FEBRUARY 9, 2017

1st NuGO ECN Online Webinar

AN INTRODUCTION TO METABONALYST

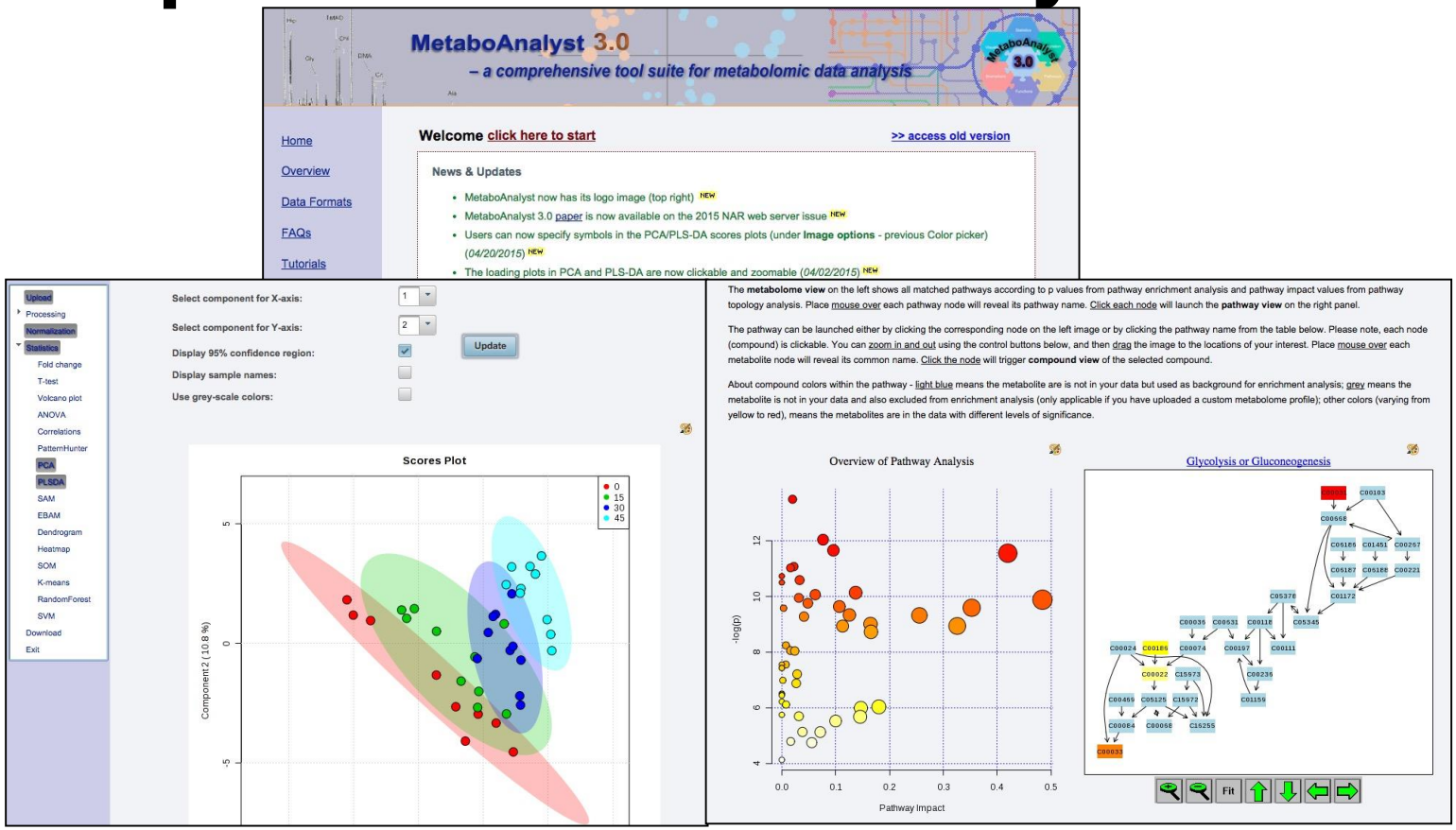
A web-based freely accessible tool for -omics data analysis and interpretation

---

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

Courtesy: Dr. David Wishart

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. cerevisae, 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 signficant 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.

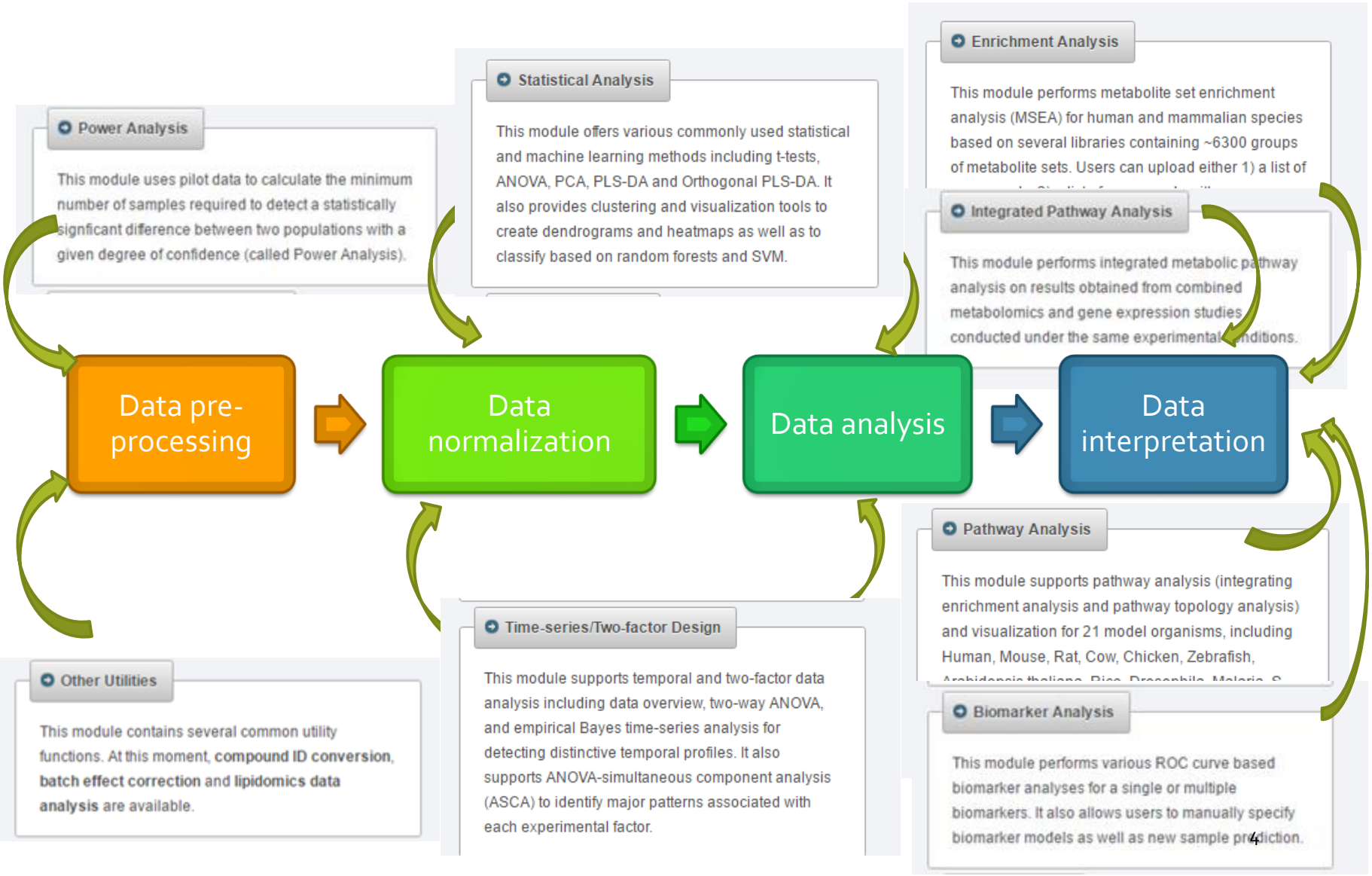
Integrated Pathway Analysis

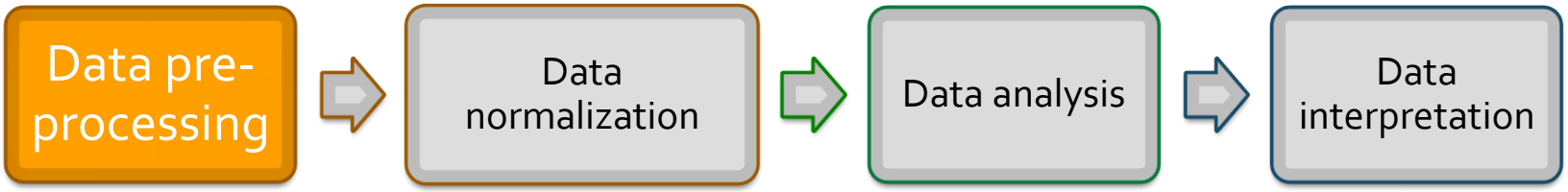
This module performs integrated metabolic pathway analysis on results obtained from combined metabolomics and gene expression studies conducted under the same experimental conditions.

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



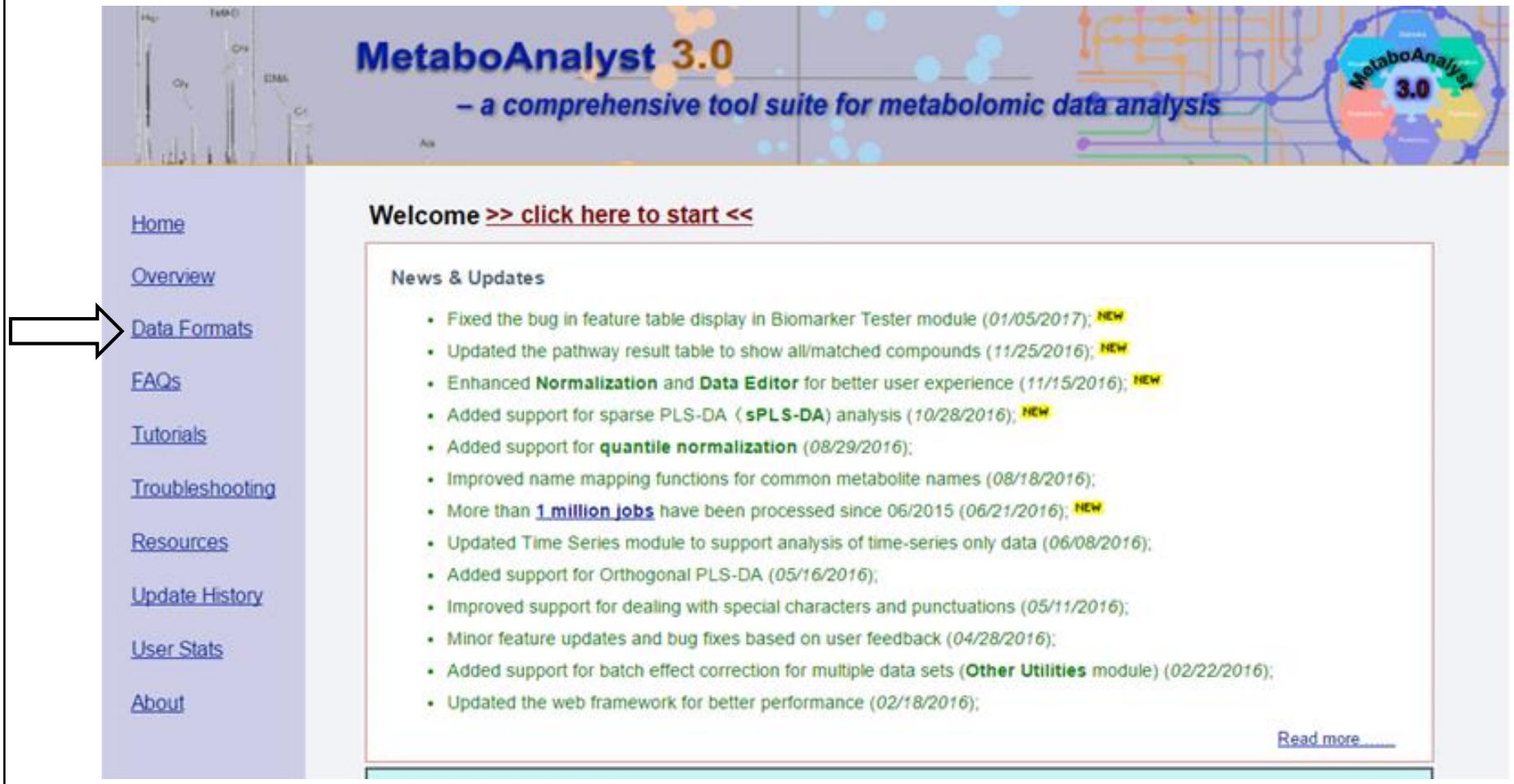


Purpose: to convert various raw data forms into data matrices suitable for statistical analysis

Supported data formats

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

Data Formats



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

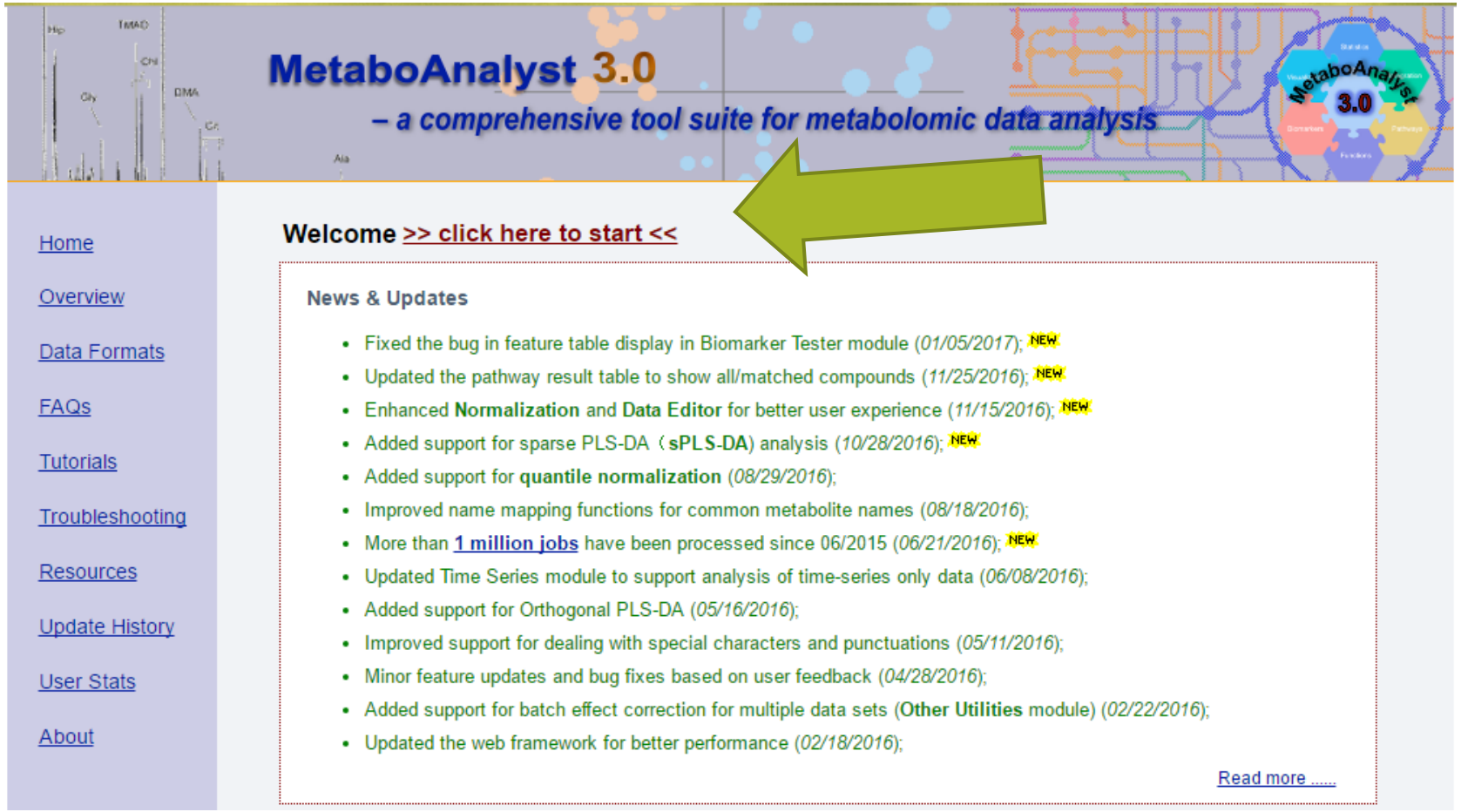
- **COMMA SEPARATED VALUES!! (.csv) or TAB DELIMITED TEXT (.txt)** → For quantitative (concentration tables) or qualitative (peak intensity or NMR/MS spectral bins).

Things to considere:

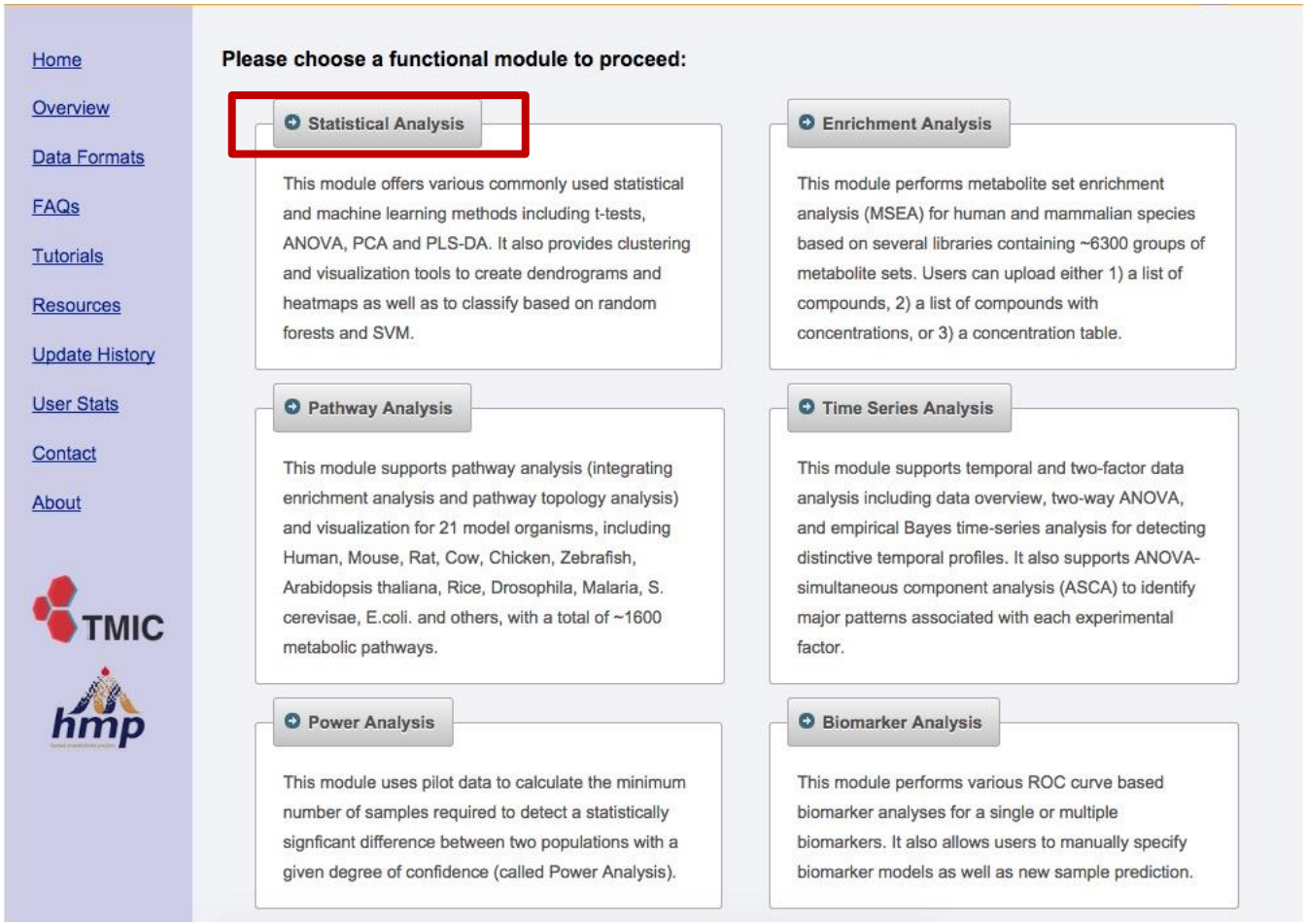
- Both samples and feature names MUST be UNIQUE. Can be combination fo letters and numbers separated by underscores [_].
- The **class label must follow immediately after the sample name (for two-factors and time series data must be two class label columns)**
- **Metaboanalyst can also support .Zip files.**

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

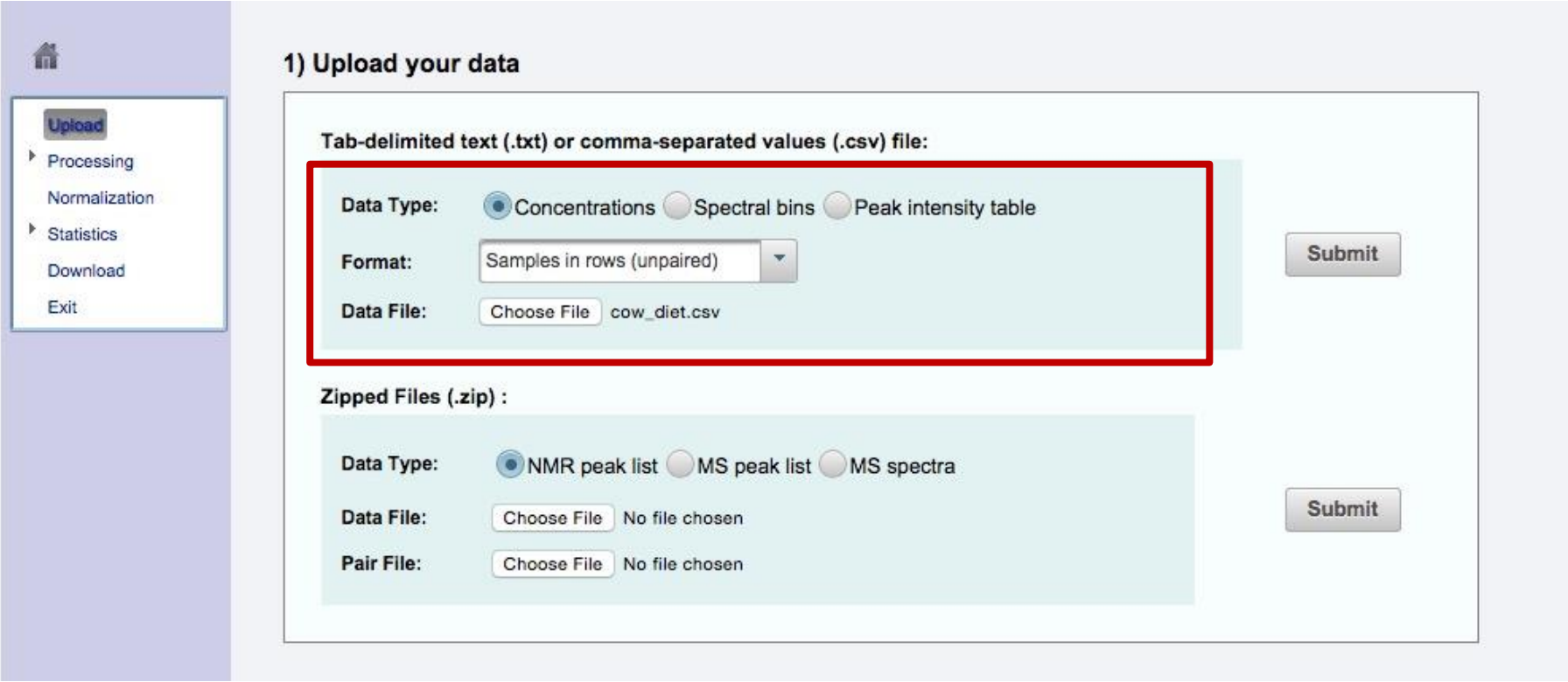
Let´s start!


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 .....

