**Supplementary Material**

1. **Case Study**

Two datasets were used to demonstrate the utility of the MetaboAnalystR package. First, to address reproducibility and flexibility, we performed statistical analysis on 1H NMR data from cancer patients that is available from the MetaboAnalyst web-server (goo.gl/KWcBFL) (Eisner *et al.*, 2011). Second, to address scalability, we used NetCDF spectra of 12 mice spinal cord samples (six wild type and six enzyme-inactivated) collected by untargeted liquid chromatography-mass spectrometry (LC-MS) (Saghatelian *et al.*, 2004). This data is also available for download from the MetaboAnalyst web-server (goo.gl/qLMXkc).

* 1. **Installation of MetaboAnalystR from Github**

Detailed instructions for installation of MetaboAnalystR on a local computer are available from <https://github.com/xia-lab/MetaboAnalystR>. To begin installation of the R package, it is first necessary to install all package dependencies. To streamline this process, an R function is available on the main page of the MetaboAnalystR github page that will automatically download any/all missing R package dependencies. Instructions to install the package on Windows/Linux/Mac directly from GitHub follow dependency installation.

* 1. **Reproducibility**

Through this example, we demonstrate how the R Command History from the MetaboAnalyst web-server can be inputted into R, and how identical results can be obtained using the MetaboAnalystR package. To begin data processing, we performed missing value imputation, using the R functions *RemoveMissingPercent* and *ImputVar*, to remove features with >50% of missing values and replace the remaining missing values with half of the minimum positive value in the original dataset. Following this, we performed log transformation of the data using the function *Normalization* (Supplementary Figure 1). To begin exploratory statistical analysis, we performed fold-change analysis (*FC.Anal.unpaired*) (Supplementary Figure 2), principal component analysis (*PCA.Anal*) (Supplementary Figure 3), significance analysis of microarray (and metabolites – *SAM.Anal*) (Supplementary Table 1), heatmap analysis (*PlotHeatMap*) (Supplementary Figure 4), and Random Forest classification (*RF.Anal*) (Supplementary Figure 5).

To demonstrate reproducibility, the R command history detailing each step was downloaded from the web-server and run in RStudio 1.1.383, with R 3.4.4 and the MetaboAnalystR package. A minor adjustment to the downloaded R code was made, where the path to the data in the function *Read.TextData* was modified to use the same 1H NMR dataset, but that is available directly from the web-server. Then, the same analysis, using the same parameters, was run. The Rscript for this analysis is available (Supplementary File 1: Rhistory\_Reproducibility.R) Results (Tables and figures) generated on the web-server and from the R package were examined and were identical, demonstrating the interchangeable functionality across all MetaboAnalyst implementations.

* 1. **Flexibility**

To demonstrate the flexibility of the R package, we utilized the same R command history downloaded from the web server as above. Within the web server, the steps for data processing follow a constricted set of steps, from uploading data, missing value imputation, data filtration, and data normalization. Many users have requested that missing value imputation be performed following data normalization, which cannot be accommodated on the web-server. However, using MetaboAnalystR, we first performed data normalization and then performed missing value imputation (see code R code below). In this case, we moved the following R command for data normalization:

R> mSet<-Normalization(mSet, "QuantileNorm", "NULL", "NULL", "PIF\_178", ratio=FALSE, ratioNum=20)

Before this command for missing value imputation:

R> mSet<-RemoveMissingPercent(mSet, percent=0.5)

The updated data analysis pipeline was then re-run in RStudio, showing the flexibility of the R package. The detailed steps for this example are available in the Supplementary File 2: Rhistory\_Flexibility.

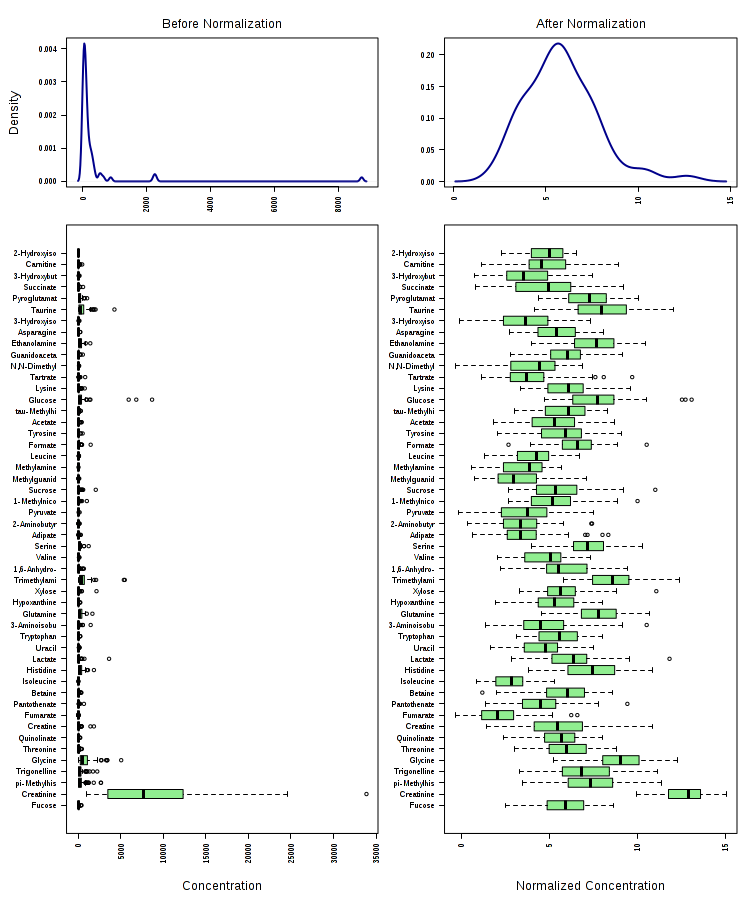
* 1. **Big Metabolomics Data and Workflow Customization**

