Using MetaboAnalyst 4.0 for Comprehensive and Integrative Metabolomics Data Analysis

Jasmine Chong,[1] David S. Wishart,[4,5,6] and Jianguo Xia[1,2,3,6]

[1]Institute of Parasitology, McGill University, Sainte-Anne-de-Bellevue, Quebec, Canada
[2]Department of Animal Sciences, McGill University, Sainte-Anne-de-Bellevue, Quebec, Canada
[3]Department of Microbiology and Immunology, McGill University, Montreal, Quebec, Canada
[4]Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada
[5]Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada
[6]Corresponding authors: *david.wishart@ualberta.ca*; *jeff.xia@mcgill.ca*

MetaboAnalyst (*https://www.metaboanalyst.ca*) is an easy-to-use web-based tool suite for comprehensive metabolomic data analysis, interpretation, and integration with other omics data. Since its first release in 2009, MetaboAnalyst has evolved significantly to meet the ever-expanding bioinformatics demands from the rapidly growing metabolomics community. In addition to providing a variety of data processing and normalization procedures, MetaboAnalyst supports a wide array of functions for statistical, functional, as well as data visualization tasks. Some of the most widely used approaches include PCA (principal component analysis), PLS-DA (partial least squares discriminant analysis), clustering analysis and visualization, MSEA (metabolite set enrichment analysis), MetPA (metabolic pathway analysis), biomarker selection via ROC (receiver operating characteristic) curve analysis, as well as time series and power analysis. The current version of MetaboAnalyst (4.0) features a complete overhaul of the user interface and significantly expanded underlying knowledge bases (compound database, pathway libraries, and metabolite sets). Three new modules have been added to support pathway activity prediction directly from mass peaks, biomarker meta-analysis, and network-based multi-omics data integration. To enable more transparent and reproducible analysis of metabolomic data, we have released a companion R package (MetaboAnalystR) to complement the web-based application. This article provides an overview of the main functional modules and the general workflow of MetaboAnalyst 4.0, followed by 12 detailed protocols: © 2019 by John Wiley & Sons, Inc.

Basic Protocol 1: Data uploading, processing, and normalization
Basic Protocol 2: Identification of significant variables
Basic Protocol 3: Multivariate exploratory data analysis
Basic Protocol 4: Functional interpretation of metabolomic data
Basic Protocol 5: Biomarker analysis based on receiver operating characteristic (ROC) curves
Basic Protocol 6: Time-series and two-factor data analysis
Basic Protocol 7: Sample size estimation and power analysis
Basic Protocol 8: Joint pathway analysis
Basic Protocol 9: MS peaks to pathway activities
Basic Protocol 10: Biomarker meta-analysis
Basic Protocol 11: Knowledge-based network exploration of multi-omics data
Basic Protocol 12: MetaboAnalystR introduction

Keywords: biomarker analysis • chemometrics • joint pathway analysis • meta-analysis • metabolic pathway analysis • metabolite set enrichment analysis • metabolomics • MS peaks to pathways • multi-omics integration

CURRENT PROTOCOLS
A Wiley Brand
Current Protocols in Bioinformatics e86, Volume 68
Published in Wiley Online Library (wileyonlinelibrary.com).
doi: 10.1002/cpbi.86

• network analysis • power analysis • reproducible data analysis • ROC curve
• web application

How to cite this article:
Chong, J., Wishart, D. S., & Xia, J. (2019). Using metaboanalyst 4.0 for comprehensive and integrative metabolomics data analysis. *Current Protocols in Bioinformatics*, *68,* e86. doi: 10.1002/cpbi.86

INTRODUCTION

The rapid development of metabolomic technologies coupled with the widespread application of metabolomics in many different fields has significantly increased the demand for user-friendly and easily accessible bioinformatics tools developed specifically for metabolomics. MetaboAnalyst is a prime example of one of the new-generation bioinformatics tools designed to meet this rapidly evolving demand. MetaboAnalyst is a comprehensive, web-based tool for metabolomic data processing, statistical analysis, interactive visualization, functional interpretation, as well as integration with other omics data. It has been carefully engineered to enable researchers with no programming skills to perform a wide array of basic as well as advanced analyses of various metabolomics data generated from both targeted and untargeted approaches. To keep up with the rapid growth and development of the field, MetaboAnalyst has continuously evolved, with four major releases over the past decade. The first version of the web server, MetaboAnalyst 1.0 (Xia, Psychogios, Young, & Wishart, 2009), was designed primarily for processing and statistical analysis of metabolomic data. The second version of the server, MetaboAnalyst 2.0 (Xia, Mandal, Sinelnikov, Broadhurst, & Wishart, 2012), incorporated additional components for functional analyses such as metabolic pathway analysis (MetPA; Xia & Wishart, 2010) and metabolite set enrichment analysis (MSEA; Xia & Wishart, 2010). The third version of the server, MetaboAnalyst 3.0 (Xia, Sinelnikov, Han, & Wishart, 2015), added four new modules including: two-factor and time-series data analysis, biomarker analysis, sample size estimation with power analysis, and joint pathway analysis. It was re-implemented using a new web framework deployed on a cloud server for improved performance. The latest version, MetaboAnalyst 4.0 (Chong et al., 2018), features three new modules including: predicting pathway activities from MS peak lists, biomarker analysis integrating multiple metabolomic datasets, and integrative analysis of multi-omics data through knowledge-based networks. The web interface has also been completely re-engineered to display underlying R commands in real time, which can be used in conjunction with our companion MetaboAnalystR (Chong & Xia, 2018; Chong, Yamamoto, & Xia, 2019) package to enable more flexible and reproducible metabolomics data analysis and batch processing.

MetaboAnalyst 4.0 consists of two main components—a data processing component to deal with different data inputs and a data analysis component containing 12 different modules. To make the complex data processing operation intuitive yet flexible, MetaboAnalyst uses a step-wise design concept to guide users through all major steps, beginning with data-type selection, formatting, "cleansing," and normalization. The 12 data analysis modules can be arranged into four general categories: (1) exploratory statistical analysis, (2) functional analysis, (3) data integration and systems biology, and (4) general utility functions. The exploratory statistical analysis category (general statistics, biomarker analysis, two-factor/time-series analysis, and power analysis) can accept data from either targeted or untargeted metabolomic data sets. The functional analysis category includes MSEA and MetPA, as well as a new module for predicting pathway activity from an MS peak list. The data integration and systems biology category

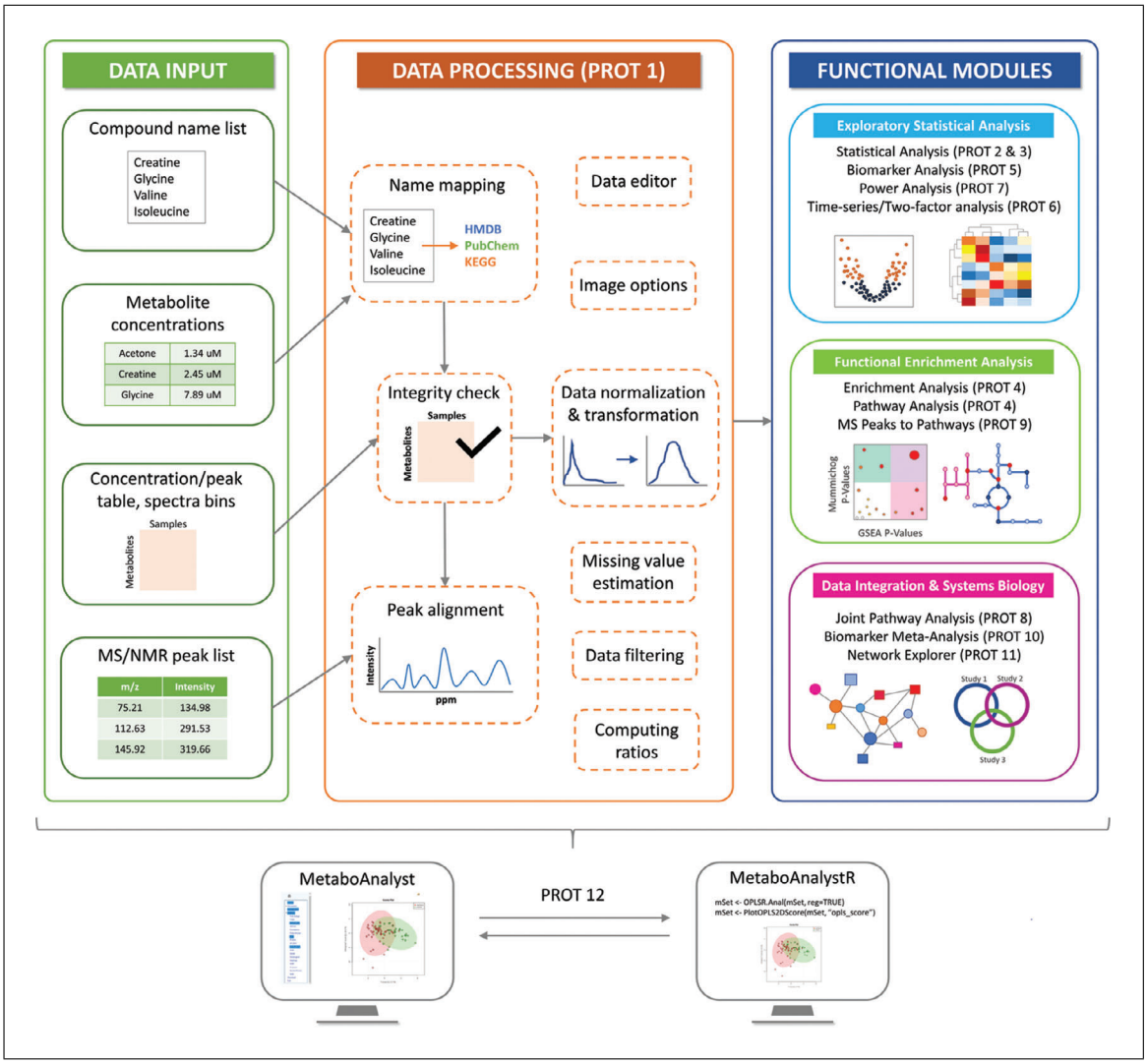


Figure 1 The MetaboAnalyst 4.0 flowchart. There are two main components: a data processing component to deal with different data inputs, and a data analysis component containing individual modules which can be categorized into "Exploratory Statistical Analysis," "Functional Enrichment Analysis," and "Data Integration and Systems Biology." PROT X refers to the relevant protocol for that component/module.

includes three modules—biomarker meta-analysis, joint pathway analysis, and network explorer. Finally, the data processing and utilities category contains common tools for compound ID conversion and batch effect correction, as well as links to three web-based tools for spectral analysis and annotation including Bayesil (Ravanbakhsh et al., 2015), GC-Autofit (*http://gc-autofit.wishartlab.com/*), and XCMS-Online (Huan et al., 2017).

Figure 1 illustrates these components as well as the general workflow of MetaboAnalyst 4.0. The colored dotted boxes indicate optional procedures or procedures applicable only for certain kinds of analyses. MetaboAnalyst has been primarily designed for the comprehensive analysis of a table or a list of features (compounds, peaks or spectral bins) commonly generated from metabolomics. Although the data processing module is shared among different modules, users need to first decide which module they want to use before they start an analysis. Making this decision early on greatly reduces potential errors as the navigation becomes much more straightforward. It also significantly improves the computational efficiency of the web server by loading only those resources needed for user-specified analysis modules.

This unit contains 12 basic protocols corresponding to the 12 main functions of MetaboAnalyst 4.0. They are summarized below.

- Basic Protocol 1—how to perform data processing, normalization, and quality checking
- Basic Protocol 2—how to identify significant features identification using various univariate methods
- Basic Protocol 3—how to perform multivariate exploratory data analysis
- Basic Protocol 4—how to perform functional interpretation using MSEA and MetPA
- Basic Protocol 5—how to perform biomarker analysis based on receiver operating characteristic (ROC) curves
- Basic Protocol 6—how to perform time series and two-factor statistical analysis
- Basic Protocol 7—how to perform sample size estimation and power analysis
- Basic Protocol 8—how to perform joint pathway analysis
- Basic Protocol 9—how to predict pathway activity from a peak list
- Basic Protocol 10—how to perform meta-analysis of multiple metabolomic data
- Basic Protocol 11—how to integrate multi-omics data using network-based approaches
- Basic Protocol 12—how to use MetaboAnalystR package for flexible and reproducible metabolomic data analysis, as well as perform raw data preprocessing of high-resolution LC-MS data.

BASIC PROTOCOL 1

DATA UPLOADING, PROCESSING, AND NORMALIZATION

MetaboAnalyst can accept either a comma separated value (`.csv`) file, a tab-delimited text (`.txt`) file, a compressed (`.zip`) file, or a list of compound names or MS peaks. A `.csv` or `.txt` file is typically used when the whole data set has already been preprocessed to a tabular format such as a compound concentration table, a spectral bin table, or a peak intensity table. The zip file format is typically used to upload multiple MS spectra or multiple peak list files. However, due to bandwidth constraints on the internet, it is generally impractical to upload a large number of raw spectra remotely to MetaboAnalyst. Rather, raw spectra should be preprocessed (peak picked and/or aligned) using locally installed software to produce the necessary (and much smaller) peak list files or peak intensity tables before uploading to MetaboAnalyst. A number of locally installable, freely available MS spectral preprocessing tools are available, including MetAlign (Lommen & Kools, 2012), OpenMS (Röst et al., 2016), MZmine (Pluskal, Castillo, Villar-Briones, & Oresic, 2010), XCMS (Smith, Want, O'Maille, Abagyan, & Siuzdak, 2006), mzMatch (Scheltema, Jankevics, Jansen, Swertz, & Breitling, 2011), and MS-DIAL (Tsugawa et al., 2015). Further, the MetaboAnalystR package has been recently updated to support raw spectral preprocessing of LC-MS data (Chong et al., 2019). For web-based tools, we recommend XCMS-Online (Tautenhahn, Patti, Rinehart, & Siuzdak, 2012) for general-purpose LC-MS spectral processing. Most NMR instruments have their own proprietary software for rapid spectral processing, binning, peak picking, and alignment. If one wishes to use a freely available alternative, we recommend Bayesil (Ravanbakhsh et al., 2015) for processing and annotating NMR spectra on human biofluids.

This protocol describes how to upload multiple NMR peak list files to MetaboAnalyst for data processing, normalization, and data quality checking. Most of the steps described can be directly applied to the processing of MS peak list files. Uploading and processing a CSV (`.csv`) file or a text (`.txt`) file are relatively straightforward and will be covered in the following protocols.

