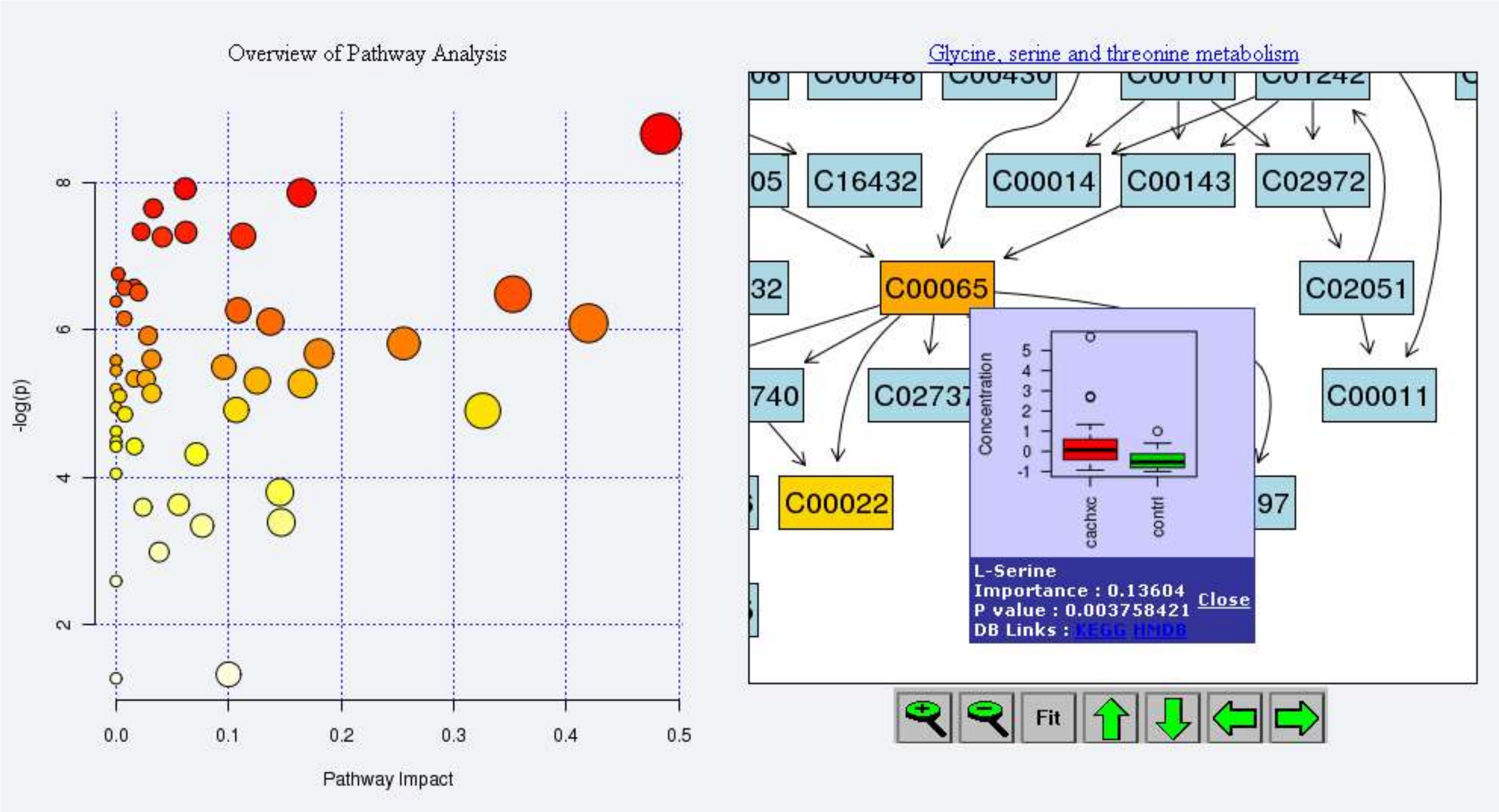
Please specify pathway analysis algorithms :

| | |
|---|---|
| Pathway Enrichment Analysis | ● Global Test<br>○ Global Ancova |
| Pathway Topology Analysis | ● Relative-betweeness Centrality ?<br>○ Out-degree Centrality ? |

# Pathway Visualization

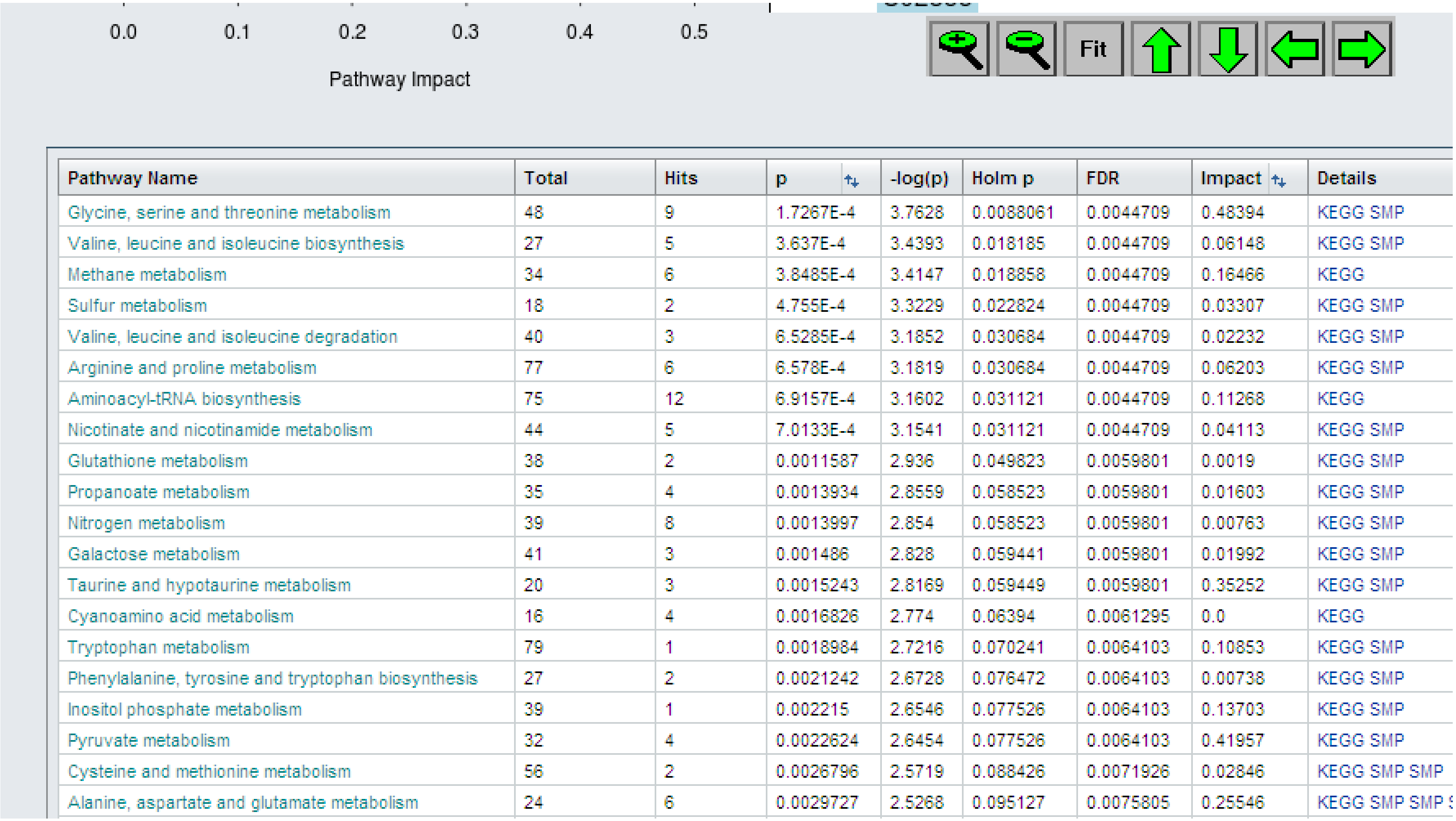


# Pathway Visualization (cont.)



L-Serine
Importance : 0.13604
P value : 0.003758421
DB Links : KEGG HMDB

# Result



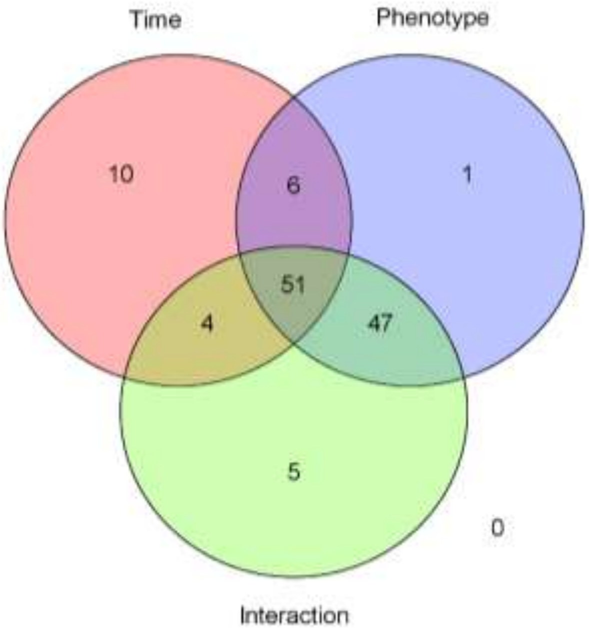
| Pathway Name | Total | Hits | p | -log(p) | Holm p | FDR | Impact | Details |
|---|---|---|---|---|---|---|---|---|
| Glycine, serine and threonine metabolism | 48 | 9 | 1.7267E-4 | 3.7628 | 0.0088061 | 0.0044709 | 0.48394 | KEGG SMP |
| Valine, leucine and isoleucine biosynthesis | 27 | 5 | 3.637E-4 | 3.4393 | 0.018185 | 0.0044709 | 0.06148 | KEGG SMP |
| Methane metabolism | 34 | 6 | 3.8485E-4 | 3.4147 | 0.018858 | 0.0044709 | 0.16466 | KEGG |
| Sulfur metabolism | 18 | 2 | 4.755E-4 | 3.3229 | 0.022824 | 0.0044709 | 0.03307 | KEGG SMP |
| Valine, leucine and isoleucine degradation | 40 | 3 | 6.5285E-4 | 3.1852 | 0.030684 | 0.0044709 | 0.02232 | KEGG SMP |
| Arginine and proline metabolism | 77 | 6 | 6.578E-4 | 3.1819 | 0.030684 | 0.0044709 | 0.06203 | KEGG SMP |
| Aminoacyl-tRNA biosynthesis | 75 | 12 | 6.9157E-4 | 3.1602 | 0.031121 | 0.0044709 | 0.11268 | KEGG |
| Nicotinate and nicotinamide metabolism | 44 | 5 | 7.0133E-4 | 3.1541 | 0.031121 | 0.0044709 | 0.04113 | KEGG SMP |
| Glutathione metabolism | 38 | 2 | 0.0011587 | 2.936 | 0.049823 | 0.0059801 | 0.0019 | KEGG SMP |
| Propanoate metabolism | 35 | 4 | 0.0013934 | 2.8559 | 0.058523 | 0.0059801 | 0.01603 | KEGG SMP |
| Nitrogen metabolism | 39 | 8 | 0.0013997 | 2.854 | 0.058523 | 0.0059801 | 0.00763 | KEGG SMP |
| Galactose metabolism | 41 | 3 | 0.001486 | 2.828 | 0.059441 | 0.0059801 | 0.01992 | KEGG SMP |
| Taurine and hypotaurine metabolism | 20 | 3 | 0.0015243 | 2.8169 | 0.059449 | 0.0059801 | 0.35252 | KEGG SMP |
| Cyanoamino acid metabolism | 16 | 4 | 0.0016826 | 2.774 | 0.06394 | 0.0061295 | 0.0 | KEGG |
| Tryptophan metabolism | 79 | 1 | 0.0018984 | 2.7216 | 0.070241 | 0.0064103 | 0.10853 | KEGG SMP |
| Phenylalanine, tyrosine and tryptophan biosynthesis | 27 | 2 | 0.0021242 | 2.6728 | 0.076472 | 0.0064103 | 0.00738 | KEGG SMP |
| Inositol phosphate metabolism | 39 | 1 | 0.002215 | 2.6546 | 0.077526 | 0.0064103 | 0.13703 | KEGG SMP |
| Pyruvate metabolism | 32 | 4 | 0.0022624 | 2.6454 | 0.077526 | 0.0064103 | 0.41957 | KEGG SMP |
| Cysteine and methionine metabolism | 56 | 2 | 0.0026796 | 2.5719 | 0.088426 | 0.0071926 | 0.02846 | KEGG SMP SMP |
| Alanine, aspartate and glutamate metabolism | 24 | 6 | 0.0029727 | 2.5268 | 0.095127 | 0.0075805 | 0.25546 | KEGG SMP SMP |

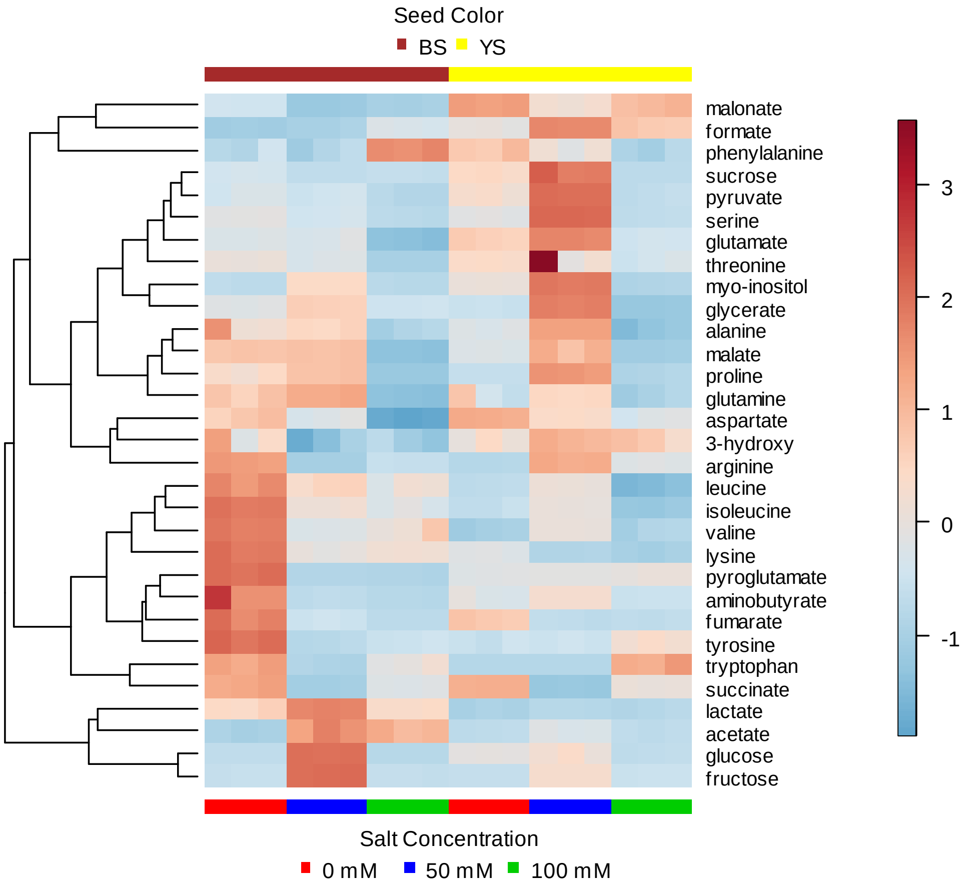
# Not Everything Was Covered

- Clustering (K-means, SOM)
- Classification (SVM, randomForests)
- Time-series data analysis
- Two -factor data analysis
- Peak searching
- ….