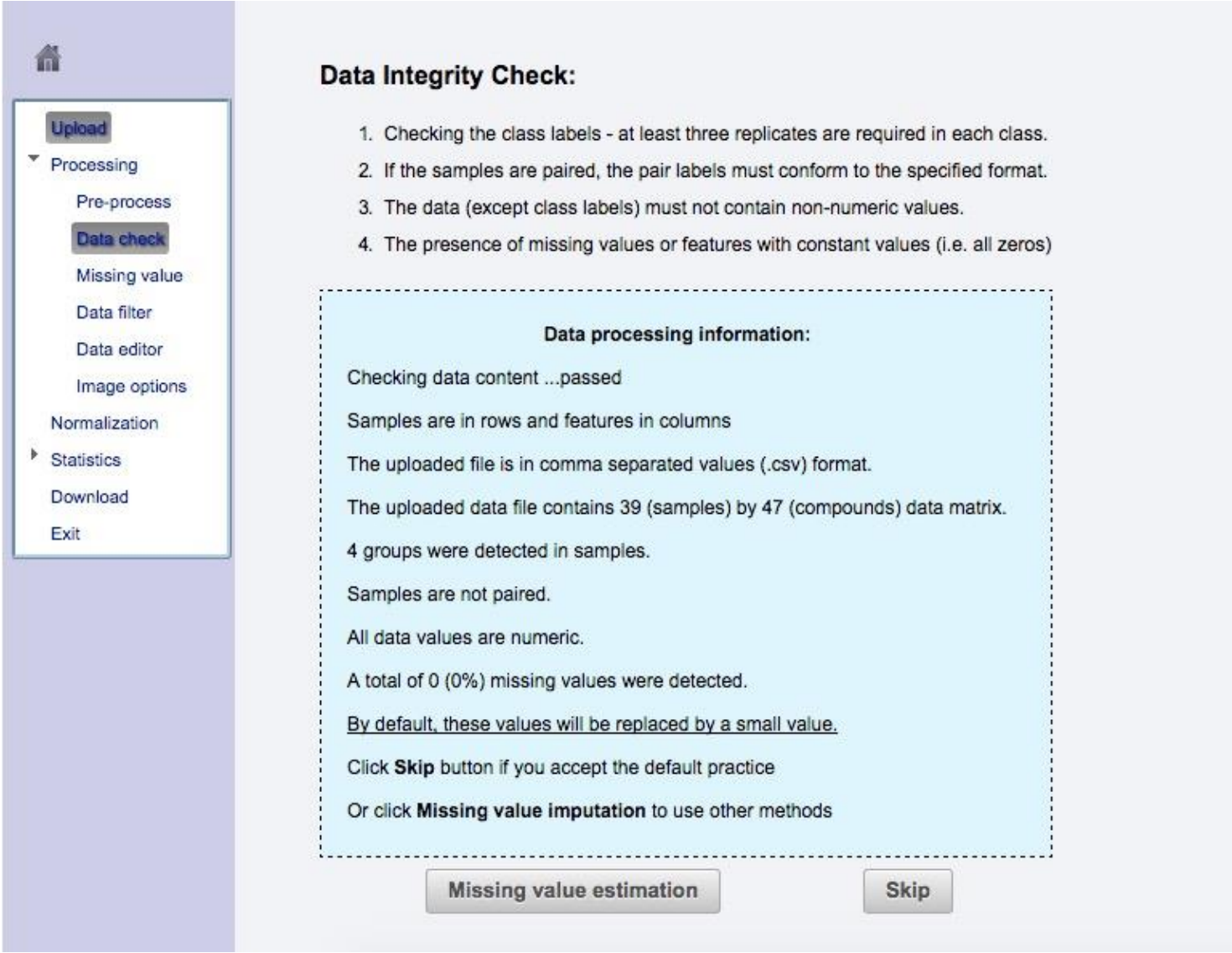
Select a Module :Statistical Analysis



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:
Samples in rows (unpaired)
Data File:
Choose File
cow_diet.csv
Submit
Zipped Files (.zip) :
Data Type:
NMR peak list
MS peak list
MS spectra
Data File:
Choose File
No file chosen
Pair File:
Choose File
No file chosen
Submit

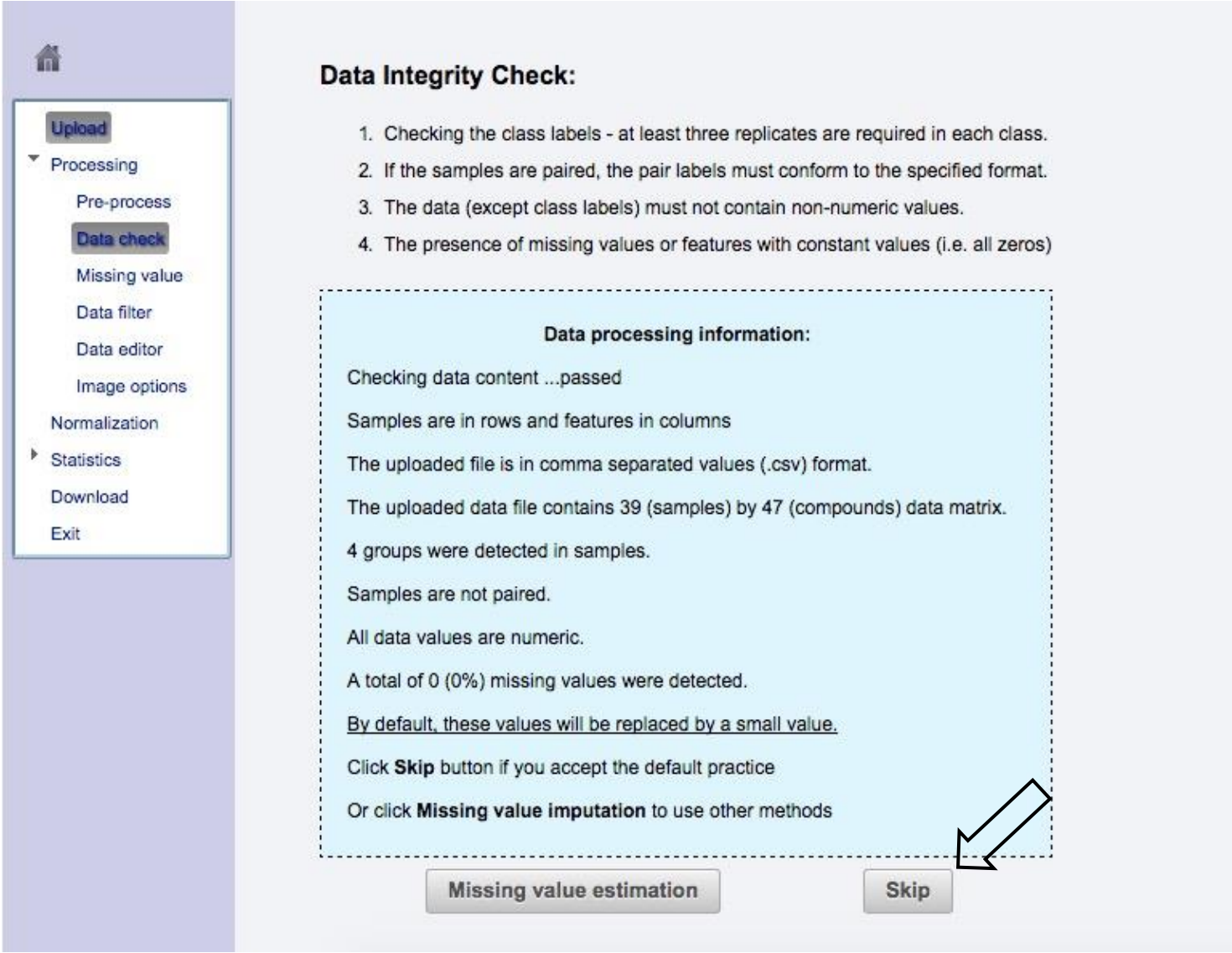
Data Integrity Check

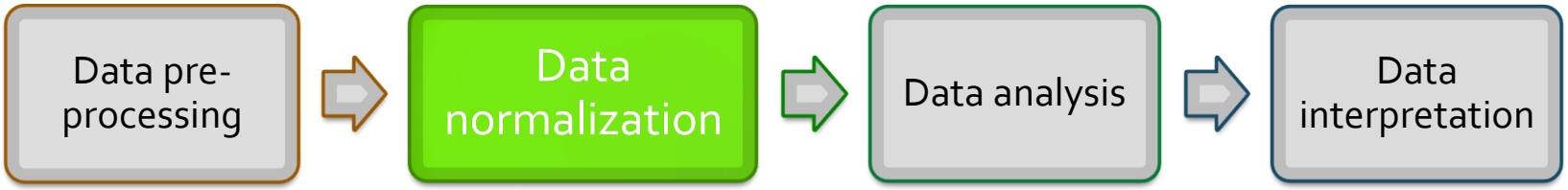


How to deal with missing values?

- 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 pre-processing
Data normalization
Data analysis
Data interpretation

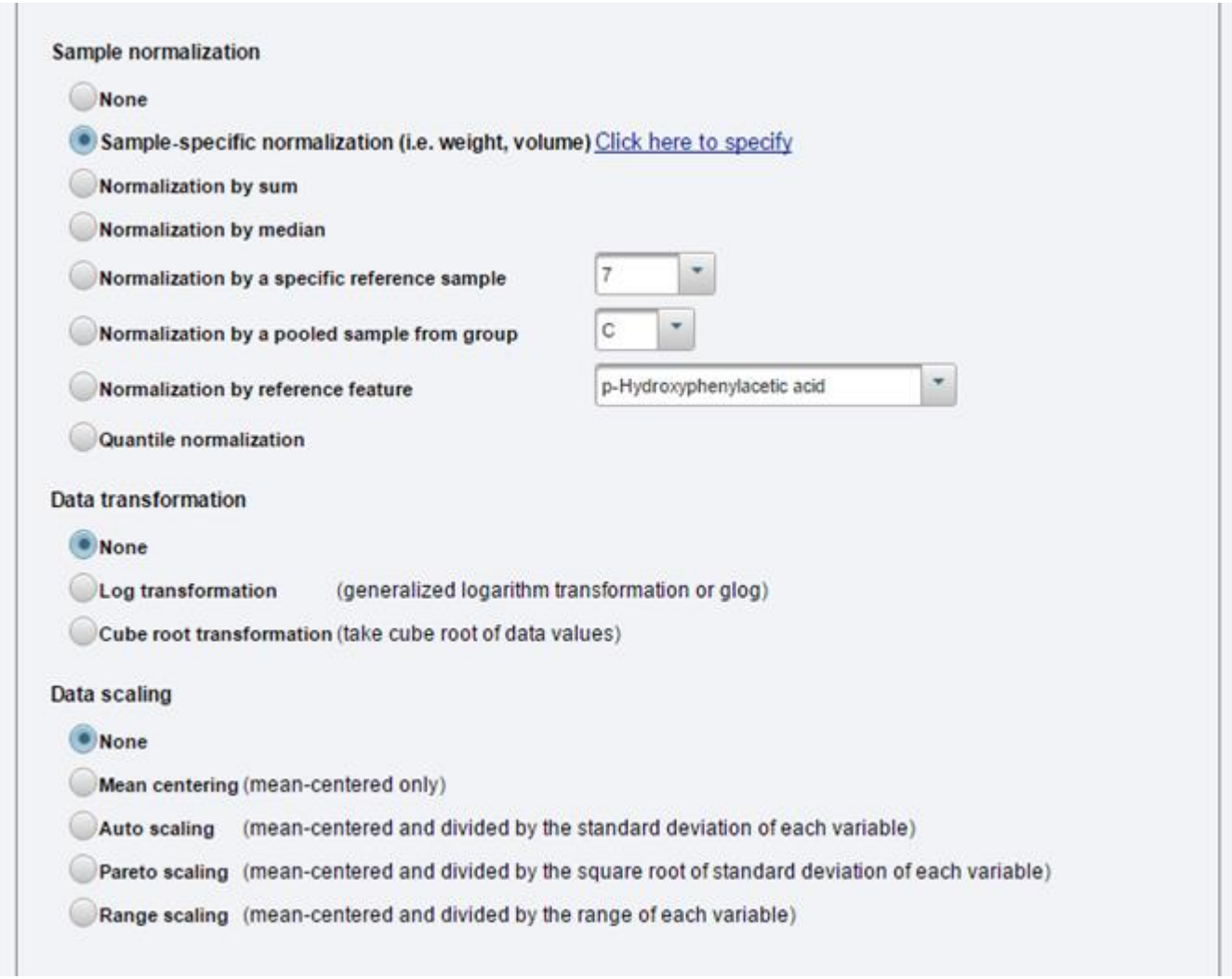
Data Normalization/Scaling

The normalization procedures are grouped into three categories.

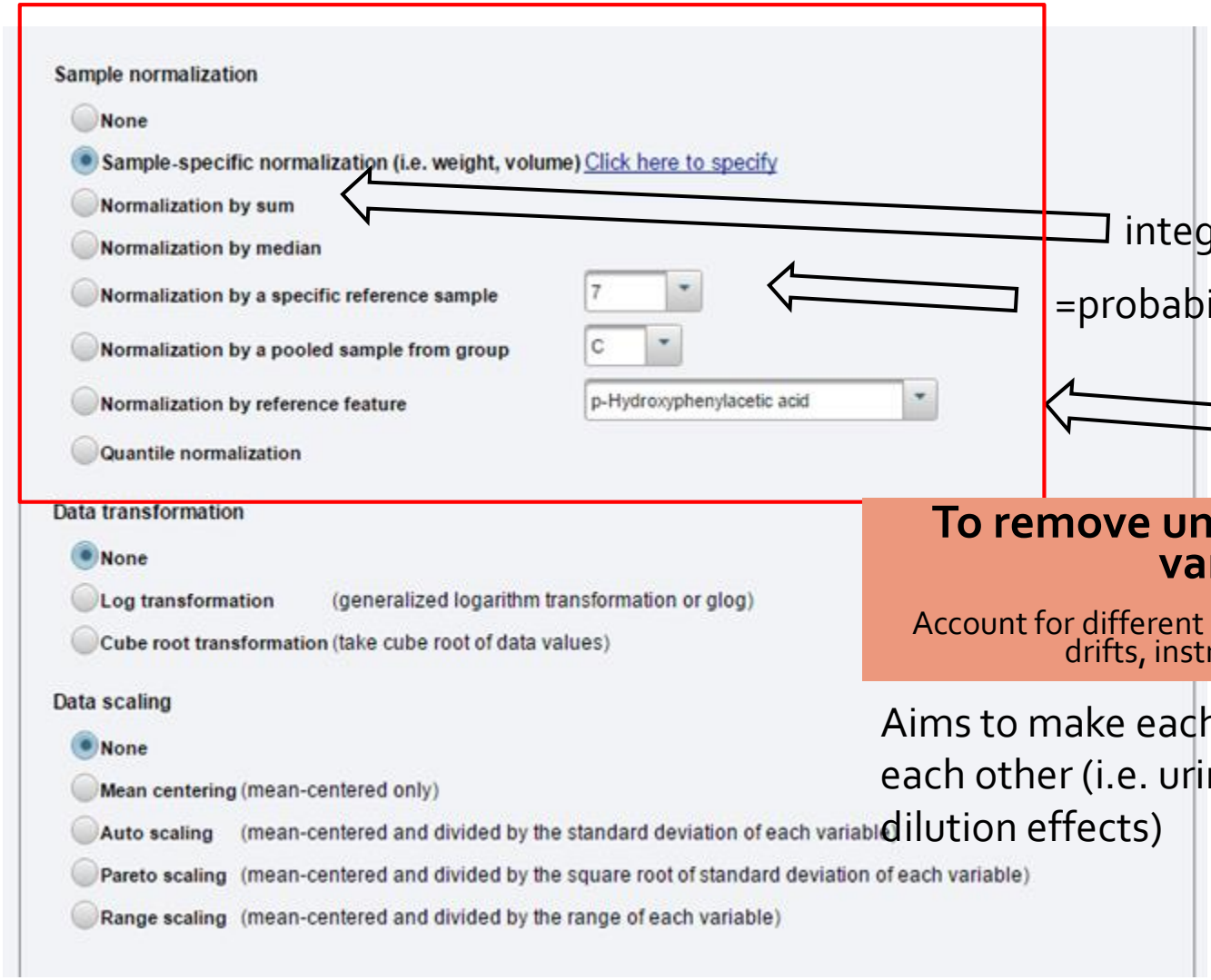
The sample normalization allows general-purpose adjustment for differences among your sample

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

You can use one or combine them



Data Normalization/Scaling



integrated area

=probabilistic quotient norm

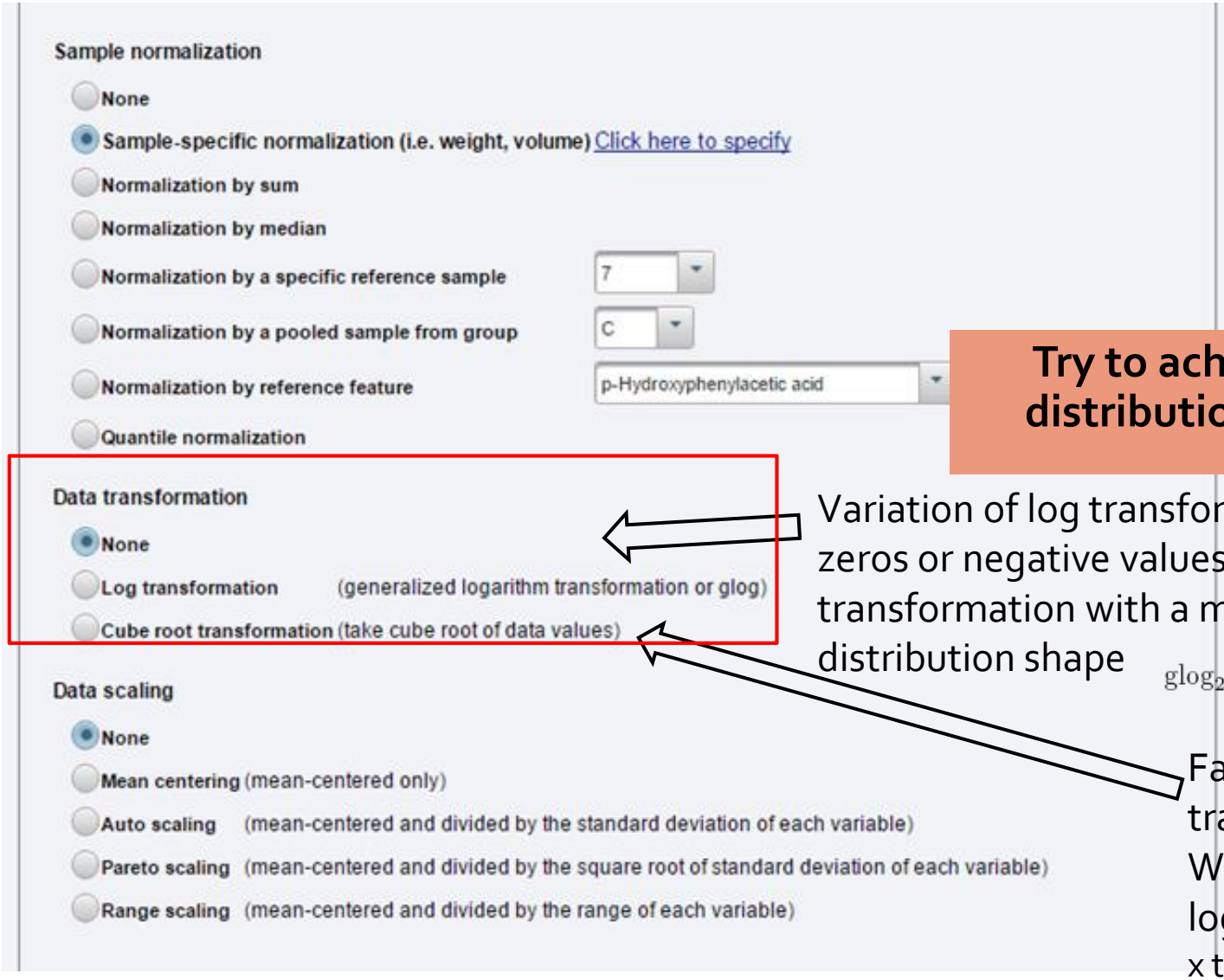
by a particular compound

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



Try to achieve a Normal distribution of your data

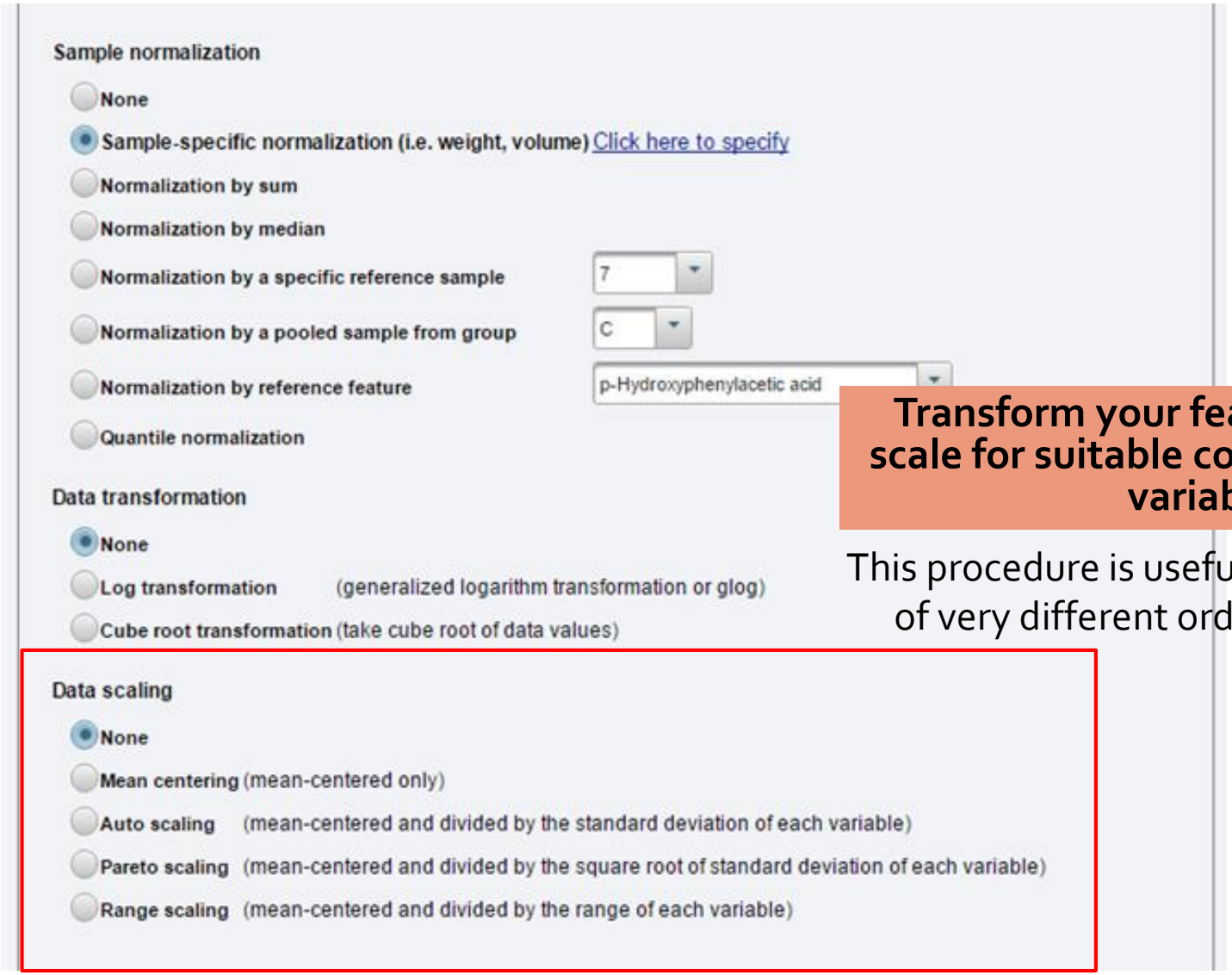
Variation of log transform. Can deal with zeros or negative values. A strong transformation with a major effect on distribution shape

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

Fairly strong transformation. Weaker than the logarithm

x to x^(1/3)

Data Normalization/Scaling



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	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\left(x_{i_{max}} - x_{i_{min}}\right)}$	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 [7]
- Normalization by a pooled sample from group [CLV]
- Normalization by reference feature [p-Hydroxyphenylacetic acid]
- Quantile normalization

Data transformation

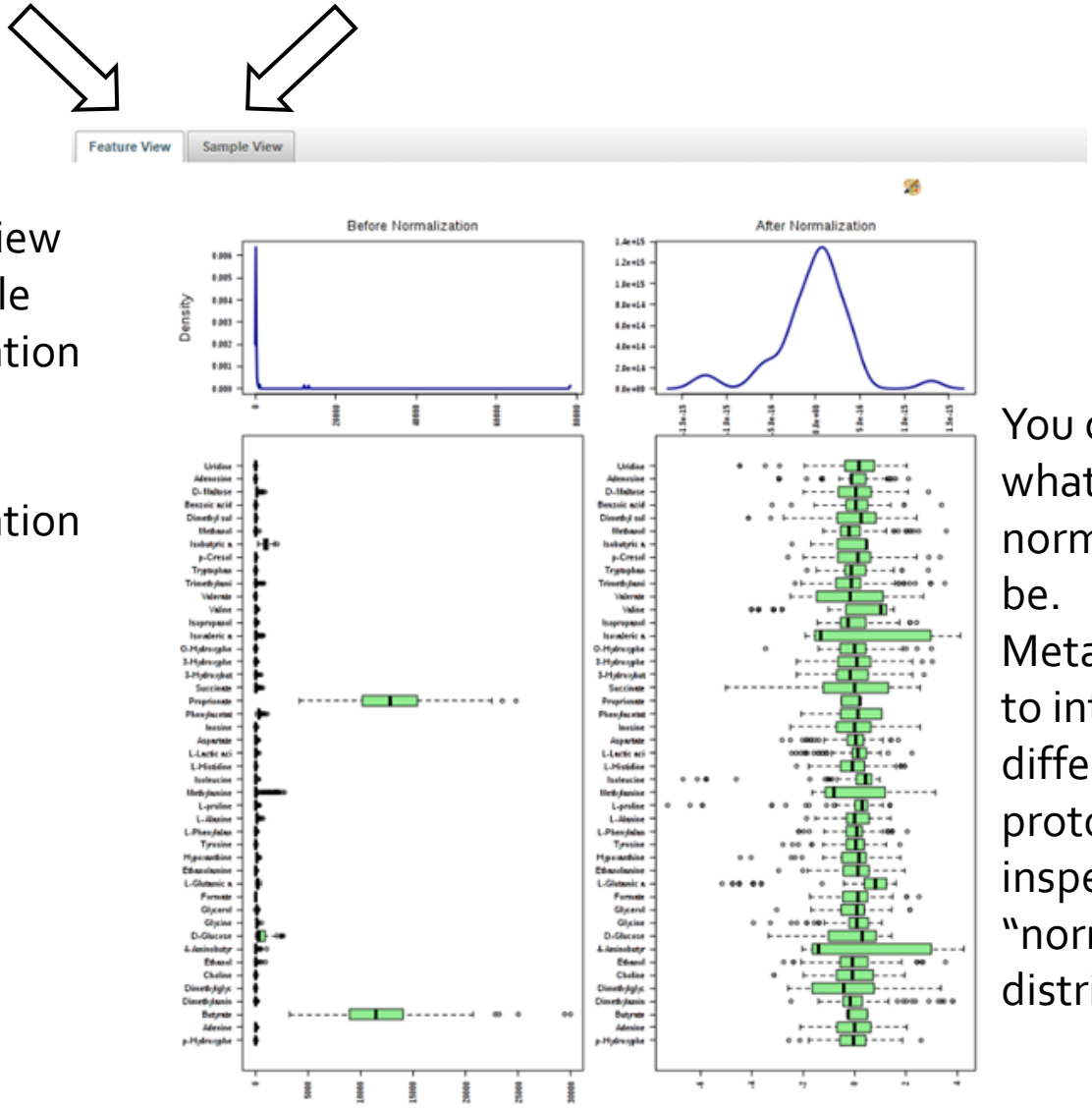
- None
- Log transformation (generalized logarithm transformation or glog)
- Cube root transformation (take cube root of data values)

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)

Normalize | View Result | Proceed

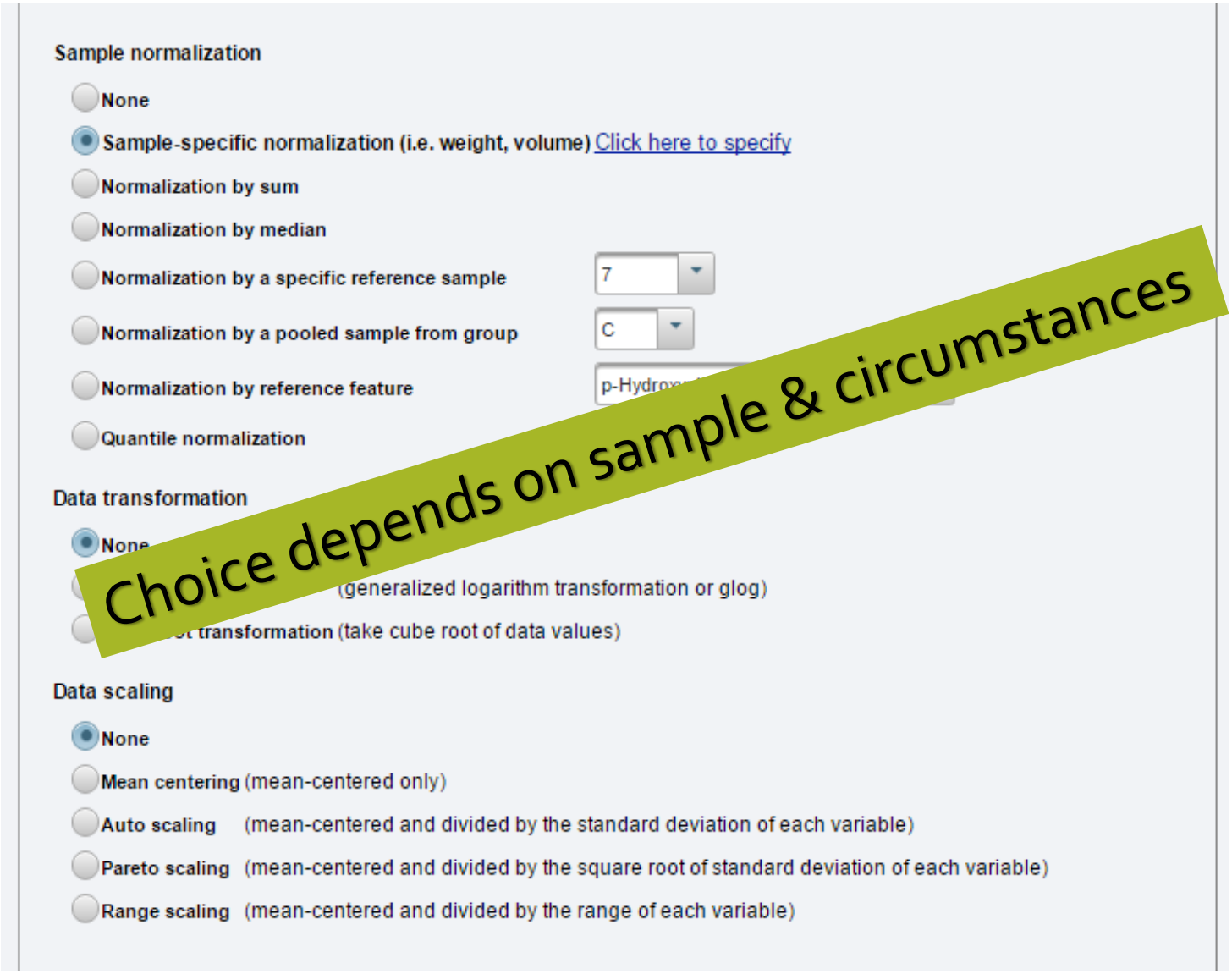
Normalization Result



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



Choice depends on sample & circumstances

Next Steps

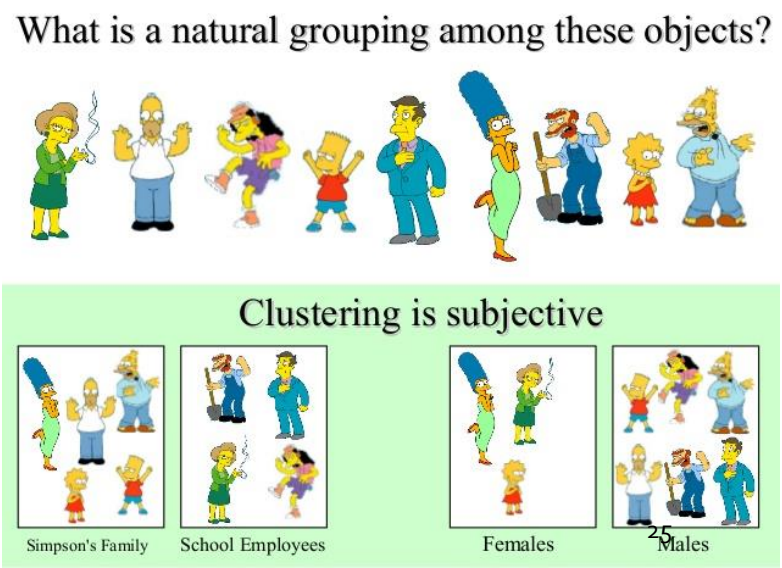
After normalization has been completed it is a good idea to look at your data a little further to identify outliers or noise that could/should be removed