Necessary Resources

Hardware

A computer with internet access

Software

An up-to-date web browser, such as Google Chrome (*http://www.google.com/chrome*), Mozilla Firefox (*http://www.mozilla.com/*), Safari (*http://www.apple.com/safari/*), or Internet Explorer (*http://www.microsoft.com/ie/*). JavaScript must be enabled in the web browser.

Files

None

Data preparation and uploading

1. In this example, we will download one of the sample datasets provided by MetaboAnalyst 4.0 and use it to illustrate MetaboAnalyst's data processing options. Go to the MetaboAnalyst 4.0 home page at *https://www.metaboanalyst.ca*. Click the "Data Formats" link on the left navigation bar to enter the "Data Formats" page.

   The "Data Formats" page provides several example data sets used by MetaboAnalyst. It contains detailed documentation describing each of these data sets as well as screenshot illustrations for all the data types supported by MetaboAnalyst. It is critical for the first-time user to get familiar with the specifications for the data type he or she intends to upload to MetaboAnalyst.

2. There are two types of Data Formats listed at the top this page. One is a set of five text files called "Sample datasets for downloading" and the other is a set of four larger files called "Zipped files (.zip) format data sets." Go to the "Zipped files (.zip) format" section and click the "download" link on the first entry in the list, which is marked with the label "NMR peak lists (2 columns—chemical shift and intensity)" and save the downloaded file as `nmr_peaks.zip`.

   These data are a subset of data collected from an ^{1}H NMR–based metabolomic study used to evaluate renal damage (glomerulonephritides) from urine samples (Psihogios et al., 2007). The data set contains 50 peak list files: 25 from renal patients and 25 from healthy controls. These files were placed into two folders labeled as Renal and Healthy, respectively. A zip file was created from these two folders. Each peak list file was saved in a CSV format containing two columns, with the first column corresponding to chemical shift peak positions (ppm) and the second column corresponding to peak intensities. The first line of each peak list file is reserved for column labels. MS peak list files can be prepared in the same way, with either two-column (mass and intensities) or three-column format (mass, retention time and intensities), but not a mixture of both.

3. Click the "Home" link at the top of the left navigation bar to return to the MetaboAnalyst home page. Then click the ">>click here to start<<" link at the top of the page to enter the "Module Overview" selection page (Fig. 2). A circular "clock" shows a total of 12 different modules. From this panel, click on the "Statistical Analysis" button located on the top left of the circle to enter the corresponding "Data Upload" page (Fig. 3).

   MetaboAnalyst uses a navigation tree (on the left side) to guide users through each of the data processing, analysis, and interpretation steps. Some links will be enabled or disabled depending on the context. The "Upload" hyperlink is highlighted in blue, representing the current step. On the right-hand side is the R Command History panel, which displays the step-by-step R command workflow of your analysis in real time. Each time the user presses a button to perform a function or generate a plot in MetaboAnalyst 4.0, the panel will display the underlying R commands. They are made visible to improve

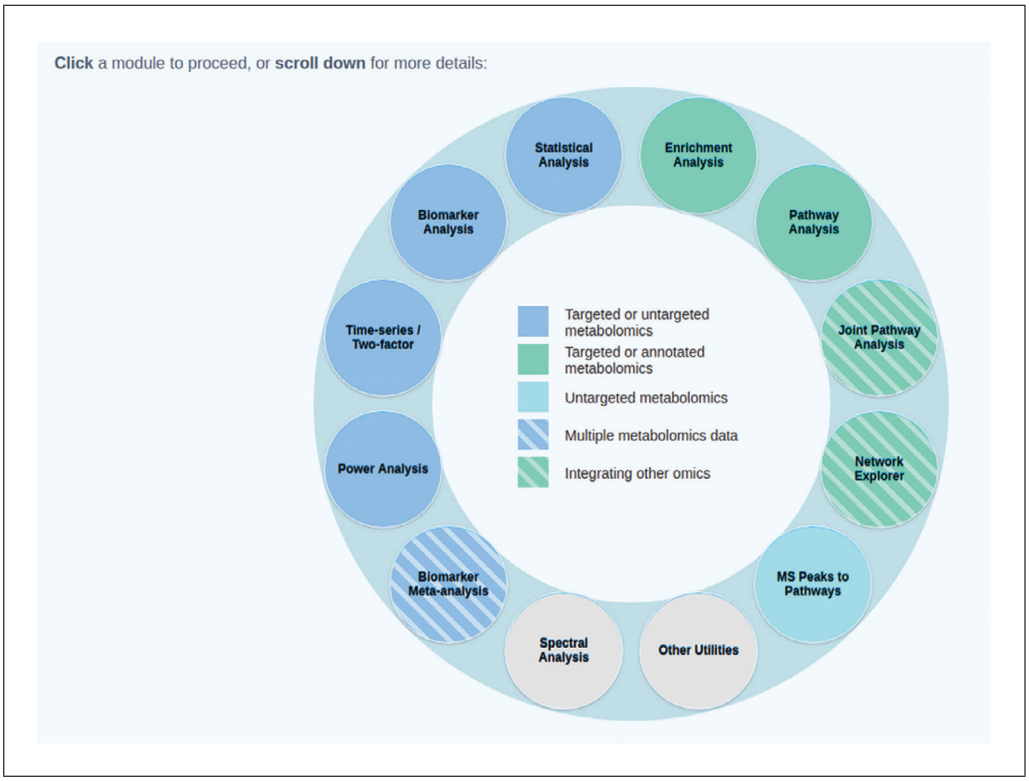


Figure 2 The "Module Overview" page. This circular panel is the entry point to the different functional modules.

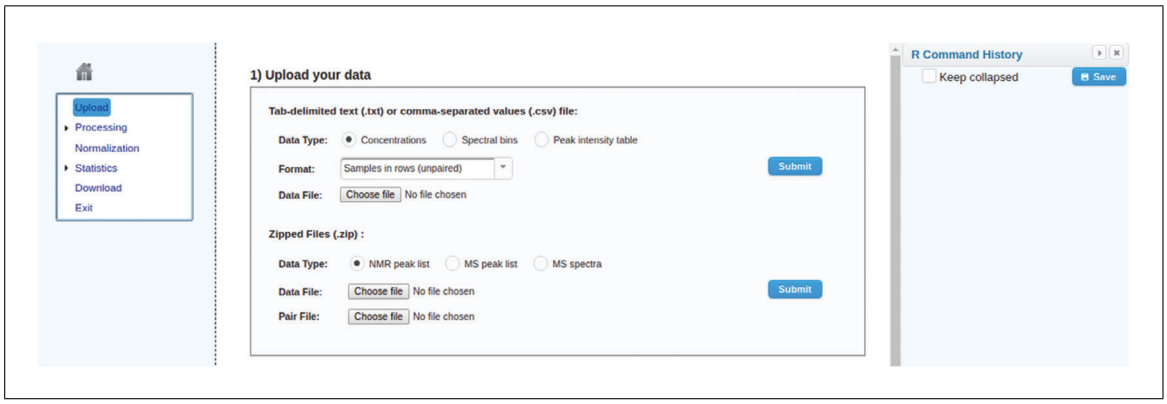


Figure 3 The "Data Upload" page for the Statistical Analysis module. Users can upload a text file in `.csv` or `.txt` format, or upload a zip file.

transparency and reproducibility in complex data analysis. The R Command History can be used to reproduce your analysis locally in R via the companion MetaboAnalystR package (further detailed in Basic Protocol 12).

4. On the "Data Upload" page, there are three types of data sets that can be uploaded (tab-delimited text, comma separated value, and zipped files). Under the "Zipped Files (.zip)," specify the data type as "NMR peak list" by clicking this radio button. Click the "Choose File" button (beside the "Data File" label) to locate the `nmr_peaks.zip` file saved on your computer, and then click "Submit" button on the right to upload the data.

There are many data-compression tools for creating zip files. Some of them use different algorithms and may not be recognized by MetaboAnalyst. Users should try to avoid the most advanced or latest compression options for creating a zip file. For paired analysis, an additional file is required to give the information for each pair of samples. Each pair

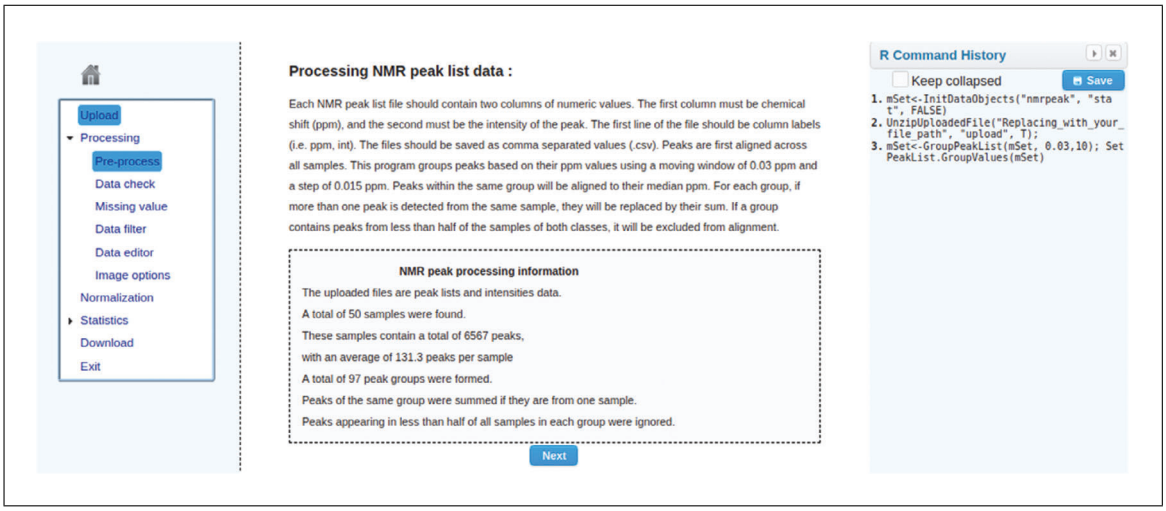


Figure 4 The "NMR Peak Processing" page showing the output from peak alignment. This page provides a description of the peak alignment algorithm as well as the results from the peak processing steps, including information on the number of samples, number of peaks, number of peaks per sample, and number of peak groups.

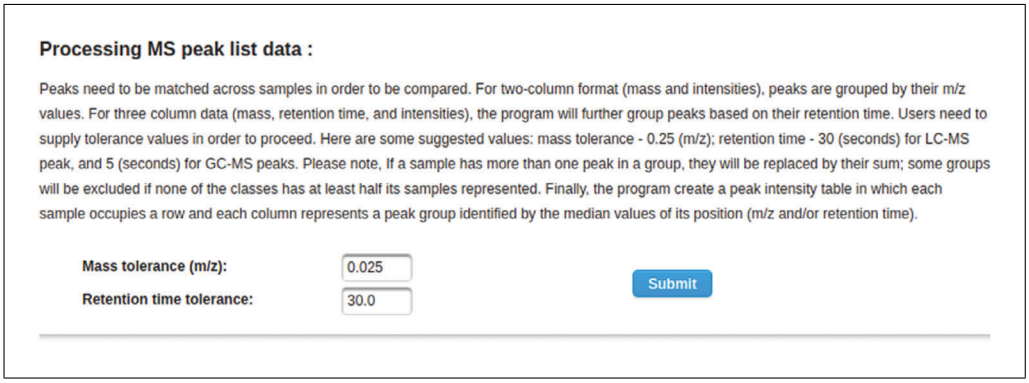


Figure 5 The "MS Peak Grouping" page showing the two parameters (mass tolerance and retention time tolerance) that can be adjusted based on the experimental setup.

must be indicated by the sample names separated by a colon, with one pair per line, and uploaded as a text (`.txt`) file.

Peak list alignment

5. After the files are uploaded, MetaboAnalyst first extracts all the data files from the zip file. It then starts to perform a peak list alignment (part of MetaboAnalyst's pre-processing and processing functions). After a few seconds, the result will be shown in Figure 4. The central view displays a summary of the data uploaded. On the right side of this page is a panel titled "R Command History" with three R commands written into the window. Take a few moments to look at the commands and notice how new commands are appended after each button click. Click the "Next" button.

   For NMR peak list alignment, peaks are aligned across all samples based on their chemical shift (ppm) values using a moving window of 0.03 ppm and a step of 0.015 ppm. Peaks within the same group will be aligned to their median chemical shift value. For each group, if more than one peak is detected from the same sample, they will be replaced by their sum. If a group contains peaks from less than half of the samples of both classes, it will be excluded from the alignment. For MS peak list alignment, users can set the parameters (mass tolerance and retention time tolerance) based on their experimental setup, as shown in Figure 5.

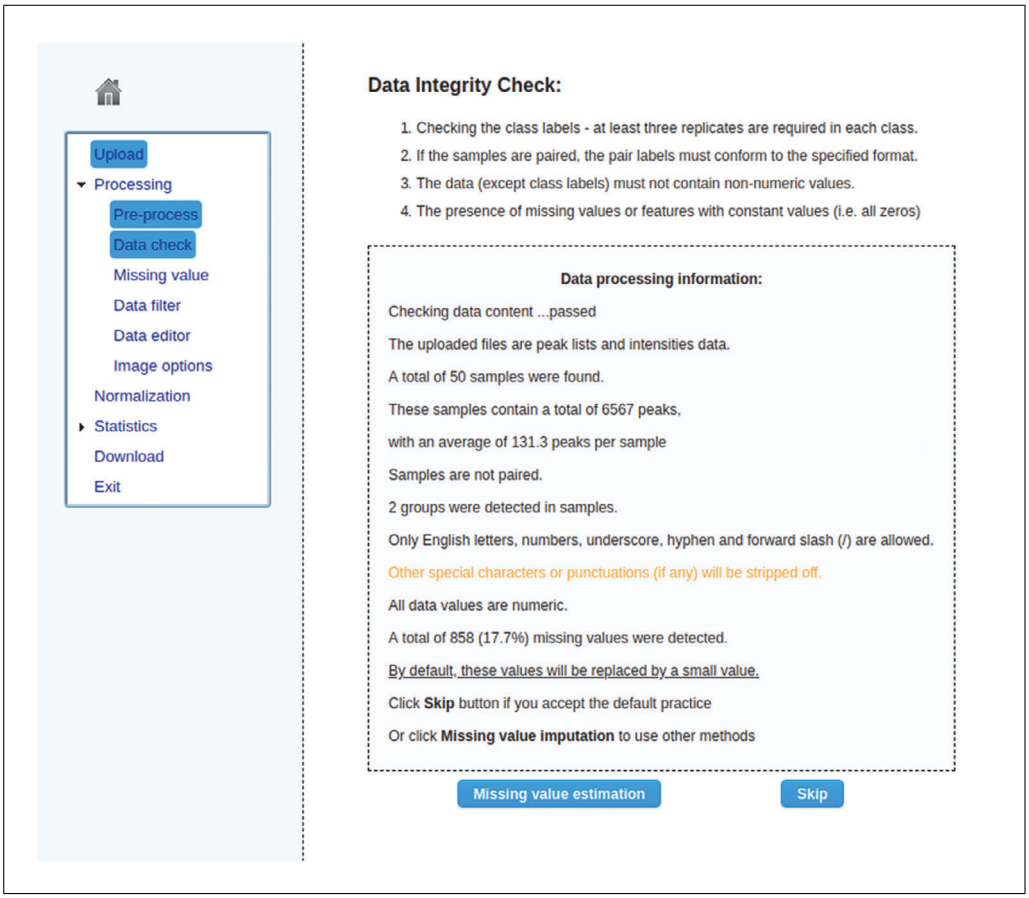


Figure 6 The "Data Integrity Check" page. This page summarizes the results of the data integrity checks. Refer to the text for further details.