# Two Factor Analysis

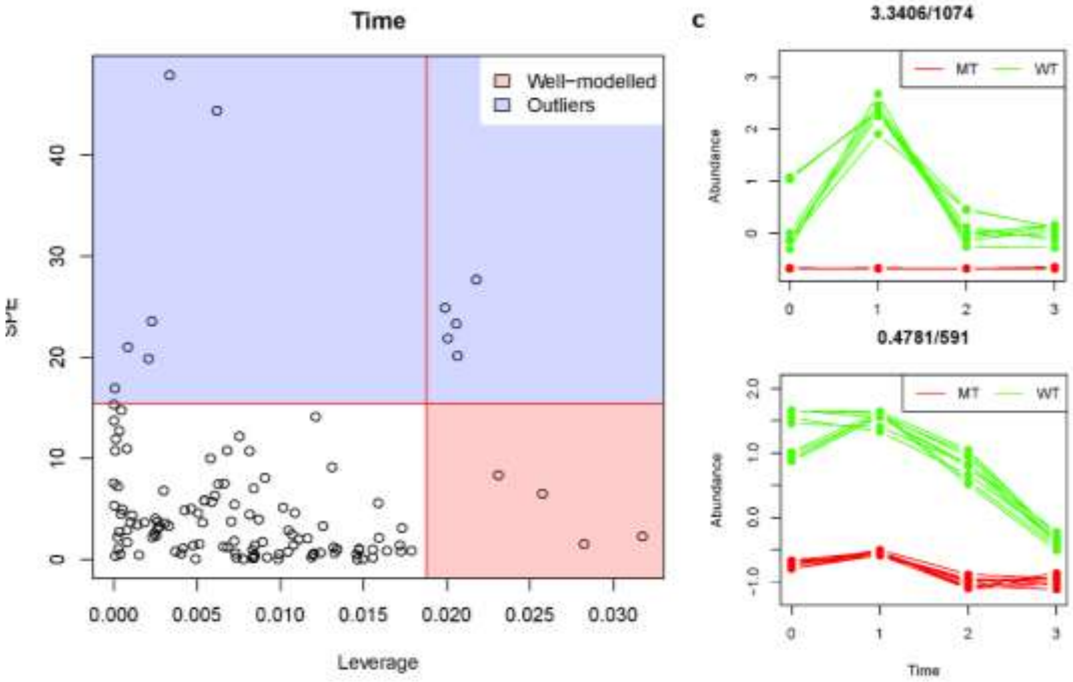
- Two – way ANOVA
- Two – way heatmap





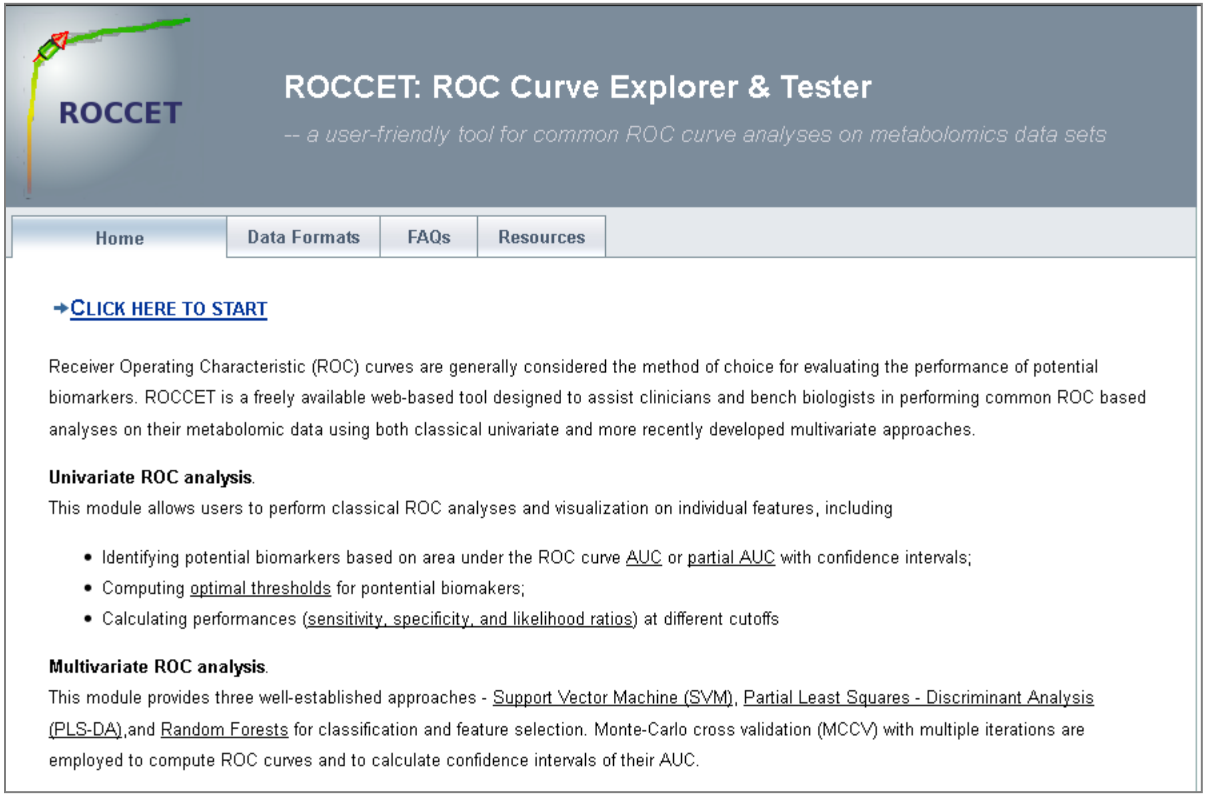
# Time series data analysis

- ANOVA-SCA
- Multivariate Empirical Bayes



# Biomarker Discovery & Performance Evaluation

# ROCCET (www.roccet.ca)



ROCCET: ROC Curve Explorer & Tester

*-- a user-friendly tool for common ROC curve analyses on metabolomics data sets*

→CLICK HERE TO START

Receiver Operating Characteristic (ROC) curves are generally considered the method of choice for evaluating the performance of potential biomarkers. ROCCET is a freely available web-based tool designed to assist clinicians and bench biologists in performing common ROC based analyses on their metabolomic data using both classical univariate and more recently developed multivariate approaches.

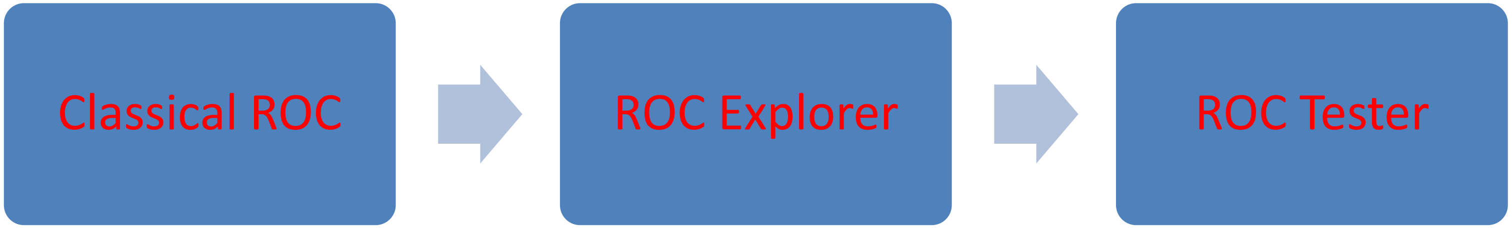
**Univariate ROC analysis**.

This module allows users to perform classical ROC analyses and visualization on individual features, including

- Identifying potential biomarkers based on area under the ROC curve AUC or partial AUC with confidence intervals;
- Computing optimal thresholds for pontential biomakers;
- Calculating performances (sensitivity, specificity, and likelihood ratios) at different cutoffs

**Multivariate ROC analysis**.

This module provides three well-established approaches - Support Vector Machine (SVM), Partial Least Squares - Discriminant Analysis (PLS-DA),and Random Forests for classification and feature selection. Monte-Carlo cross validation (MCCV) with multiple iterations are employed to compute ROC curves and to calculate confidence intervals of their AUC.



Classical ROC

ROC Explorer

ROC Tester

# Sensitivity, Specificity & ROC curve

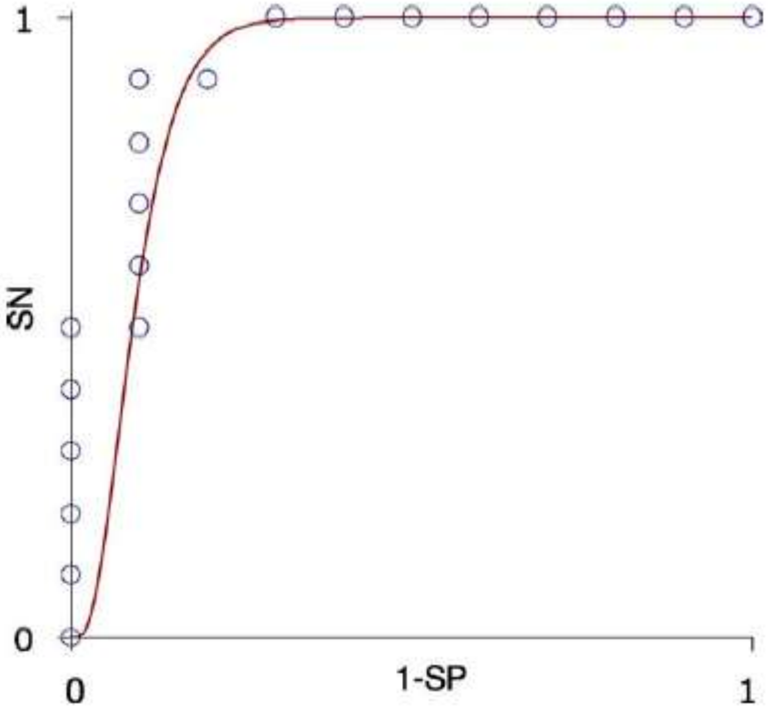
- Two important performance measures in a diagnostic tests
  - Sensitivity (true positive rate)
  - Specificity (true negative rate)
- Cutoff dependent
  - Increase cutoff, will improve specificity, decease sensitivity
- ROC curves integrate these two measures

# How to construct ROC curves

- Input: a score on a univariate scale
  - A test gives continuous value (i.e. blood *Glucose* level)
  - A classifier that produces a continuous score (i.e. likelihood, probabilities)

# Classical ROC curve

| 2-h plasma glucose (mmol/L) | |
|---|---|
| Healthy | Diseased |
| | |
| 4.86 | |
| 5.69 | |
| 6.01 | |
| 6.06 | |
| 6.27 | |
| 6.37 | |
| 6.55 | |
| 7.29 | 7.29 |
| 7.82 | |
| | 9.22 |
| | 9.79 |
| | 11.28 |
| | 11.83 |
| 12.06 | |
| | 18.48 |
| | 18.50 |
| | 20.49 |
| | 22.66 |
| | 26.01 |