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

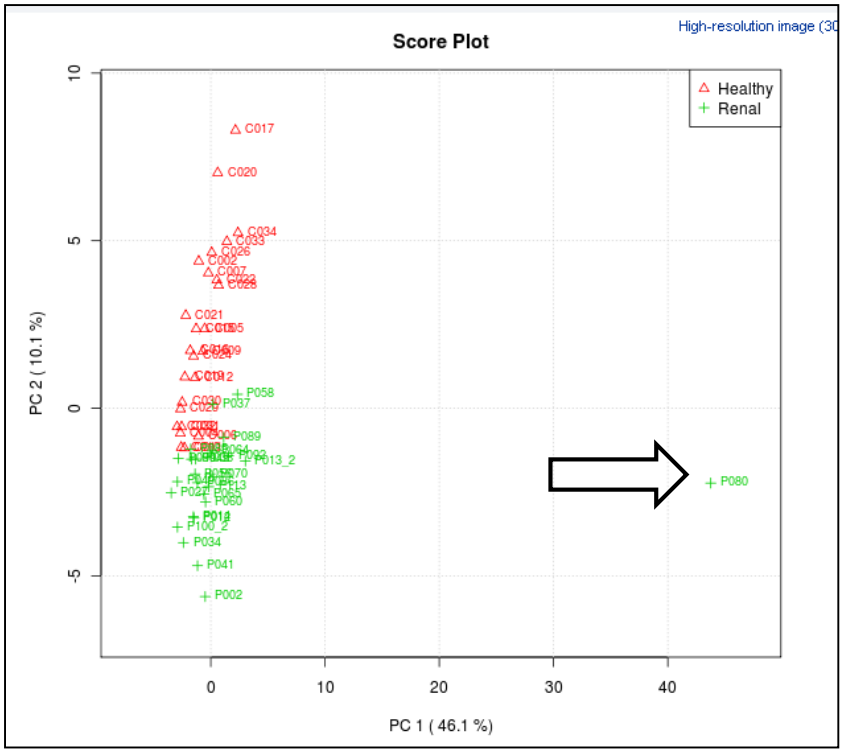


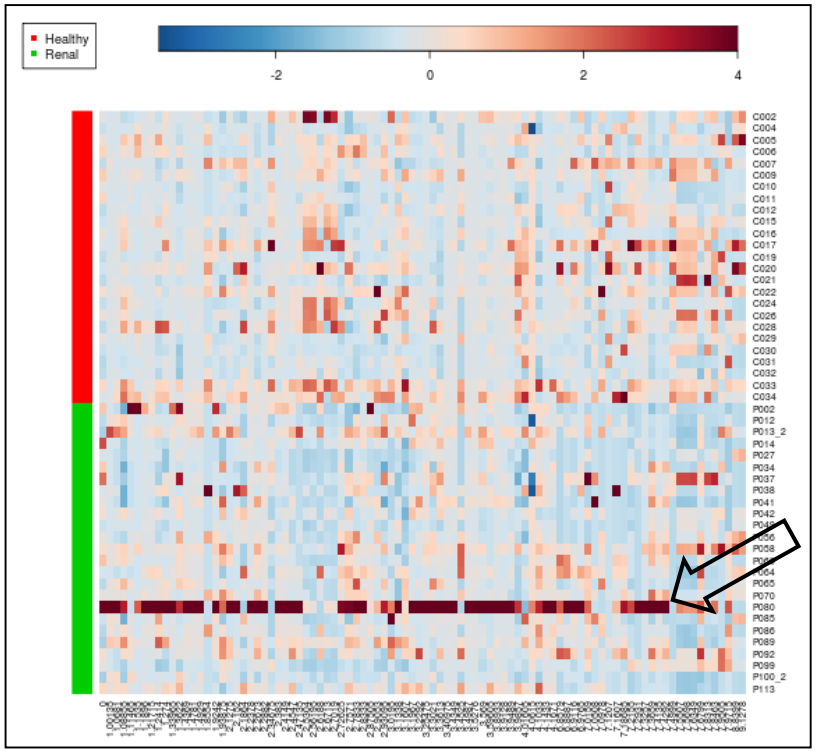
Quality Control

- 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?





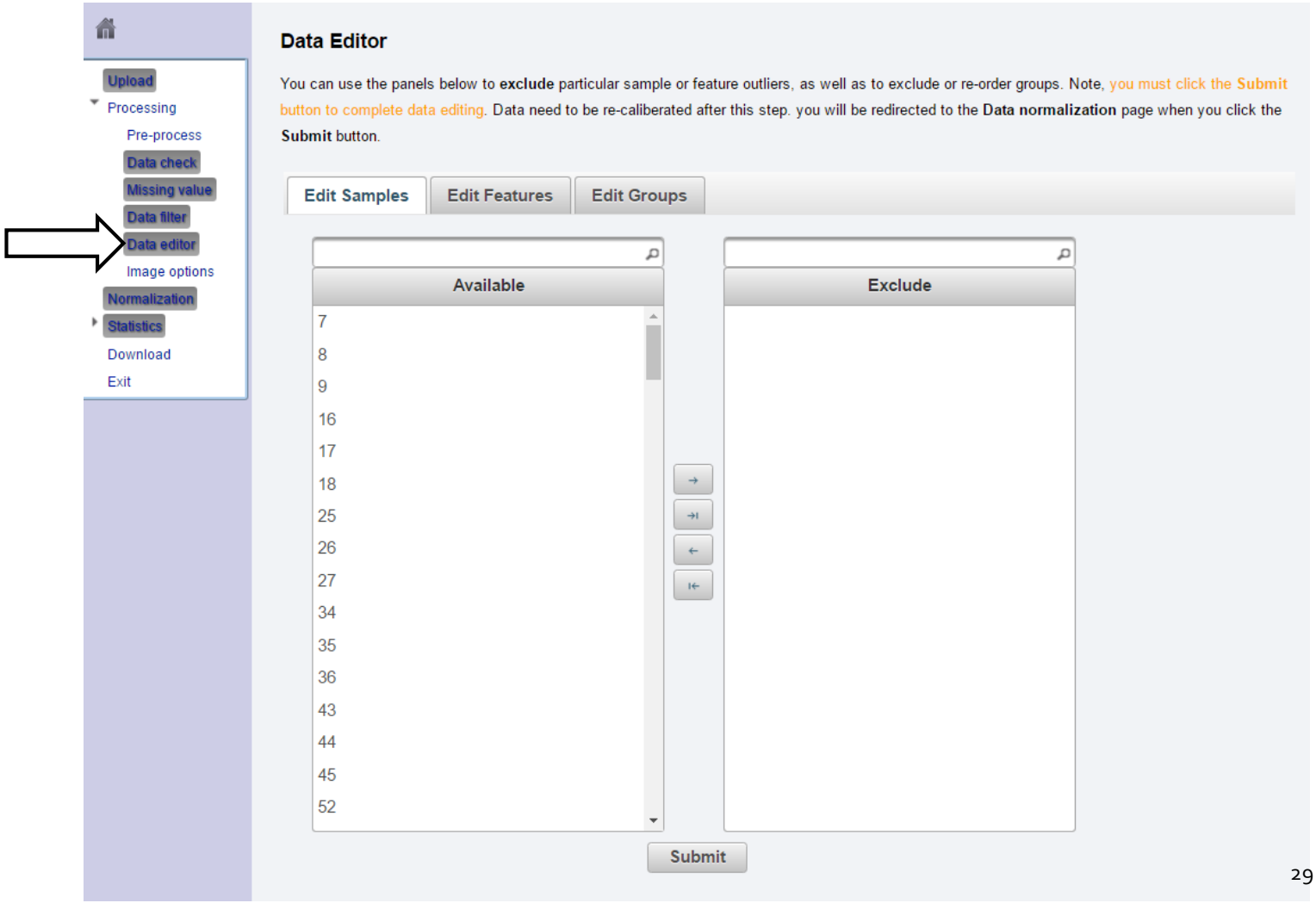
Finding outliers via PCA

Finding outliers via Heatmap

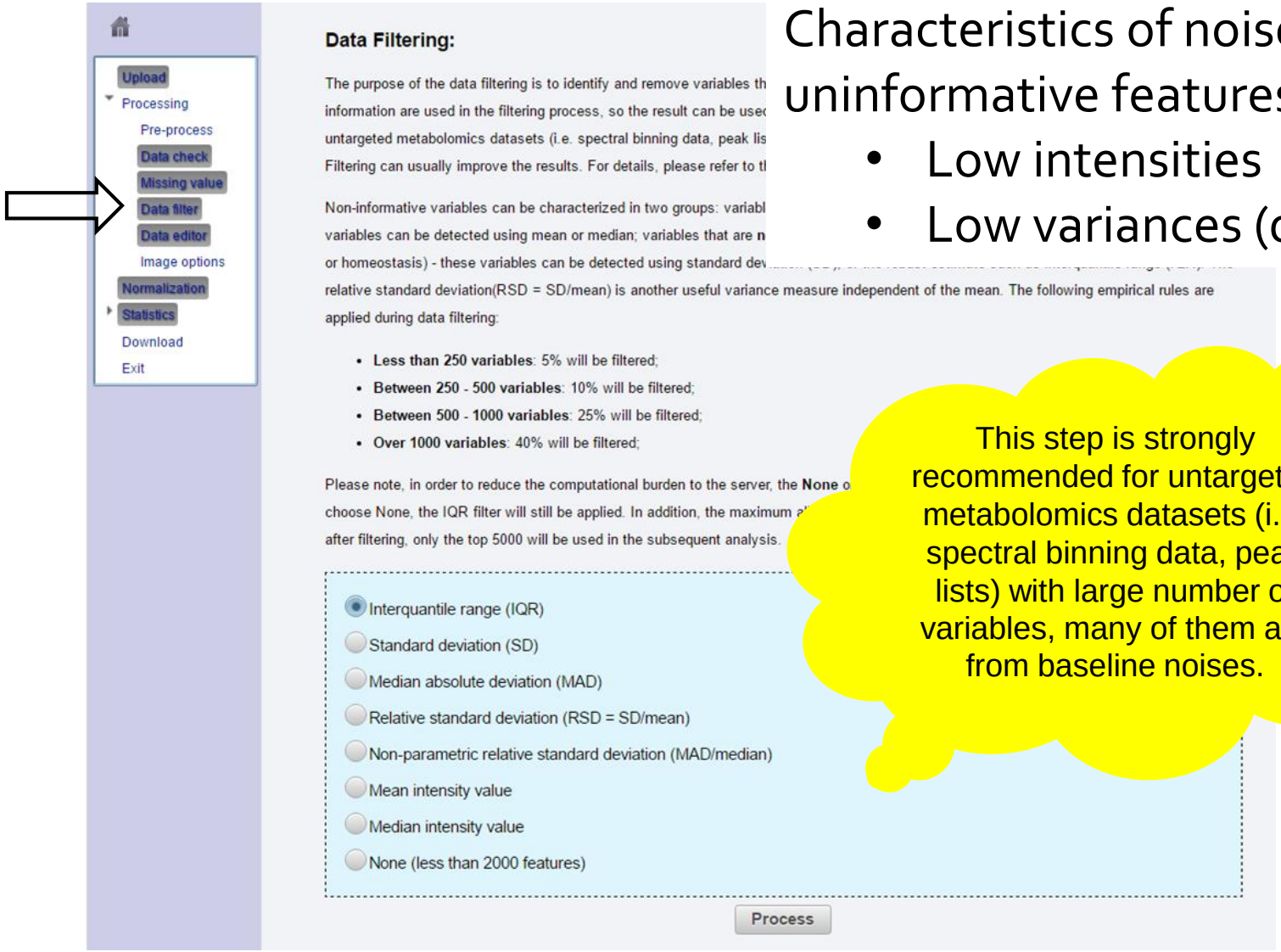
How to detect and deal with outlier?

- 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)



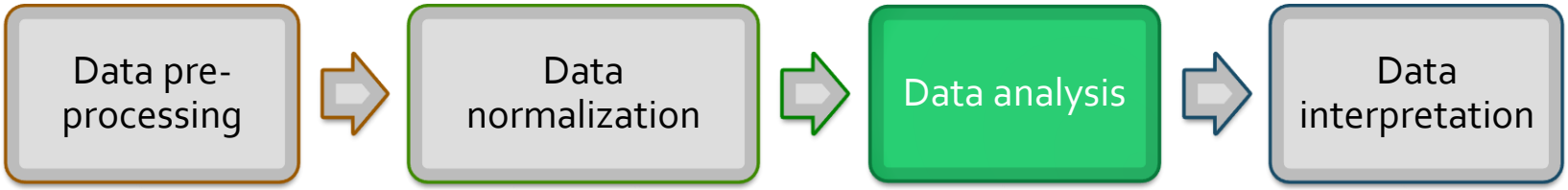
Noise Reduction (Data Filtering)



Characteristics of noise & uninformative features

- Low intensities
- Low variances (default)

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 several statistical analysis 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

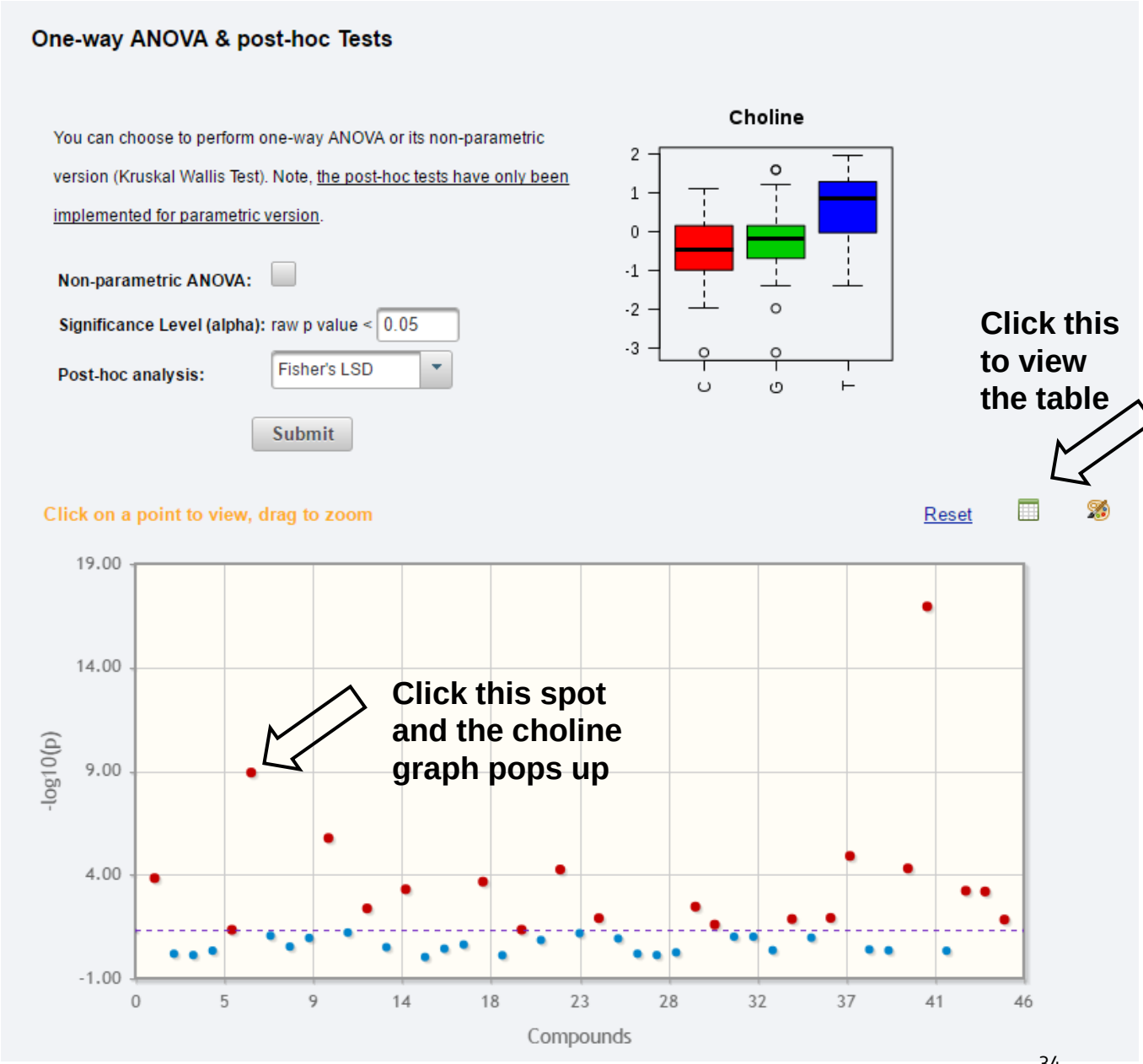
Partitional Clustering: K-means Self Organizing Map (SOM)

Classification & Feature Selection

Random Forest

Support Vector Machine (SVM)

ANOVA



Courtesy: Dr. David Wishart

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

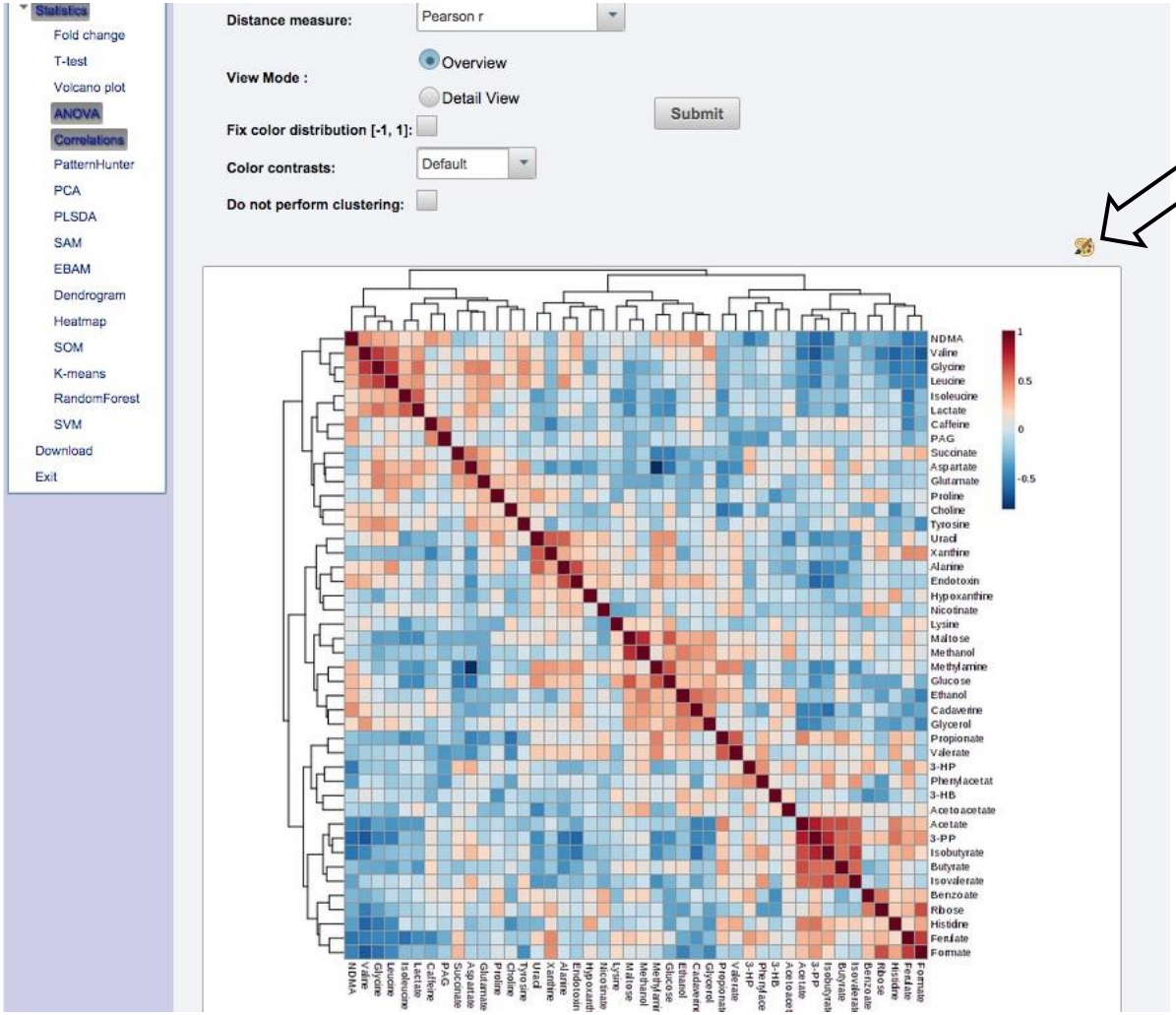
Partitional Clustering: K-means Self Organizing Map (SOM)

Classification & Feature Selection

Random Forest

Support Vector Machine (SVM)

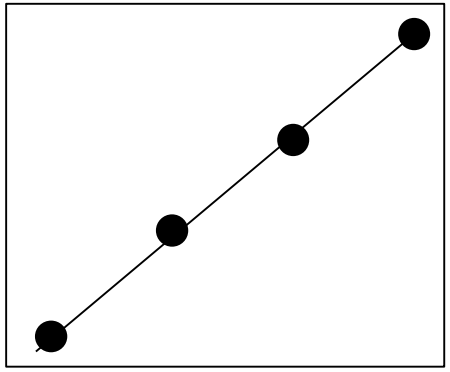
Overall Correlation Pattern

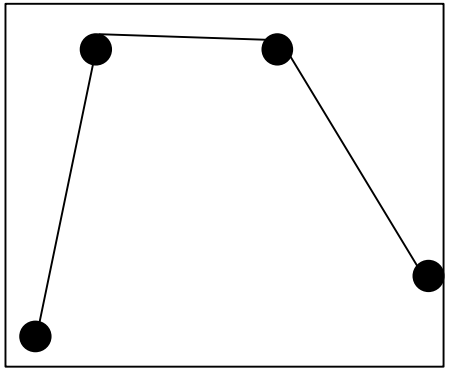


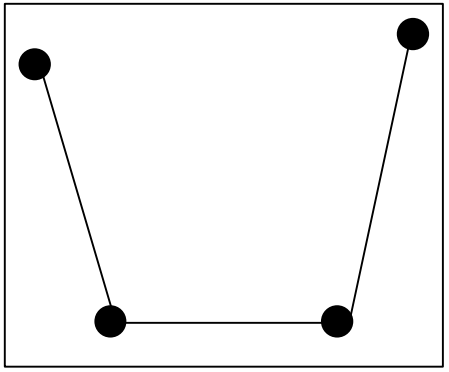
Click this to save a high res. image

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







Courtesy: Dr. David Wishart

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)

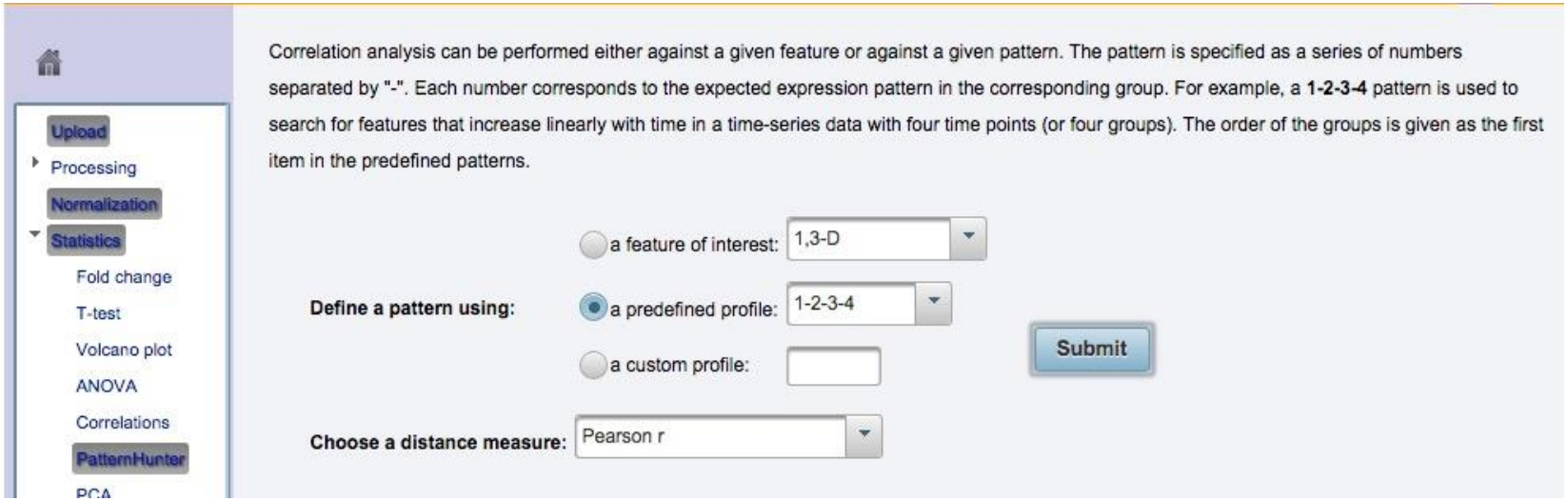
Classification & Feature Selection

Random Forest

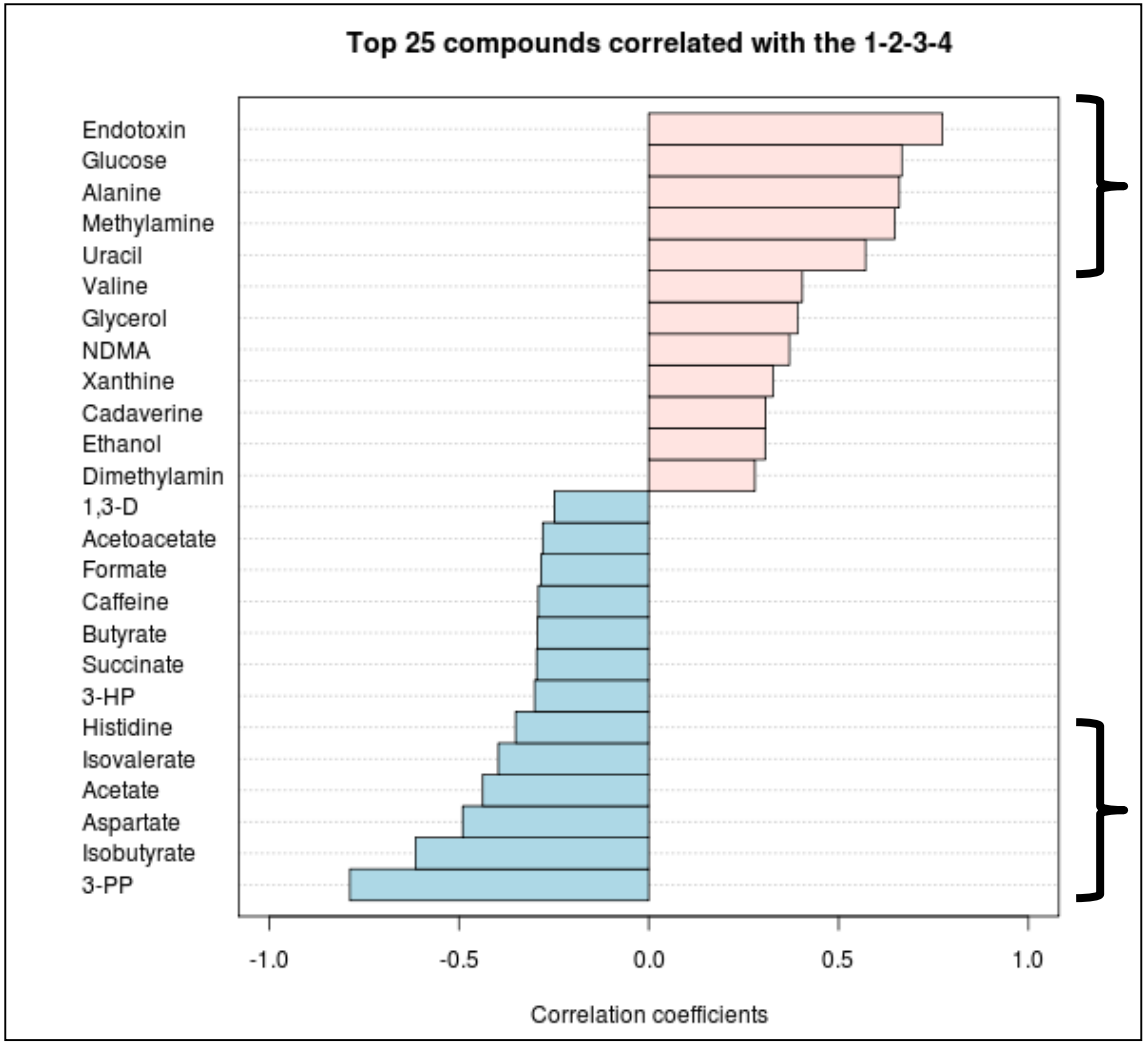
Support Vector Machine (SVM)