Data integrity check and missing value estimation

6. The "Data Integrity Check" page should now be displayed. The data integrity check is performed to ensure that the data meets the basic requirements for meaningful downstream analysis. The results are shown in Figure 6.

   MetaboAnalyst checks for three common problems with metabolomic data: (1) The sample and variable names must be unique and contain no special characters (e.g., Latin or Greek letters), (2) at least three replicates are required for each group, and (3) all data values except phenotype labels must be numeric. Missing values are allowed and can be indicated as blanks or marked as NA (without quotes). For paired analysis, MetaboAnalyst also checks if the data pairs conform to the specified format.

7. The results indicate that 17.7% of the data values in this data set are missing. By default, missing values are considered to be caused by signals below the detection limit and will be replaced by a very small value (half of the minimum positive value found in the data set). In this case, accept the default option by clicking the "Skip" button to go to the next page.

   MetaboAnalyst provides more advanced procedures to deal with missing values. Click the "Missing value imputation" button to open the corresponding page (Fig. 7). For example, users can automatically exclude features with too many missing values, or use various methods to perform missing value estimation such as replacing missing values using the mean/median, Probabilistic PCA (PPCA), Bayesian PCA (BPCA), or Singular Value Decomposition (SVD) (Stacklies, Redestig, Scholz, Walther, & Selbig, 2007).

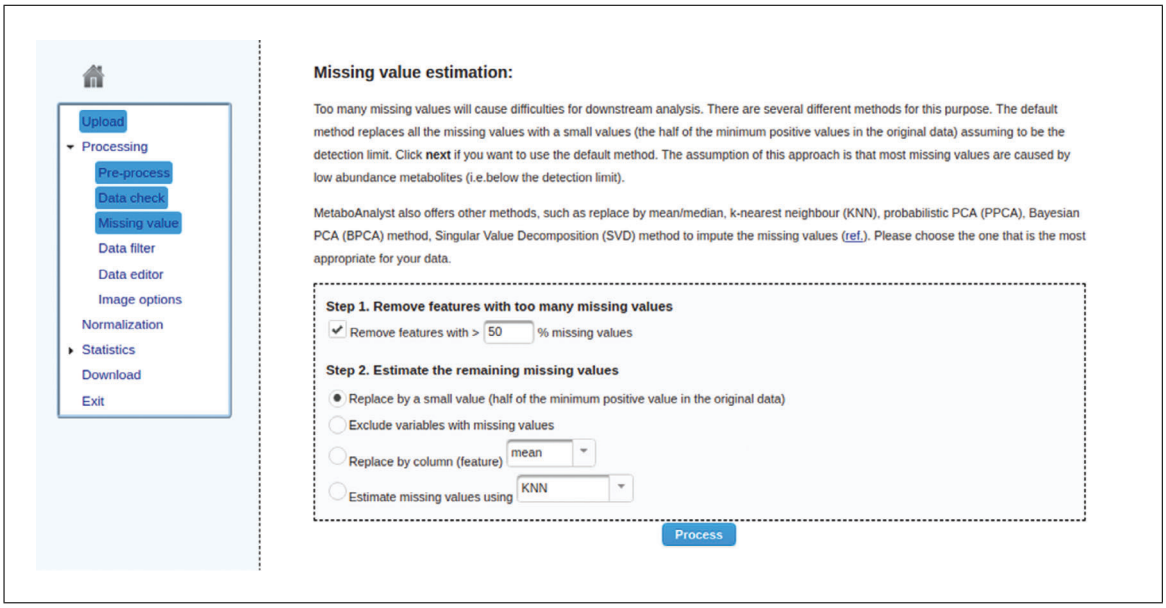


Figure 7 The "Missing Value Imputation" page. This screenshot illustrates the options available for missing value estimation.

Low quality data filtering

8. After completing the data integrity check, the next step involves navigating to the "Data Normalization" page. For untargeted metabolomics data (i.e., peak lists, spectral bins), an extra step on data filtering will be shown (Fig. 8). In this case, select the default "Interquantile range (IQR)" and then "Submit" to perform the data filtering. Next, click "Proceed" to move to the "Data Normalization" page.

   The purpose of data filtering is to remove noise or non-informative variables in a given data set. Three types of variables are usually considered to be of low quality: (a) baseline noise (b) non-informative variables, and (c) variables showing low repeatability. Baseline noise variables are characterized by very small or close-to-zero values. They can be detected by comparing their mean or median values. Non-informative variables are characterized by near-constant values throughout the experimental conditions. They can be detected by comparing their standard deviations (SD) or the robust estimate interquantile ranges (IQR). Variables with low repeatability can be detected by comparing against QC samples using the relative standard deviation (RSD = SD/mean). Those with high RSD will be removed from the subsequent analysis. As shown in Figure 8, the total number of variables removed is proportional to the total variable size. Please note that the effect of data filtering is cumulative. For instance, when there are significant levels of noise in the data, users can first filter the data by intensity (i.e., median), then click the "Data filter" hyperlink on the navigation tree to go back to this page and filter the data once more by variance. The data filtering methods in MetaboAnalyst were implemented mainly based on the non-specific filtering (nsFilter) approach used by the Bioconductor genefilter package (Gentleman et al., 2004). Data filtering usually leads to improved statistical power in the downstream data analysis (Hackstadt & Hess, 2009).

Data normalization

9. The "Data Normalization" page is shown in Figure 9. In this case, select "Normalization by sum" for sample normalization, "None" for data transformation and "Auto scaling" for data scaling. After selecting these options, click the "Normalize" button.

   The purpose of data normalization is to reduce any systematic bias and to improve overall data consistency so that meaningful biological comparisons can be made. At this point, the data has been transformed to a matrix with the samples in rows and the variables (compounds/peaks/bins) in columns. MetaboAnalyst offers three types of

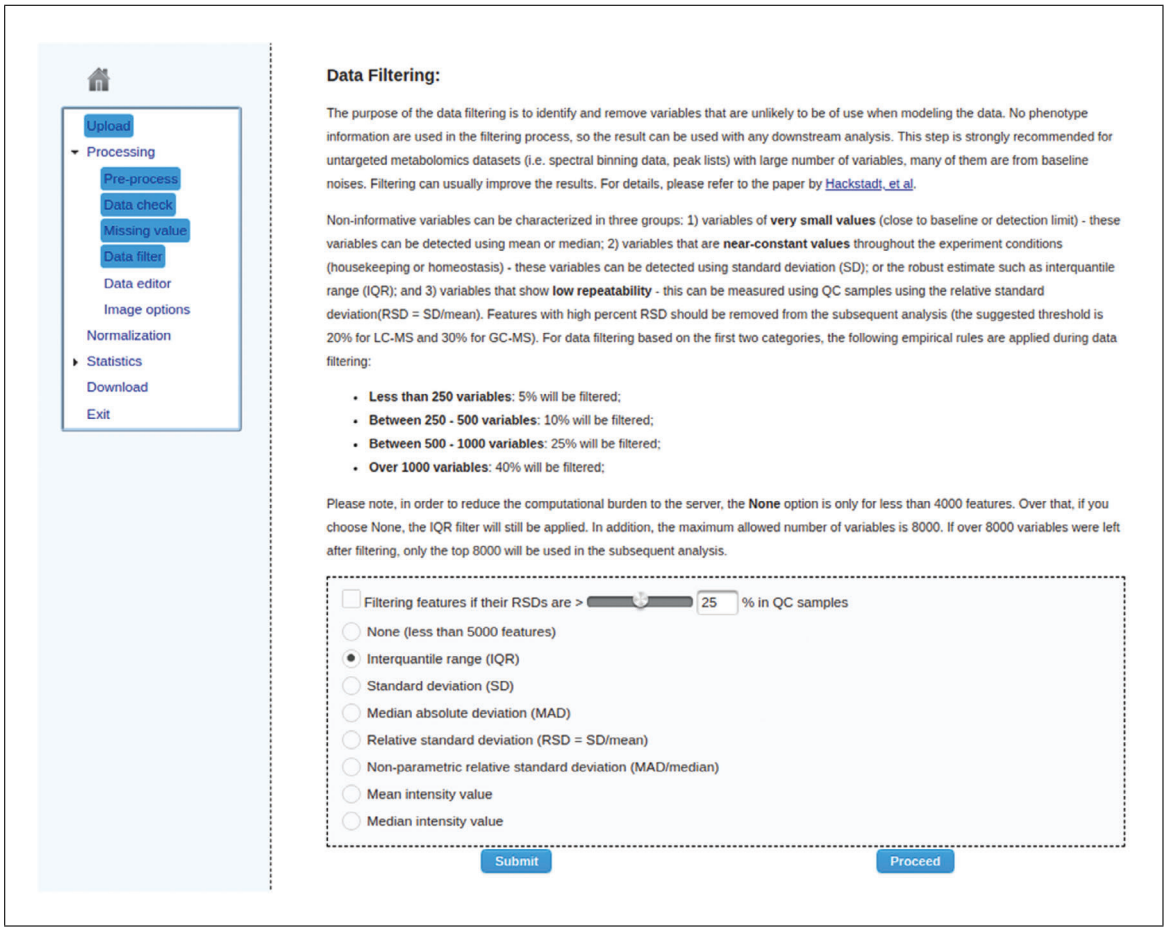


Figure 8 The "Data Filtering" page. This page demonstrates the methods and empirical rules used for data filtering in MetaboAnalyst. Interquantile range (IQR) is selected in this case.

normalization: (1) sample normalization, (2) data transformation, and (3) data scaling. The aim of sample normalization is to make each sample (row) comparable to each other. For instance, normalization can adjust metabolite concentration levels in urine samples with different dilution effects due to a patient's water intake. A common route for urine normalization is to divide metabolite concentrations by the urine's creatinine concentration (normalization by a reference feature). MetaboAnalyst offers five options for sample normalization—no normalization, normalization by sum, normalization by median (Hendriks et al., 2007), normalization by a reference sample (Dieterle, Ross, Schlotterbeck, & Senn, 2006), normalization by a reference feature, and sample-specific normalization. Meanwhile, data transformation is a method to transform variables so that they exhibit a more normal or Gaussian distribution as opposed to a skewed or logarithmic distribution. Three different data transformation methods are offered: (1) no transformation, (2) log transformation, and (3) cube-root transformation. Finally, data scaling aims to make each variable comparable to each other. This procedure is useful when variables are of very different orders of magnitude (some metabolites are at micromolar levels while other metabolites are at millimolar levels). Four methods have been implemented for this purpose—no scaling, auto scaling, Pareto scaling, and range scaling. Users should be aware that currently there is no consensus on which methods work best for different types of metabolomic data. As a result, users are encouraged to conduct a bit of trial and error testing by trying a transformation or normalization step and visually assessing how Gaussian (or "bell-shaped") the data distribution appears (see Fig. 10). For a more detailed discussion on the effects of various normalization and scaling procedures, please refer to the two excellent papers by Craig and van den Berg (Craig, Cloarec, Holmes, Nicholson, & Lindon, 2006; van den Berg, Hoefsloot, Westerhuis, Smilde, & van der Werf, 2006).

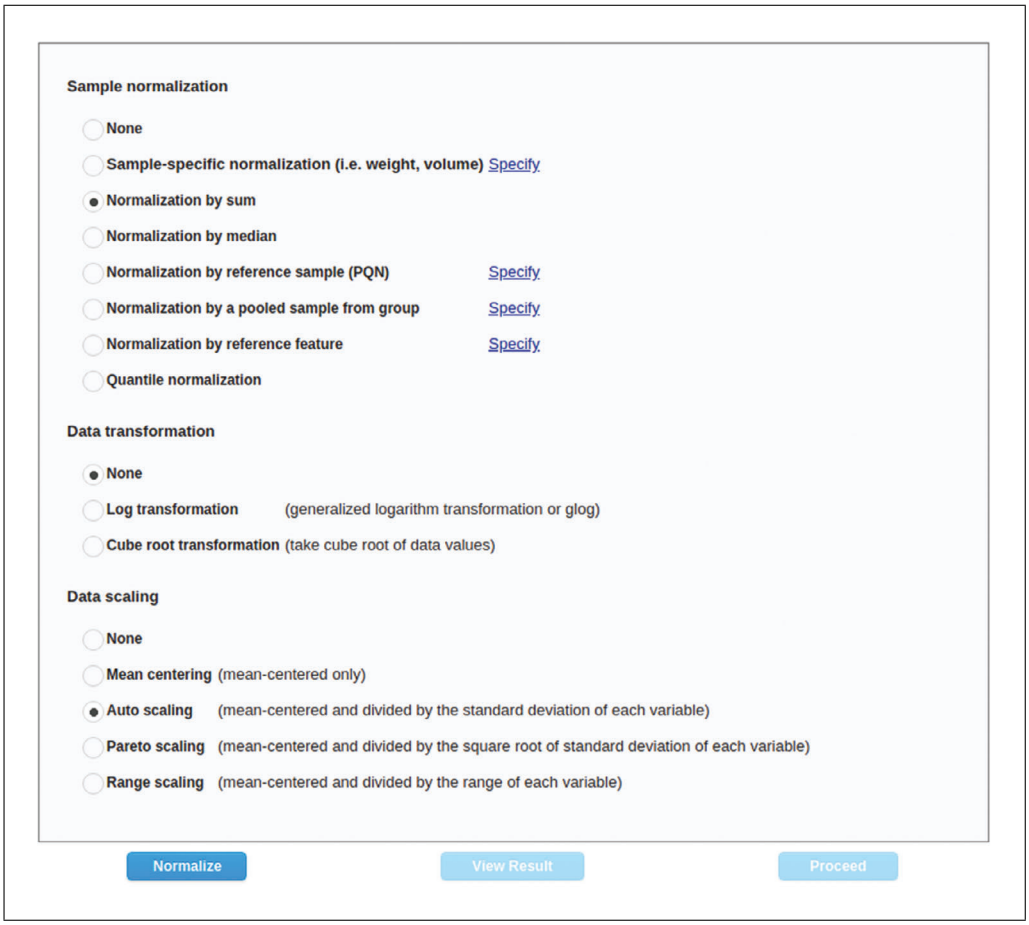


Figure 9 The "Data Normalization" page showing all available options.