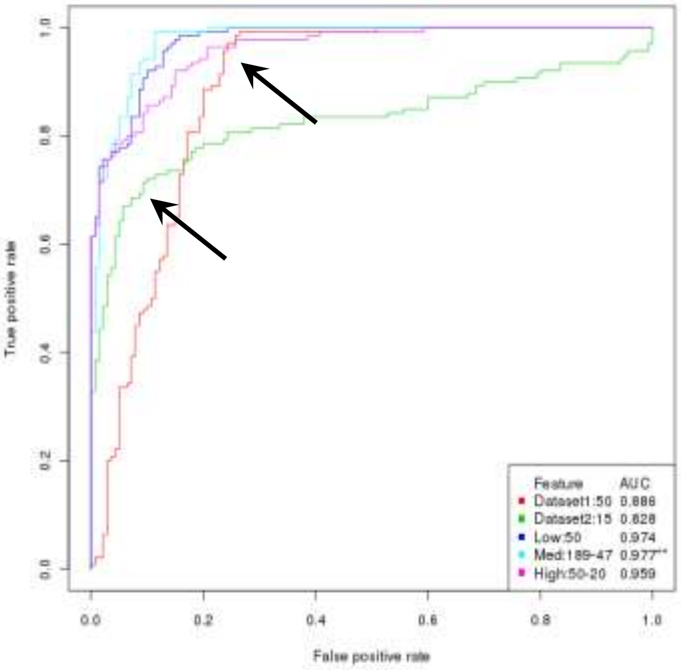


| $1 - SP$ | $SN$ |
|---|---|
| 1.00 | 1.00 |
| 0.90 | 1.00 |
| 0.80 | 1.00 |
| 0.70 | 1.00 |
| 0.60 | 1.00 |
| 0.50 | 1.00 |
| 0.40 | 1.00 |
| 0.30 | 1.00 |
| 0.20 | 0.90 |
| 0.10 | 0.90 |
| 0.10 | 0.80 |
| 0.10 | 0.70 |
| 0.10 | 0.60 |
| 0.10 | 0.50 |
| 0.00 | 0.50 |
| 0.00 | 0.40 |
| 0.00 | 0.30 |
| 0.00 | 0.20 |
| 0.00 | 0.10 |
| 0.00 | 0.00 |

*TA Lasko, et al (2005)*

# Explore ROC space

- The ROC curve itself (visualization)
- Compare different ROC curves
  - Area under the curve
    - AUC
  - When two curves cross
    - Partial AUC (pAUC)
  - Confidence Intervals
    - Empirical ROC curves are based on samples



# Understand AUC

- Area under an ROC curve (AUC)

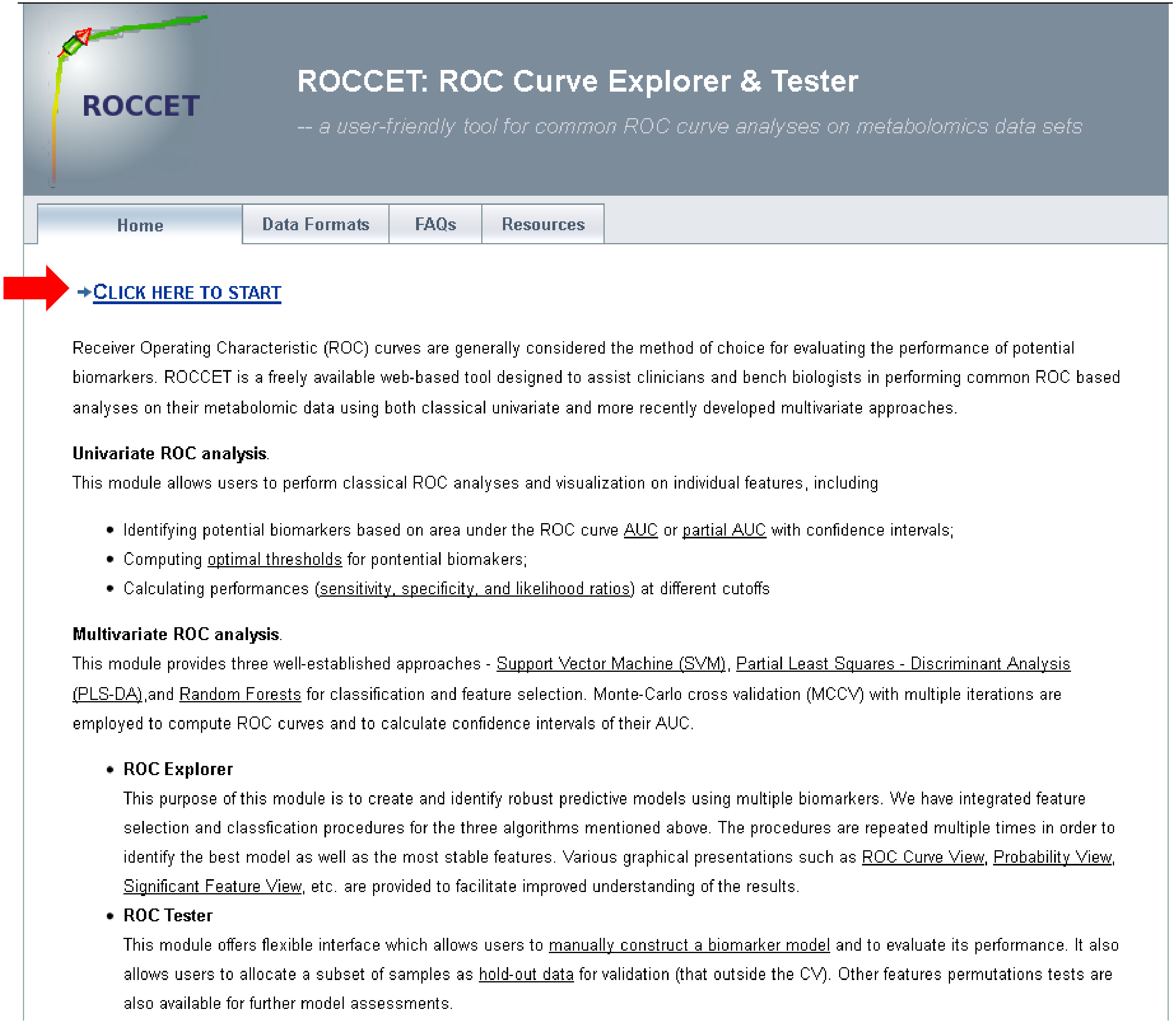
    a. The probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one

    a. The average specificity across all values of sensitivity

    b. The average sensitivity across all values of specificity

# ROCCET

- ROC curves – based biomarker discovery and performance evaluation
  - Classical ROC Curve Analysis for individual biomarker
  - Multivariate biomarker model creation & assessment (automatic / manual mode)
    - PLSDA, Linear SVM, Random Forests
  - Calculate AUC & partial AUC with confidence intervals
  - Other supporting utilities



# ROCCET: ROC Curve Explorer & Tester

*-- a user-friendly tool for common ROC curve analyses on metabolomics data sets*

→CLICK HERE TO START

Receiver Operating Characteristic (ROC) curves are generally considered the method of choice for evaluating the performance of potential biomarkers. ROCCET is a freely available web-based tool designed to assist clinicians and bench biologists in performing common ROC based analyses on their metabolomic data using both classical univariate and more recently developed multivariate approaches.

**Univariate ROC analysis**.

This module allows users to perform classical ROC analyses and visualization on individual features, including

- Identifying potential biomarkers based on area under the ROC curve AUC or partial AUC with confidence intervals;
- Computing optimal thresholds for pontential biomakers;
- Calculating performances (sensitivity, specificity, and likelihood ratios) at different cutoffs

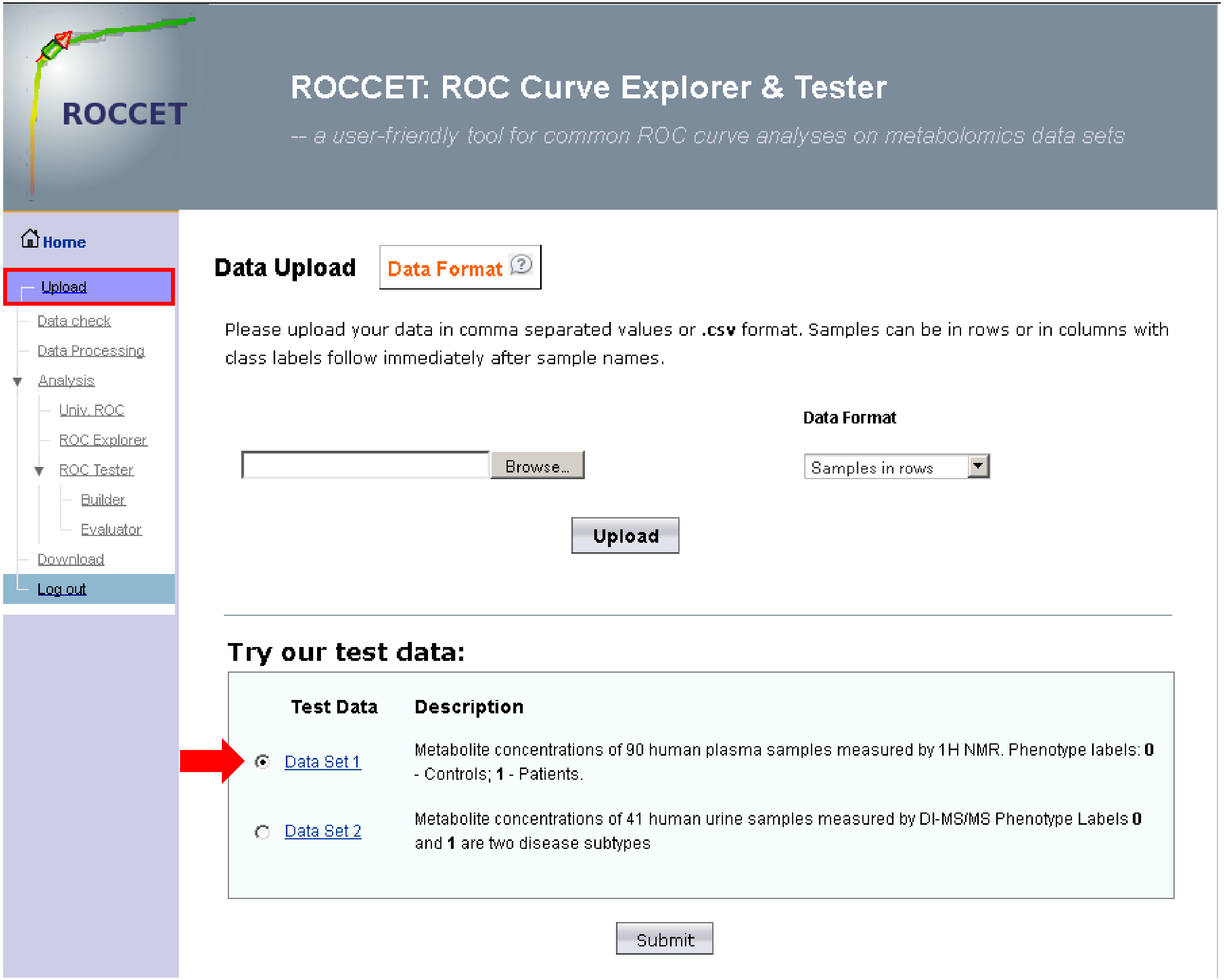
**Multivariate ROC analysis**.

This module provides three well-established approaches - Support Vector Machine (SVM), Partial Least Squares - Discriminant Analysis (PLS-DA),and Random Forests for classification and feature selection. Monte-Carlo cross validation (MCCV) with multiple iterations are employed to compute ROC curves and to calculate confidence intervals of their AUC.

- **ROC Explorer**

  This purpose of this module is to create and identify robust predictive models using multiple biomarkers. We have integrated feature selection and classfication procedures for the three algorithms mentioned above. The procedures are repeated multiple times in order to identify the best model as well as the most stable features. Various graphical presentations such as ROC Curve View, Probability View, Significant Feature View, etc. are provided to facilitate improved understanding of the results.
- **ROC Tester**

  This module offers flexible interface which allows users to manually construct a biomarker model and to evaluate its performance. It also allows users to allocate a subset of samples as hold-out data for validation (that outside the CV). Other features permutations tests are also available for further model assessments.



# ROCCET: ROC Curve Explorer & Tester

*-- a user-friendly tool for common ROC curve analyses on metabolomics data sets*

## Data Upload

Data Format

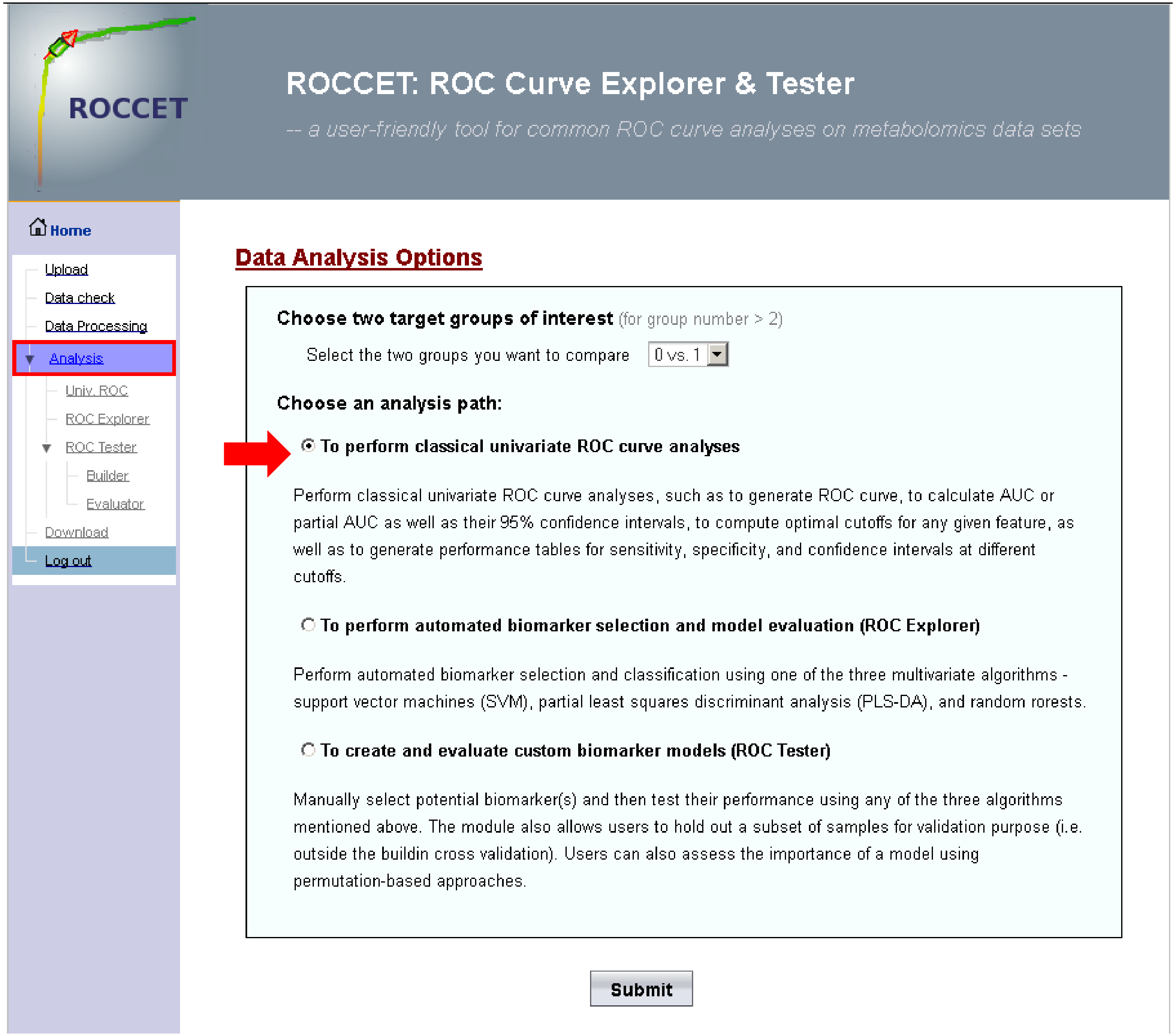
Please upload your data in comma separated values or **.csv** format. Samples can be in rows or in columns with class labels follow immediately after sample names.