Pattern Searching

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



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

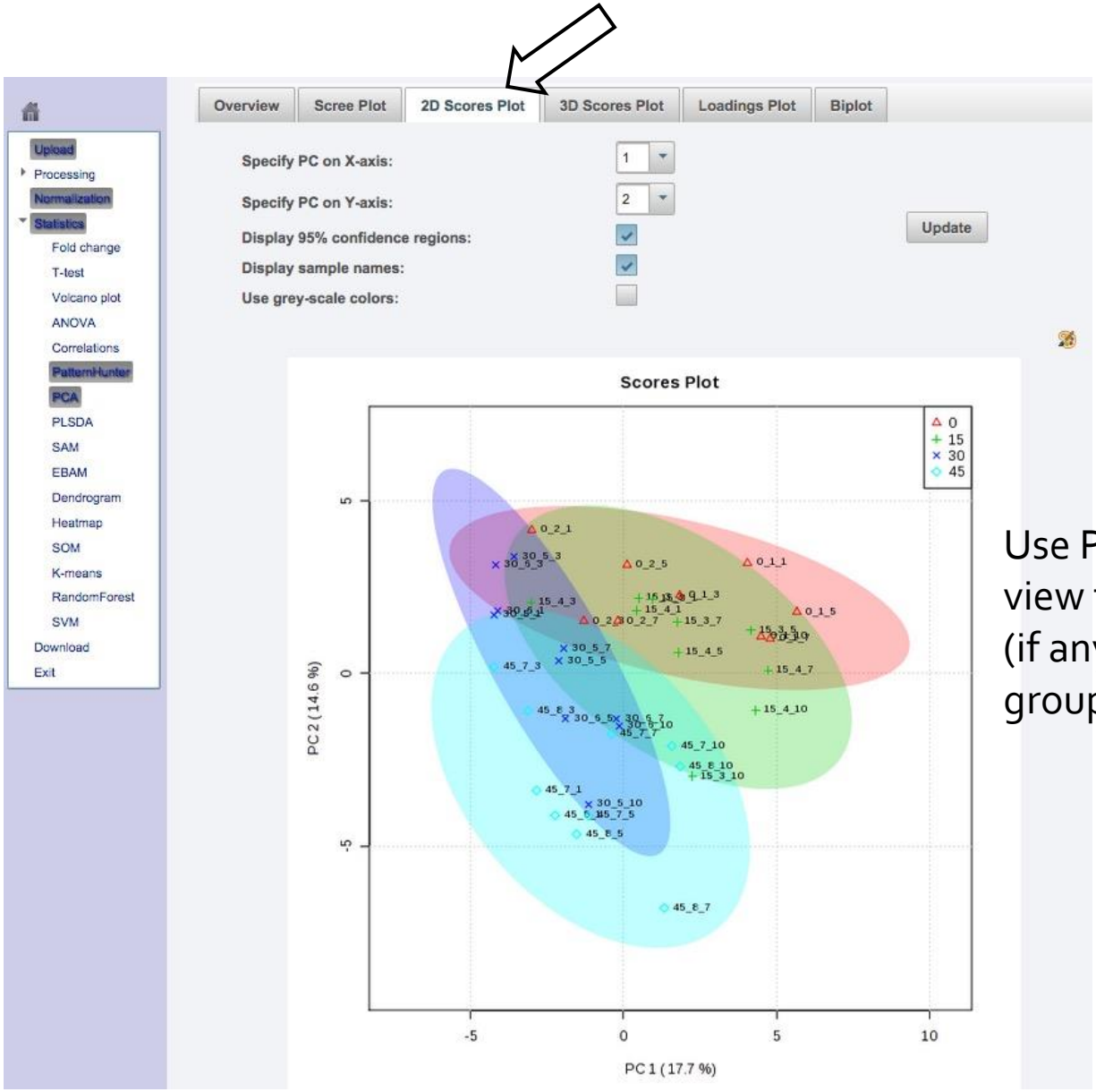
Partitional Clustering: K-means Self Organizing Map (SOM)

Classification & Feature Selection

Random Forest

Support Vector Machine (SVM)

PCA Scores Plot



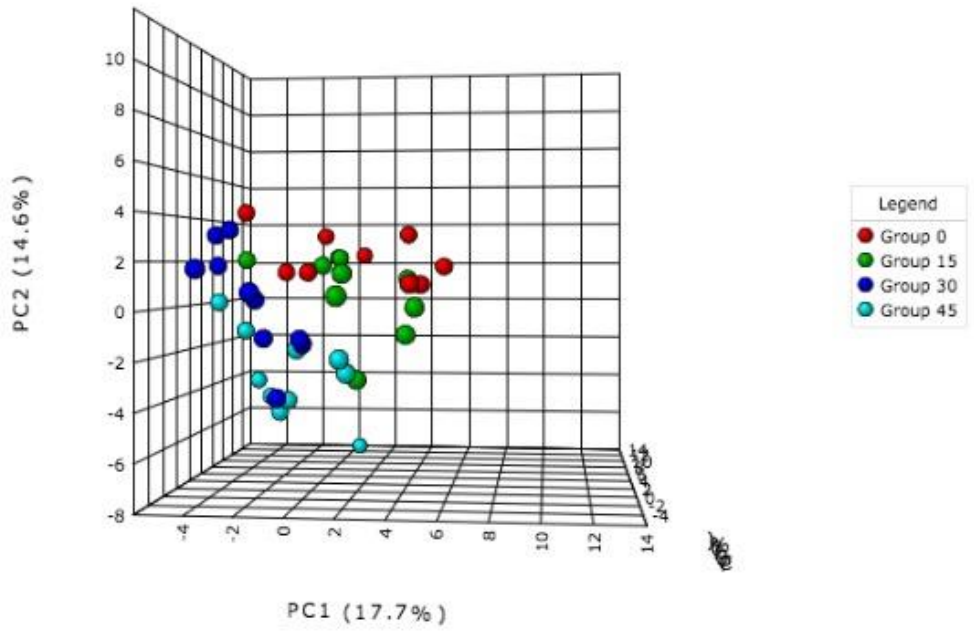
Use PCA option to view the separation (if any) in the groups

3D Score Plot

Overview | Scree Plot | 2D Scores Plot | 3D Scores Plot | Loadings Plot | Biplot

Click **Update** to load the plot; **Dragging** to rotate the view around the axis; **Clicking** on any data point to view a summary of the corresponding sample; **Scrolling** to zoom in and out; use the **Export** button below to export the current view as a PNG image.

Update | Export



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

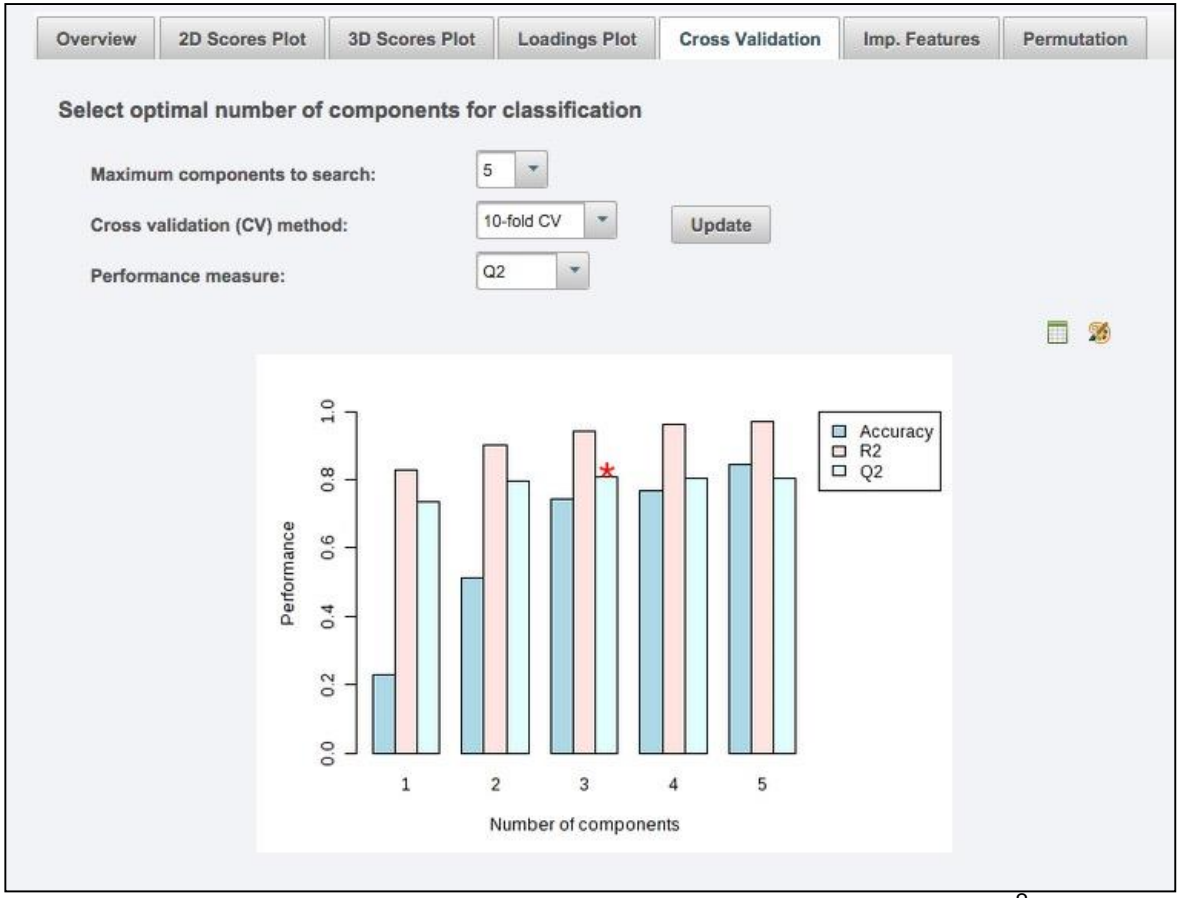
- Use PLS-DA option to view the separation of the (labeled) groups
- PLS-DA "rotates" the PCA axes to maximize separation
- Look at the 2D PLS Scores Plot
- Look at the Q^2 and R^2 (Cross Validation) values
- Use the VIP plot to ID important metabolites

PLS-DA Score Plot

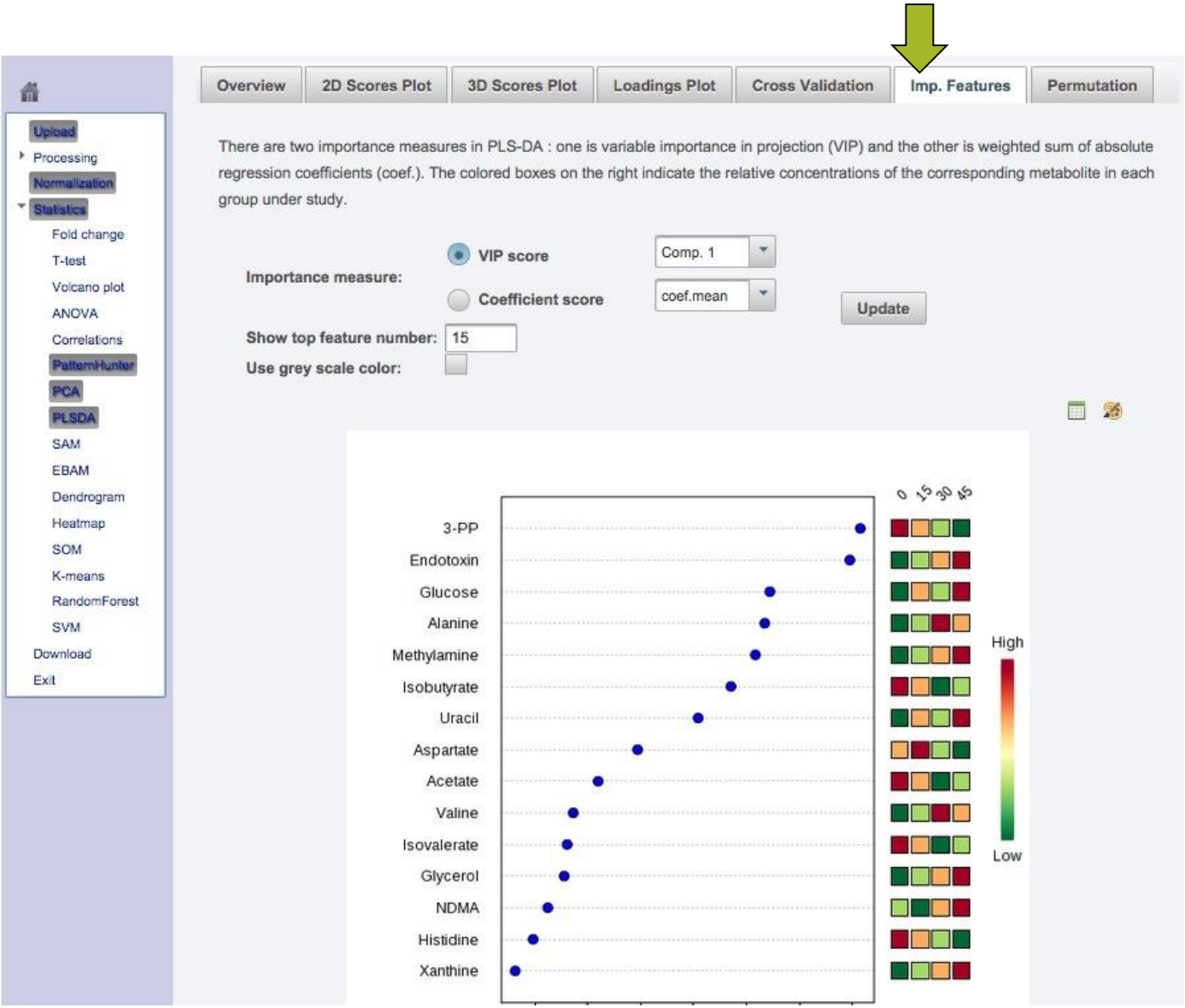
- Use PLS-DA option to view the separation of the (labeled) groups
- PLS-DA "rotates" the PCA axes to maximize separation
- Look at the 2D PLS Scores Plot
- Look at the Q^2 and R^2 (Cross Validation) values
- Use the VIP plot to ID important metabolites

Evaluation of PLS-DA Model

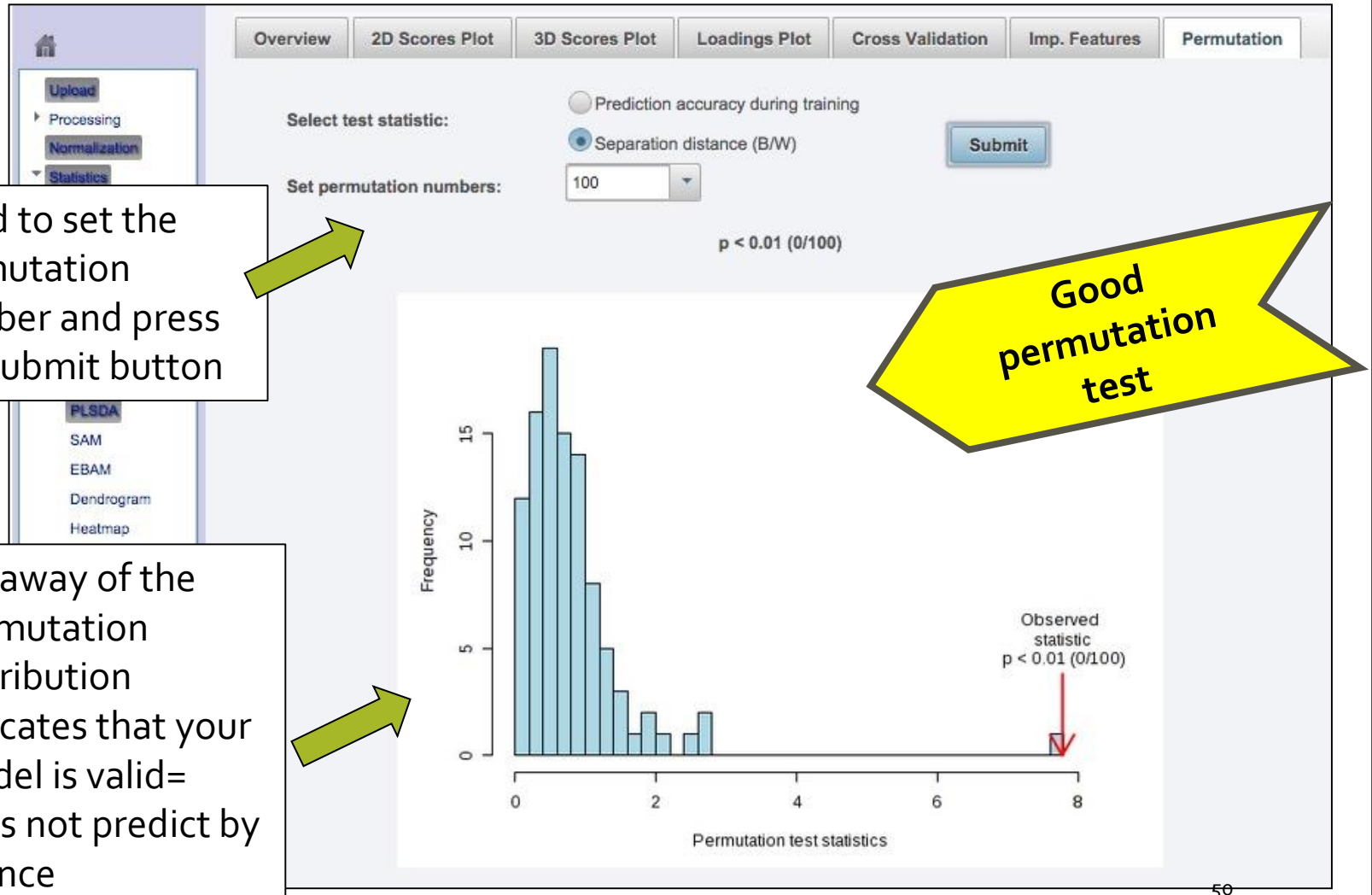
- PLS-DA Model evaluated by cross validation of Q^2 and R^2
- Using too many components can over-fit
- 3 component model seems to be a good compromise here
- Better R^2 and Q^2 as closer to 1



Important Compounds in the Model



Model Validation



Need to set the permutation number and press the submit button

Far away of the permutation distribution indicates that your model is valid= does not predict by chance

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

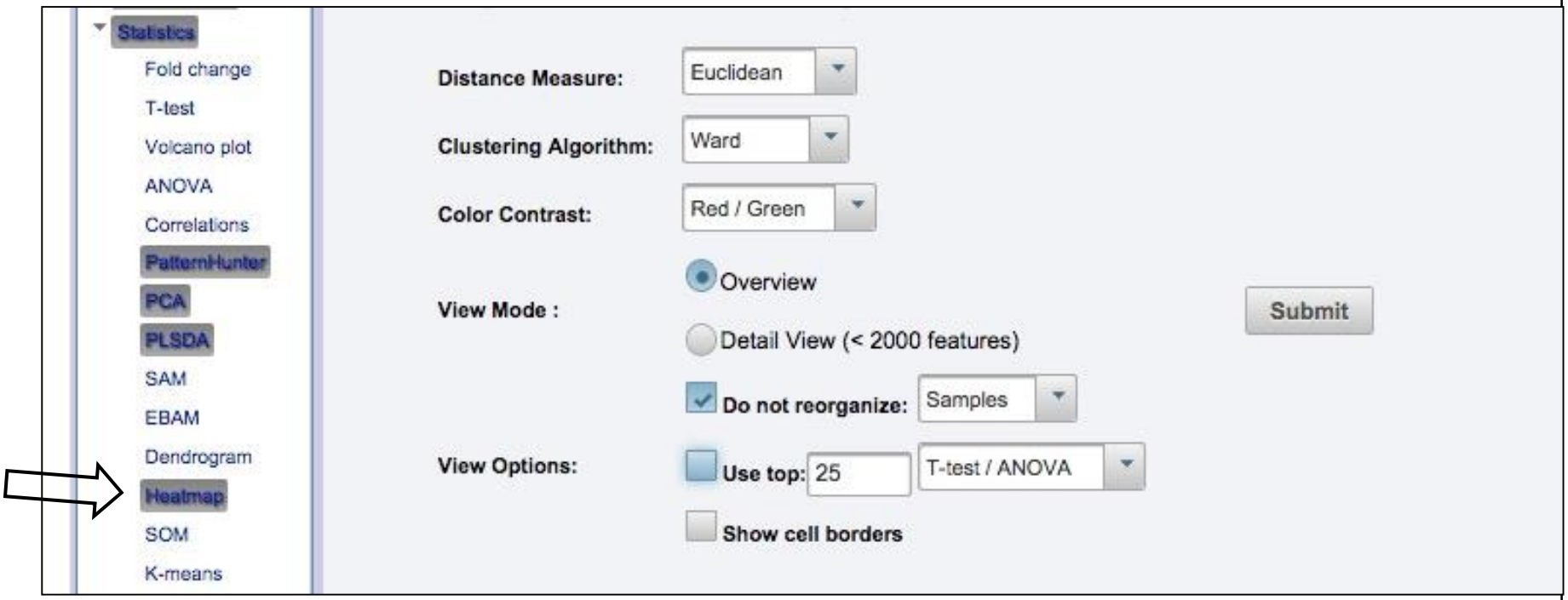
Partitional Clustering: K-means Self Organizing Map (SOM)

Classification & Feature Selection

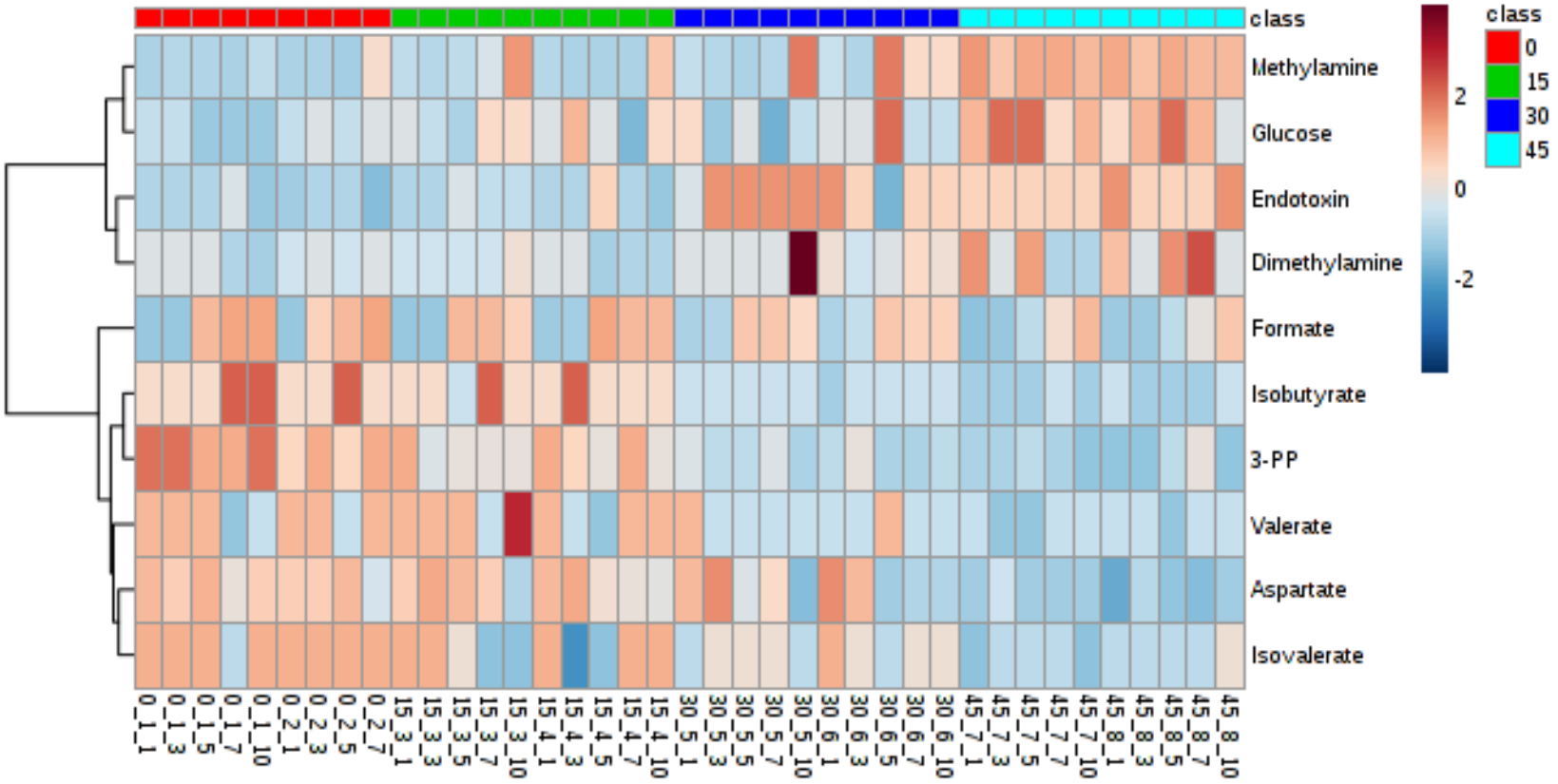
Random Forest

Support Vector Machine (SVM)

Heatmap Visualization



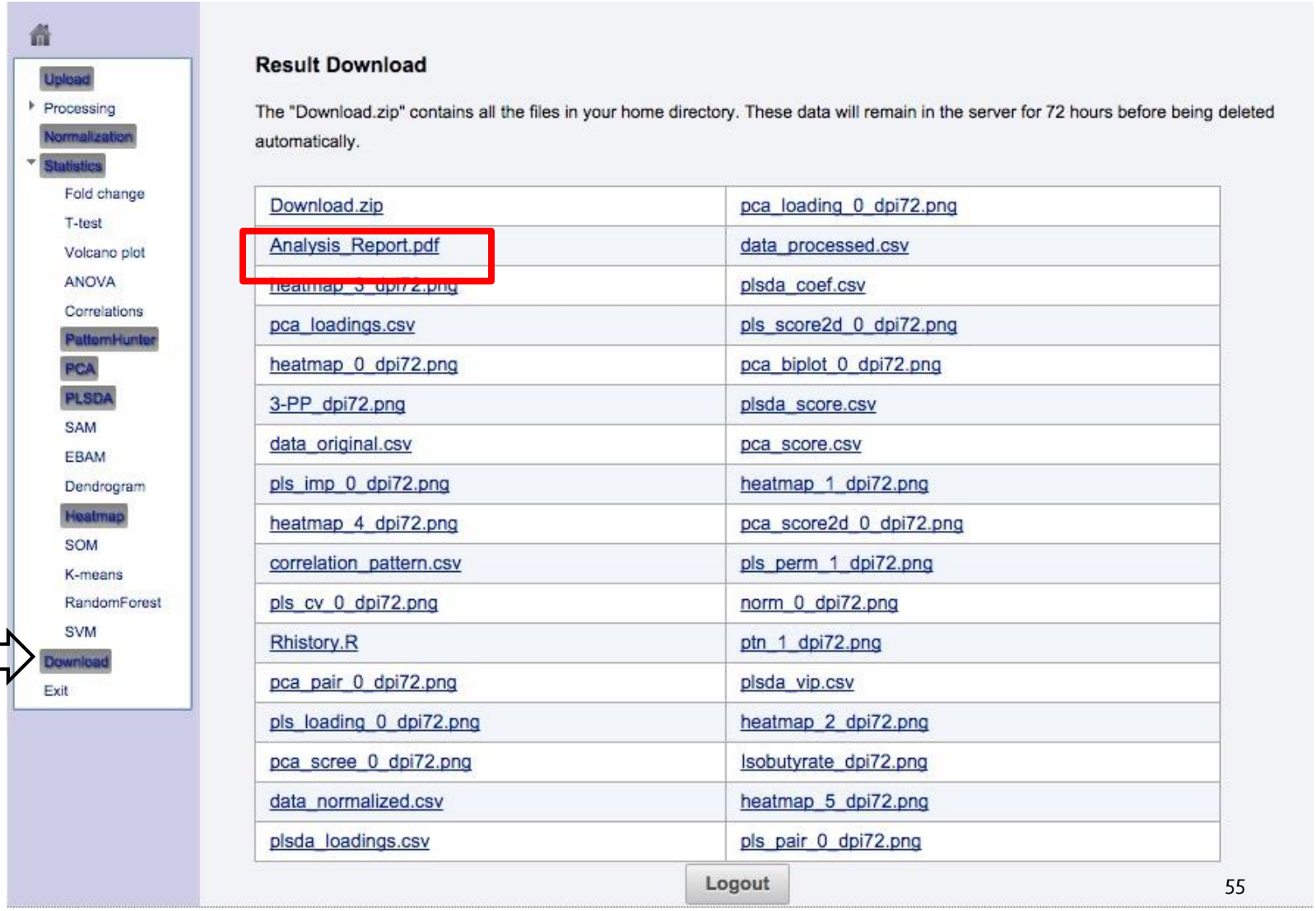
Heatmap Visualization (cont.)