10. Click "View Result" to view a graphical summary of the effect of data normalization on the data (Fig. 10). The top plots show the overall data distribution based on kernel density estimation (before normalization—on the left, and after normalization—on the right), while the two horizontal box plots on the bottom show the distributions of individual variables or metabolite concentrations (before and after normalization). Users should compare the graphical summary on the left and the right (before and after normalization) to guide them toward choosing the methods that work best with their data.

    Our initial choice of normalization, scaling and transformation operations has "fortuitously" produced a set of data that appears to be properly normalized. This is because of the appearance of the characteristic "bell-shaped" distribution of the top right kernel density graph. However, if the shape of this graph were highly skewed or still looked like the distribution shown on the left, then more normalization, scaling or transformation adjustments would need to be made. To make these data adjustments, users should click on the close button (the circled "X" on the upper right corner of the pop-up window) to return to the "Data Normalization" page. A different selection of normalization, transformation, and scaling operations could then be selected (try normalization by median, log transformation and Pareto scaling). After making these changes the user would have to press the "Normalize" button at the bottom again, followed by selecting the "View Results" button to pop up a new window. Inspection of the new results would reveal a kernel density distribution (on the upper right) that is too narrow and slightly skewed. Therefore, clearing the pop-up window and selecting the original normalization functions (normalization by sum, no transformation, and auto-scaling) and repeating the normalization process would be the best option.

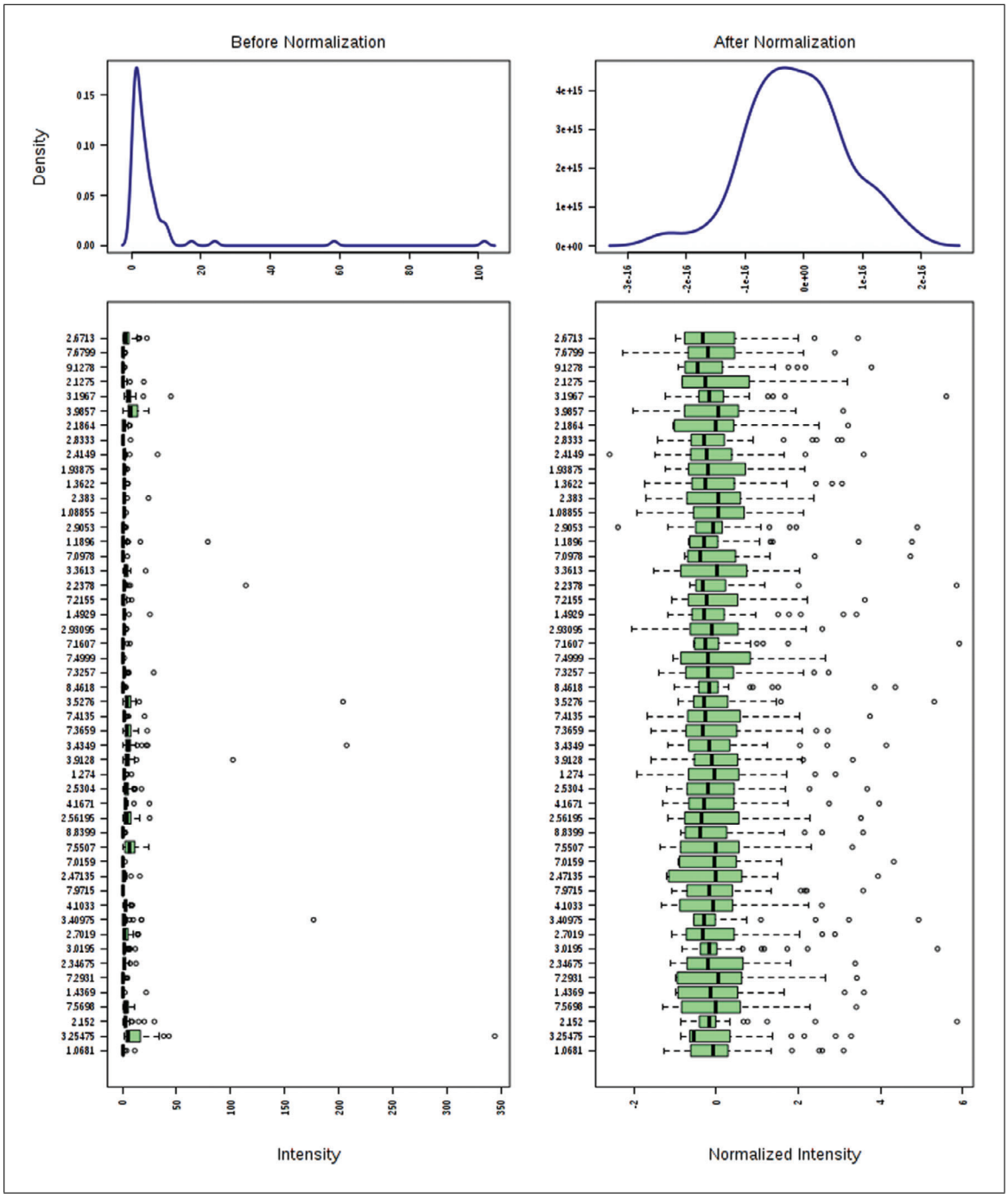


Figure 10 The "Data Normalization Result" page showing a graphical summary of the data before and after the normalization procedure(s).

11. Once a user is satisfied with the shape and skewness of the upper right curve (looking like a symmetric Gaussian curve), click the "Proceed" button.

Data quality check, outlier detection, and exclusion

12. Prior to conducting any kind of data analysis in metabolomics, it is important to assess the overall data quality and check for any obvious outliers. As there are no obvious outliers based on the above procedures, we will redo our data normalization to create some "artificial" outliers—for illustration purposes only. Click the "Normalization" hyperlink on the navigation tree (on the left side of the window) to return to the "Data Normalization" page. Make sure to set the Sample

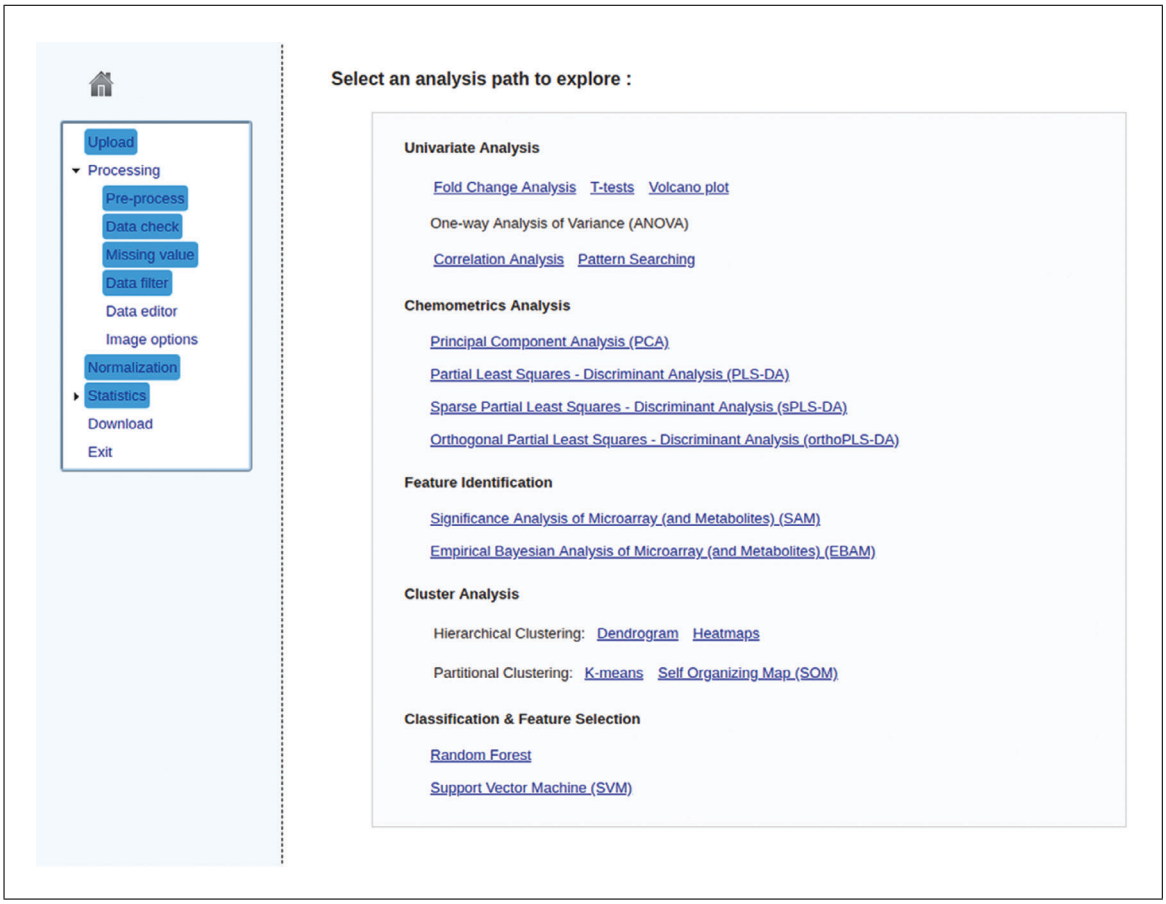


Figure 11 A screenshot showing an overview of the available statistical methods organized under five main categories for exploratory analysis.

normalization to "none," choose "Auto scaling" for Data scaling, and click the "Normalize" button. The previously normalized data is now overwritten by the new data containing several artificial outliers.

In most metabolomic studies, urine samples need to be adjusted for dilution effects by a method called sum normalization (when using chemometric methods), by creatinine concentration (when using quantitative approaches), or by a reference sample [also known as probabilistic quotient normalization (Dieterle et al., 2006)]. If we omit this normalization step, some samples with very different dilution effects will stand out as distinct outliers.

13. At this point, the "Data Normalization Result" page should be displayed. Click the "Proceed" button to move on to the "Data Analysis Overview" page (Fig. 11). At this stage, the "Statistics" hyperlink on the navigation tree is highlighted in blue. On this page, five major analysis categories are presented: (1) Univariate Analysis, (2) Chemometrics Analysis, (3) Feature Identification, (4) Cluster Analysis, and (5) Classification and Feature Selection. Listed under each of these five categories are many other subcategories containing individual methods. Two approaches are particularly useful for obtaining a data overview and performing outlier detection—principal component analysis (PCA), which is listed under Chemometrics Analysis, and heatmaps, which is listed under Cluster Analysis. These two methods will be more fully explored in Basic Protocol 3 when we perform multivariate data analysis. Here, the focus is on their abilities to generate a data overview and to perform outlier detection.

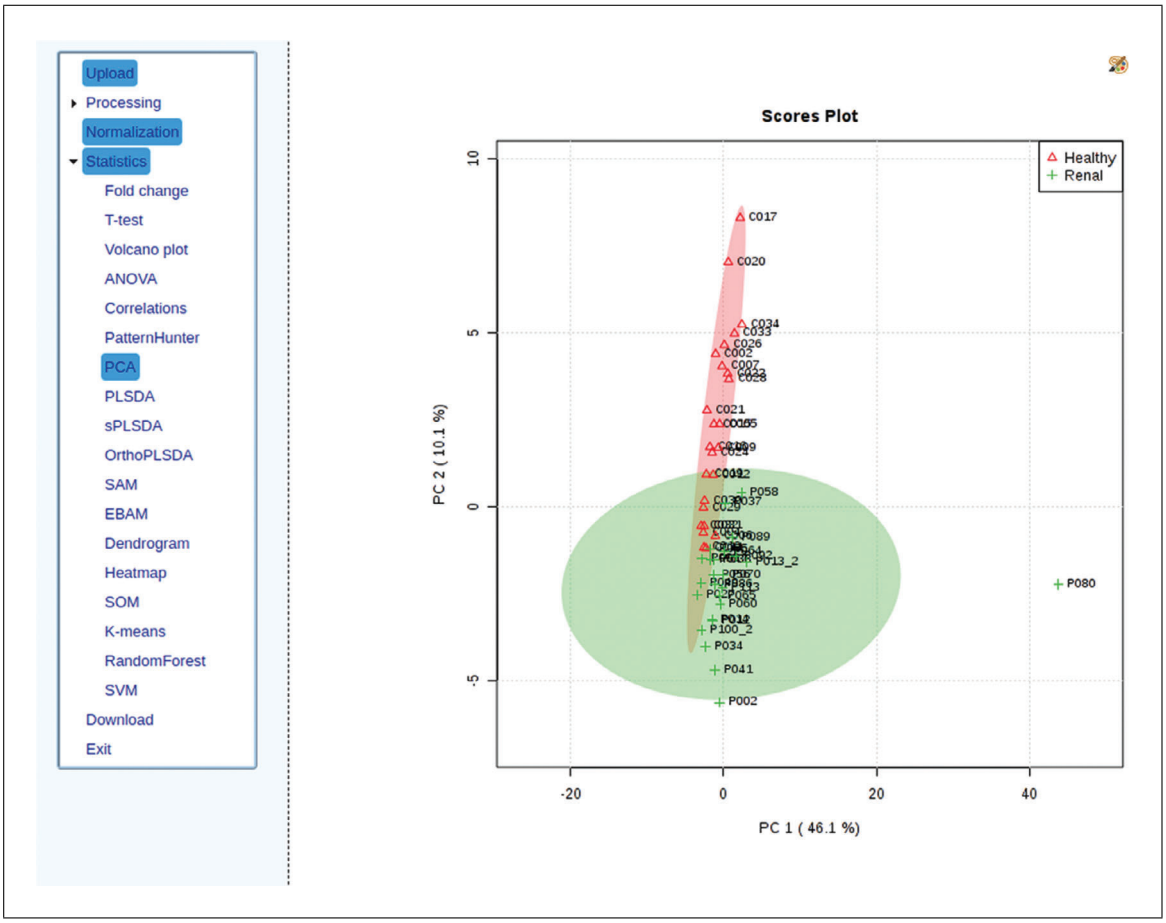


Figure 12 A screenshot showing outlier detection with the PCA 2D score plot. Sample P080 stands out as a potential outlier.

14. Click the "Principal Component Analysis" hyperlink under "Chemometrics Analysis." After a few seconds, the PCA results should be presented in multi-tab panels. Click the "2D Scores Plot" tab to view the score plot between PC1 and PC2. The default 2D Scores Plot draws 95% confidence regions for each group. The result is shown in Figure 12. The sample labeled *P080* from the Renal group clearly stands out as being distinct from most other data points. It is a potential outlier, and consequently more investigation should be performed for this sample.