**Data Format**

Samples in rows

## Try our test data:

| Test Data | Description |
|---|---|
| ● Data Set 1 | Metabolite concentrations of 90 human plasma samples measured by 1H NMR. Phenotype labels: **0** - Controls; **1** - Patients. |
| ○ Data Set 2 | Metabolite concentrations of 41 human urine samples measured by DI-MS/MS Phenotype Labels **0** and **1** are two disease subtypes |



ROCCET: ROC Curve Explorer & Tester

*-- a user-friendly tool for common ROC curve analyses on metabolomics data sets*

## Data Analysis Options

**Choose two target groups of interest** (for group number > 2)

Select the two groups you want to compare 0 vs. 1

**Choose an analysis path:**

● **To perform classical univariate ROC curve analyses**

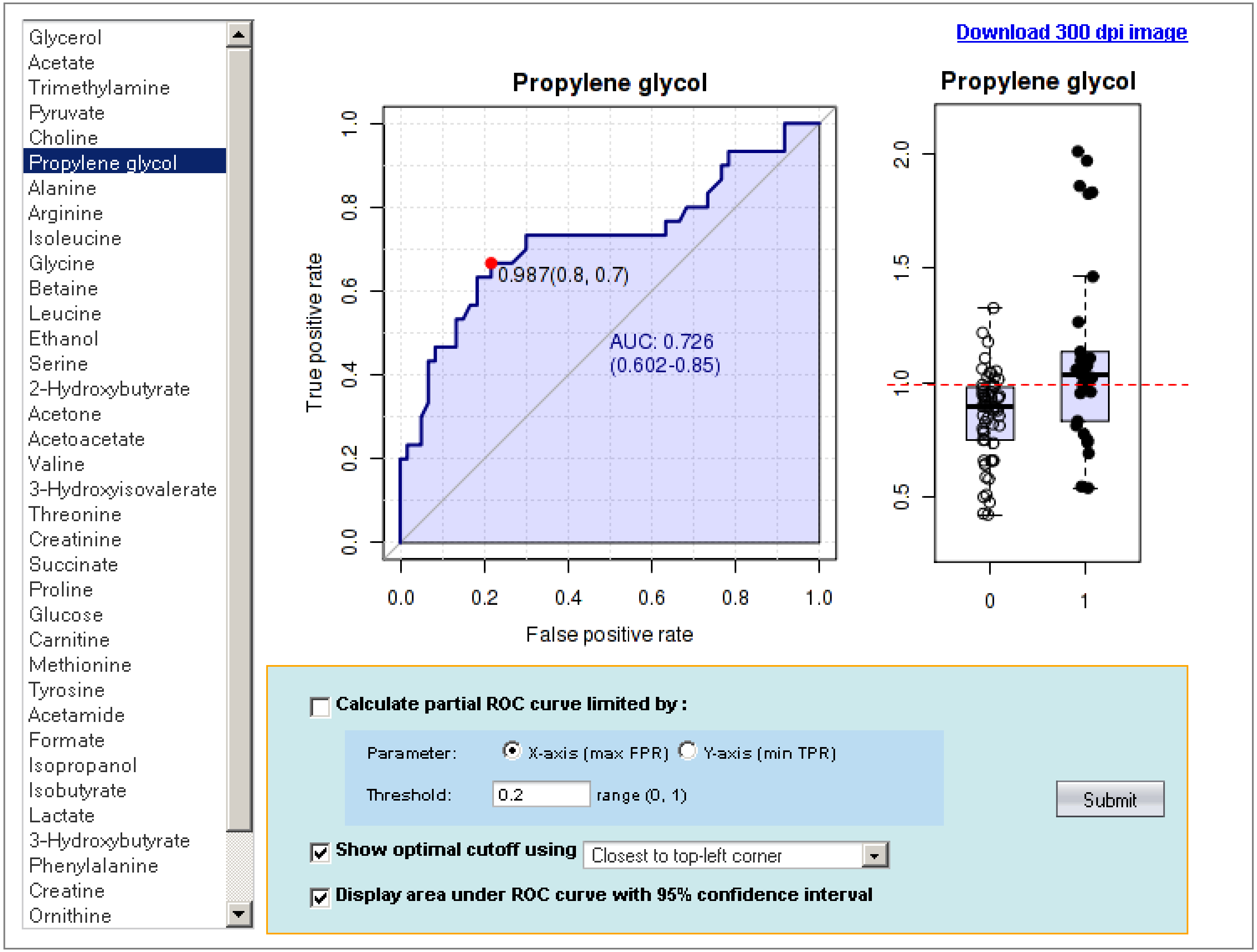
Perform classical univariate ROC curve analyses, such as to generate ROC curve, to calculate AUC or partial AUC as well as their 95% confidence intervals, to compute optimal cutoffs for any given feature, as well as to generate performance tables for sensitivity, specificity, and confidence intervals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

○ **To perform automated biomarker selection and model evaluation (ROC Explorer)**

Perform automated biomarker selection and classification using one of the three multivariate algorithms - support vector machines (SVM), partial least squares discriminant analysis (PLS-DA), and random rorests.

○ **To create and evaluate custom biomarker models (ROC Tester)**

Manually select potential biomarker(s) and then test their performance using any of the three algorithms mentioned above. The module also allows users to hold out a subset of samples for validation purpose (i.e. outside the buildin cross validation). Users can also assess the importance of a model using permutation-based approaches.



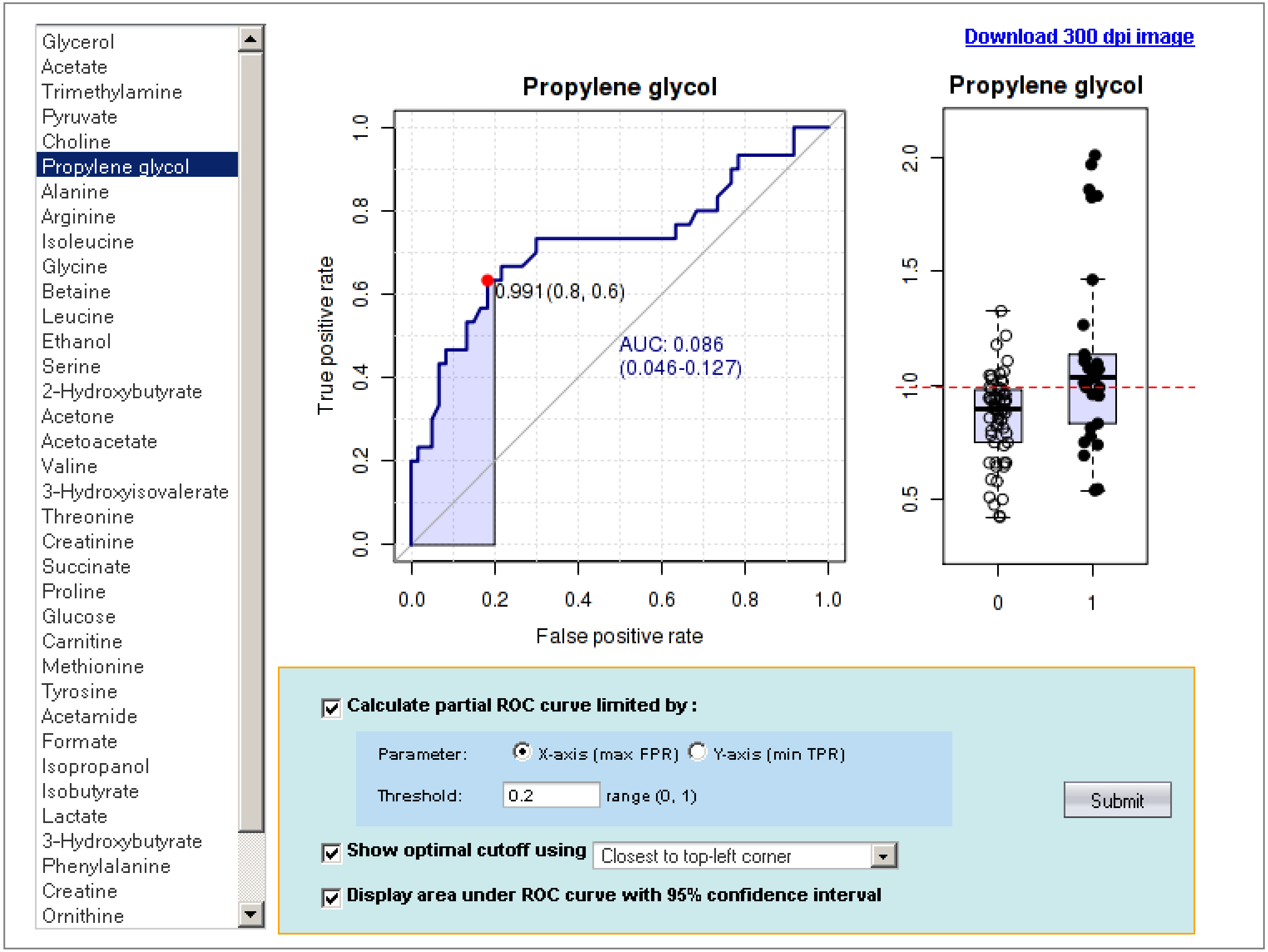
☐ **Calculate partial ROC curve limited by :**

Parameter: ● X-axis (max FPR) ○ Y-axis (min TPR)

Threshold: 0.2 range (0, 1)

☑ **Show optimal cutoff using** Closest to top-left corner

☑ **Display area under ROC curve with 95% confidence interval**



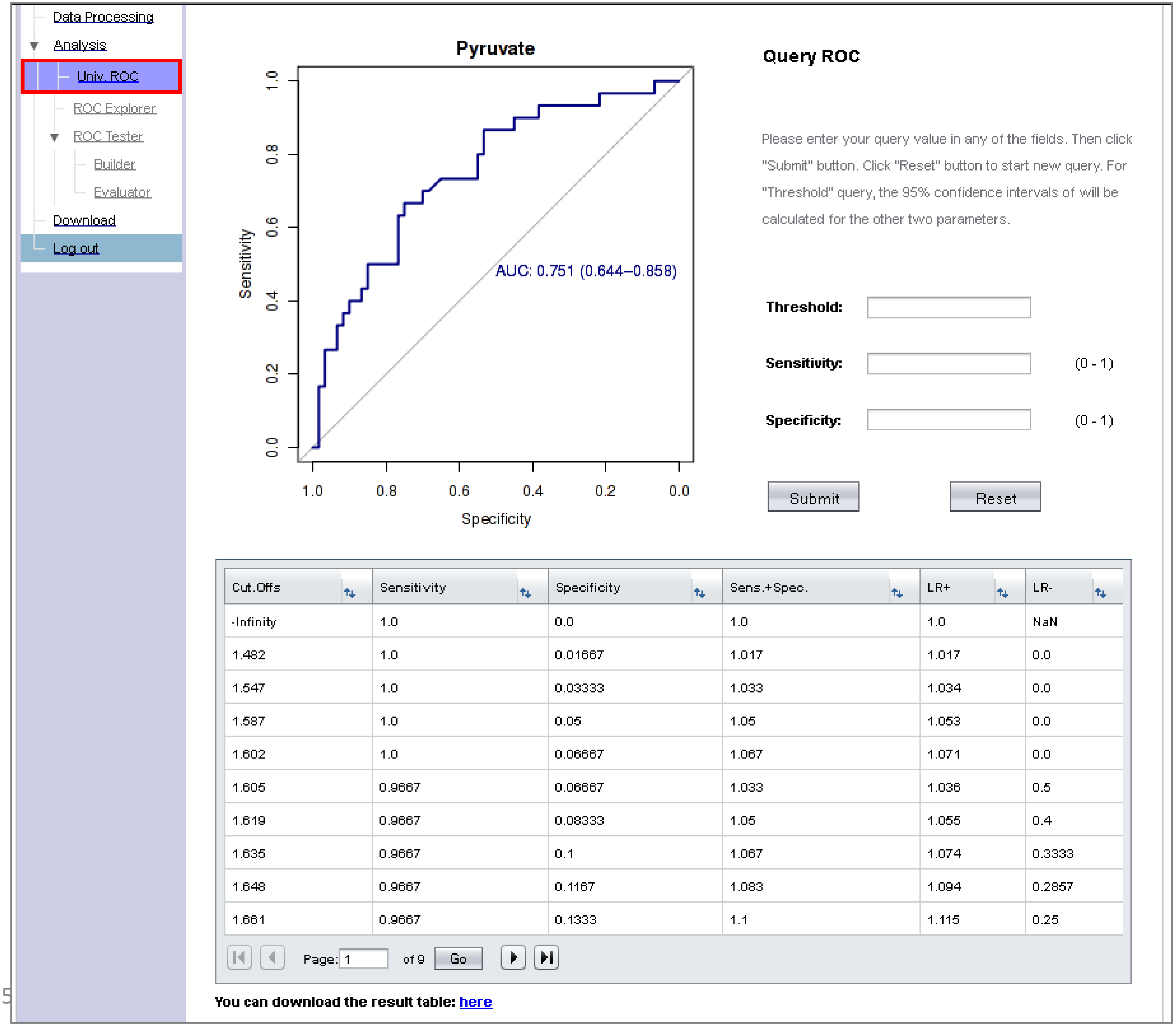
☑ **Calculate partial ROC curve limited by :**

Parameter: ● X-axis (max FPR) ○ Y-axis (min TPR)

Threshold: 0.2 range (0, 1)

☑ **Show optimal cutoff using** Closest to top-left corner

☑ **Display area under ROC curve with 95% confidence interval**



## Query ROC

Please enter your query value in any of the fields. Then click "Submit" button. Click "Reset" button to start new query. For "Threshold" query, the 95% confidence intervals of will be calculated for the other two parameters.