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



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	-0.81292	19332	3.4067e-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	0.00024011	0.0025416
5	3-HB	-0.41943	14024	0.007862	0.041097
6	Isovalerate	-0.39861	13818	0.011096	0.056193
7	Lysine	-0.24401	12291	0.13439	0.30381
8	Methanol	-0.24257	12277	0.13878	0.30381
9	Ferulate	-0.22929	12145	0.16028	0.32753
10	Fumarate	-0.21966	12050	0.17906	0.33398
11	Histidine	-0.2169	12023	0.18474	0.33398
12	Propionate	-0.21015	11956	0.19912	0.34861
13	Maltose	-0.2003	11859	0.22148	0.37177
14	Acetoacetate	-0.17772	11638	0.27907	0.39748
15	Choline	-0.11886	11054	0.47111	0.65124
16	Tyrosine	-0.10657	10933	0.51847	0.67689
17	PAG	-0.079788	10668	0.62921	0.76957
18	3-HP	-0.074918	10620	0.65033	0.80438
19	Formate	-0.051347	10387	0.75823	0.84626
20	Aspartate	-0.031981	10198	0.84674	0.86515
21	Caffeine	0.011541	9763	0.94297	0.94297
22	Ribose	0.038043	9495.1	0.81357	0.85054
23	1,3-D	0.043168	9443.8	0.79419	0.84834
24	Succinate	0.04904	9435	0.78542	0.84834
25	Glucose	0.057544	9311.8	0.72787	0.83439
26	Cadaverine	0.060643	9280.8	0.71382	0.83439
27	Phenylacetate	0.063742	9250.2	0.69986	0.83439
28	Hypoxanthine	0.10011	8802	0.50847	0.67689
29	Ethanol	0.18304	8071.8	0.26471	0.3888
30	NDMA	0.18492	8053	0.25975	0.3888
31	Proline	0.18713	8031.2	0.25399	0.3888
32	Glutamate	0.19354	7969.8	0.23829	0.38819
33	Benzoate	0.21978	7718.8	0.17854	0.33398
34	Valerate	0.23936	7515.1	0.14221	0.30381
35	Glycerol	0.26901	7213.3	0.096869	0.23888
36	Glycine	0.28064	7107.3	0.083833	0.21812
37	Nicotinate	0.28412	7083	0.078811	0.21708
38	Methylamine	0.28908	7024.2	0.07431	0.21708
39	Isoleucine	0.30355	6881	0.060303	0.18896
40	Xanthine	0.30654	6861.3	0.058565	0.18896
41	Dimethylamine	0.33298	6590.1	0.038326	0.13856
42	Leucine	0.35142	6407.9	0.028264	0.11066
43	Valine	0.3509	6116.7	0.016744	0.071541
44	Lactate	0.42354	5692.5	0.0071709	0.041097
45	Uracil	0.45172	5417	0.0038928	0.026137
46	Endotoxin	0.50141	4926.1	0.0011471	0.0089853
47	Alanine	0.62026	3751.8	2.5337e-05	0.00080058

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.

Hierachical 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.

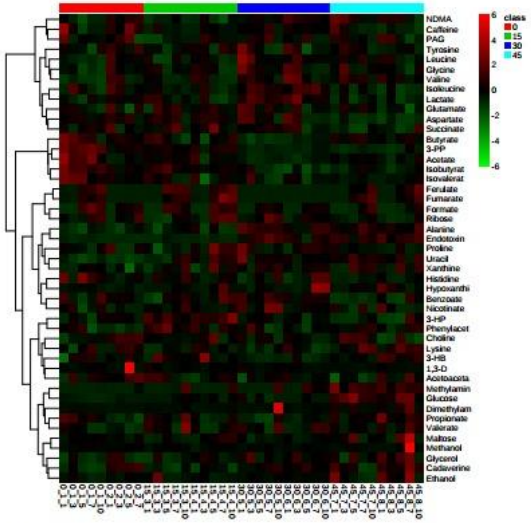
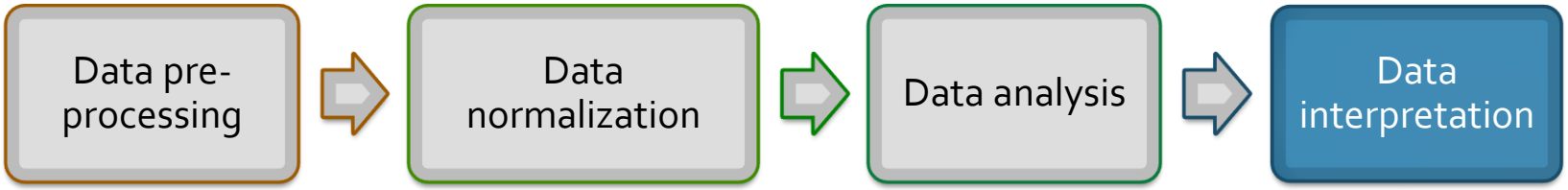


Figure 15: Clustering result shown as heatmap (distance measure using `euclidean`, and clustering algorithm using `ward`).

18

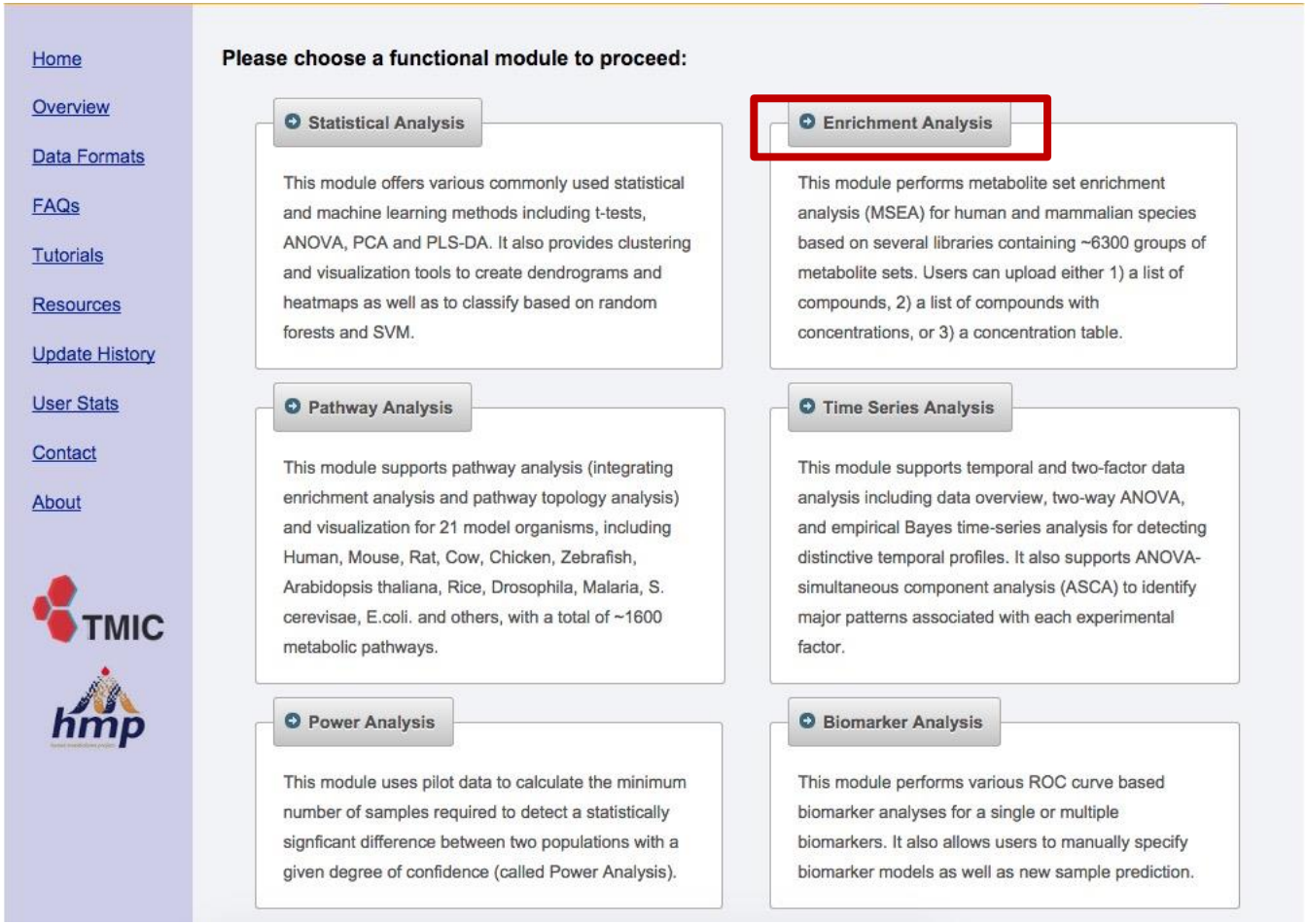


Metabolite enrichment analysis

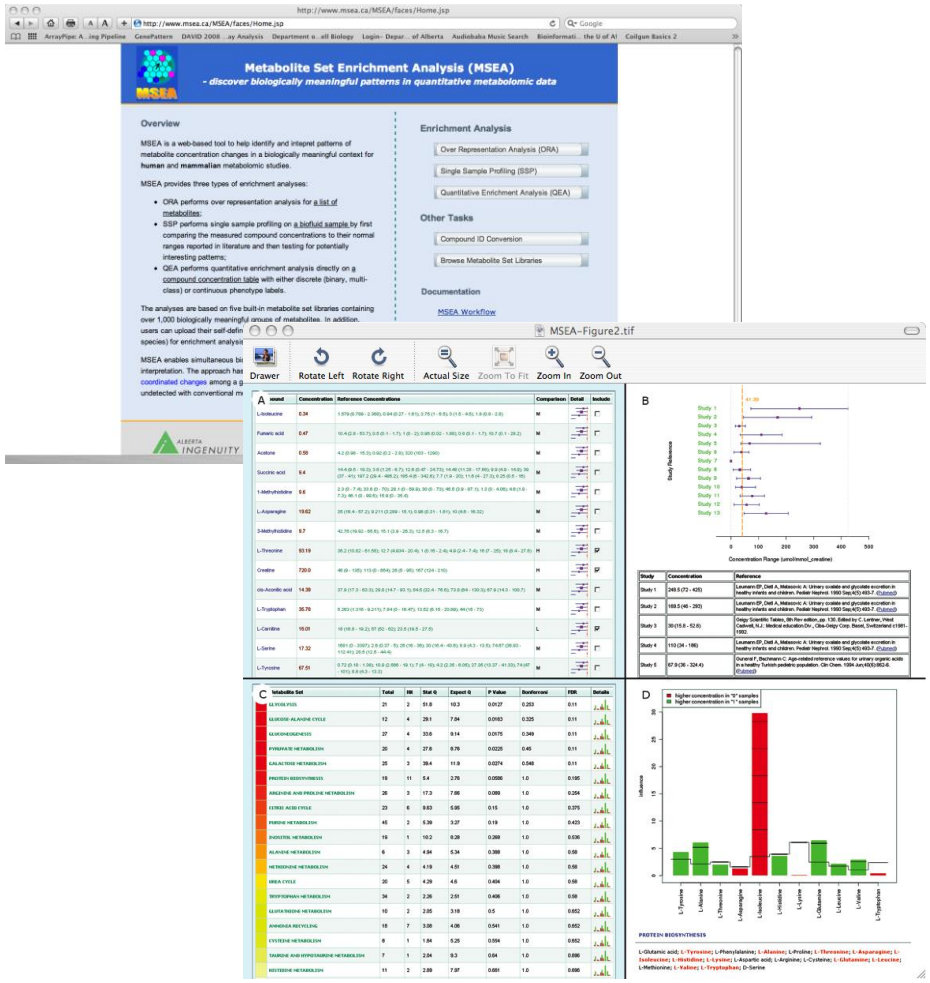
Pathway analysis

Biomarker Analysis

Select a Module (Enrichment Analysis)



Metabolite Set Enrichment Analysis (MSEA)



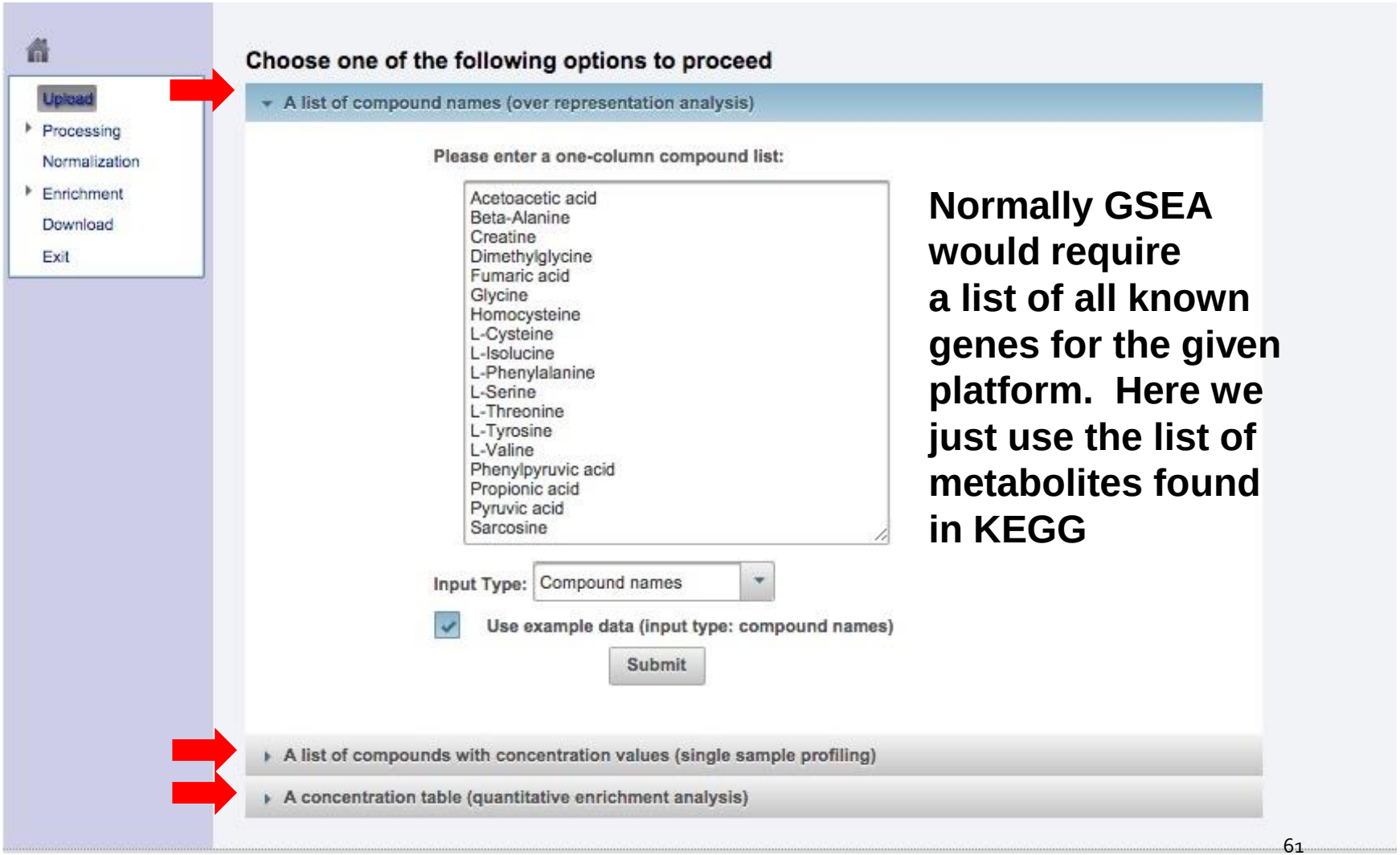
- Designed to handle lists of metabolites (with or without concentration data)
- Modeled after Gene Set Enrichment Analysis (GSEA)
- Supports over representation analysis (ORA), single sample profiling (SSP) and quantitative enrichment analysis (QEA)
- Contains a library of 6300 pre-defined metabolite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

Courtesy: Dr. David Wishart

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



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:

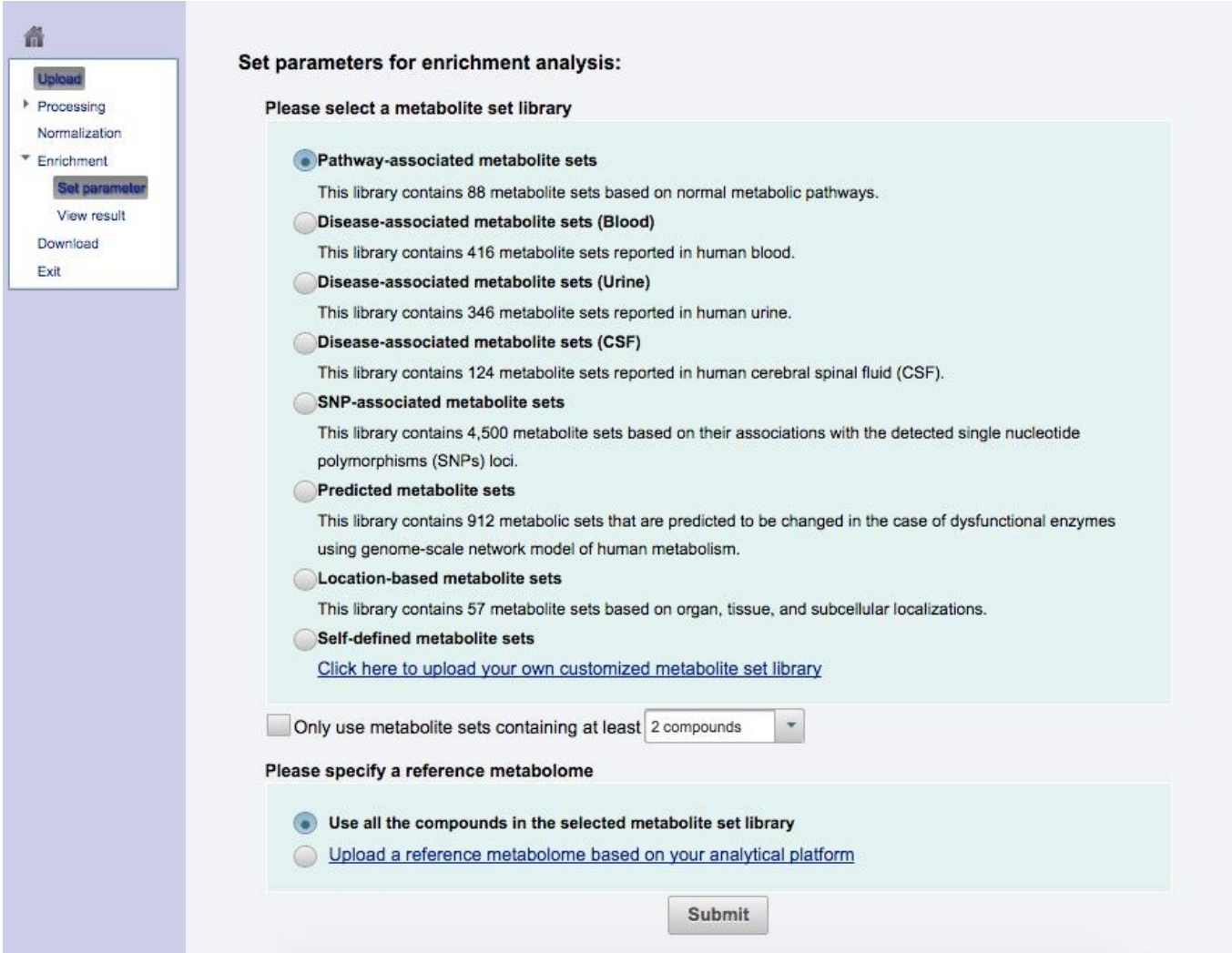
- Query names in normal white indicate exact match - marked by "1" in the download file;
- Query names highlighted in red indicate **no match** - marked by "0" in the downloaded file;
- For compound name mapping, the no match query names will be highlighted in yellow indicate **no exact match found**. 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	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	444972	C00122	
Glycine	Glycine	HMDB00123	750	C00037	
Homocysteine	Homocysteine	HMDB00742	778	C05330	
L-Cysteine	L-Cysteine	HMDB00574	5862	C00097	
L-Isolucine		-	-	-	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	6287	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	

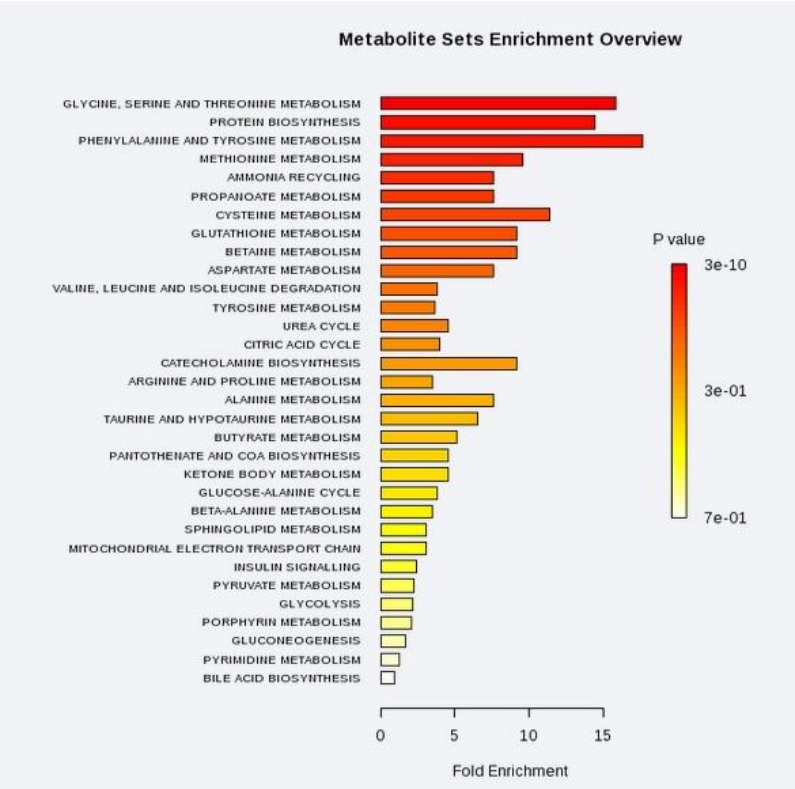
You can download the result here

Submit

Select a Metabolite Set Library



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	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
	INSULIN SIGNALLING	19	1	0.415	0.345	1.0	1.0	View
	PYRUVATE METABOLISM	20	1	0.436	0.36	1.0	1.0	View
	GLYCOLYSIS	21	1	0.458	0.374	1.0	1.0	View
	PORPHYRIN METABOLISM	22	1	0.48	0.388	1.0	1.0	View
	GLUCONEOGENESIS	27	1	0.589	0.454	1.0	1.0	View
	PYRIMIDINE METABOLISM	36	1	0.785	0.556	1.0	1.0	View
	BILE ACID BIOSYNTHESIS	49	1	1.07	0.672	1.0	1.0	View

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	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 METAB						8	View
	AMMONIA RECYCLIN						4	View
	PROPANOATE METAI						4	View
	CYSTEINE METABOL							View
	GLUTATHIONE META							View
	BETAINE METABOLIS							View
	ASPARTATE METABC							View
	VALINE, LEUCINE AN							View
	TYROSINE METABOL							View
	UREA CYCLE							View
	CITRIC ACID CYCLE							View
	CATECHOLAMINE BI							View
	ARGININE AND PROL							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

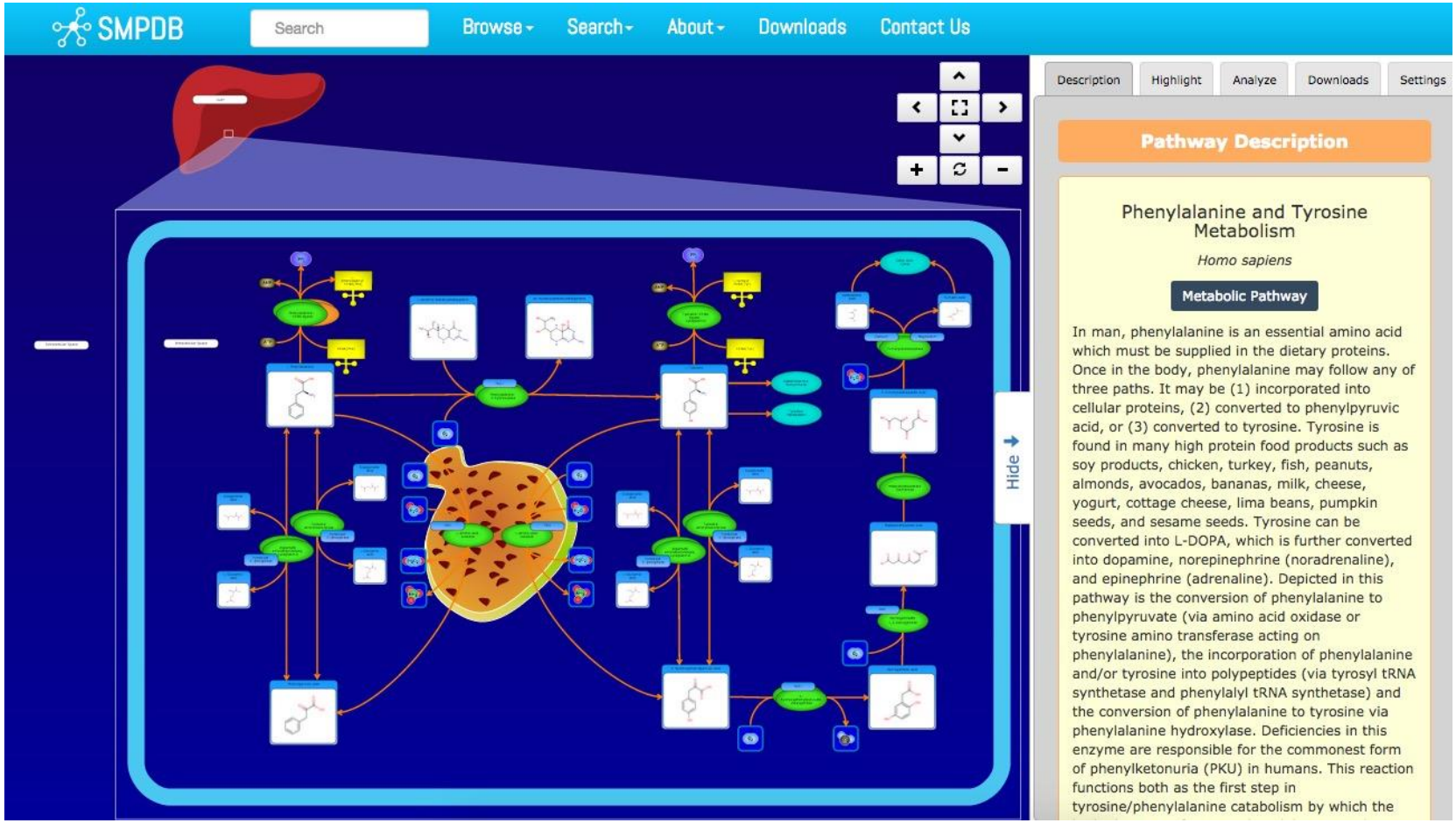
Current metabolite set:

Set Name	Metabolites	References
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	SMPDB

OK

Click on SMPDB to see more information

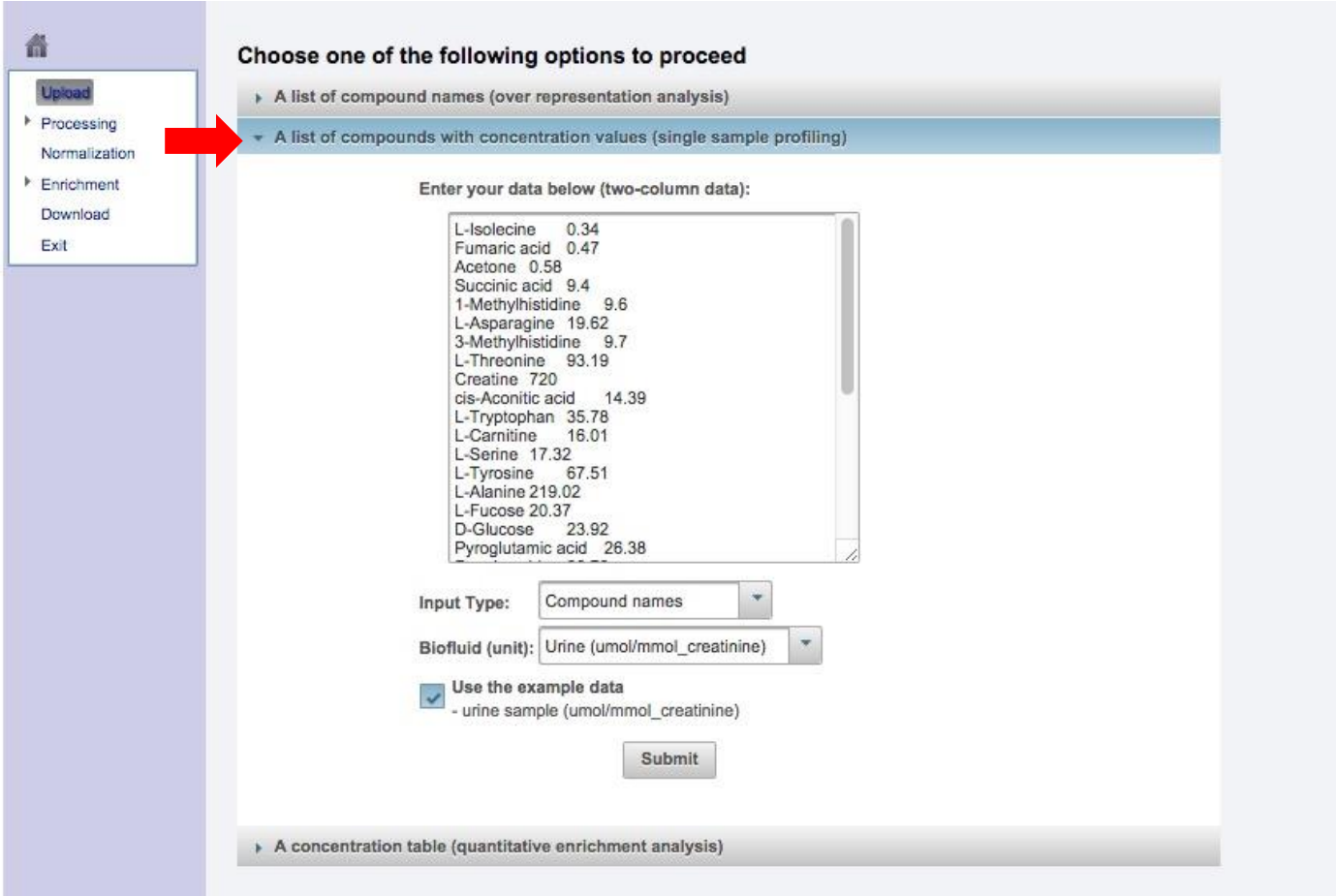
Metabolism of the compound of interest in SMPDB



Single Sample Profiling (SSP)

Basically used by doctors to analyze a patient

Aim: compare to normal references



Concentration Comparison

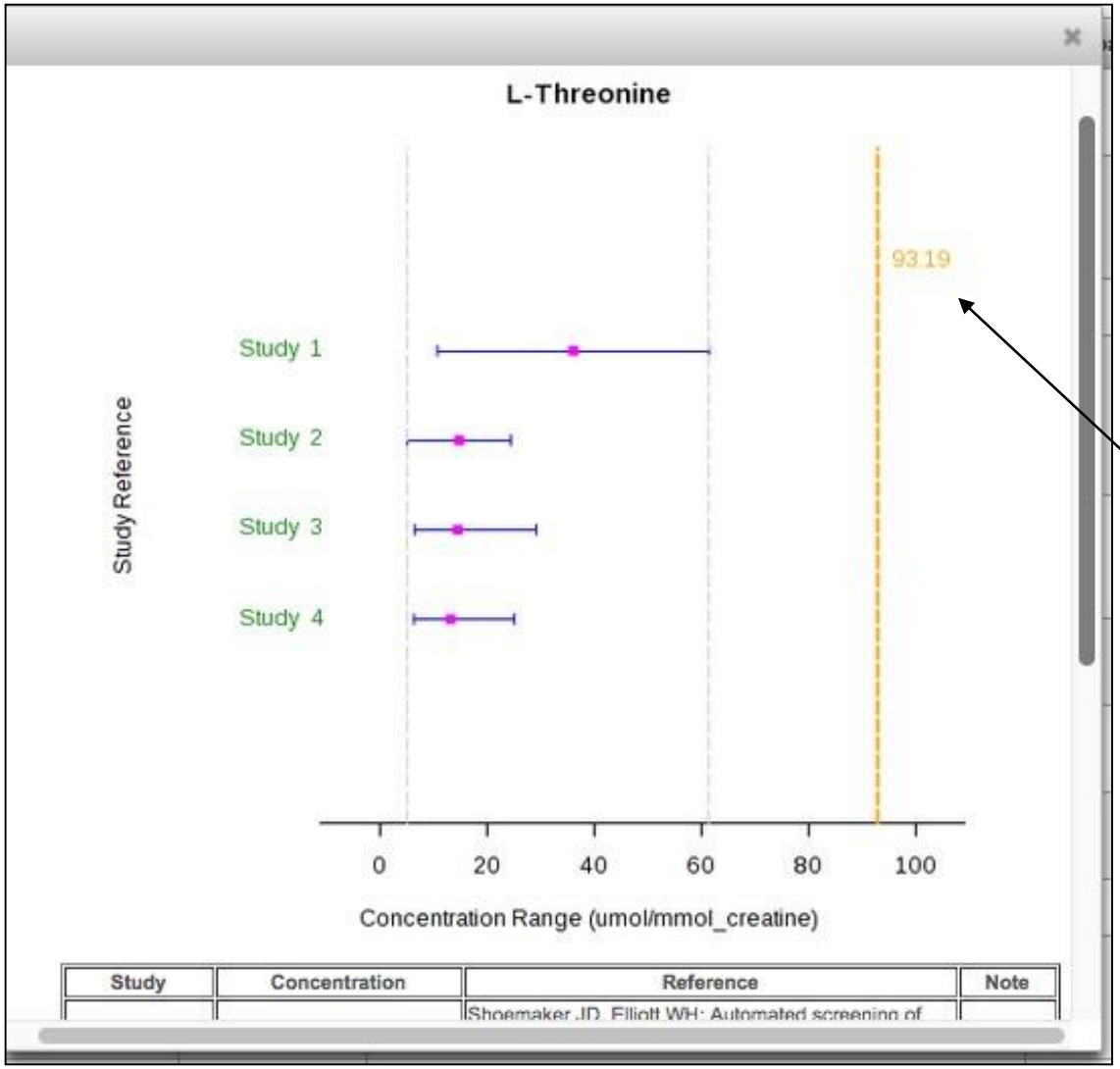
- Upload
- Processing
  - Name check
  - Conc. check
  - Data check
  - Missing value
  - Data filter
  - Data editor
  - Image options
- Normalization
- Enrichment
- Download
- Exit

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.

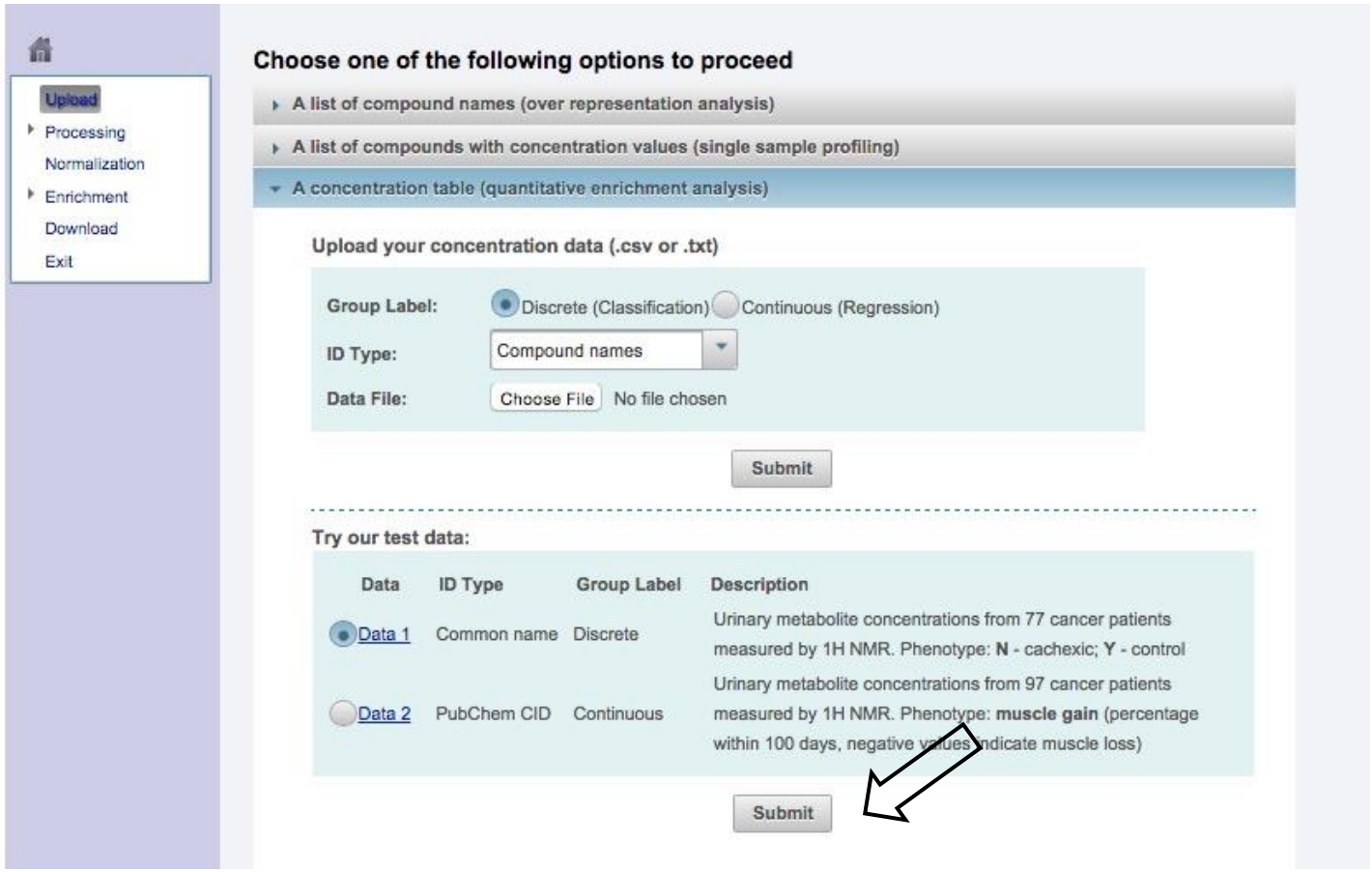
Compound	Concentration	Reference concentrations	Comparison	Detail	Include
L-Isoleucine	0.34	3.75 (1 - 6.5); 3.55 (1.7 - 5.4); 0.02125 (0.0086 - 0.0339); 1.3 (0.5 - 2.7); 1.3 (0.4 - 2.6)	M	View	☐
Fumaric acid	0.47	0.95 (0.02 - 1.88); 0.4 (0.2 - 0.8); 10.4 (2.8 - 53.7); 0.5 (0.1 - 1.7); 10.7 (0.1 - 28.2); 0.1 (0.1 - 1.7); 0.25 (0.1 - 0.4); 0.7 (0.2 - 1.7)	M	View	☐
Acetone	0.58	2.24 (0 - 6.37); 3.9 (0.8 - 17.6)	M	View	☐
Succinic acid	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	View	☐
1-Methylhistidine	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	View	☐
L-Asparagine	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	View	☐
3-Methylhistidine	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	View	☐
L-Threonine	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	View	☑
Creatine	720	113 (0 - 654); 113 (0 - 654); 46 (3 - 448)	H	View	☑
cis-Aconitic acid	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	View	☐
L-Tryptophan	35.78	13.52 (6.15 - 20.89); 5.6 (9.3 - 2.1); 6.3 (3.4 - 11.1)	H	View	☑
L-Carnitine	16.01	4.5 (0.62 - 15.2); 5 (0.7 - 16.4)	M	View	☐

Concentration Comparison (cont.)

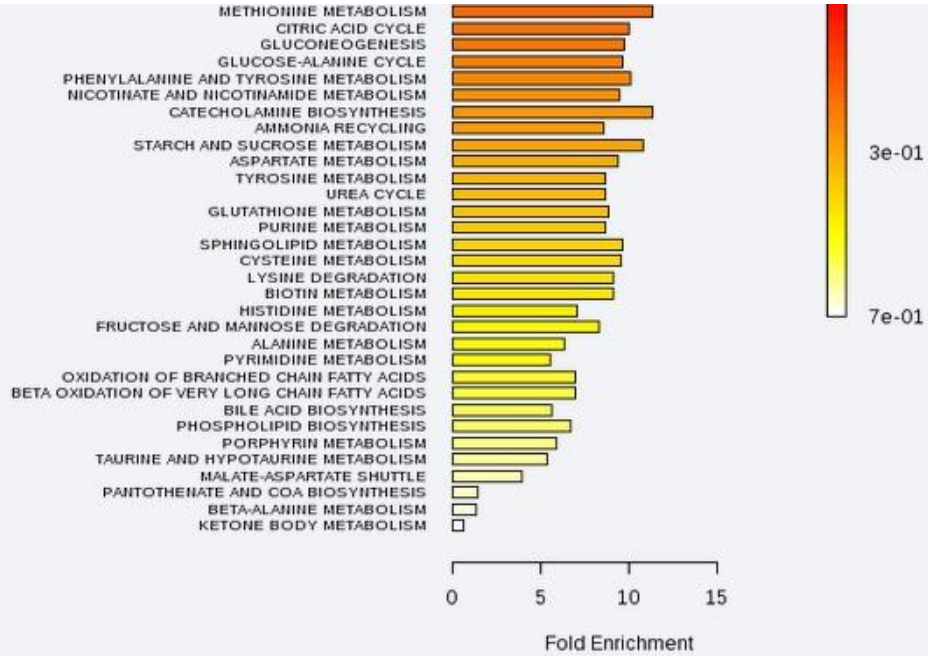


Your value. It´s out of the normal range

Quantitative Enrichment Analysis (QEA)



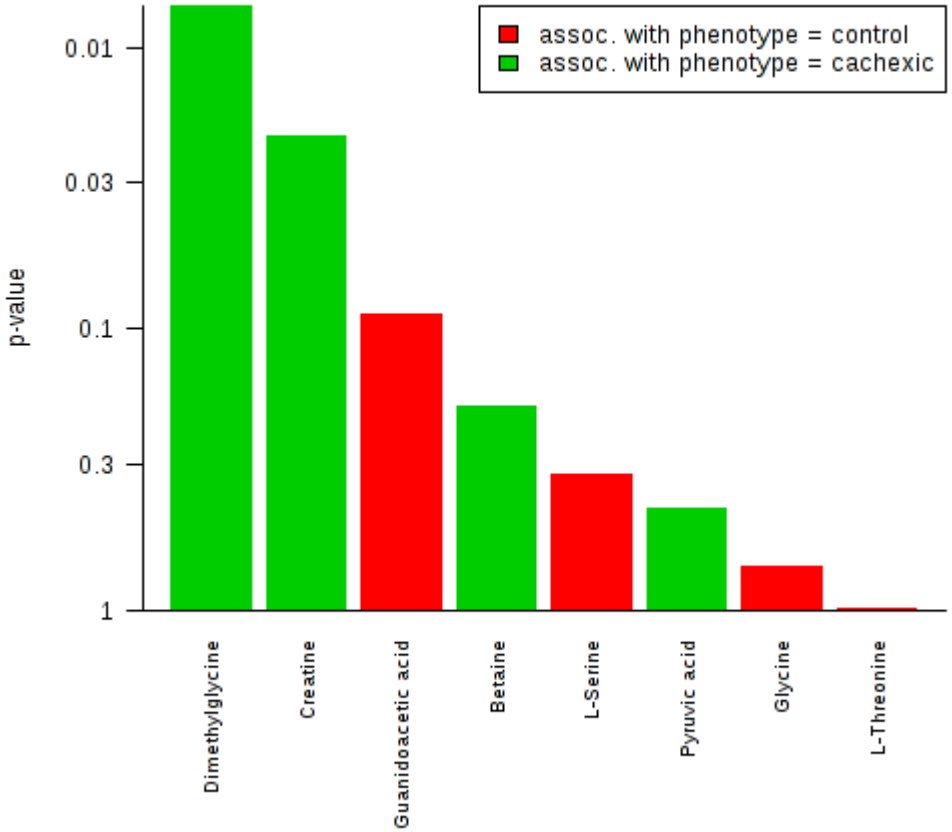
Result



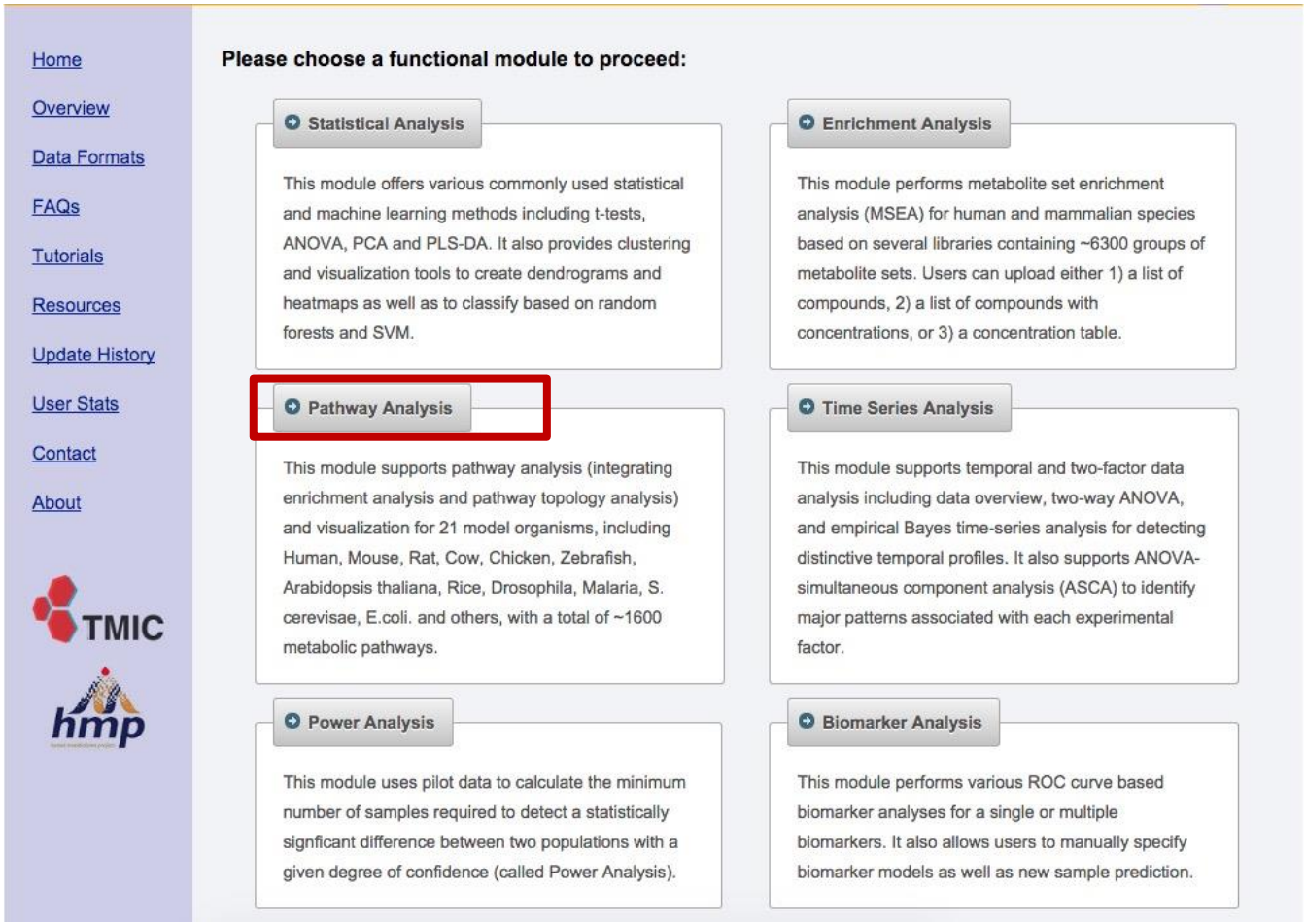
	Metabolite Set	Total	Hits	Statistic	Expected	P value	Holm P	FDR	Details
	GALACTOSE METABOLISM	25	3	18.866	1.3158	1.4154E-6	6.5107E-5	6.5107E-5	View
	TRYPTOPHAN METABOLISM	34	1	25.111	1.3158	3.4524E-6	1.5536E-4	7.9406E-5	View
	VALINE, LEUCINE AND ISOLEUCINE DEGRADATION	36	2	21.24	1.3158	1.569E-5	6.9038E-4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

Click on details to see more

The Matched Metabolite Set



Pathway Analysis Module

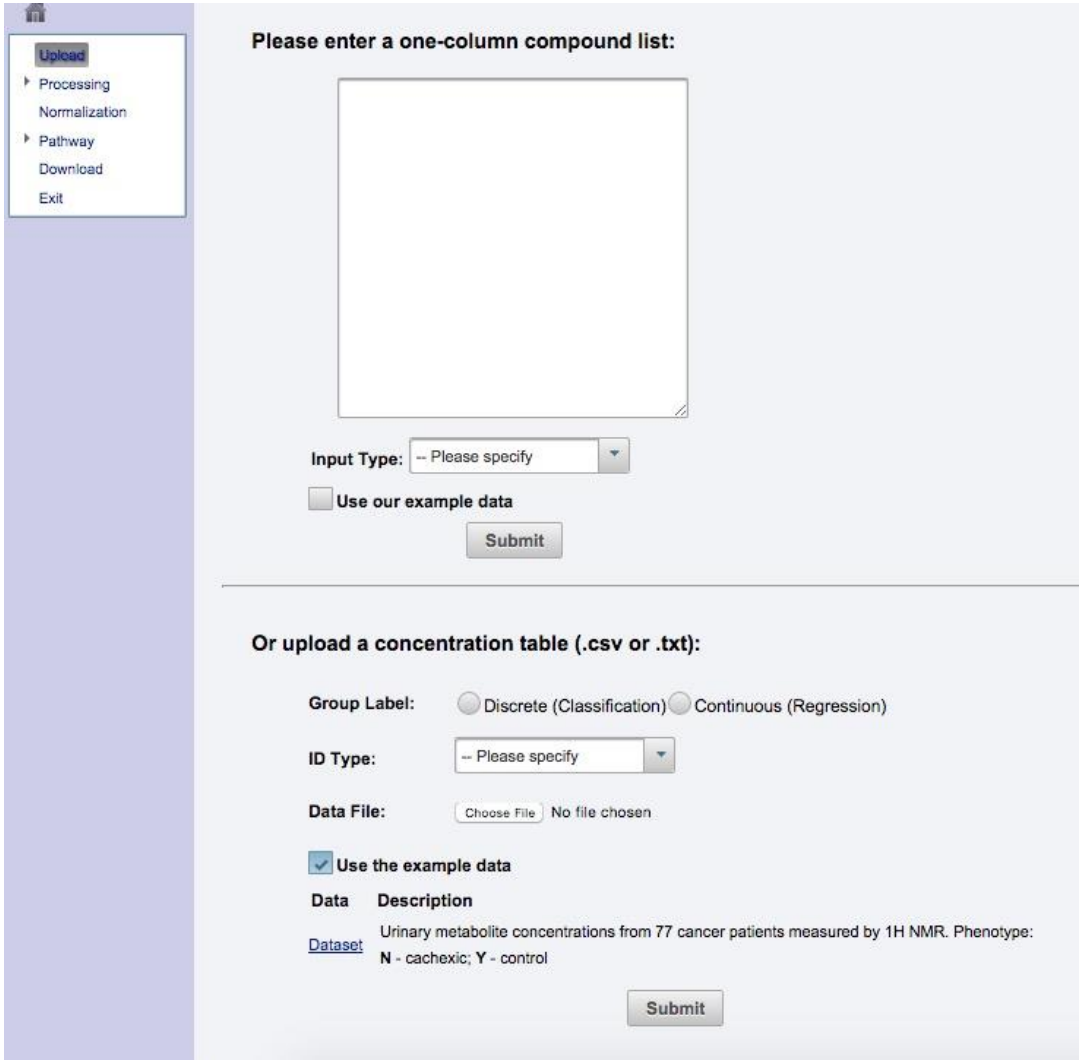


Pathway Analysis

- 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:
Choose File
No file chosen
Use the example data
Data
Description
Dataset
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: PIF_178
  - ( ) Create a pooled average sample from group: cachexic
- ( ) Normalization by reference feature: 1,6-Anhydro-beta-D-glucose