15. Now the left navigation bar has expanded to show all the methods available under the Statistics analysis module. This allows users to navigate easily to the different statistics methods without having to constantly return to the "Data Analysis Overview" page. Click the "Heatmap" hyperlink on the navigation tree. After a few seconds, a heatmap should be displayed. A heatmap is a visualization technique that converts a numerical table into a corresponding 2D color map (ranging from "hot" to "cold" colors) to provide an intuitive overview of the data values. Heatmaps are often used together with hierarchical clustering techniques to reorganize the rows/columns of data with dendrograms plotted on the corresponding side(s). This is the default behavior of MetaboAnalyst. In this case, the goal is to obtain an overview of the data without any kind of clustering. Check the "Do not reorganize" check box and change the default "Samples" option to "Both" (rows and columns) from the drop-down menu. Leave all other options in their default mode and click the "Submit" button. The result is shown in Figure 13. Again, sample *P080* clearly stands out. The heatmap shows that many peak intensity values coming from this sample are extraordinarily high compared to those of other samples.

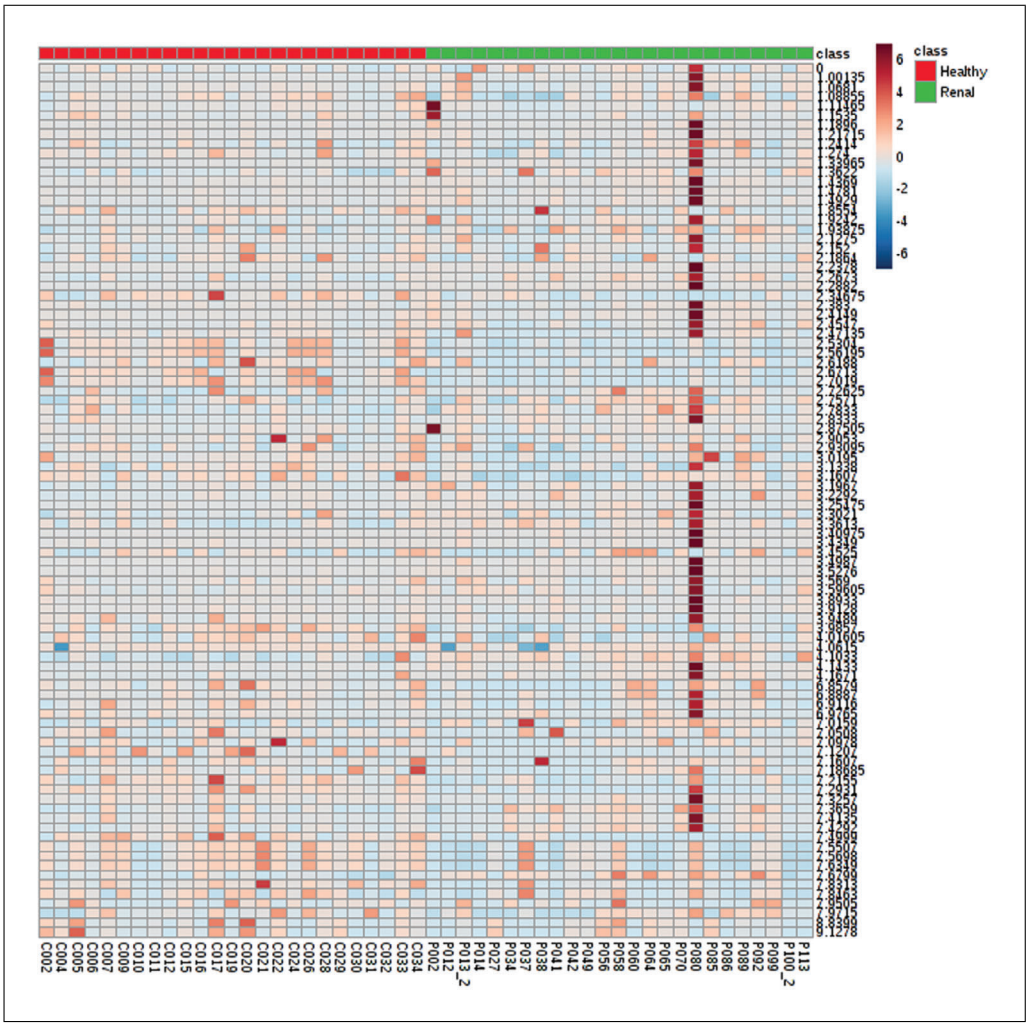


Figure 13 A screenshot showing outlier detection using heatmap visualization. Samples are in columns and features are in rows. The colors vary from deep blue to dark brown to indicate data values that change from very low (cold) to extremely high (hot). Note that most data points from Sample P080 exhibit very high values.

In some cases, the sample labels may become difficult to read. Users can select the "Detail View" option instead of the default "Overview" to re-generate a larger heatmap to show names for all features and samples.

16. To exclude this sample from being used in the analysis, return to the navigation tree on the left. Click the "Processing" hyperlink (the triangle) to expand all of its branches. Locate the "Data Editor" hyperlink and click on it. The "Data Editor" page is shown in Figure 14. It contains three tabs labeled "Edit Samples," "Edit Features," and "Edit Groups," with the Edit Samples tab shown as default. Scroll down through the list of sample names to locate sample *P080*. Select it and then click the button with the right arrow (add) to move it to the "Exclude" list. Sample *P080* has now been removed from the left and appears in the right-hand "Exclude" panel. Click the "Submit" button. Users will be re-directed to the "Data Normalization" page.

Removing a sample will affect the data structure, and consequently normalization needs to be re-performed on the edited data set. After normalization, users should still conduct a data overview and check for any data outliers or abnormalities, as described in the previous steps. The process is iterative. Note that any previously removed samples can be restored by using the left arrow buttons on the "Data Editor" page.

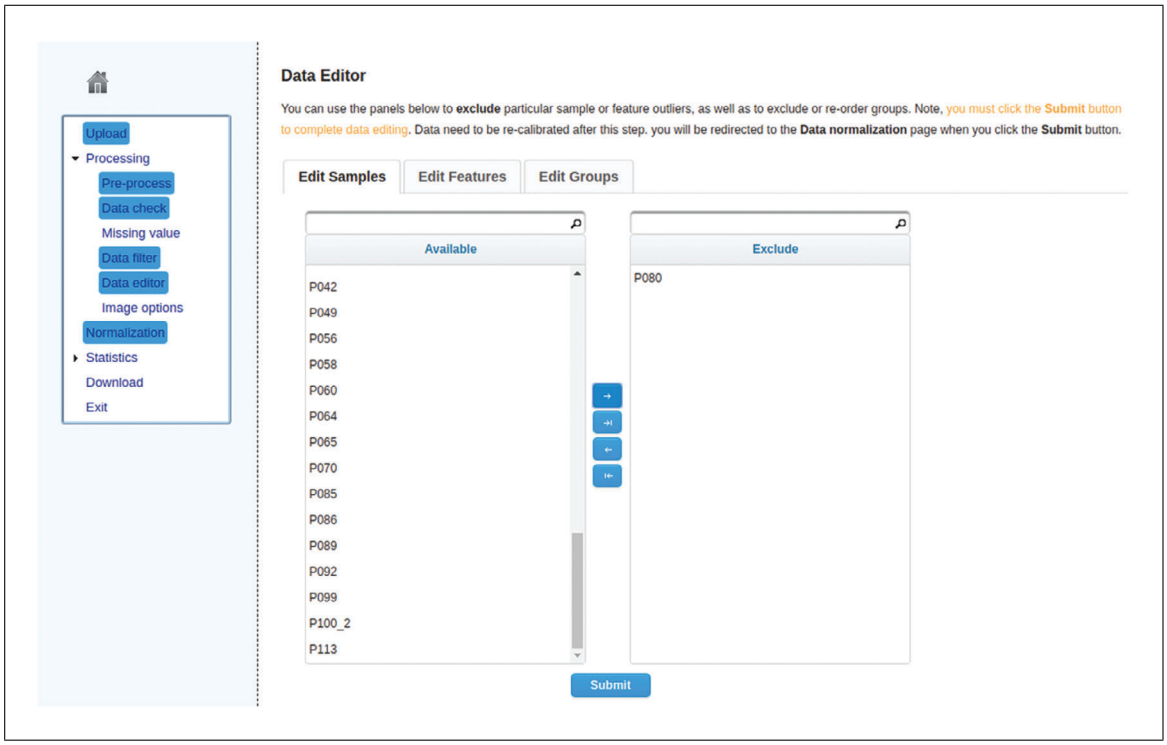


Figure 14 A screenshot showing how the Data Editor is used to exclude outlier(s). Here, the Sample Editor is used to exclude sample *P080*. Similar procedures can be performed to exclude variables using the Feature Editor or to exclude a group from multiple-group data using the Group Editor. Users must click the "Submit" button after completing the operation.

Data download

17. Click the "Download" hyperlink on the bottom of the navigation tree. The download page displays a table showing all the figures and result tables, as well as the R Command History file generated during the analysis session. Click the "Generate Report" button to create a PDF report that describes all the steps performed, with necessary background information, followed by the results in the form of bulleted lists, tables, and images as appropriate. At this moment, you should download the analysis report and the zip file, and click "Exit" to complete this session.

BASIC PROTOCOL 2

IDENTIFICATION OF SIGNFICANT VARIABLES

In metabolomics studies, it is often assumed that most observed changes in metabolite concentrations or spectral profiles are a result of normal physiological variations (background noise) and that only a small proportion of these changes are associated with the experimental condition of interest. Identifying these "key" features is typically the first step toward finding useful biomarkers or understanding the biological processes involved in the condition under investigation. A variety of approaches have been developed for these tasks, with the majority based on classical univariate statistical methods. MetaboAnalyst 4.0 supports three common feature selection approaches: (a) identifying variables that are significantly different among different conditions, (b) identifying variables that show particular patterns of change under different conditions, and (c) identifying variables that are significantly associated with other known biomarkers or features of interests. This protocol describes how to use MetaboAnalyst 4.0 to perform these three types of analyses.

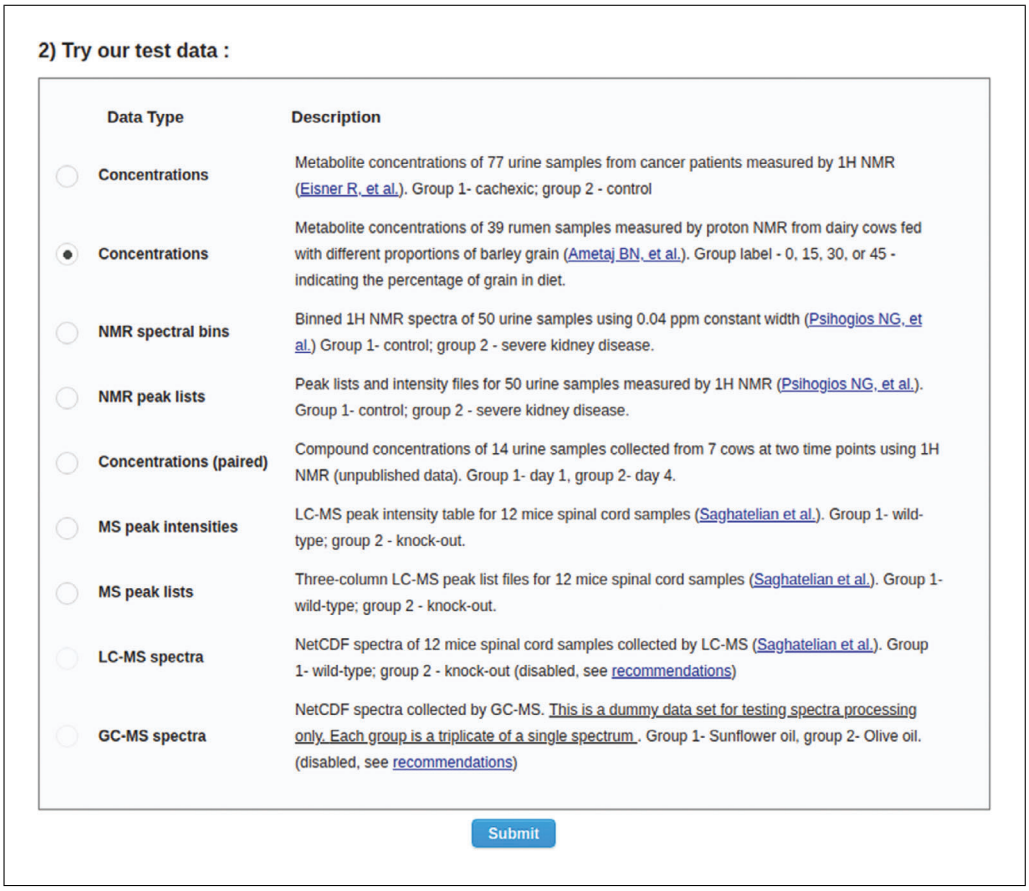


Figure 15 A screenshot showing example datasets available for users to explore various functions for the Statistical Analysis Module.

Necessary Resources

Hardware

A computer with internet access

Software

An up-to-date web browser, such as Google Chrome (*http://www.google.com/chrome*), Mozilla Firefox (*http://www.mozilla.com/*), Safari (*http://www.apple.com/safari/*), or Internet Explorer (*http://www.microsoft.com/ie/*). JavaScript must be enabled in the web browser.

Files

None

Data upload and processing

1. Go to the MetaboAnalyst 4.0 home page (*https://www.metaboanalyst.ca*). Click "click here to start" to enter the "Module Overview" page. Click the "Statistical Analysis" circular button to enter the corresponding "Data Upload" page. Scroll down the page to locate the "Try our test data" section on the bottom half of the page. Select the second concentration dataset (rumen samples) and click the "Submit" button at the bottom of the page (Fig. 15).

   Alternatively, users can follow the steps as described in Basic Protocol 1 (steps 1 to 4) to download the specific data set needed for this protocol and then upload it to MetaboAnalyst. To do this, go to the "Data Formats" page and download the data set labeled as "Compound concentration data—cow, four groups" in CSV format. On the

"Data Upload" page, make sure that the "Concentrations" radio button is checked under the "Tab delimited text (`.txt`) or comma separated values (`.csv`)" section. Make sure "samples in rows (unpaired)" is selected in the drop-down menu for this data format. Click on the "Choose File" button, browse through the computer's file manager to locate and select the downloaded data, and then click "Submit" (located on the right) to upload the data. This data set contains metabolite concentrations of 39 rumen samples measured by ^{1}H NMR from dairy cows fed with different proportions of barley grain (Ametaj et al., 2010). There are four groups—0, 15, 30, and 45—indicating the percentage of grain in the diet.

2. The "Data Integrity Check" page indicates that the data passes all the data integrity checks and no missing values have been detected. Click the "Skip" button to go to the "Data Normalization" page.