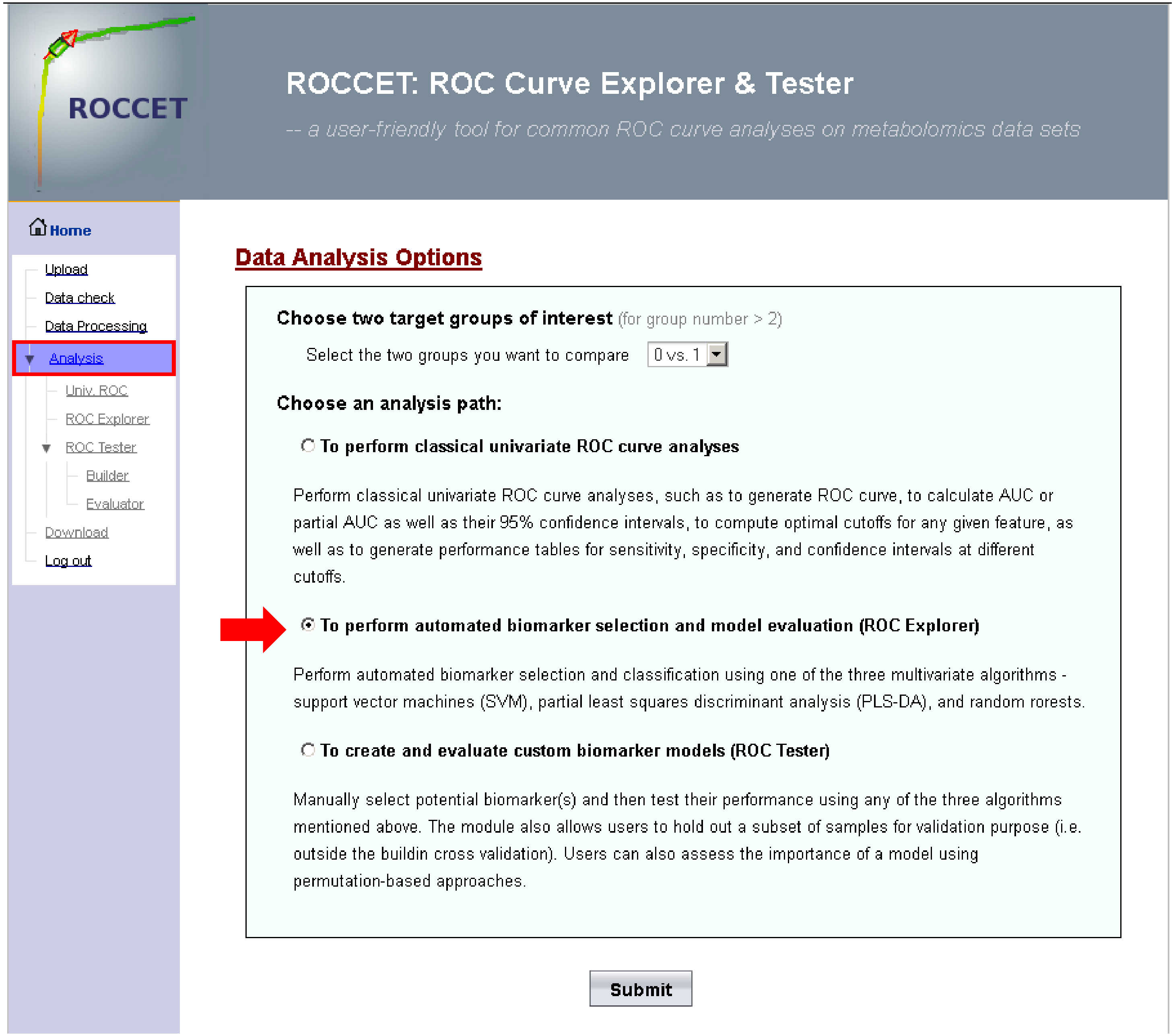
**Threshold:**

**Sensitivity:** (0 - 1)

**Specificity:** (0 - 1)

| Cut.Offs | Sensitivity | Specificity | Sens.+Spec. | LR+ | LR- |
|---|---|---|---|---|---|
| -Infinity | 1.0 | 0.0 | 1.0 | 1.0 | NaN |
| 1.482 | 1.0 | 0.01667 | 1.017 | 1.017 | 0.0 |
| 1.547 | 1.0 | 0.03333 | 1.033 | 1.034 | 0.0 |
| 1.587 | 1.0 | 0.05 | 1.05 | 1.053 | 0.0 |
| 1.602 | 1.0 | 0.06667 | 1.067 | 1.071 | 0.0 |
| 1.605 | 0.9667 | 0.06667 | 1.033 | 1.036 | 0.5 |
| 1.619 | 0.9667 | 0.08333 | 1.05 | 1.055 | 0.4 |
| 1.635 | 0.9667 | 0.1 | 1.067 | 1.074 | 0.3333 |
| 1.648 | 0.9667 | 0.1167 | 1.083 | 1.094 | 0.2857 |
| 1.661 | 0.9667 | 0.1333 | 1.1 | 1.115 | 0.25 |

**You can download the result table: here**



ROCCET: ROC Curve Explorer & Tester

-- a user-friendly tool for common ROC curve analyses on metabolomics data sets

Data Analysis Options

**Choose two target groups of interest** (for group number > 2)

Select the two groups you want to compare: 0 vs. 1

**Choose an analysis path:**

○ **To perform classical univariate ROC curve analyses**

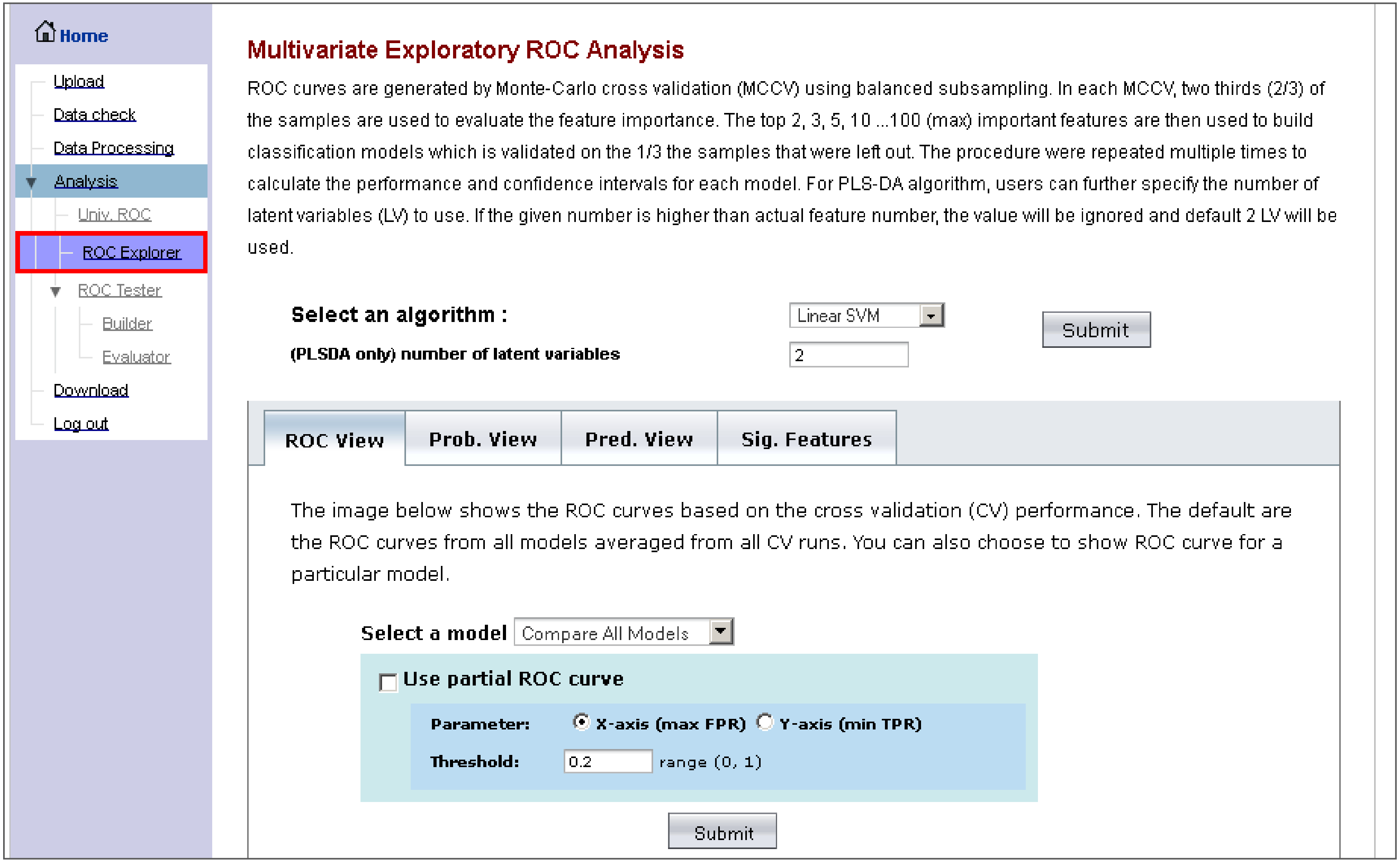
Perform classical univariate ROC curve analyses, such as to generate ROC curve, to calculate AUC or partial AUC as well as their 95% confidence intervals, to compute optimal cutoffs for any given feature, as well as to generate performance tables for sensitivity, specificity, and confidence intervals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

● **To perform automated biomarker selection and model evaluation (ROC Explorer)**

Perform automated biomarker selection and classification using one of the three multivariate algorithms - support vector machines (SVM), partial least squares discriminant analysis (PLS-DA), and random rorests.

○ **To create and evaluate custom biomarker models (ROC Tester)**

Manually select potential biomarker(s) and then test their performance using any of the three algorithms mentioned above. The module also allows users to hold out a subset of samples for validation purpose (i.e. outside the buildin cross validation). Users can also assess the importance of a model using permutation-based approaches.



## Multivariate Exploratory ROC Analysis

ROC curves are generated by Monte-Carlo cross validation (MCCV) using balanced subsampling. In each MCCV, two thirds (2/3) of the samples are used to evaluate the feature importance. The top 2, 3, 5, 10 ...100 (max) important features are then used to build classification models which is validated on the 1/3 the samples that were left out. The procedure were repeated multiple times to calculate the performance and confidence intervals for each model. For PLS-DA algorithm, users can further specify the number of latent variables (LV) to use. If the given number is higher than actual feature number, the value will be ignored and default 2 LV will be used.

**Select an algorithm :** Linear SVM

**(PLSDA only) number of latent variables** 2

The image below shows the ROC curves based on the cross validation (CV) performance. The default are the ROC curves from all models averaged from all CV runs. You can also choose to show ROC curve for a particular model.