Data transformation

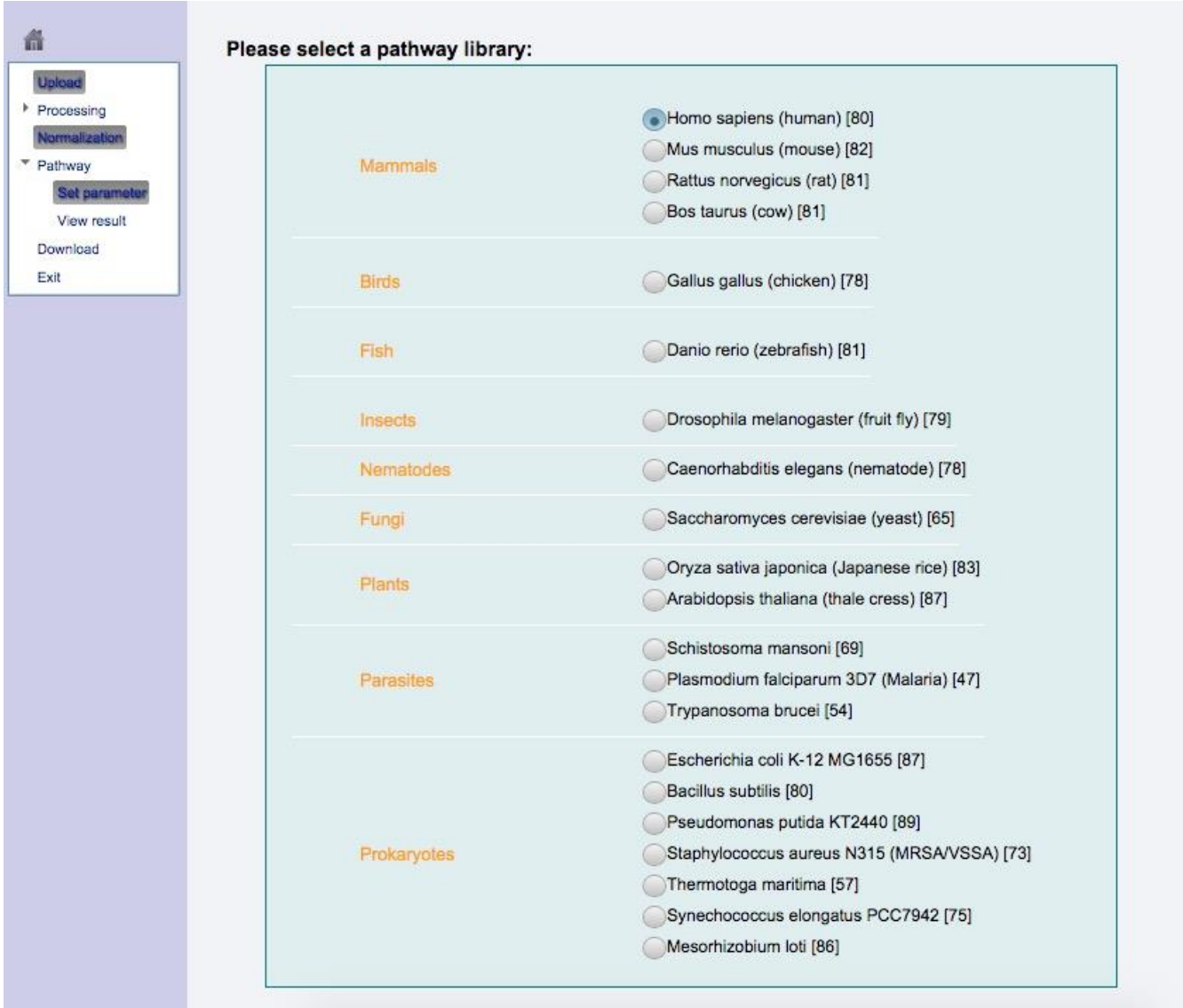
- (●) None
- ( ) Log transformation (generalized logarithm transformation or glog)
- ( ) Cube root transformation (take cube root of data values)

Data scaling

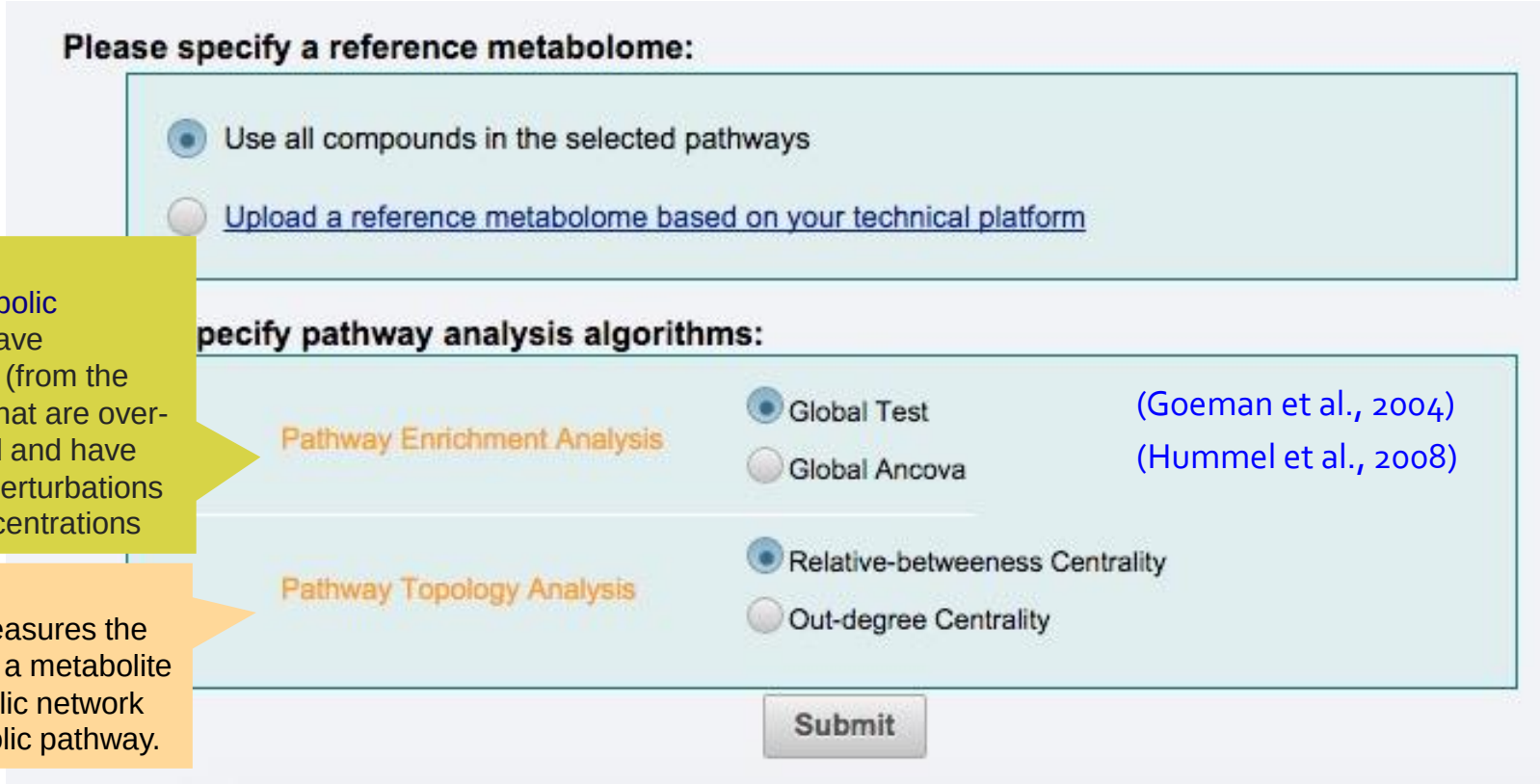
- ( ) None
- (●) Auto scaling (mean-centered and divided by the standard deviation of each variable)
- ( ) Pareto scaling (mean-centered and divided by the square root of standard deviation of each variable)
- ( ) Range scaling (mean-centered and divided by the range of each variable)

Submit

Select Pathway Libraries



Perform Network Topology Analysis

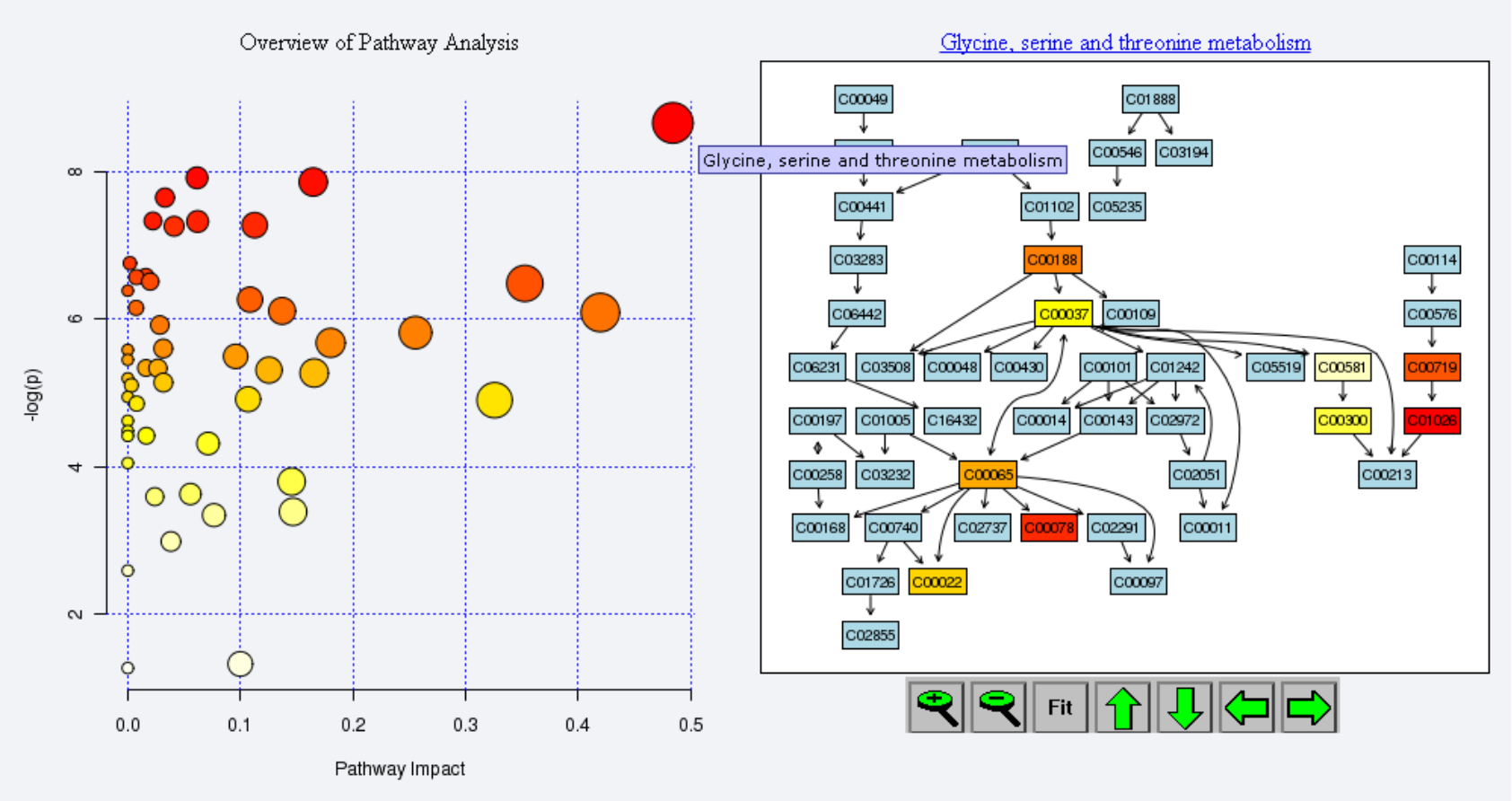


Identifies which metabolic pathways have compounds (from the input lists) that are over-represented and have significant perturbations to their concentrations

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

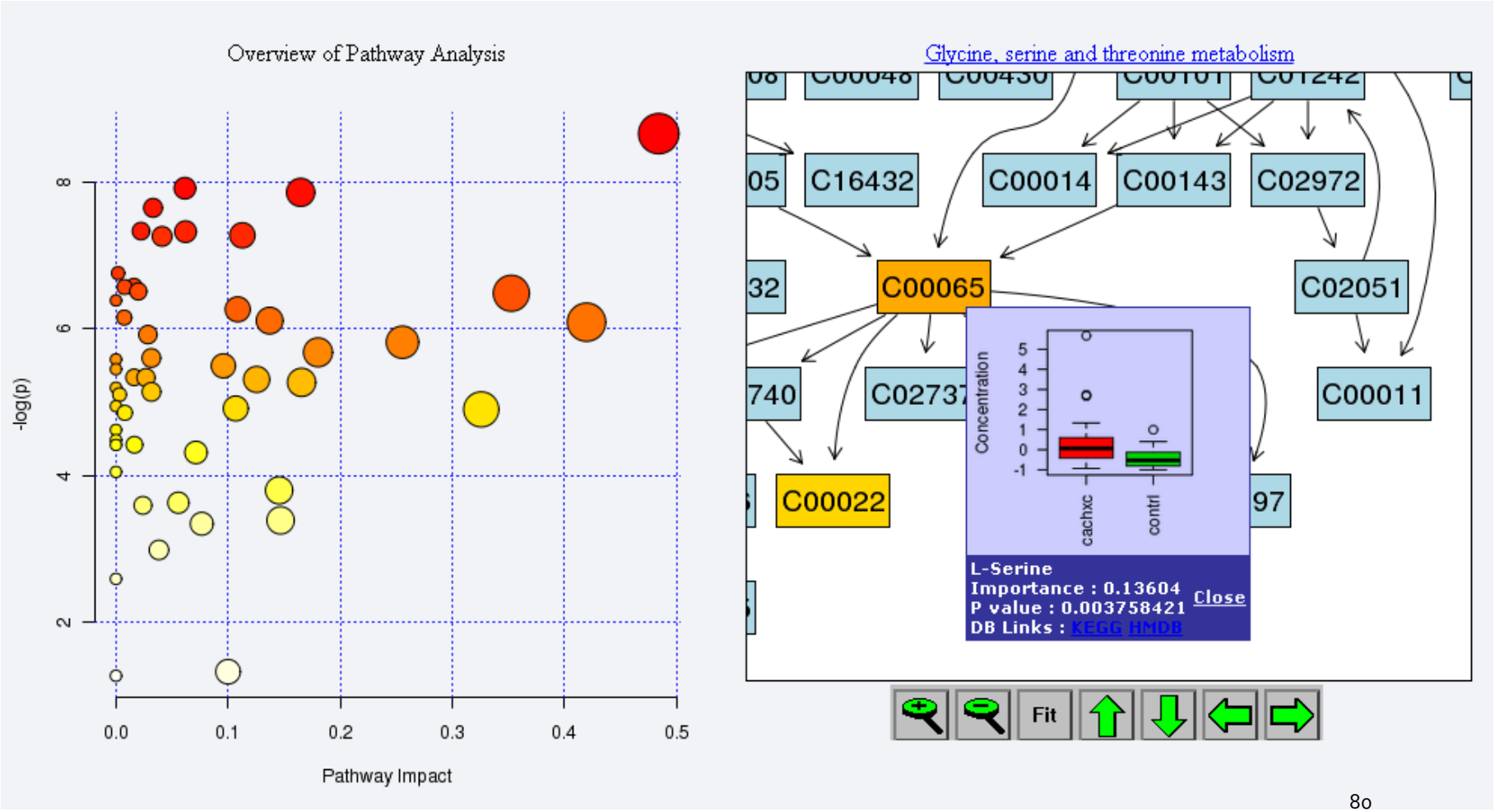
MetPA's pathway topological analysis is based on the centrality measures of a metabolite in a given metabolic network. Centrality is a local quantitative measure of the position of a node relative to the other nodes, and is often used to estimate a node's relative importance or role in network organization.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



Result

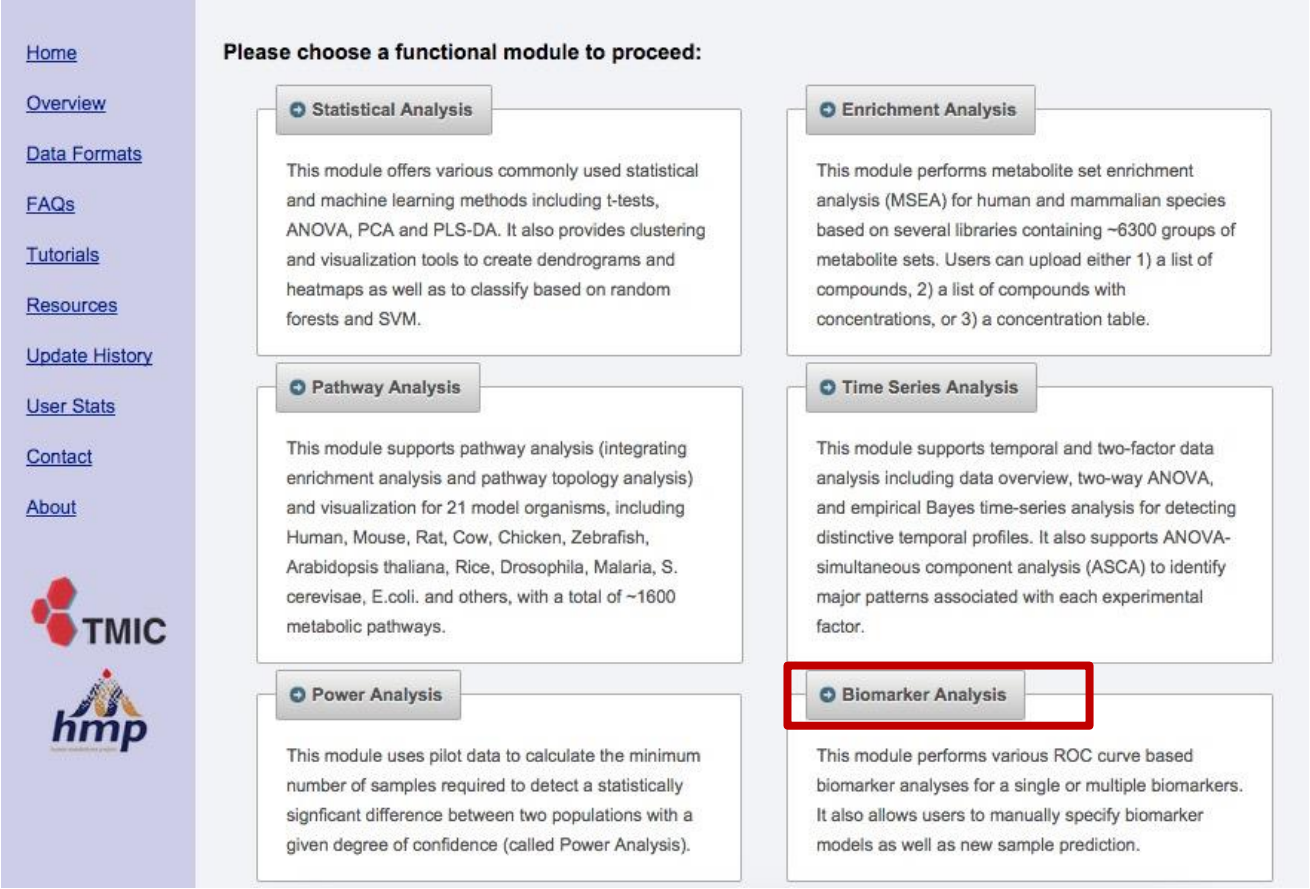
1 2 3

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

1 2 3

Submit

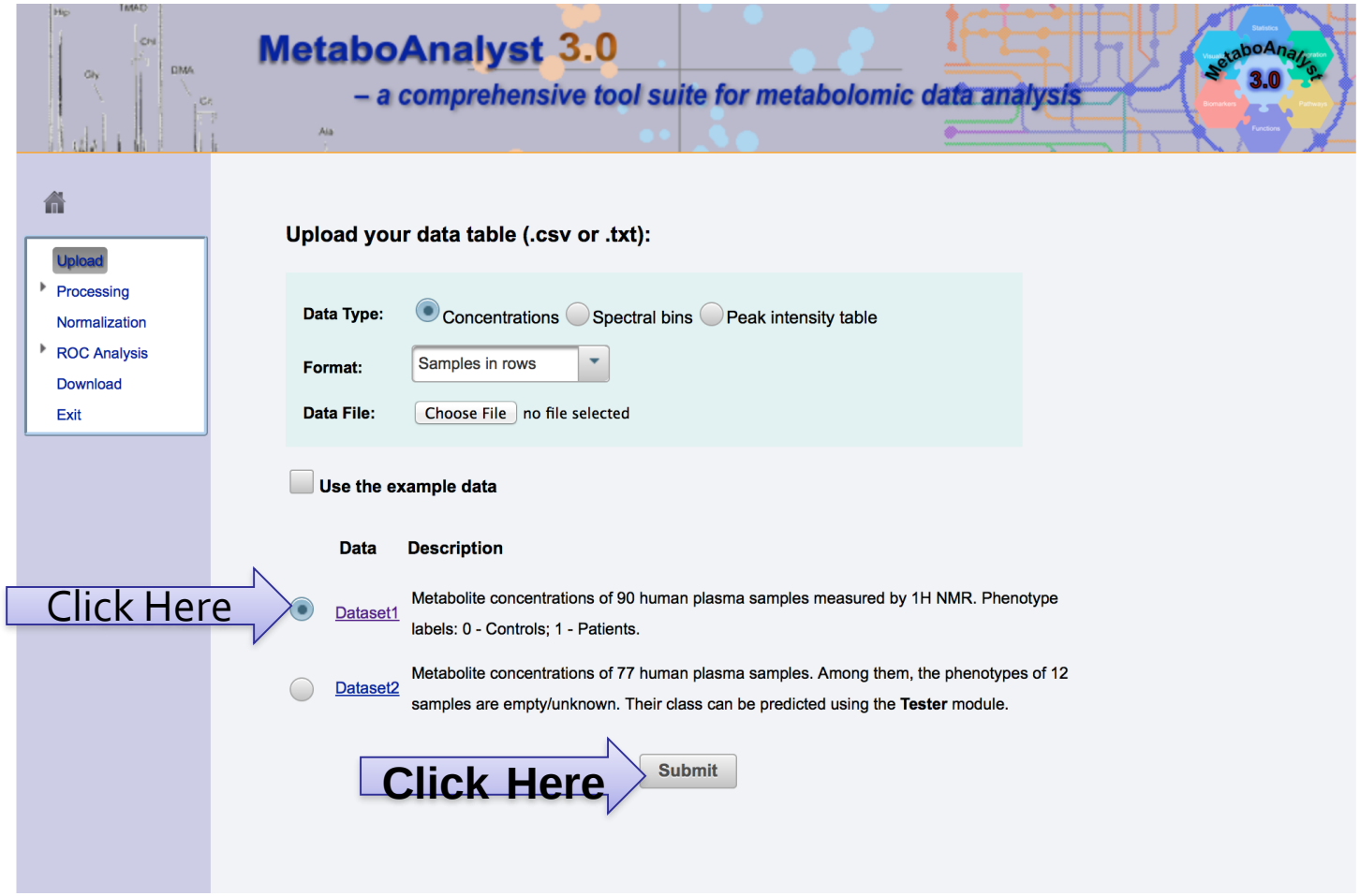
Select a Module (Biomarker Analysis)