   Targeted metabolomics data sets are usually of high quality and require little or no pre-processing. The low-quality data-filtering step is not performed by default. However, users can click the "Data filter" hyperlink on the navigation tree to run this procedure if they wish.

3. The "Data Normalization" page should be displayed (see Fig. 9). In this case, select the "Normalization by a pooled average sample from group" and then click "Specify." A dialog box will appear to select a reference group; ensure that group "0" is selected from the drop-down menu. Leave the Data transformation set to "none" and select "Auto scaling" for Data scaling. Click the "Normalize" button at the bottom of the page to process the data.

   In this study, the group labeled as "0" served as the baseline control for the other groups. We can therefore use it to normalize other groups to remove any systematic bias. Alternatively, users can specify a representative sample (usually a typical sample from the control group with the smallest number of missing values) to perform the normalization via reference sample. Both methods are based on the concept of probabilistic quotient normalization (PQN; Dieterle et al., 2006).

4. Click the "View Result" button to visualize the graphical summary, which shows that the data look reasonably "bell-shaped" after our normalization procedure. Click the "Proceed" button at the bottom of the page. The "Data Analysis Overview" page should be shown (see Fig. 11). MetaboAnalyst contains 16 statistical methods carefully selected for their utility in routine metabolomic data analysis. These statistical methods are organized into five categories—Univariate Analysis, Chemometrics Analysis, Feature Identification, Cluster Analysis, and Classification and Feature Selection. Univariate statistical analysis and feature identification will be the focus of this protocol.

   Please note that some methods are automatically disabled such as T-tests, Fold Change Analysis, Volcano plots, and SVM analysis. These methods are designed only for two group analyses and are therefore inappropriate for this four-group data analysis.

Identification of variables significantly different among experimental conditions

5. Click the "One-way Analysis of Variance (ANOVA)" hyperlink near the top of the "Data Analysis Overview" page to proceed. The ANOVA results based on the default parameters are shown. The points highlighted in red are the significant compounds selected based on the default *p*-value threshold (0.05), which is marked by a dashed line. Click on one of the points to see a boxplot showing its concentration values in each group (Fig. 16).

   For each variable, ANOVA tests whether there are any statistically significant differences between the means of three or more groups. A positive result indicates that at least two of the groups differ significantly. The post-hoc analysis is used in conjunction with ANOVA to further test which means are significantly different from one another. MetaboAnalyst

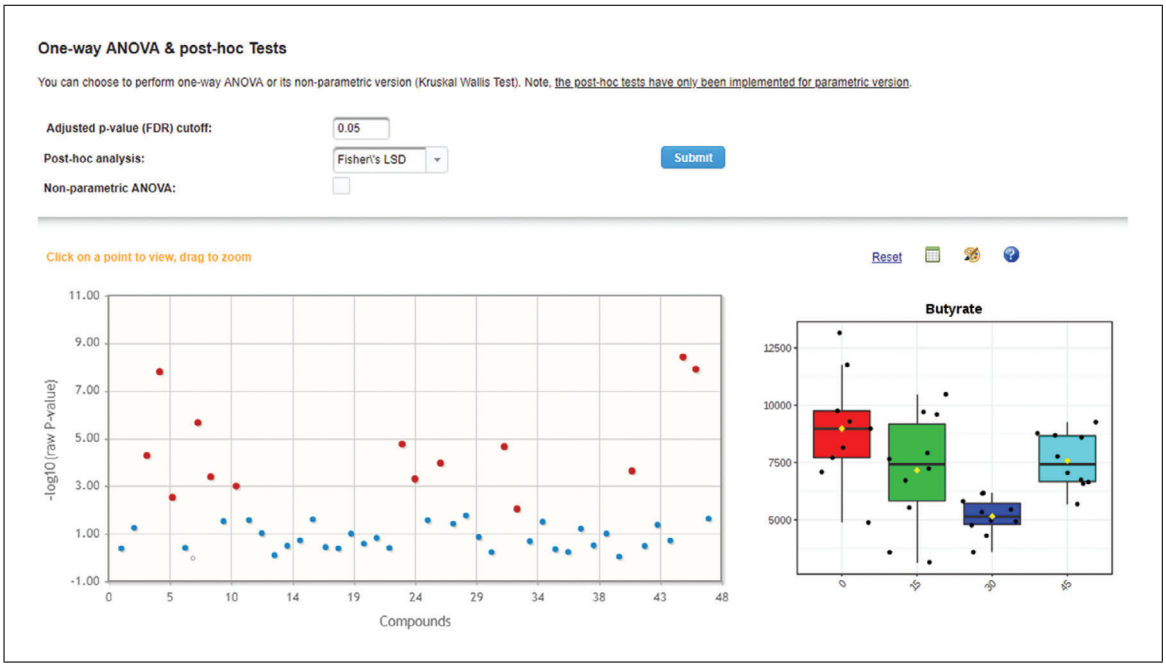


Figure 16 A screenshot showing the ANOVA results with default parameters. Clicking on any dot will create a boxplot of concentrations of that selected feature per group.

provides two commonly used post-hoc tests—Fisher's least significant difference (LSD, default) and Tukey's honestly significant difference (HSD).

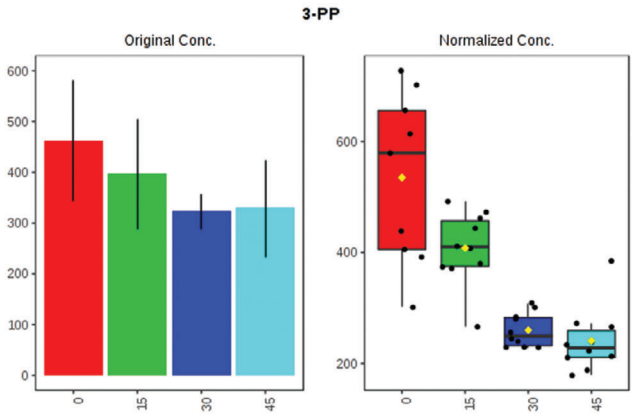
6. There are two small icons that appear near the top right corner of the ANOVA graph. Click on the icon that looks like a multi-column table. This allows one to see the "Feature Details View" of the ANOVA result table (Fig. 17). The first column shows the feature names. Click on a name, for instance *3-PP*, to view boxplot summaries using both original and normalized concentrations.

   The table shown in Fig. 17 gives numerical details from the ANOVA and post-hoc analysis. Sortable lists of p-values and the $-\log_{10}$ of the p-values are given along with the false discovery rate (FDR) adjusted p-values (based on Benjamini–Hochberg procedure). The FDR was calculated to adjust for multiple tests. The post-hoc results are reported in group pairs separated by semicolons to indicate that the corresponding groups are significantly different. For example, the post-hoc analysis for the compound 3-PP (3-phenylpropionate) shows "0—15; 0—30; 0—45; 15—30; 15—45," meaning that each group is significantly different from other groups except groups 30 and 45, based on the Fisher's LSD. Clicking "3-PP" on the leftmost column (under Name) will generate a corresponding bar graph and a box plot (Fig. 17), which shows a declining concentration of 3-PP with an increasing grain percentage in the diet. Note that the difference between group 30 and group 45 is very small and is not statistically significant.

7. Another very useful function for the identification of significantly different variables is the Significance Analysis of Microarray (SAM) approach, originally designed for microarray data analysis (Tusher, Tibshirani, & Chu, 2001). To use SAM, click the "SAM" hyperlink on the navigation tree (about midway down the menu) to enter the "SAM Analysis" page. At the top of the page, users can set the parameters for SAM analysis, and underneath is the SAM plot (Fig. 18). The current parameters are set to a default value of 0.6. Click the "View details" hyperlink next to the delta value and a dialog box will appear that shows two plots which allow users to decide a proper delta value to control FDR and the total number of significant features identified. For instance, increase the delta value to 1.0 in the box marked "Delta value (FDR control)," press the "Submit" button, and then click the "View details" link. The FDR (on the Y axis) will be close to zero (left plot) when the Delta (on

Feature Details View

Click the feature name to view a graphical summary. The bar plots on the left show the original values (mean +/- SD). The box and whisker plots on the right summarize the normalized values. Note, positive infinite numbers are represented as 999999, and negative infinite numbers -999999.



You can download the table: here (Right click and "Save Link as....")

Name	f.value	p.value	-log10(p)	FDR	Post-hoc tests
Endotoxin	26.598	3.8169E-9	8.4183	1.7939E-7	30 - 0; 45 - 0; 30 - 15; 45 - 15
Glucose	24.076	1.2409E-8	7.9063	2.4798E-7	45 - 0; 45 - 15; 45 - 30
3-PP	23.576	1.5828E-8	7.8006	2.4798E-7	0 - 15; 0 - 30; 0 - 45; 15 - 30; 15 - 45
Alanine	14.811	2.1736E-6	5.6628	2.554E-5	30 - 0; 45 - 0; 30 - 15; 45 - 15

Figure 17 A screenshot showing the detailed results of the ANOVA and post-hoc analysis. The columns with arrow icons can be sorted by clicking the corresponding icon. Clicking a compound name in the first column will generate a boxplot summary of its concentrations in different groups.

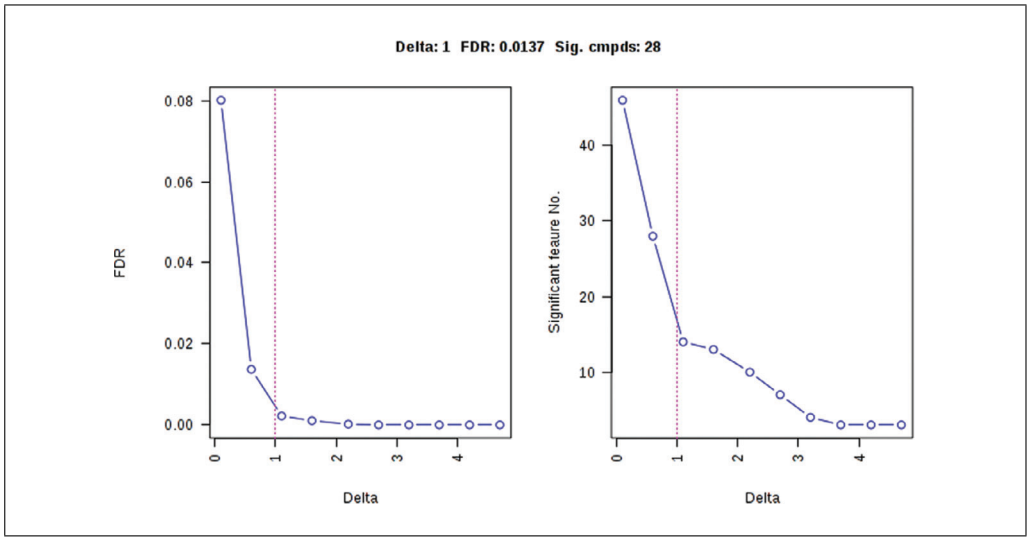


Figure 18 A screenshot showing the parameters for SAM analysis and the Delta plots. The left plot shows Delta versus the false discovery rate (FDR) and the right plot shows Delta versus the number of identified compounds.

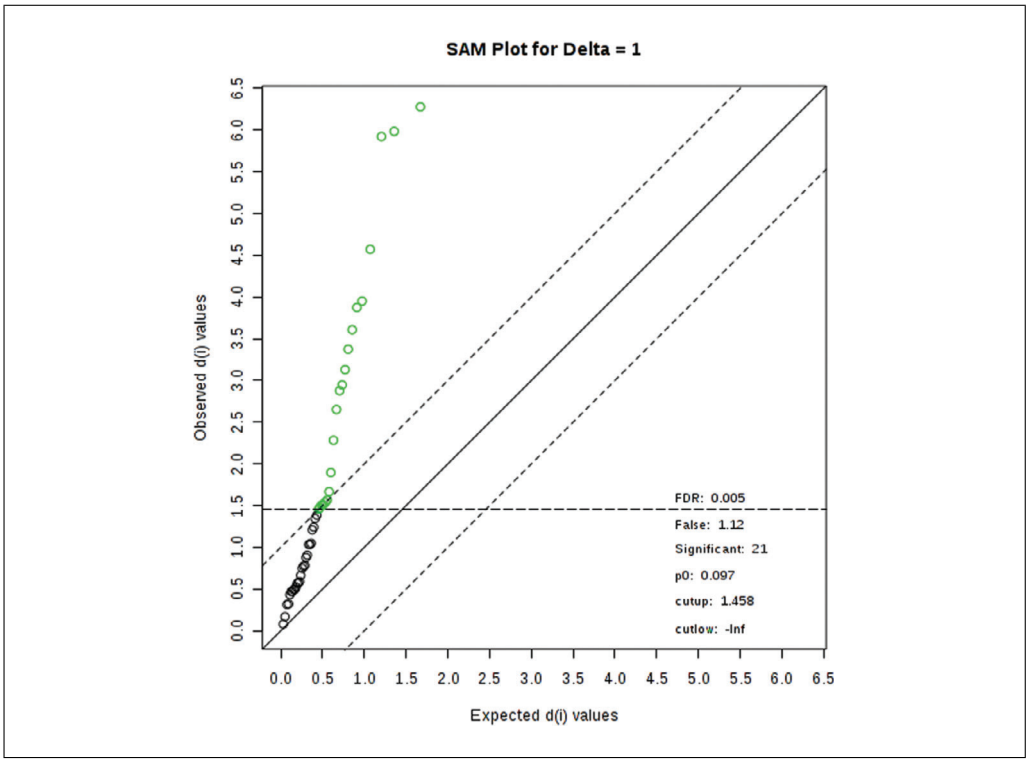


Figure 19 A screenshot showing the SAM result with a Delta value of 1.0. The SAM plot is a scatter plot of the observed relative difference versus the expected relative difference estimated by data permutation. The solid diagonal line indicates where these two measures are the same. The dotted lines are drawn at a distance of Delta from the solid line. The significant variables are highlighted in green.

the X axis) is set to 1.0 (Fig. 19). Looking at the right plot, which shows the number of significant features (on the Y axis) and the corresponding Delta (on the X axis), this choice of delta permits the identification of ~15 significant compounds (right plot). Close the dialog box and return to the SAM plot (Fig. 19). It indicates that with the given delta (1.0), 21 significant compounds were identified with an FDR 0.05 or one (1.12) false positive. There are two small icons near the top right corner of the SAM plot. Click on the table icon to access the "Feature Details View" on the ranked list of important compounds identified by SAM, which are almost identical to those identified by ANOVA.