The MetaboAnalyst web-server is unaccommodating for the analysis of big metabolomics data, with a file size limit of 50M on the public server. This can be frustrating for users as with rapid advances in metabolomics technologies, comes increasingly larger datasets. In this case, the MetaboAnalyst web-application can be utilized to perform a pilot study on a subset of data, and then the R command history can be altered, changing only the data uploading steps, to run big metabolomics data analysis (or batch processing). In this example, we uploaded a zipped file containing a subset of the mice samples, three wild-type and three enzyme-inactivated (Supplementary Data 1: pilot\_ms\_netcdf.zip). As this is raw spectra data, MetaboAnalyst requires several data pre-processing steps before it is able to create a usable data matrix. In particular, it utilizes the XCMS R package to filter and identify peaks, match peaks, and fill in missing peak data (Smith *et al.*, 2006). To accomplish this, the R functions *Read.MSspec*, *MSspec.rtCorrection*, and *MSspec.fillPeaks* are used in successive order. Then, the function*SetupMSdataMatrix* is used to create a data matrix suitable for further downstream analysis. Following data pre-processing, we performed missing value imputation and data normalization, and then fold-change (FC) analysis and Empirical Bayesian Analysis of Microarray (and Metabolites) (EBAM).

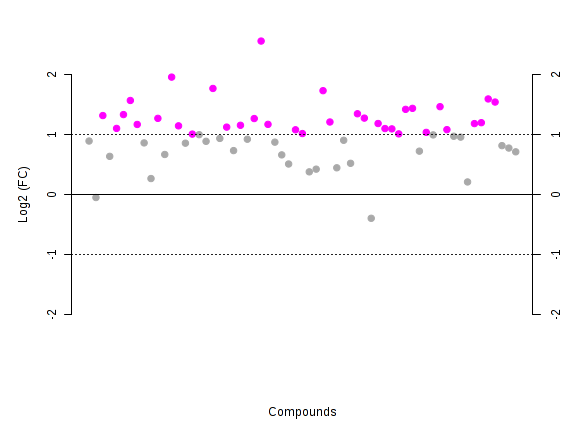
The R history (Supplementary File 3: Rhistory\_Original.R) was then downloaded from the web-server to perform batch processing on the remaining data. With the R command history file, we must load MetaboAnalystR, and also adjust the file path in the function *UnzipUploadedFile*, to the zip file containing all 12 mice samples (goo.gl/qLMXkc). We also adjust the file path in the function *Read.MSspec* to the “upload” folder, which will contain the unzipped files in the working directory. Then, we used the code below to execute batch processing of the LC-MS data, which performed data uploading, data processing, FC, and EBAM analyses using the exact same parameters.

> R CMD BATCH --restore --save Rhistory\_Batch.R batchoutput.R

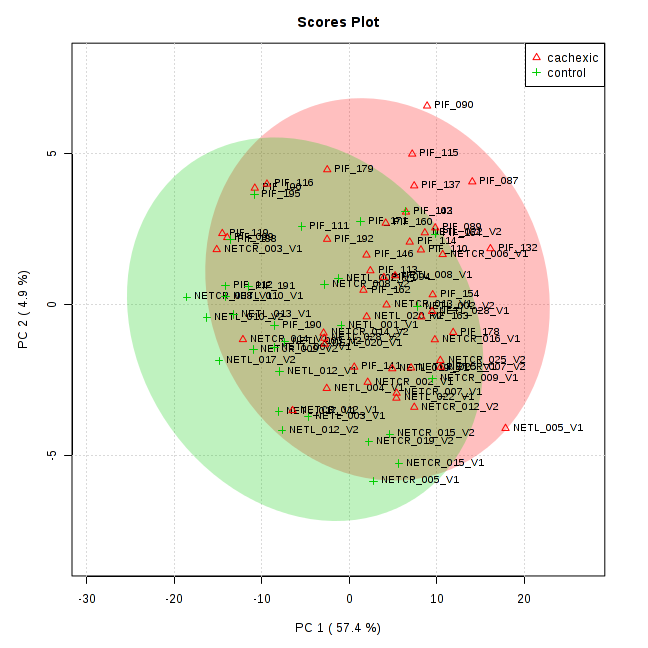
In the above command, the Rhistory\_Batch.R file contains the code to be executed, and batchoutput.R is the name of the output file (Supplementary File 4: batchoutput.R). We save the results at the end of the session in a .Rdata file using --save. The RScript for this example, with detailed steps, is available in the supplementary material (Supplementary File 5: Rhistory\_Batch).