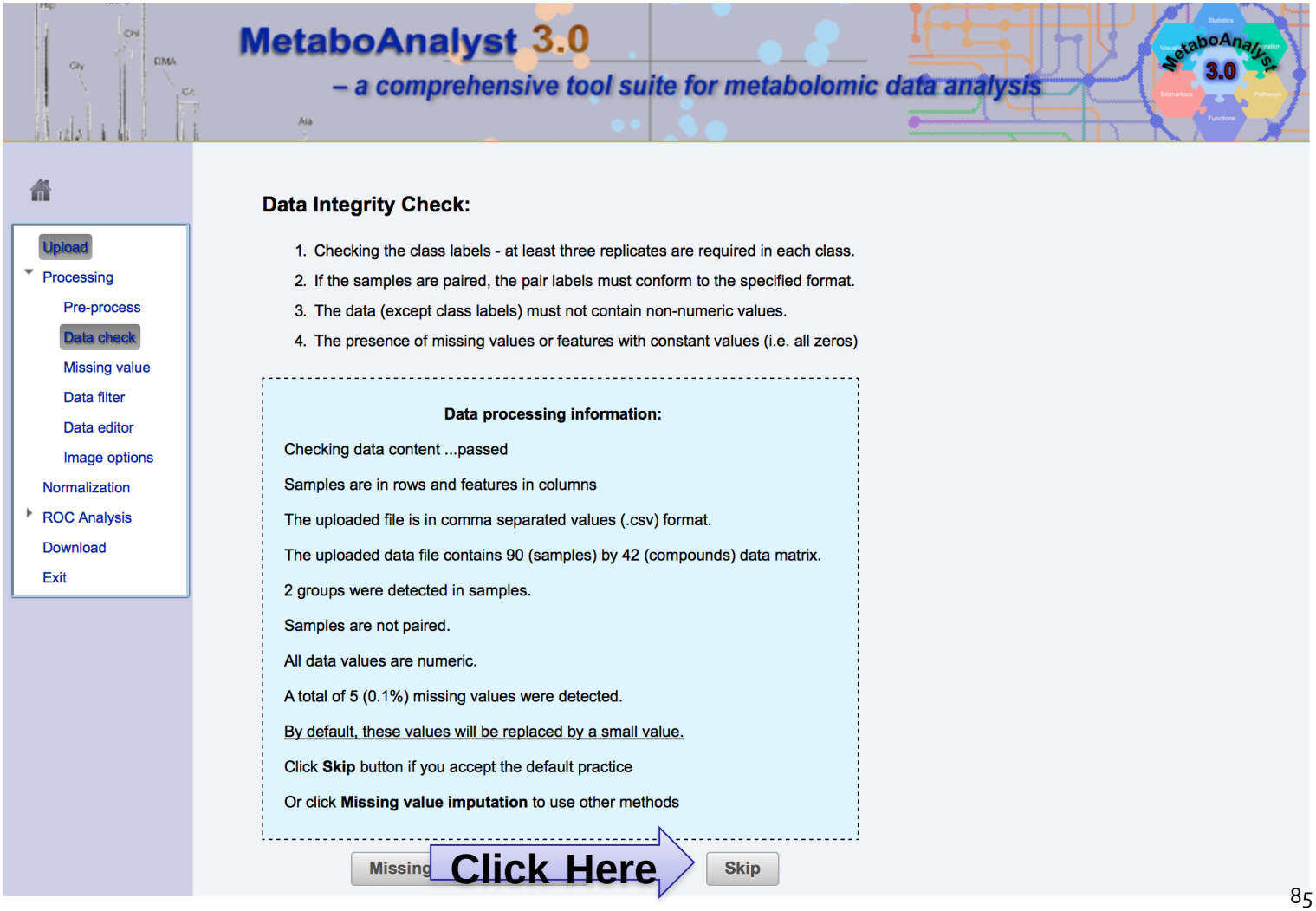
Biomarker Analysis

- 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

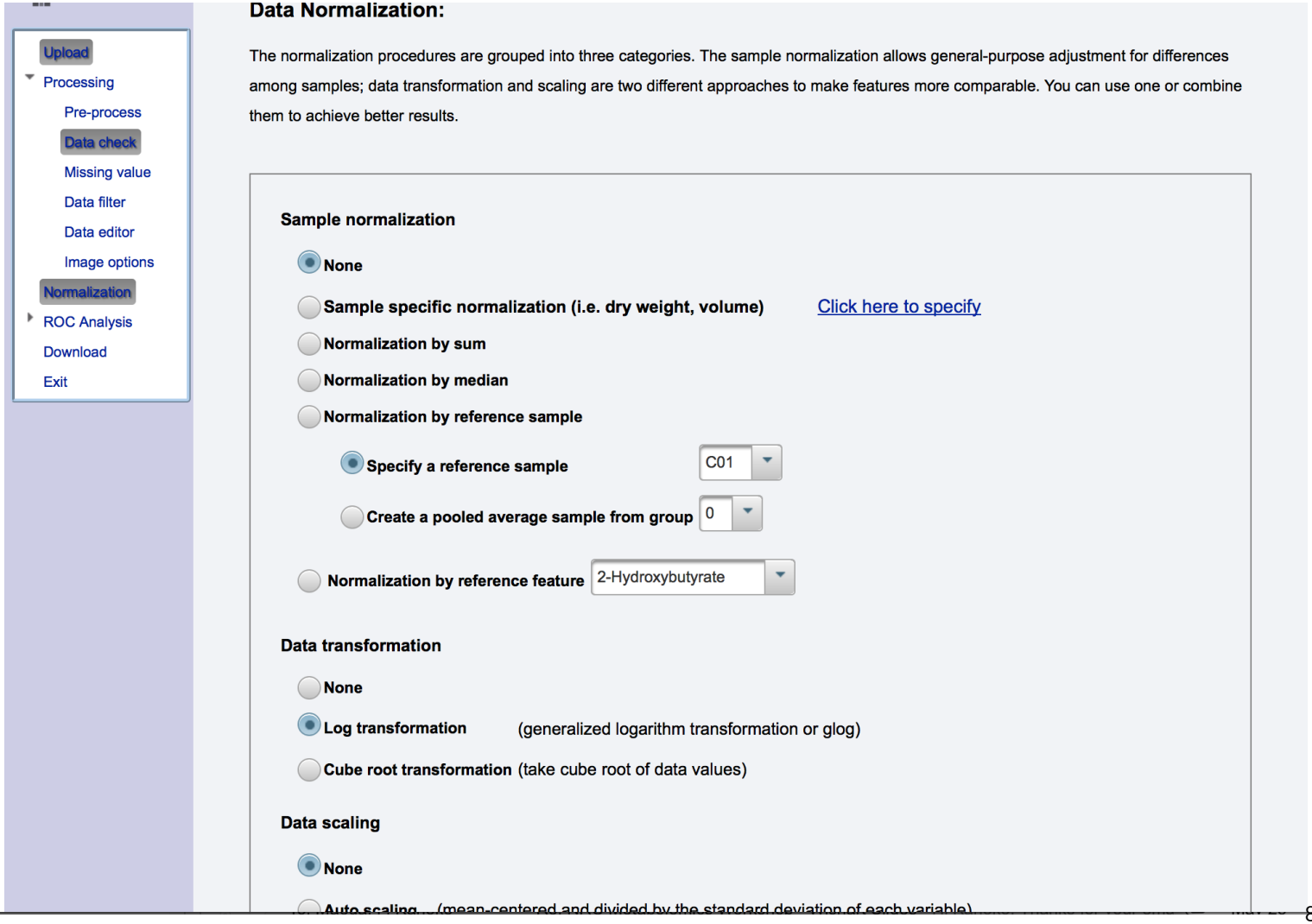
Select Test Data Set 1



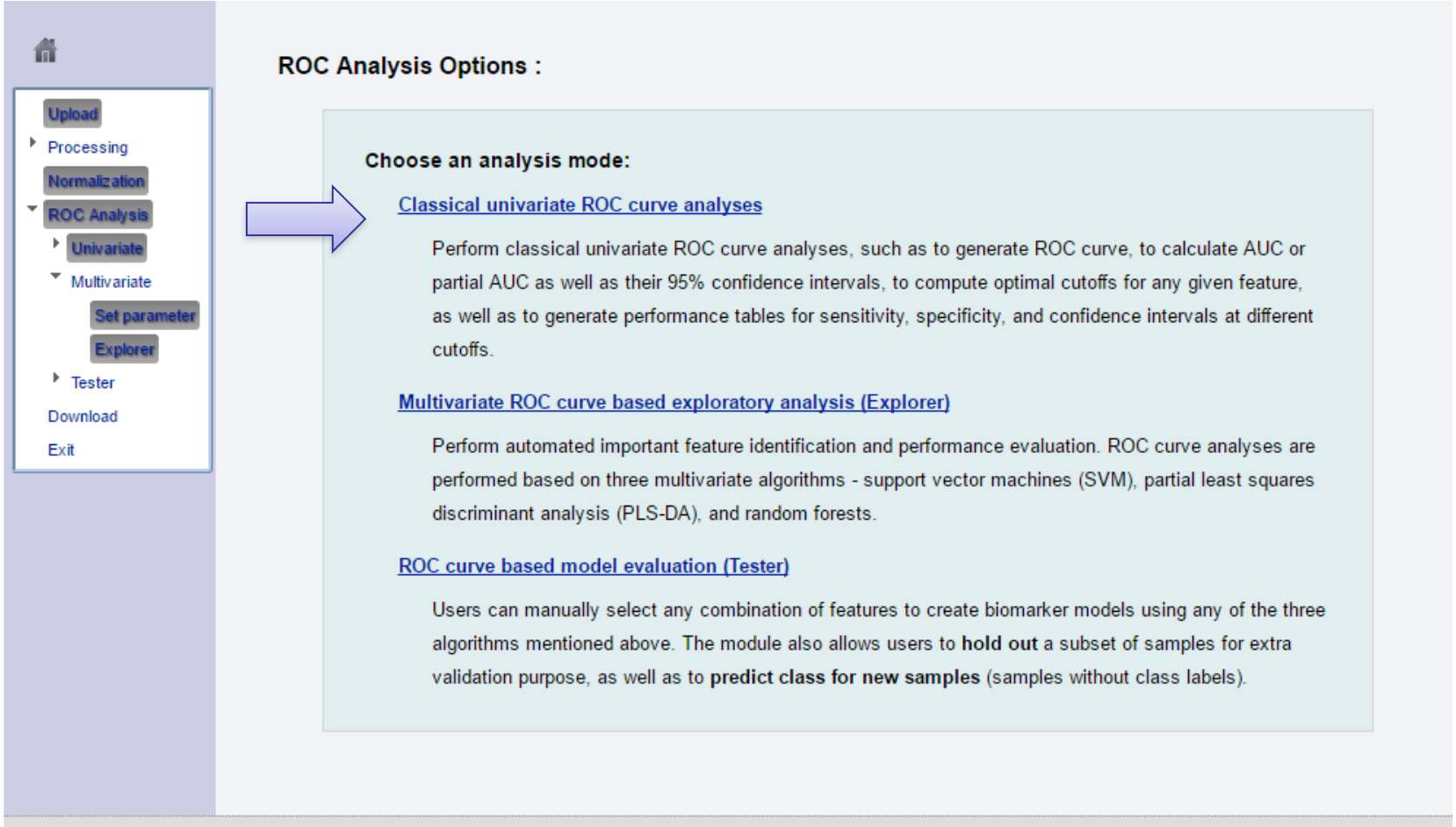
Perform Data Integrity Check



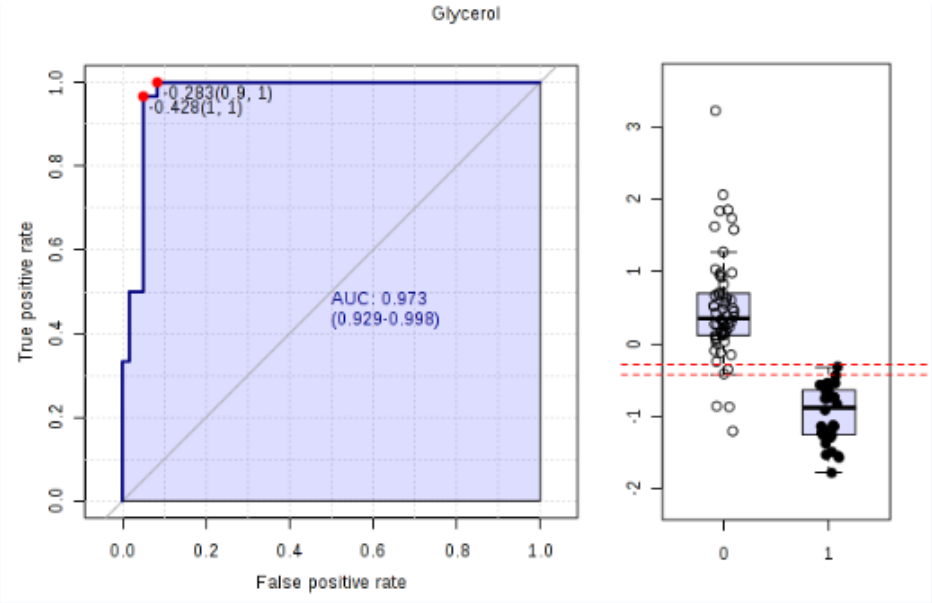
Perform Normalization



Select Multivariate Option



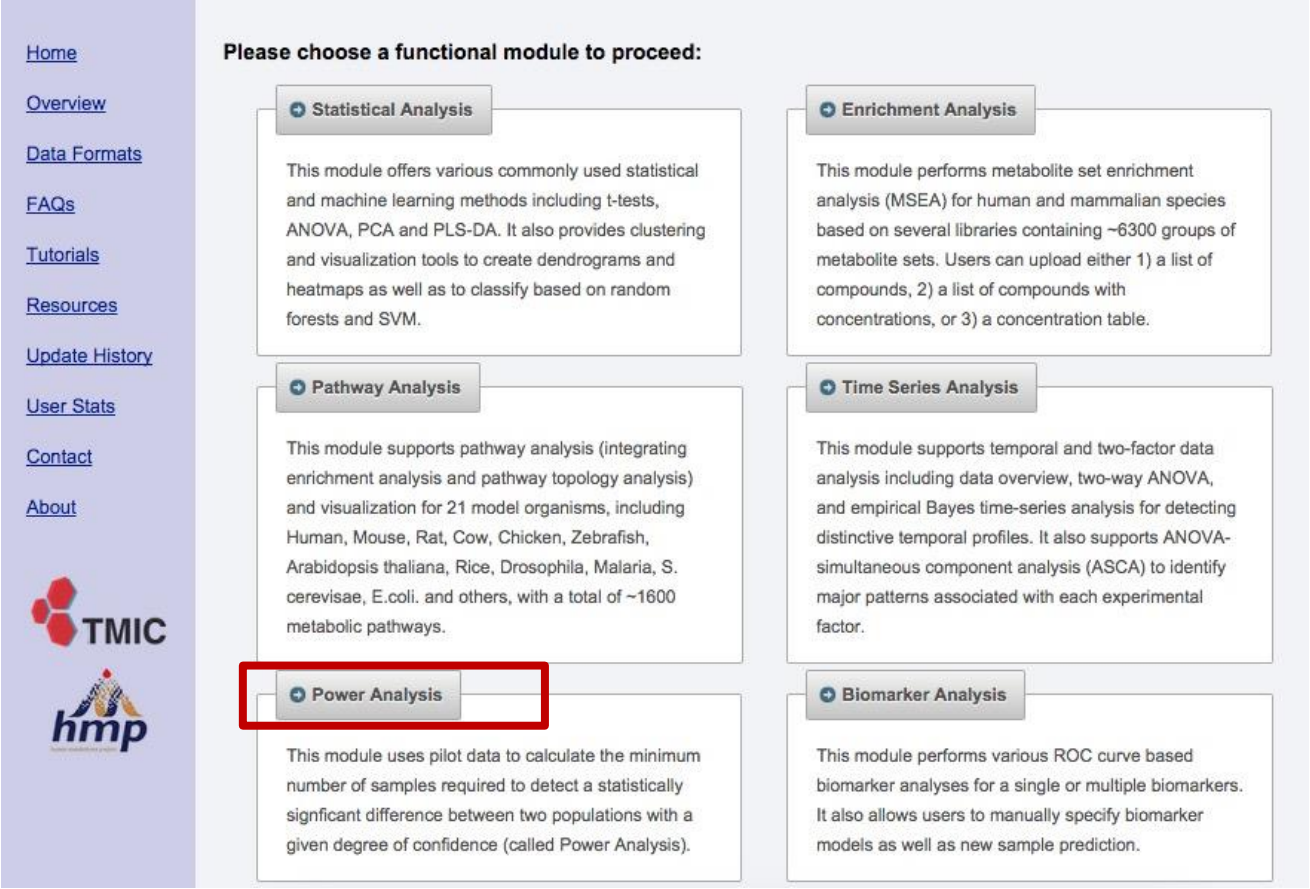
ROC curve analysis



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

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

Select a Module (Power Analysis)



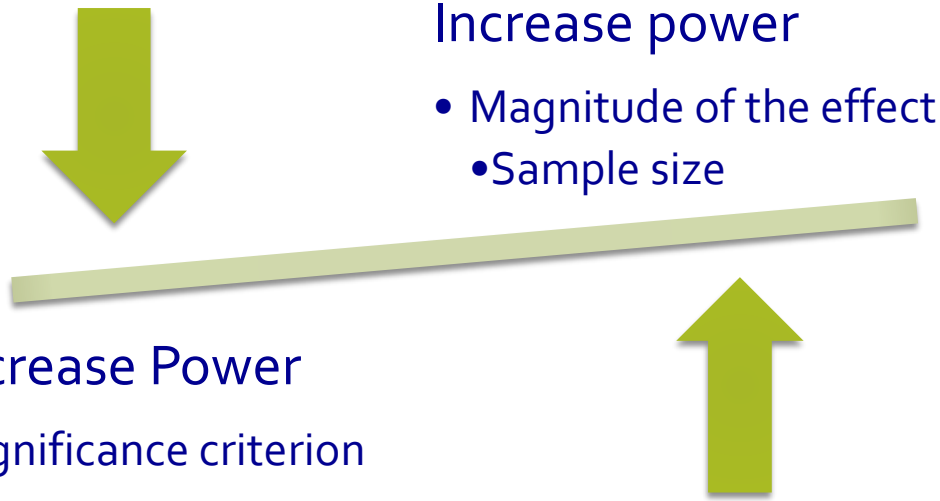
Statistical Power

- 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?

Courtesy: Dr. David Wishart

Statistical Power (cont.)

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



Increase power

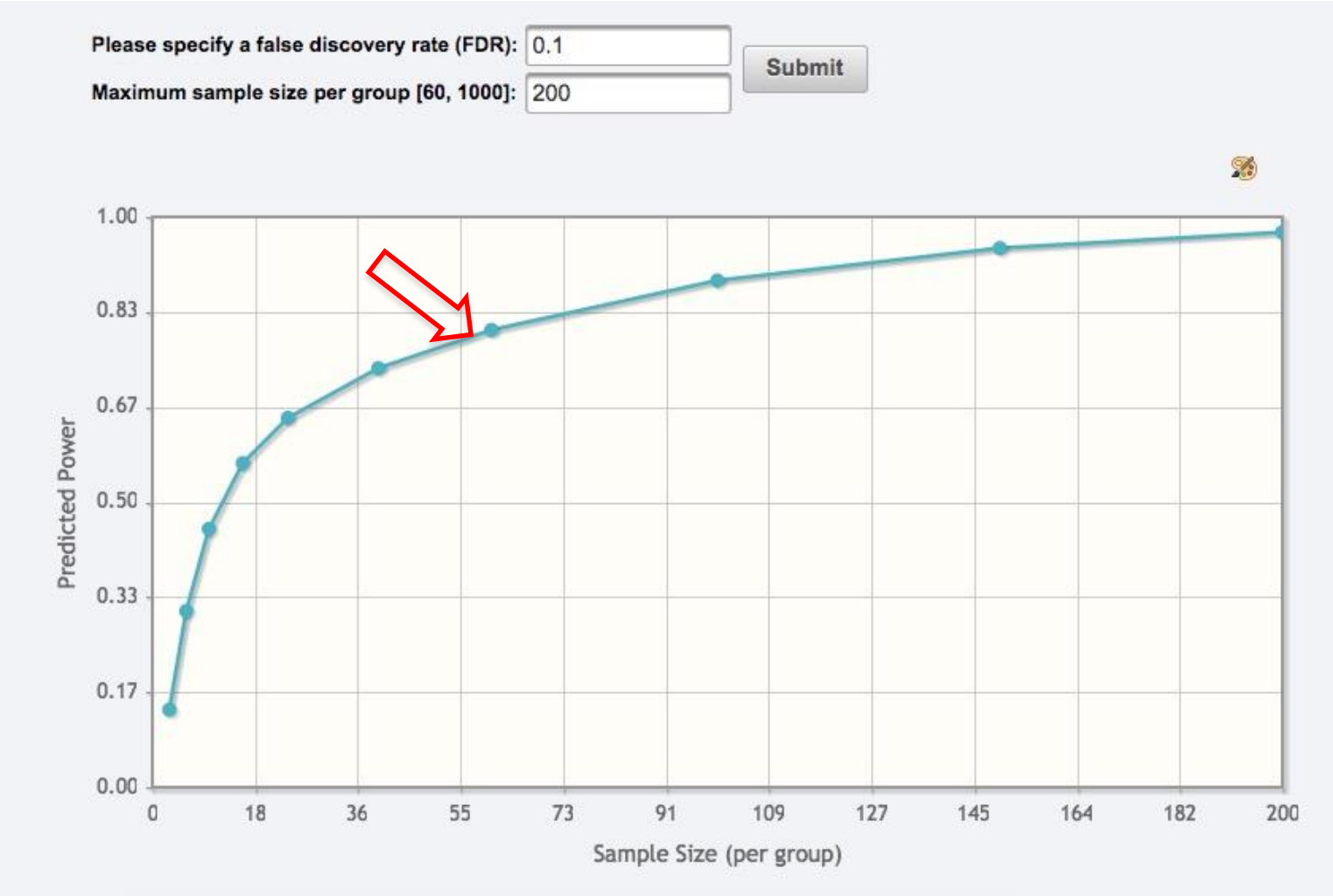
- Magnitude of the effect
  - Sample size

Decrease Power

- Significance criterion

Courtesy: Dr. David Wishart

Power vs. Sample size curve

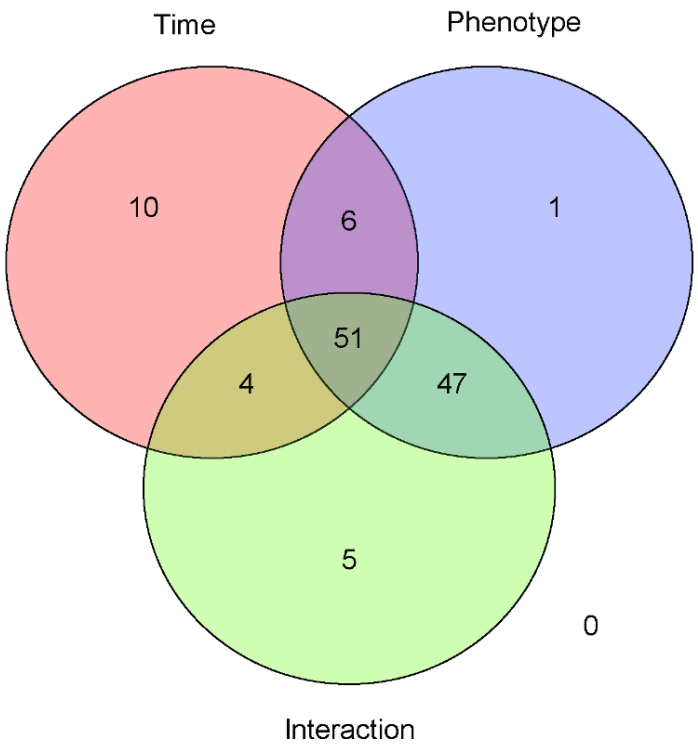


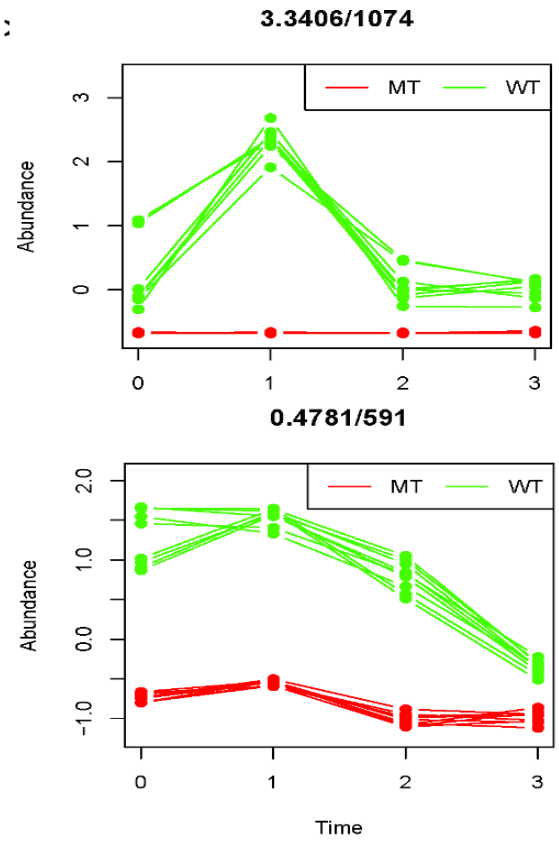
At least 60 samples/group will needed to get a power of 0.8

Not Everything Was Covered

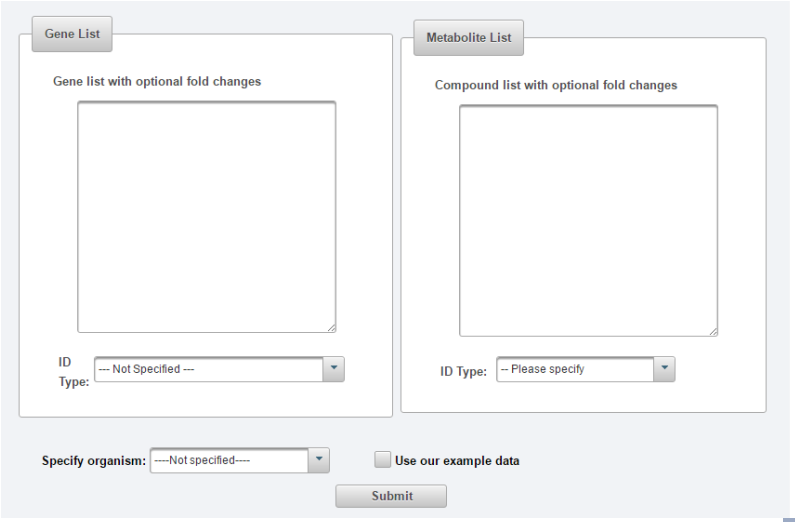
- Clustering Methods (K-means, SOM)
- Classification Methods (SVM, Random Forests)
- SAM and EDAM – (used for identification of differentially expressed genes in microarray experiments)
- Time-series data analysis & Two factor data analysis
- Integrative pathway analysis (gene and metabolite)
- Batch 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

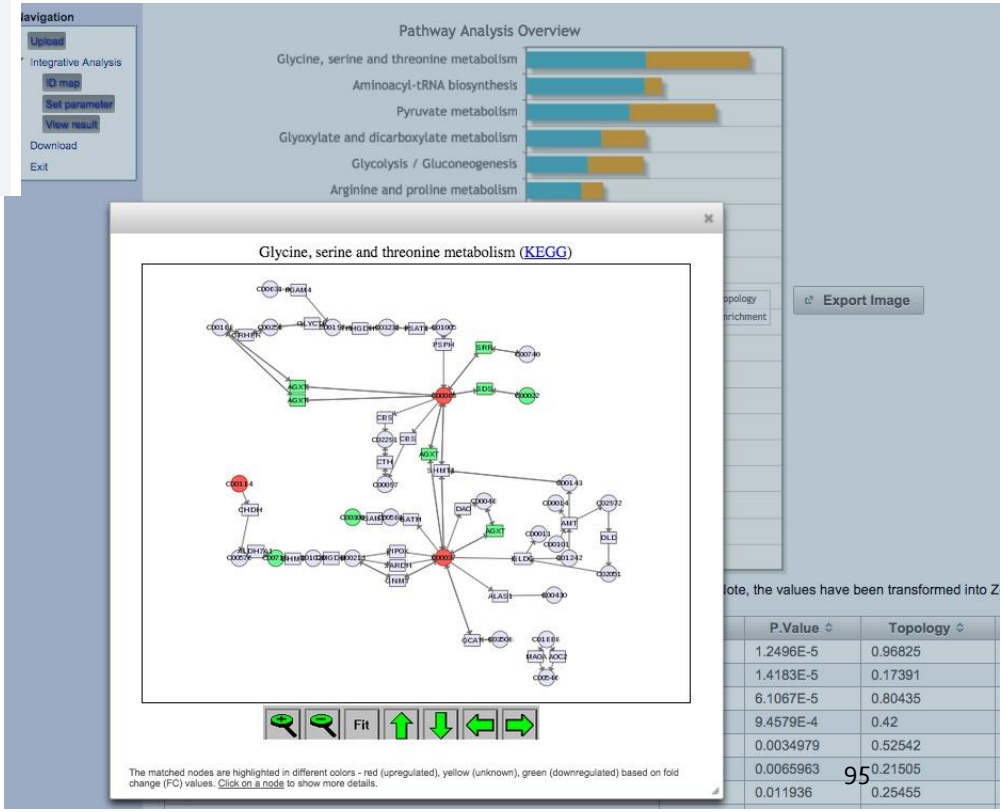
Time Series Analysis in MetaboAnalyst



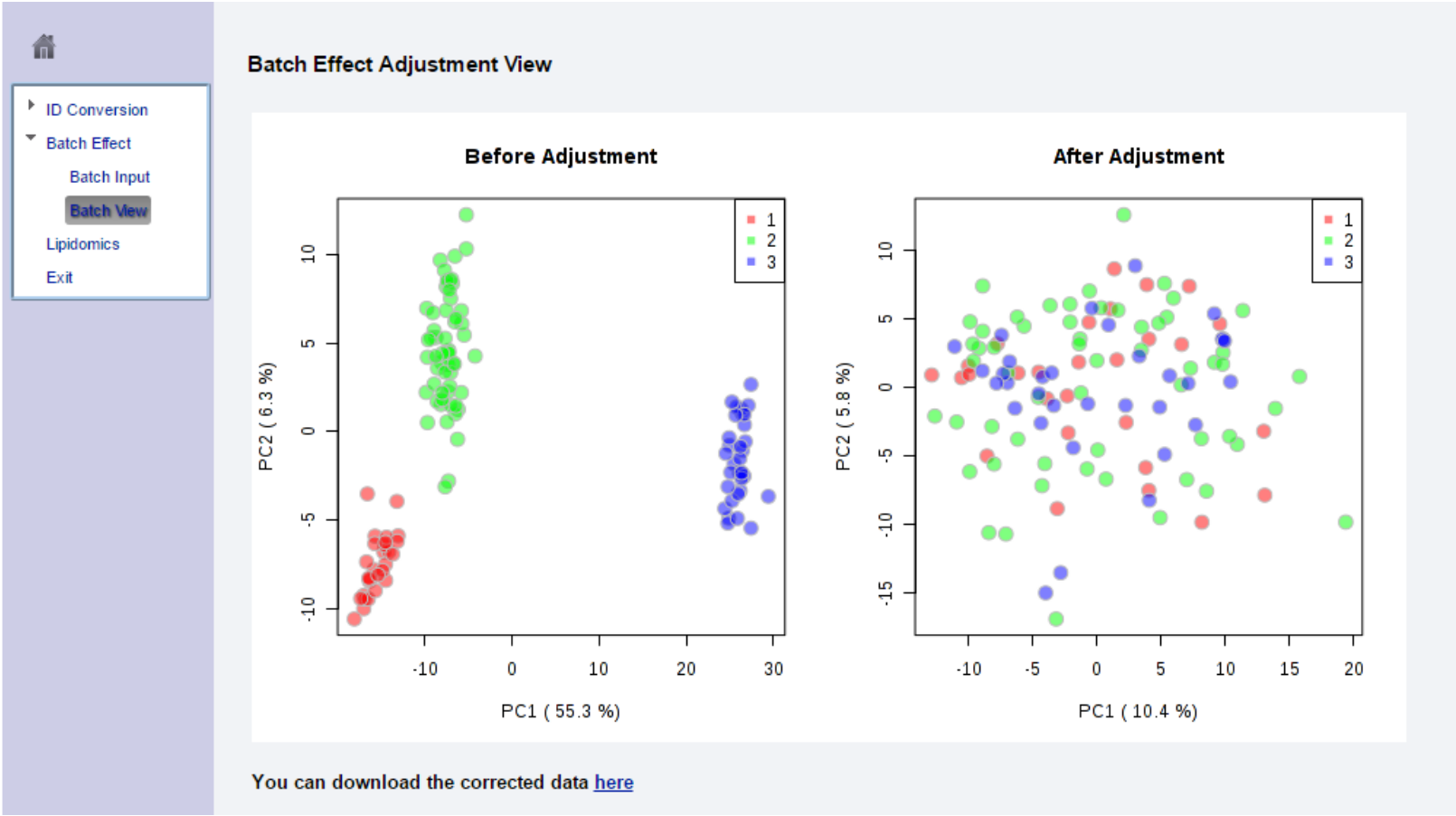


Integrative Pathway Analysis





Batch Adjustment



Lipidomics

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

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

Choose File No file chosen

Submit

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


Thank
you!!