SAM was designed to address the issue that in high-dimensional data analysis, the estimate of the variance tends to be unstable when the sample size is small. For a small sample size (3 to ~8 per group), the use of SAM or EBAM [Empirical Bayesian Analysis of Microarrays (Efron, Tibshirani, Storey, & Tusher, 2001), also available in MetaboAnalyst] is advised. When more samples are available (>10 per group), the variance estimate is fairly stable, and one can reliably use standard t-tests or ANOVA methods.

Identification of variables showing a particular pattern of change

8. Now click the "PatternHunter" hyperlink on the navigation tree (near the middle of the list) to enter the page for PatternHunter analysis. In this case, we are interested in compounds with concentration changes that follow a specific pattern. This task can be accomplished using the pattern match function based on a predefined pattern or a user-defined pattern.

MetaboAnalyst's pattern-matching method was implemented based on the template matching algorithm originally designed to select genes showing strain- or region-dependent patterns of expression (Pavlidis & Noble, 2001). It can be adapted to search

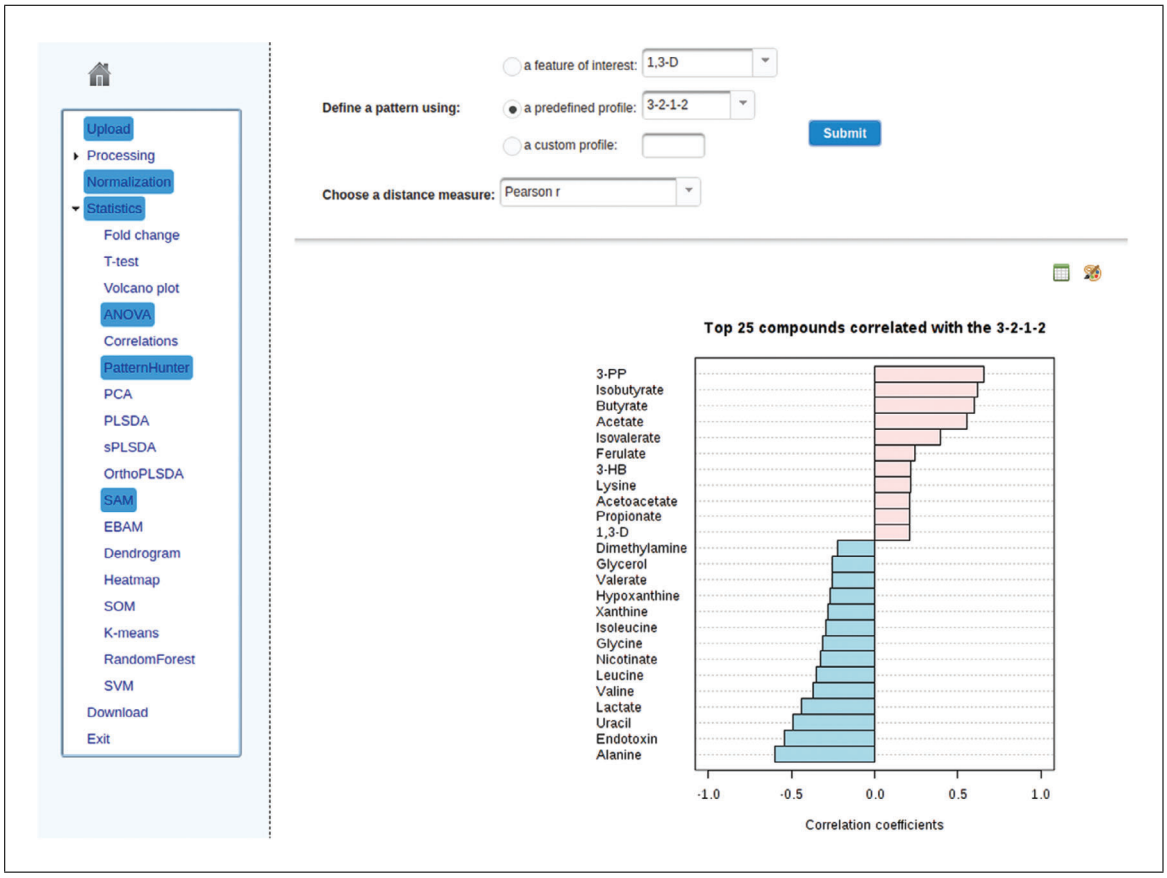


Figure 20 A screenshot showing the results from a pattern search using the given pattern "3-2-1-2" based on the *Pearson r* distance measure (light pink: positive correlation; light blue: negative correlation).

for very complex patterns within the data. The pattern is specified as a series of numbers separated by "-" with each number corresponding to the expected change in the corresponding group. For example, a pattern of "1-2-3-4" can be used to search for variables with linearly increasing values across the four corresponding groups.

9. To search for compounds exhibiting a concentration decrease in the first three groups (0%, 15%, 30% grain), followed by an increase in the last group (45% grain), we can use a pre-defined pattern from MetaboAnalyst. There are three radio buttons located beside the text "Define a pattern using." Select the radio button titled "a predefined profile" and then pattern "3-2-1-2" from the drop-down list. Keep all other values as default and click the "Submit" button. The result is shown in Figure 20. The image shows both positively correlated (in light pink) and negatively correlated (in light blue) compounds displayed in a bar graph. Note that the compound *3-PP* (3-phenylpropionate) is ranked as the most positively correlated with the pattern. However, based on the previous result (Fig. 17), it is known that 3-PP concentration decreases across the four groups. The mean concentration in group 45 is somewhat lower than in group 30, which does not match the last part of the given pattern. This is because the concentration changes of 3-PP in the first three groups match very closely with the initial part of the pattern (3-2-1), despite the mismatch in the last part. To identify compounds with concentrations that increase in the last group, it is necessary to adjust some parameters to obtain the desired result. There are two ways to achieve this purpose, as discussed in the next two steps.

10. The previous result was based on the default distance measure "Pearson *r*" (the Pearson correlation coefficient) which calculates the strength of linear dependence between the two variables. It is sensitive to outliers, unequal variances, and

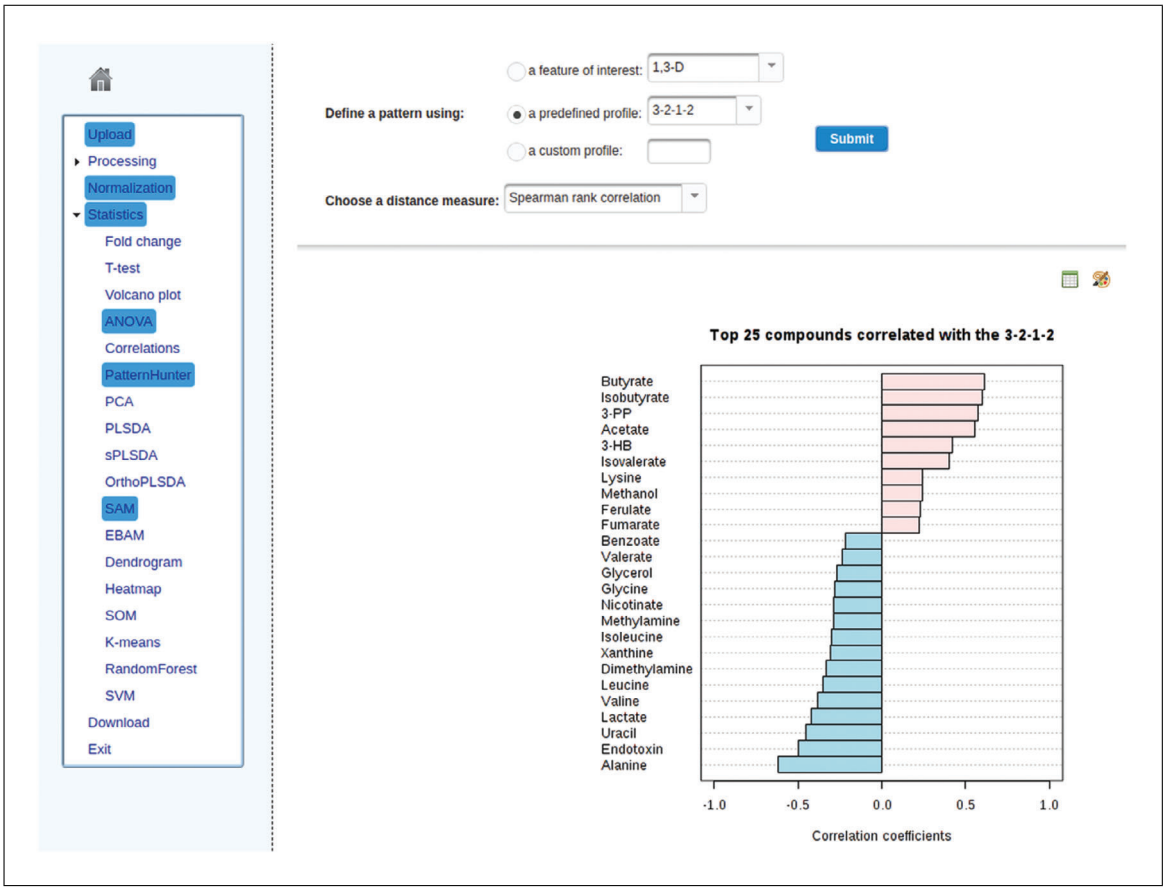
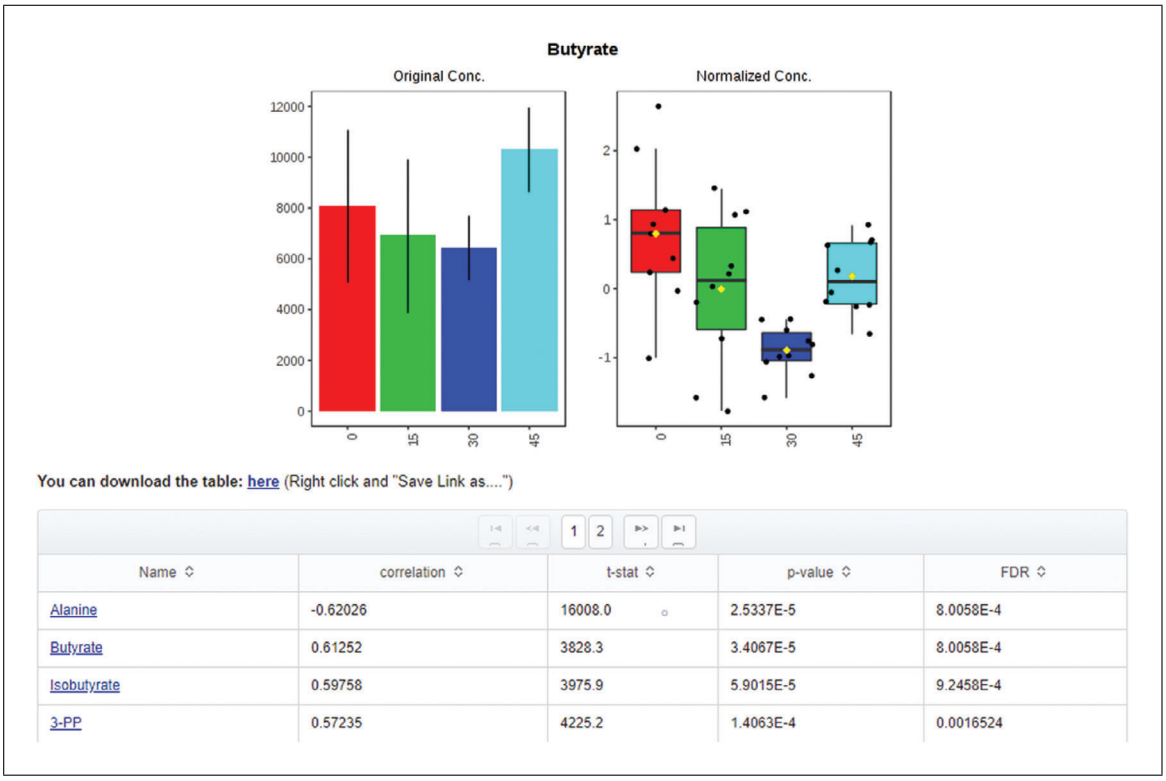


Figure 21 A screenshot showing compounds that match the given pattern "3-2-1-2" using the *Spearman rank correlation* as a distance measure.

non-normality. For those more interested in the direction of the change rather than the amplitude of the change, Spearman rank correlation should be used. Spearman rank correlation is similar to the Pearson correlation coefficient, but the correlation is calculated between the ranks rather than actual values of the variables. To use the Spearman test, go to the text marked "Choose a distance measure" and select the "Spearman rank correlation" from the drop-down menu. Choose the predefined pattern "3-2-1-2" and then click the "Submit" button right next to the option. The result is shown in Figure 21. In this case, the top-ranked compound is *Butyrate*. Click on the table icon near the top right corner of the plot to access the "Feature Details View" on the ranked list of important compounds identified by the PatternHunter Analysis (Fig. 22). Click "Butyrate" to view a bar graph and a box plot summary of its concentration change. Indeed, its concentration decreases in the first three groups and increases in the last group.

Figure 22 gives the detailed results from the PatternHunter analysis. The first column lists the names of the compounds and the second column lists the correlation coefficients that indicate the strength of the match. The third column shows the p-values calculated using a t-test on the correlation coefficients. The p-values in the fourth column indicate the significance of the match. Note that the table was initially ordered by p-values. To change the order in any column, click on the up/down arrow icons at the top of each column header.

11. Users can also specify their own patterns to give more weight to the last part of the pattern. Click on the PatternHunter hyperlink on the left menu to return to the PatternHunter page. With the distance measure set to "Pearson *r*," select the radio button marked as "a custom profile" and enter the pattern "3-2-1-**3**" (note that the



Name	correlation	t-stat	p-value	FDR
Alanine	-0.62026	16008.0	2.5337E-5	8.0058E-4
Butyrate	0.61252	3828.3	3.4067E-5	8.0058E-4
Isobutyrate	0.59758	3975.9	5.9015E-5	9.2458E-4
3-PP	0.57235	4225.2	1.4063E-4	0.0016524

Figure 22 A screenshot showing the details from a correlation analysis. A box plot of *Butyrate* concentration is shown on top and was generated by clicking *Butyrate* in the Name column of the table.

expected change of the last group is increased by one), then click the "Submit" button right next to the option. The result shows that *Butyrate* is indeed ranked at the top of the list.

Using a custom profile is a very flexible approach that allows users to search for any arbitrary profile within the data. For example, one can enter "2-4-8-16" as the specified profile to search for compounds with concentrations that are exponentially increasing (with base 2) rather than linearly. In this case, make sure the distance measure selected is "Pearson r," not the "Spearman rank correlation." This is because the "1-2-3-4" and "2-4-8-16" have the same ranks.

Identification of variables that significantly associate with a known biomarker

12. Based on a review of the literature, it is known that elevated levels of *Endotoxin* are associated with certain inflammatory responses. *Endotoxin* is also top-ranked by the ANOVA test used in this protocol (step 6, Fig. 17). To identify other metabolites that are significantly associated with this compound, click the "PatternHunter" hyperlink on the navigation tree (left side) to return to the page for PatternHunter analysis.

