

# Metabolomic Data Analysis with MetaboAnalyst 3.0

User ID: guest7566235459750502455

April 11, 2016

## 1 Data Processing and Normalization

### 1.1 Reading and Processing the Raw Data

MetaboAnalyst accepts a variety of data types generated in metabolomic studies, including compound concentration data, binned NMR/MS spectra data, NMR/MS peak list data, as well as MS spectra (NetCDF, mzXML, mzDATA). Users need to specify the data types when uploading their data in order for MetaboAnalyst to select the correct algorithm to process them. Table 1 summarizes the result of the data processing steps.

#### 1.1.1 Reading Peak Intensity Table

The peak intensity table should be uploaded in comma separated values (.csv) format. Samples can be in rows or columns, with class labels immediately following the sample IDs.

Samples are in columns and features in rows. The uploaded file is in comma separated values (.csv) format. The uploaded data file contains 55 (samples) by 546 (peaks(mz/rt)) data matrix.

#### 1.1.2 Data Integrity Check

Before data analysis, a data integrity check is performed to make sure that all the necessary information has been collected. The class labels must be present and contain only two classes. If samples are paired, the class label must be from  $-n/2$  to  $-1$  for one group, and  $1$  to  $n/2$  for the other group ( $n$  is the sample number and must be an even number). Class labels with same absolute value are assumed to be pairs. Compound concentration or peak intensity values should all be non-negative numbers. By default, all missing values, zeros and negative values will be replaced by the half of the minimum positive value found within the data (see next section)

#### 1.1.3 Missing value imputations

Too many zeroes or missing values will cause difficulties for downstream analysis. MetaboAnalyst offers several different methods for this purpose. The default method replaces all the missing and zero values with a small values (the half of the minimum positive values in the original data) assuming to be the detection limit. The assumption of this approach is that most missing values are caused by low abundance metabolites (i.e. below the detection limit). In addition, since zero values may cause problem for data normalization (i.e. log), they are also replaced with this small value. User can also specify other methods, such as replace by mean/median, or use K-Nearest Neighbours, Probabilistic PCA (PPCA), Bayesian PCA (BPCA) method, Singular Value Decomposition (SVD) method to impute the missing values <sup>1</sup>. Please choose the one that is the most appropriate for your data.

---

<sup>1</sup>Stacklies W, Redestig H, Scholz M, Walther D, Selbig J. *pcaMethods: a bioconductor package, providing PCA methods for incomplete data.*, Bioinformatics 2007 23(9):1164-1167

0 variables were removed for threshold 50 percent. Missing variables were imputed using KNN

### 1.1.4 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 can usually improves the results. Data filter is strongly recommended for datasets with large number of variables ( $> 250$ ) datasets contain much noise (i.e.chemometrics data). Filtering can usually improve your results<sup>2</sup>.

*For data with number of variables  $< 250$ , this step will reduce 5% of variables; For variable number between 250 and 500, 10% of variables will be removed; For variable number btween 500 and 1000, 25% of variables will be removed; And 40% of variabled will be removed for data with over 1000 varaibles.*

No data filtering was performed.

Table 1: Summary of data processing results

	Features (positive)	Missing/Zero	Features (processed)
CtrlBKKS.2202	546	0	546
CtrlBKKS.2203	546	0	546
CtrlBKKS.2206	546	0	546
CtrlBKKS.2211	546	0	546
HF_BKKS.2214	546	0	546
HF_BKKS.2216	546	0	546
HF_BKKS.2217	546	0	546
HF_BKKS.2218	546	0	546
CtrlBKKS.db+.2225	546	0	546
CtrlBKKS.db+.2226	546	0	546
CtrlBKKS.db+.2227	546	0	546
CtrlBKKS.db+.2231	546	0	546
HF_BKKS.db+.2239	546	0	546
HF_BKKS.db+.2241	546	0	546
HF_BKKS.db+.2242	546	0	546
HF_BKKS.db+.2243	546	0	546
CtrlBL6.2249	546	0	546
CtrlBL6.2250	546	0	546
CtrlBL6.2251	546	0	546
CtrlBL6.2252	546	0	546
HF_BL6.2260	546	0	546
HF_BL6.2261	546	0	546
HF_BL6.2262	546	0	546
HF_BL6.2266	546	0	546
CtrlBL6.db+.2273	546	0	546
CtrlBL6.db+.2274	546	0	546
CtrlBL6.db+.2275	546	0	546
CtrlBL6.db+.2279	546	0	546
HF_BL6.db+.2286	546	0	546
HF_BL6.db+.2287	546	0	546
HF_BL6.db+.2288	546	0	546
HF_BL6.db+.2289	546	0	546
CtrlBTBR.2292	546	0	546
CtrlBTBR.2293	546	0	546
CtrlBTBR.2297	546	0	546
CtrlBTBR.2298	546	0	546
HF_BTBR.2306	546	0	546
HF_BTBR.2307	546	0	546
HF_BTBR.2308	546	0	546
HF_BTBR.2309	546	0	546
CtrlBTBR.ob+.2316	546	0	546
CtrlBTBR.ob+.2317	546	0	546
CtrlBTBR.ob+.2318	546	0	546
CtrlBTBR.ob+.2320	546	0	546
HF_BTBR.ob+.2329	546	0	546
HF_BTBR.ob+.2331	546	0	546
HF_BTBR.ob+.2333	546	0	546
HF_BTBR.ob+.2334	546	0	546
Test PoolNegative_01	546	0	546
Test PoolNegative_02	546	0	546
Test PoolNegative_03	546	0	546
Test PoolNegative_04	546	0	546
Test PoolNegative_05	546	0	546
Test PoolNegative_06	546	0	546
Test PoolNegative_07	546	0	546

<sup>2</sup>Hackstadt AJ, Hess AM. *Filtering for increased power for microarray data analysis*, BMC Bioinformatics. 2009; 10: 11.

Error occurred during normalization of your data .... Fail to proceed. Please check if the data format you uploaded is correct. For further information, please contact [jeff.xia@mcgill.ca](mailto:jeff.xia@mcgill.ca) with a description of the problem. Your feedback is highly appreciated!

## 2 Statistical and Machine Learning Data Analysis

MetaboAnalyst offers a variety of methods commonly used in metabolomic data analyses. They include:

1. Univariate analysis methods:
  - Fold Change Analysis
  - T-tests
  - Volcano Plot
  - One-way ANOVA and post-hoc analysis
  - Correlation analysis
2. Multivariate analysis methods:
  - Principal Component Analysis (PCA)
  - Partial Least Squares - Discriminant Analysis (PLS-DA)
3. Robust Feature Selection Methods in microarray studies
  - Significance Analysis of Microarray (SAM)
  - Empirical Bayesian Analysis of Microarray (EBAM)
4. Clustering Analysis
  - Hierarchical Clustering
    - Dendrogram
    - Heatmap
  - Partitional Clustering
    - K-means Clustering
    - Self-Organizing Map (SOM)
5. Supervised Classification and Feature Selection methods
  - Random Forest
  - Support Vector Machine (SVM)

Please note: some advanced methods are available only for two-group sample analysis.

### 3 Data Annotation

Please be advised that MetaboAnalyst also supports metabolomic data annotation. For NMR, MS, or GC-MS peak list data, users can perform peak identification by searching the corresponding libraries. For compound concentration data, users can perform metabolite set enrichment analysis and metabolic pathway analysis.

---

The report was generated on Mon Apr 11 22:30:21 2016 with R version 3.2.2 (2015-08-14). Thank you for using MetaboAnalyst! For suggestions and feedback please contact Jeff Xia (*jeff.xia@mcgill.ca*).