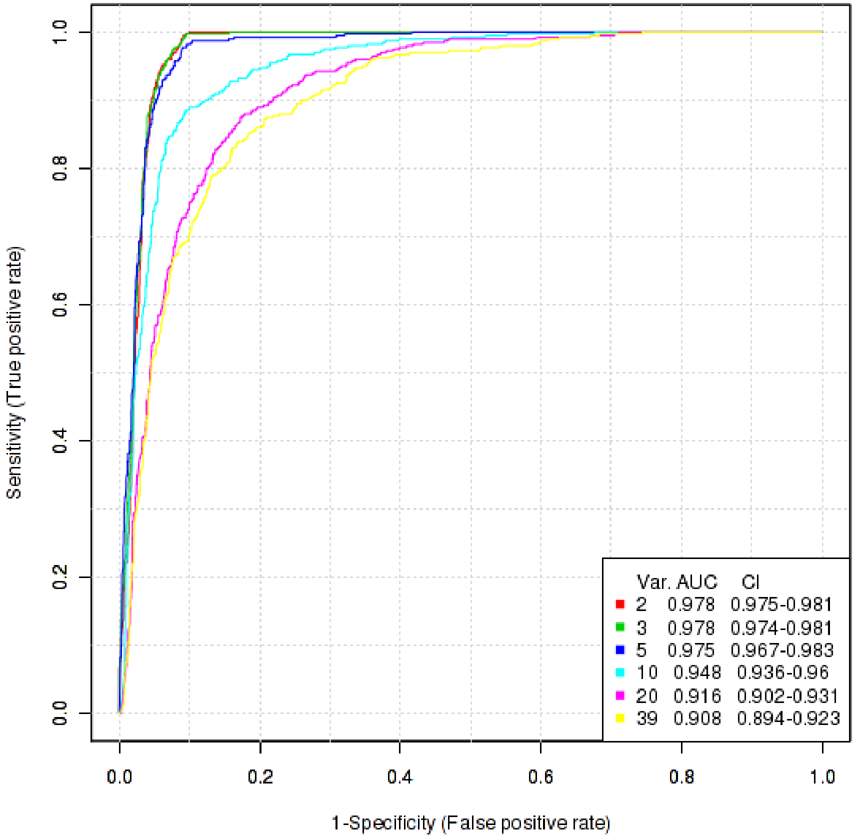
**Select a model** Compare All Models

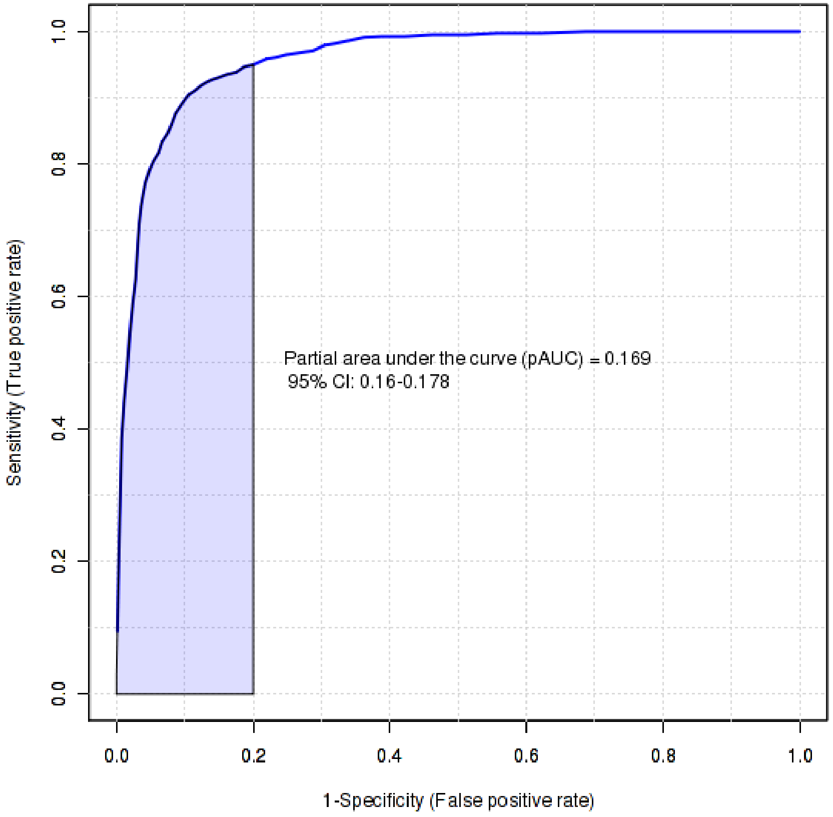
□ **Use partial ROC curve**

**Parameter:** ● **X-axis (max FPR)** ○ **Y-axis (min TPR)**

**Threshold:** 0.2 range (0, 1)

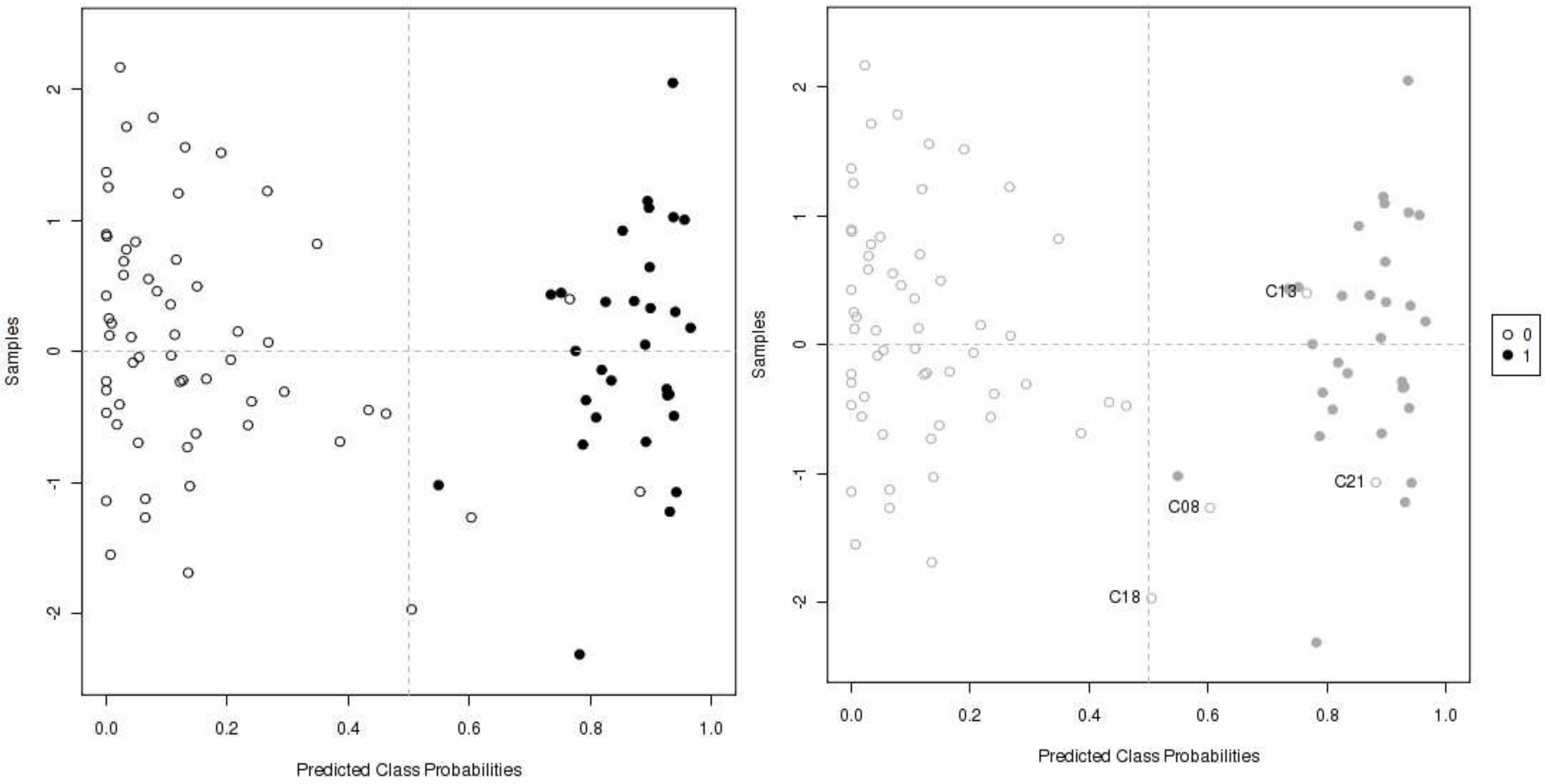
# AUC, pAUC & CI



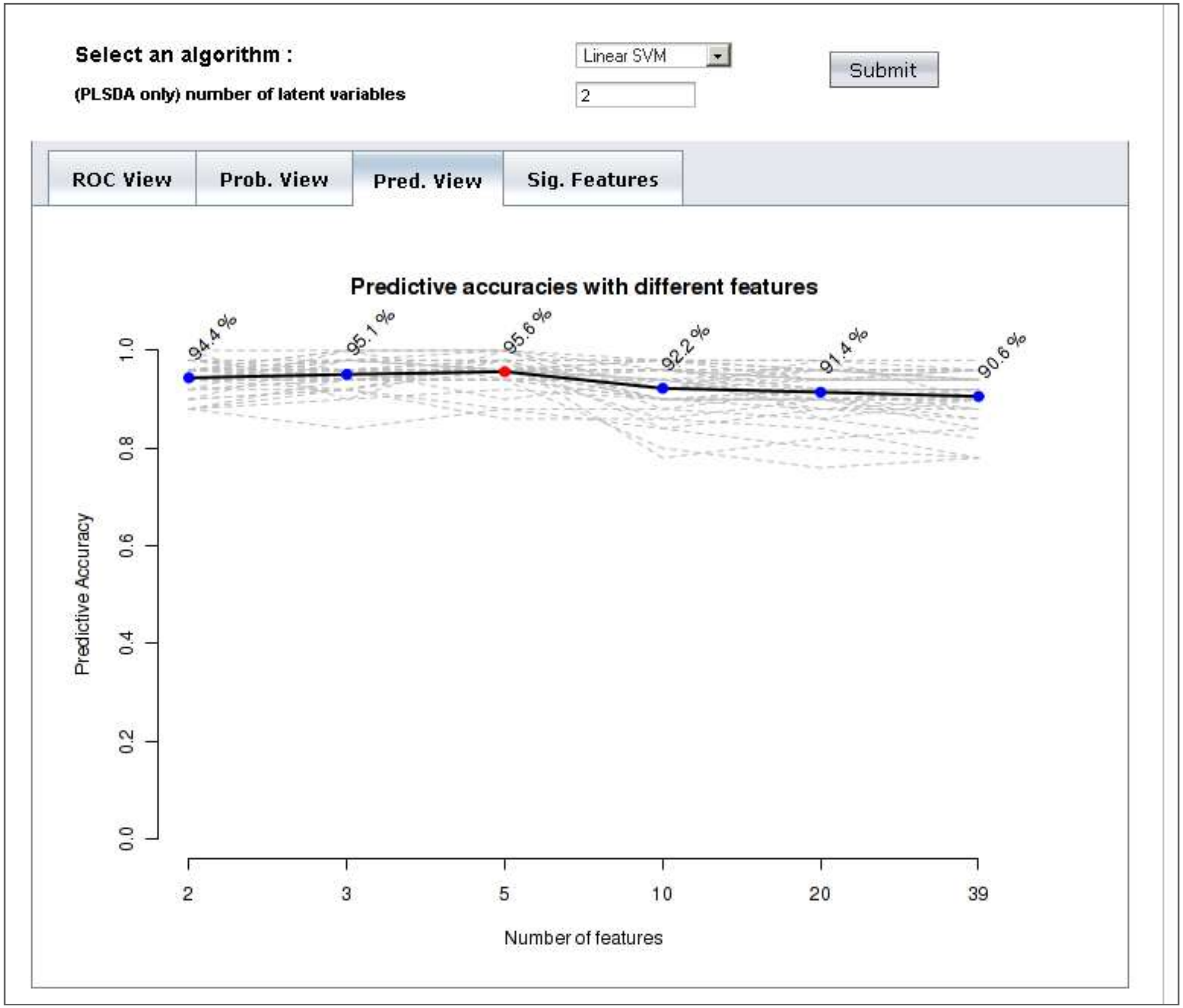


Partial area under the curve (pAUC) = 0.169
95% CI: 0.16-0.178

# Posterior probabilities

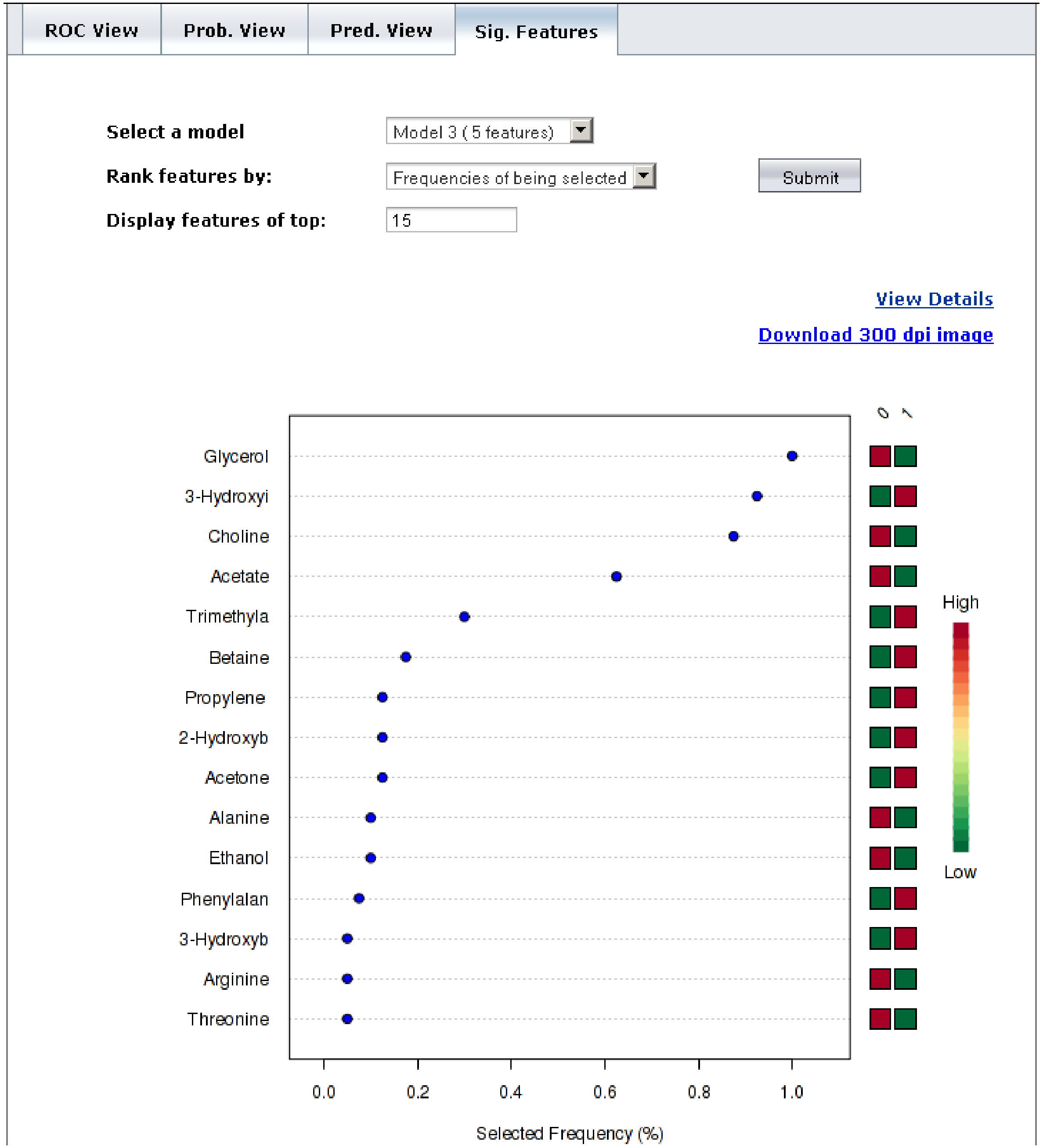


# Accuracies



Select an algorithm : Linear SVM

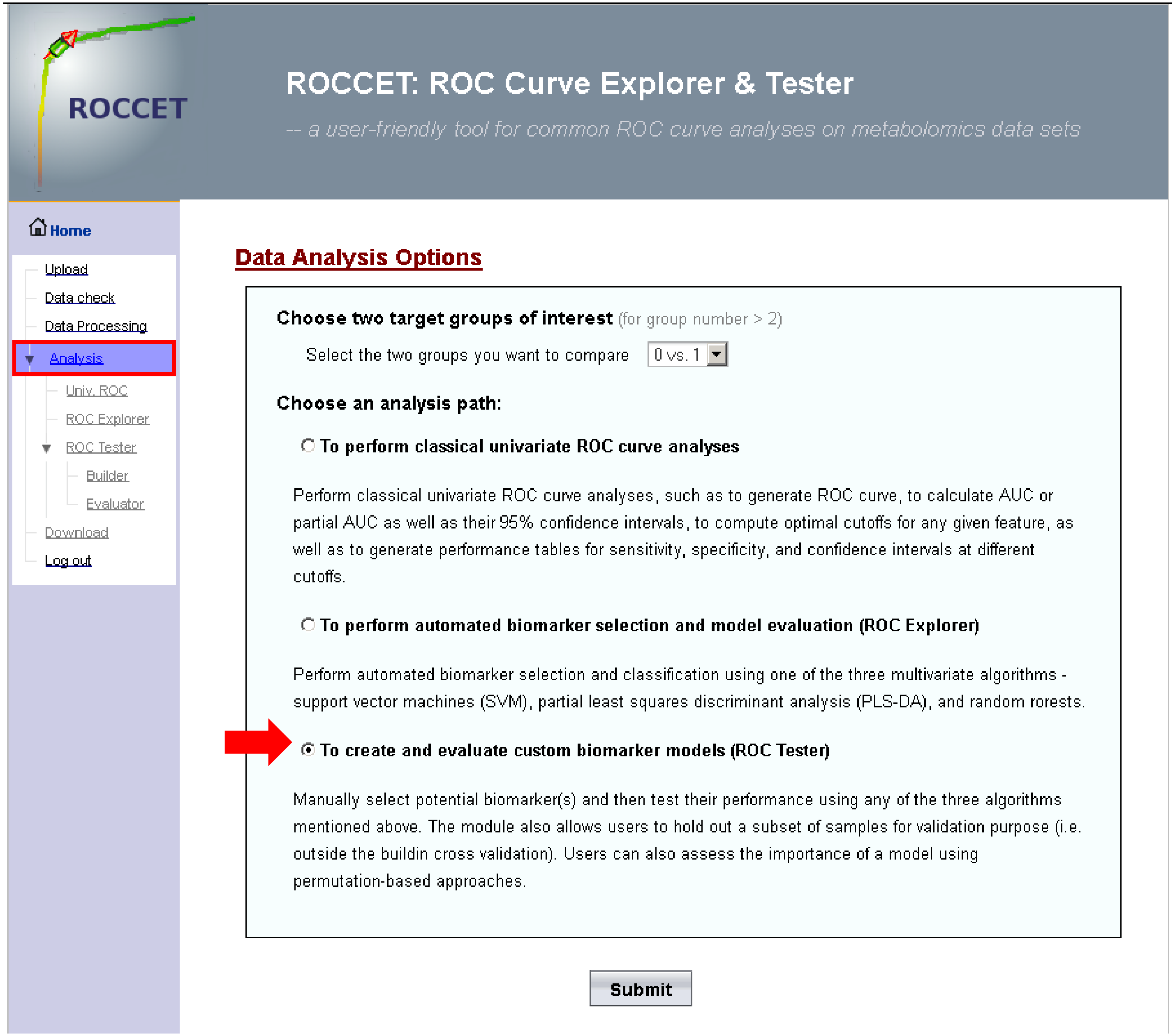
(PLSDA only) number of latent variables 2



Select a model: Model 3 ( 5 features)

Rank features by: Frequencies of being selected

Display features of top: 15



# ROCCET: ROC Curve Explorer & Tester

-- a user-friendly tool for common ROC curve analyses on metabolomics data sets

## Data Analysis Options

**Choose two target groups of interest** (for group number > 2)

Select the two groups you want to compare 0 vs. 1

**Choose an analysis path:**

○ **To perform classical univariate ROC curve analyses**

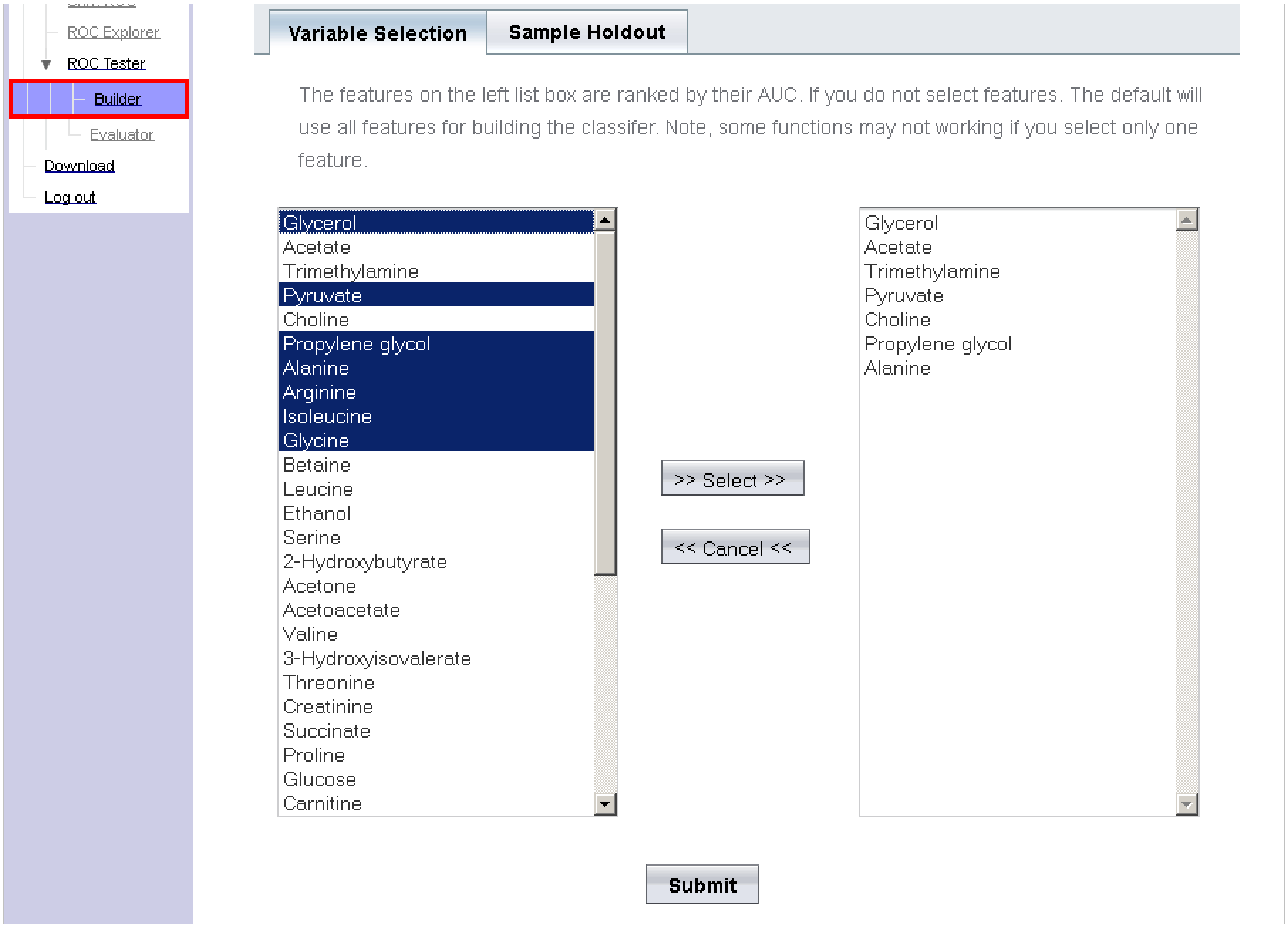