****

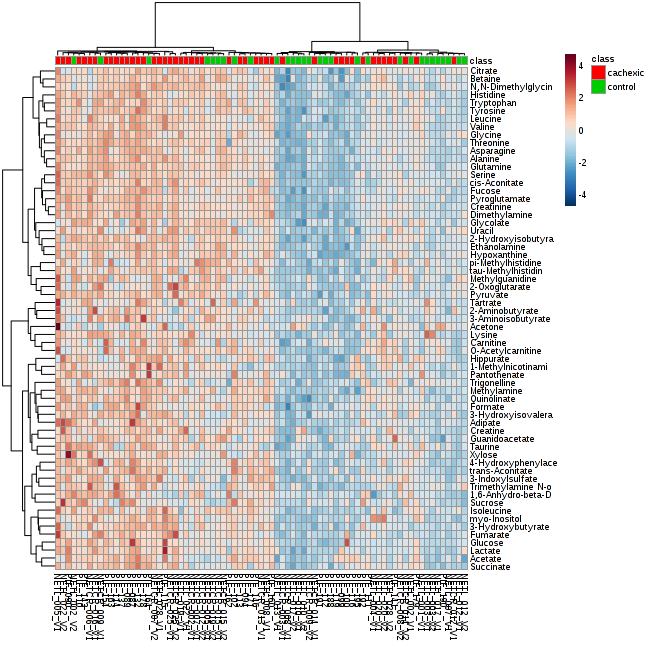
**Supplementary Figure 1.** Box-plot and kernel density plots of the first 50 features in the example dataset before and after log-transformation of data (using the Normalization function).



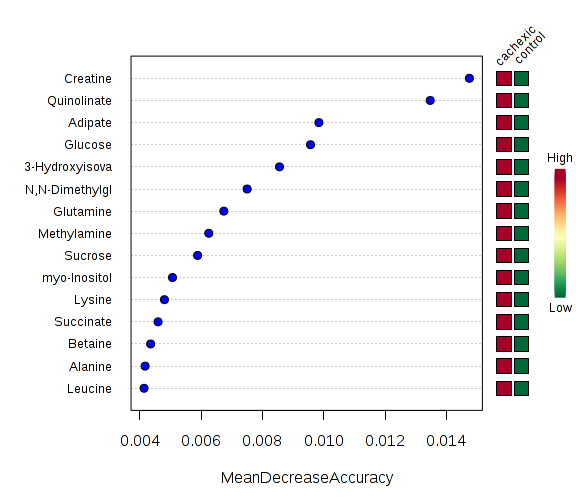
**Supplementary Figure 2.** Plot of important features selected using fold-change analysis with a threshold of 2. The pink circles represent features above the FC threshold of 2. The values are on the log-scale to both up-and down-regulated features can be plotted symmetrically.



**Supplementary Figure 3.** Score plot of the principle components (PC) 1 and 2. PC1 explains about 57.4% of the observed variance, whereas PC2 explains only 4.9% of observed variance.



**Supplementary Figure 4.** Heatmap of the clustering analysis using Euclidean distance and the Ward clustering algorithm.



**Supplementary Figure 5.** Plot of significant features as identified using Random Forest classification. The features are ranked by their mean decrease in classification accuracy following 500 permutations.

**Supplementary Table 1.** The top five important features identified by SAM.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Compounds | d-value | Standard Deviation | Raw p-value | q-value |
| Quinolinate | -5.0149 | 0.24692 | 0 | 0 |
| Glucose | -4.6062 | 0.35119 | 0 | 0 |
| 3-Hydroxyisovalerate | -4.5699 | 0.33082 | 0 | 0 |
| Leucine | -4.5598 | 0.25194 | 0 | 0 |
| Succinate | -4.4571 | 0.39956 | 0 | 0 |

**References**

Eisner, R. *et al.* Learning to predict cancer-associated skeletal muscle wasting from 1H-NMR profiles of urinary metabolites. *Metabolomics* 2011;7(1):25-34.

Saghatelian, A. *et al.* Assignment of endogenous substrates to enzymes by global metabolite profiling. *Biochemistry* 2004;43(45):14332-14339.

Smith, C.A. *et al.* XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical chemistry* 2006;78(3):779-787.