Perform classical univariate ROC curve analyses, such as to generate ROC curve, to calculate AUC or partial AUC as well as their 95% confidence intervals, to compute optimal cutoffs for any given feature, as well as to generate performance tables for sensitivity, specificity, and confidence intervals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

○ **To perform automated biomarker selection and model evaluation (ROC Explorer)**

Perform automated biomarker selection and classification using one of the three multivariate algorithms - support vector machines (SVM), partial least squares discriminant analysis (PLS-DA), and random rorests.

● **To create and evaluate custom biomarker models (ROC Tester)**

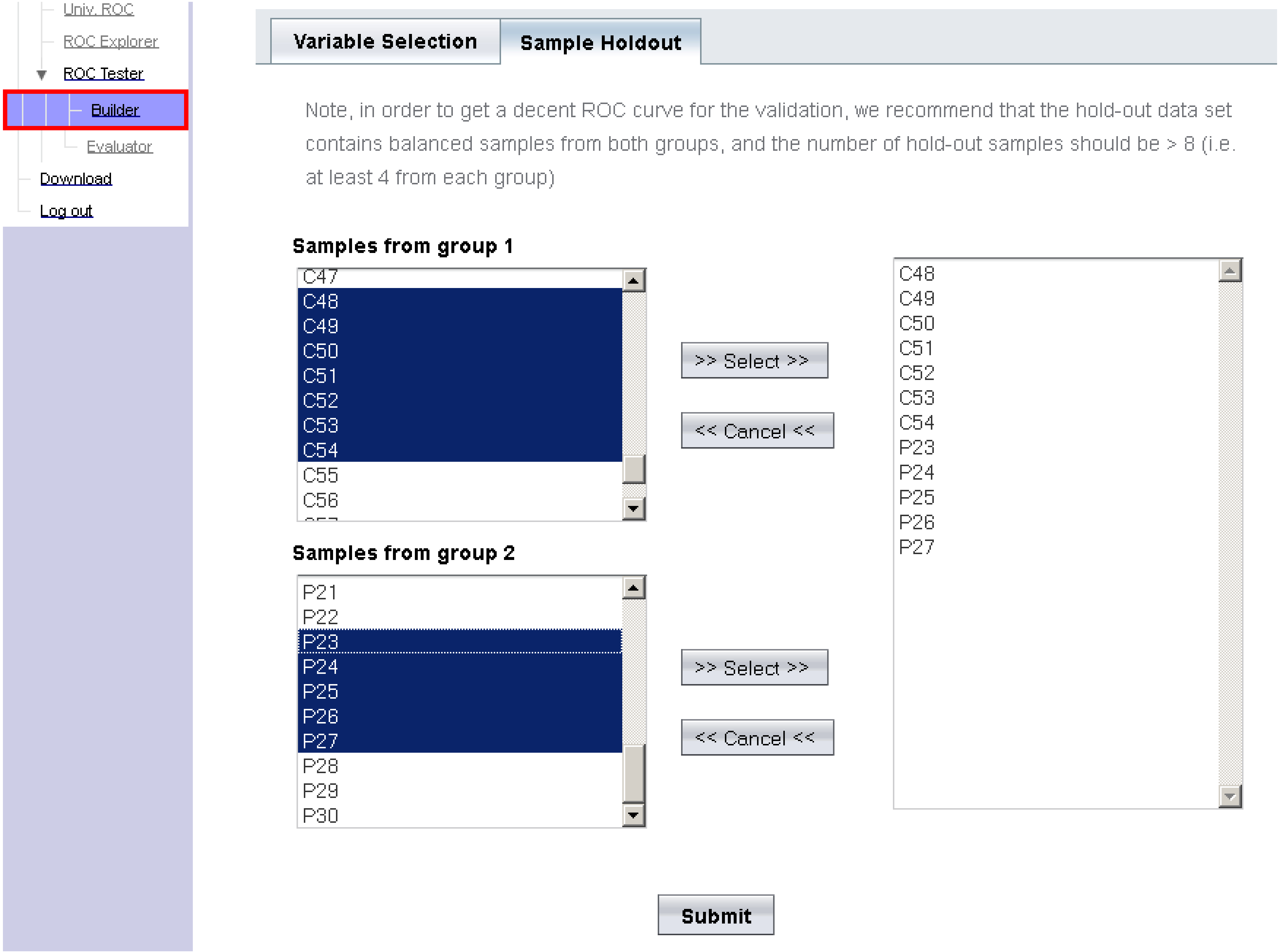
Manually select potential biomarker(s) and then test their performance using any of the three algorithms mentioned above. The module also allows users to hold out a subset of samples for validation purpose (i.e. outside the buildin cross validation). Users can also assess the importance of a model using permutation-based approaches.



The features on the left list box are ranked by their AUC. If you do not select features. The default will use all features for building the classifer. Note, some functions may not working if you select only one feature.

Glycerol
Acetate
Trimethylamine
Pyruvate
Choline
Propylene glycol
Alanine
Arginine
Isoleucine
Glycine
Betaine
Leucine
Ethanol
Serine
2-Hydroxybutyrate
Acetone
Acetoacetate
Valine
3-Hydroxyisovalerate
Threonine
Creatinine
Succinate
Proline
Glucose
Carnitine

Glycerol
Acetate
Trimethylamine
Pyruvate
Choline
Propylene glycol
Alanine



Note, in order to get a decent ROC curve for the validation, we recommend that the hold-out data set contains balanced samples from both groups, and the number of hold-out samples should be > 8 (i.e. at least 4 from each group)

**Samples from group 1**

C47
C48
C49
C50
C51
C52
C53
C54
C55
C56

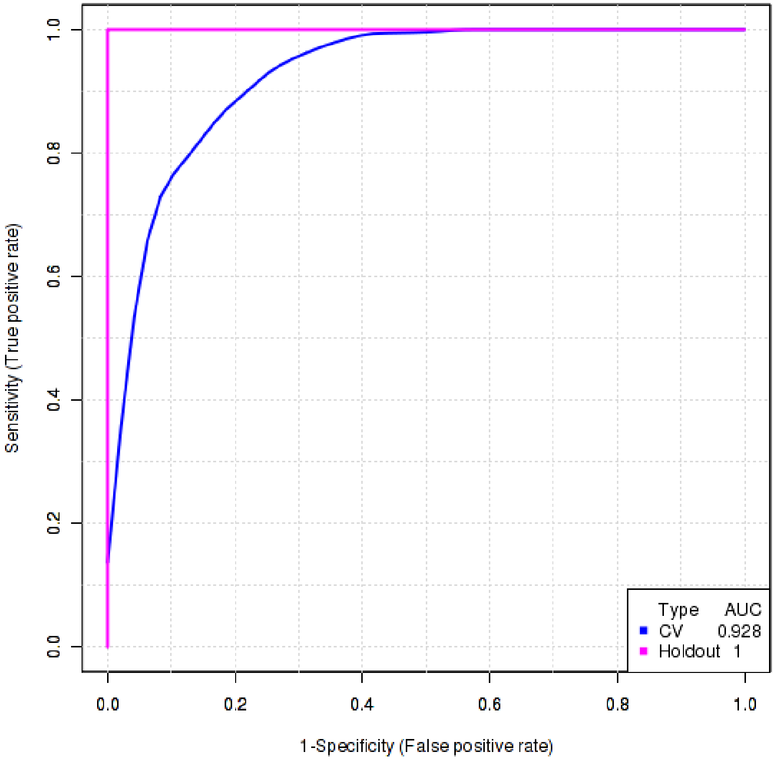
**Samples from group 2**

P21
P22
P23
P24
P25
P26
P27
P28
P29
P30

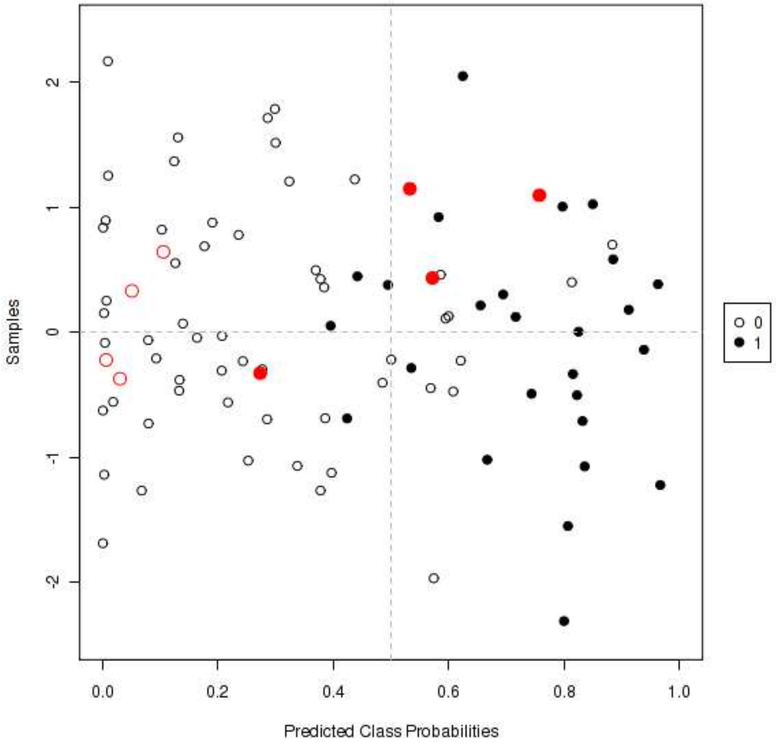
C48
C49
C50
C51
C52
C53
C54
P23
P24
P25
P26
P27

# ROC & Posterior probabilities (with hold-out)

- AUC = 1

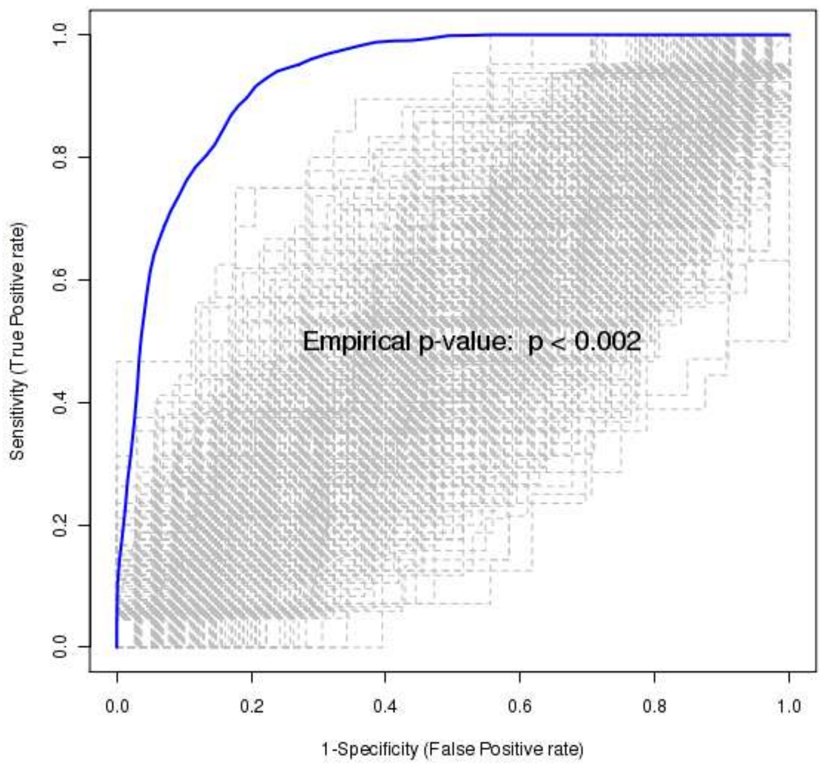


- Accuracy = 7/8

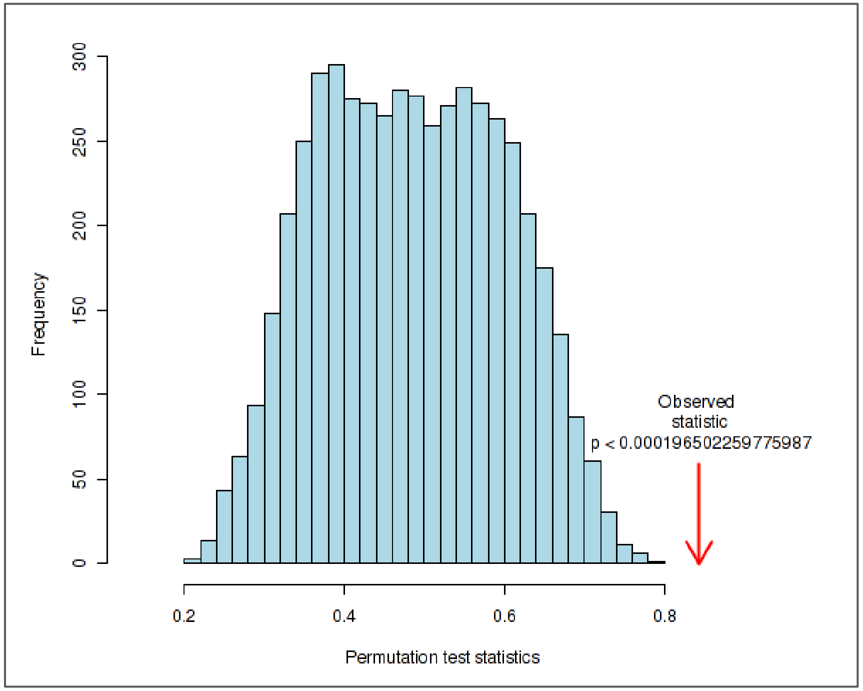


# Permutations

## Based on AUC

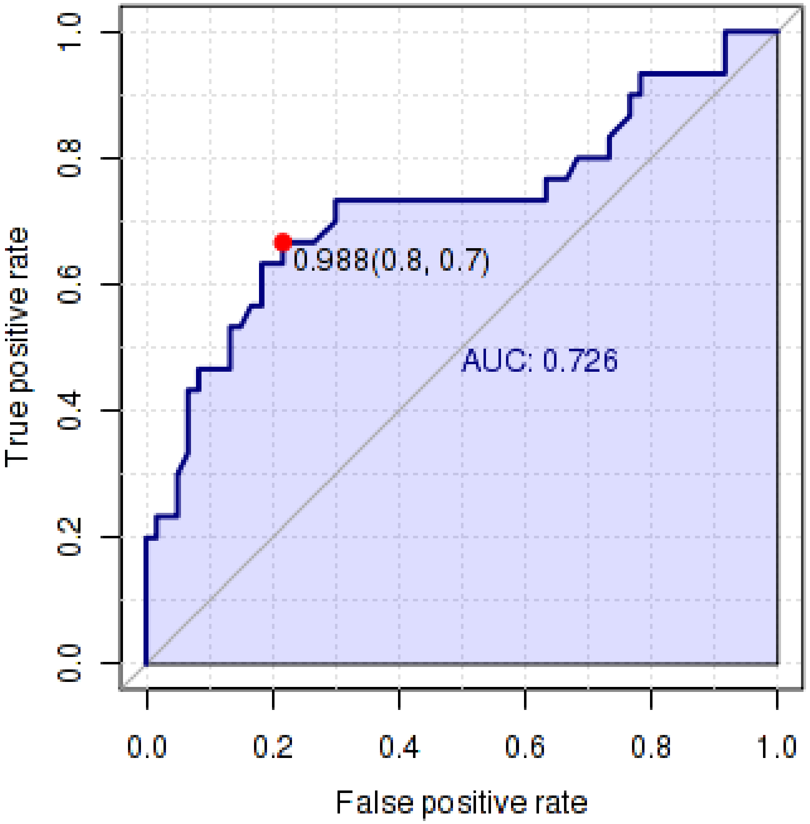


## Based on accuracy

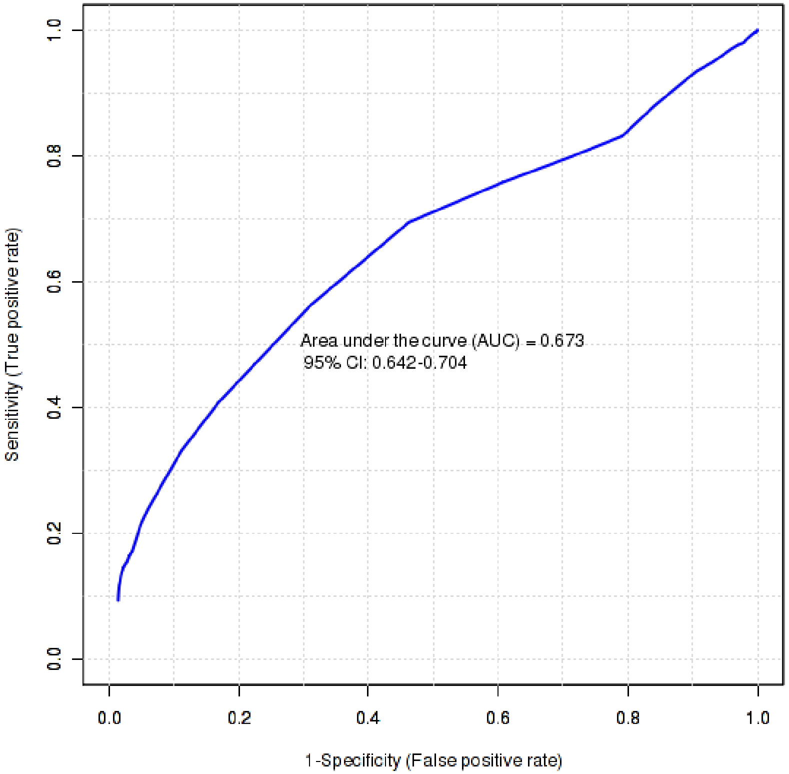


# Over-estimation

**Classical : 0.726**



**CV-based: 0.673**



# Some Technical Details (1)

- Calculate AUC
  - Empirical or non-parametric method
    - Connecting data points with straight lines
    - Trapezoid rules
- Calculate CI
  - Bootstrapping  (classical univariate)
  - Repeated random sampling & cross validation

# Some Technical Details (2)

- Biomarker selection
  - Classical univariate
    - AUC/pAUC
  - Multivariate MCCV-based
    1. Feature selection
       - PLSDA (VIP score)
       - RandomForest (mean decrease accuracy)
       - Linear SVM (feature weight)
    2. Model Selection
       - AUC/pAUC

# Acknowledgements

- Dr. David Wishart
- Dr. David Broadhurst
- Dr. Rupa Mandal
- The Metabolomic Innovation Center (TMIC)
- University of Alberta, Canada

UNIVERSITY OF ALBERTA

THE METABOLOMICS INNOVATION CENTRE