13. To identify compounds that follow the same diet-dependent trends as seen with *Endotoxin* (the direction, but not the amplitude of the change is the main concern), go to the "Choose a Distance Measure" option and select "Spearman rank correlation" from the drop-down menu. Click on the radio button labeled "a feature of interest," scroll down the list of features on the drop-down menu, and select "Endotoxin," which is near the bottom of the list. Click the "Submit" button. As shown in Figure 23, *Alanine* is most positively correlated with *Endotoxin* levels, while *3-PP* is most negatively correlated with *Endotoxin* levels.

Data download

14. Click the "Download" hyperlink on the navigation tree. The download page (Fig. 24) displays a table showing all the figures, result tables, as well as the R

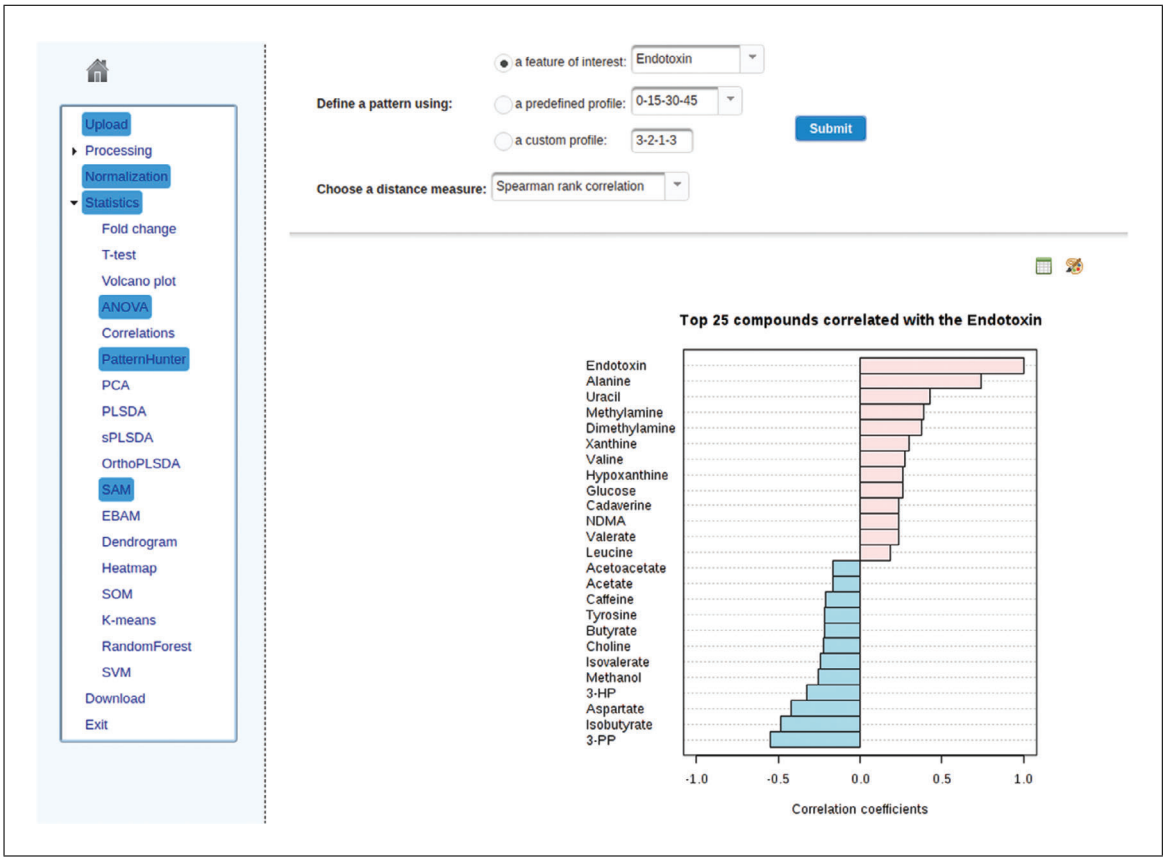


Figure 23 A screenshot showing compounds that are associated with *Endotoxin* using *Spearman rank correlation* as a distance measure.

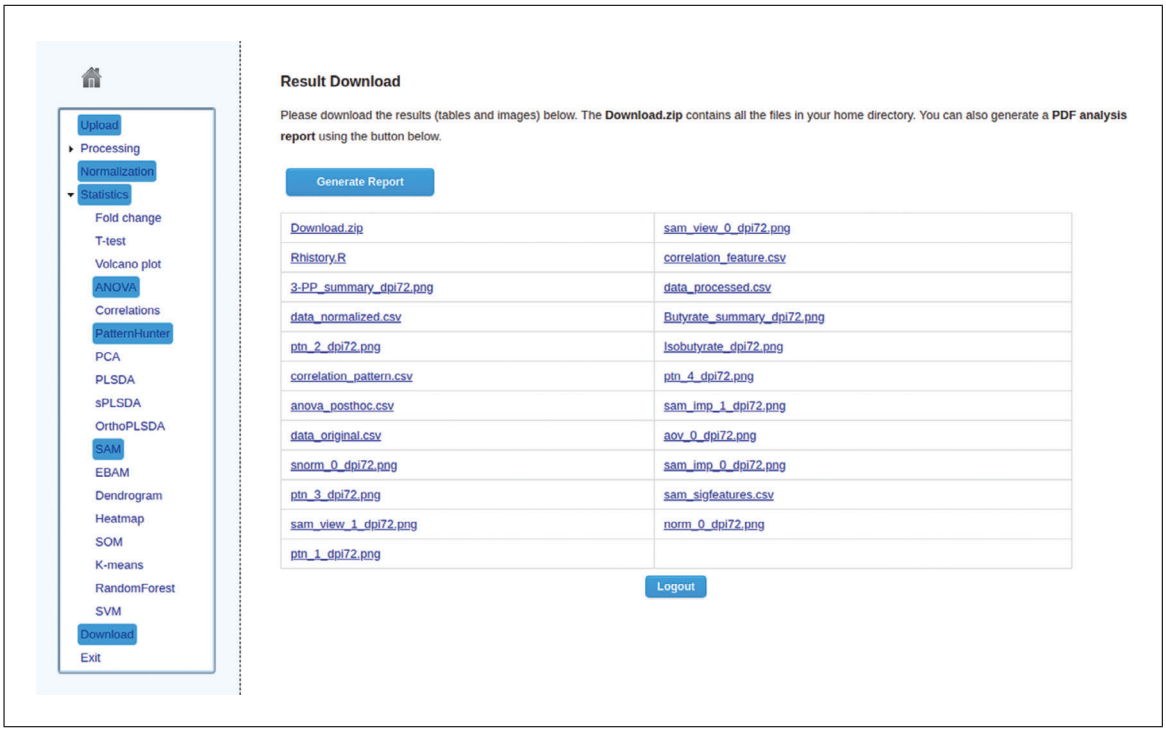


Figure 24 A screenshot showing the “Result Download” page. The files were produced during the analysis and can be downloaded in a single zip file (`Download.zip`). Users need to click the “Generate Report” button to create the PDF report and download it separately.

Command History file generated during the analysis session. Click the "Generate Report" button to create a PDF report that describes all steps performed with the necessary background information, followed by the results in the form of bulleted lists, tables, and images as appropriate. At this moment, download the analysis report and the zip file, and click "Exit" to complete this session.

The three files, `anova_posthoc.csv`, `correlation_pattern.csv`, *and* `correlation_feature.csv`, *contain all the features selected by the ANOVA test, the pattern match analysis, and the feature match analysis, respectively. Users can open these files using a spreadsheet program to view all the numerical details.*

BASIC PROTOCOL 3

MULTIVARIATE EXPLORATORY DATA ANALYSIS

Univariate methods (such as *t*-tests and ANOVA) are simple to use, and the results are usually easy to understand. They are widely used in metabolomics studies for selecting important features from metabolomic data. However, univariate approaches are often considered suboptimal, as they ignore correlations that are known to be present among variables (i.e., peaks or metabolites). Multivariate methods, which simultaneously take all variables into consideration, are generally considered more suitable for high-dimensional "omics" data analysis. This protocol will give detailed instructions on how to use several multivariate approaches implemented in MetaboAnalyst for comprehensive metabolomic data analysis. They include two unsupervised methods, PCA and hierarchical clustering with heatmap, and two supervised methods, partial least squares discriminant analysis (PLS-DA) and random forests classification.

Necessary Resources

Hardware

A computer with internet access

Software

An up-to-date web browser, such as Google Chrome (*http://www.google.com/chrome*), Mozilla Firefox (*http://www.mozilla.com/*), Safari (*http://www.apple.com/safari/*), or Internet Explorer (*http://www.microsoft.com/ie/*). JavaScript must be enabled in the web browser.

Files

None

Data upload and processing

1. This protocol will use the same data set as the previous protocol. Please follow the steps (1 to 4) described in Basic Protocol 2 to perform the data upload, processing, and normalization. Once completed, make sure to navigate to the "Data Analysis Overview" page.

Data overview and pattern discovery with PCA

2. Click the "principal component analysis" hyperlink on the main page or the "PCA" hyperlink on the left navigation tree. After a few seconds, the PCA results should be presented in a multi-panel page. The default panel shows a pair-wise scores plot between the first five principal components (PCs). The variance explained by each PC is shown on the corresponding diagonal cell.

PCA is an unsupervised clustering or dimensionality-reduction method which projects the data into a new coordinate system such that most of the data variance lies in the first few principal components or PCs. A simplified analogy of PCA can be made as follows: consider a 3-dimensional object such as a donut. Using PCA, one would like to generate a set of 2D projections that could explain what the 3D shape of the donut

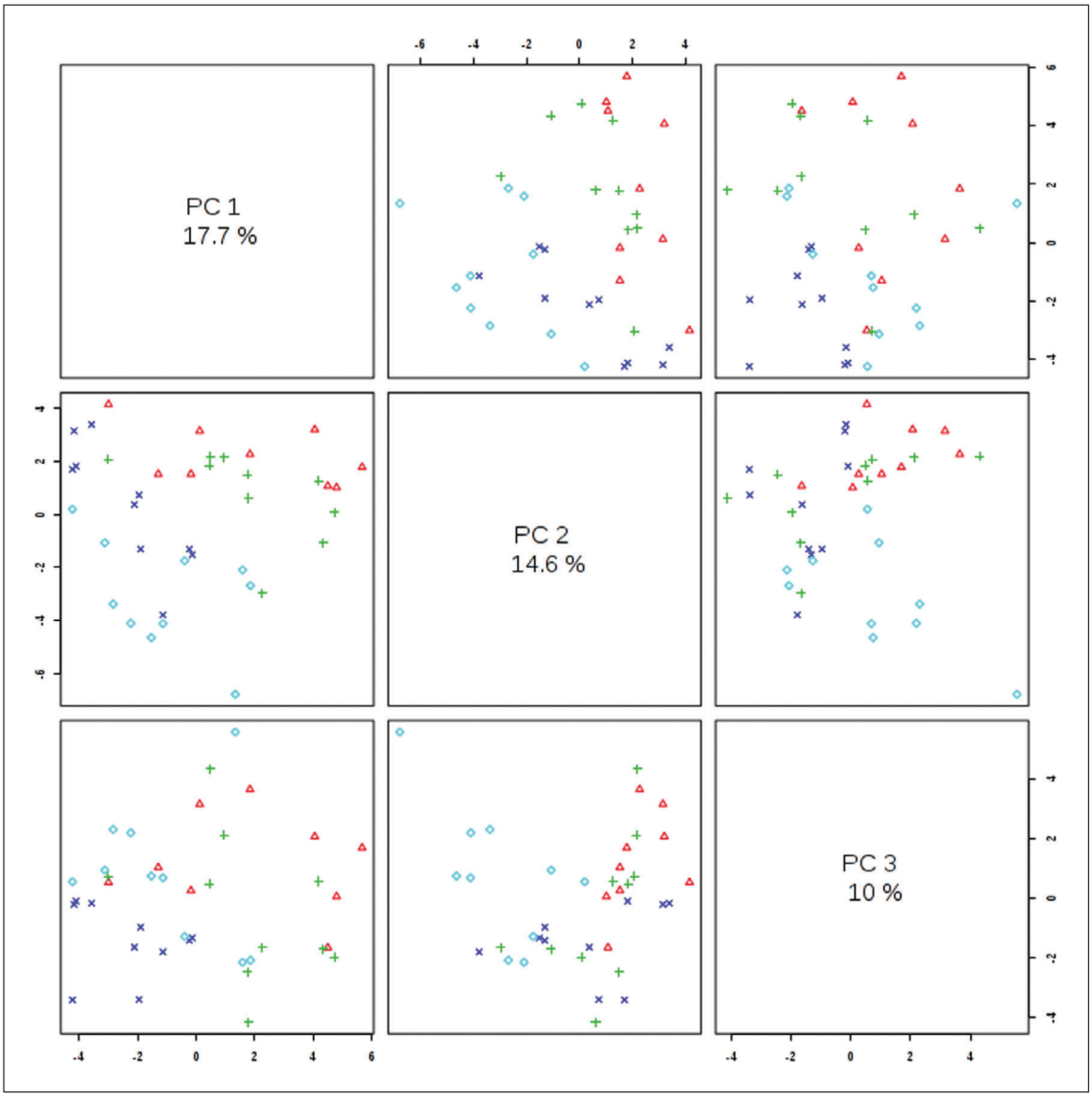


Figure 25 A screenshot showing the pair-wise score plots between the top three principal components (PCs). Their explained variances are plotted in the corresponding diagonal cells.

is. Obviously, the first most informative 2D projection is the one showing the donut's characteristic "O" shape. The second most informative 2D projection would be the projection of the donut lying on its side (looking like a hotdog bun). Other projections could be possible (the donut at a 45^o angle, say), but the combination of the first two orthogonal projections would be the most informative way of describing the donut. While this example shows how a 3D object can be reduced to a series of 2D projections, the strength of PCA is that it can do the same with any high-dimensional object just as easily. As a dimensionality-reduction technique, PCA is particularly appealing, as it allows one to visually detect sample patterns or groupings. In practice, PCA is most commonly used to identify how one sample is different from another, and which variables contribute most to this difference. The results of PCA are usually discussed in terms of scores and loadings. The scores represent the original data in the new coordinate system, and the loadings are the weights applied to the original data during the projection process.

3. As the first three PCs account for similar levels of variances in the data, they will be the focus of the remainder of the analysis. Change the number of displayed PCs from 5 to 3, and then click the "Update" button. The result is shown in Figure 25. These data points are drawn in different colors and shapes based on their group memberships (users can click the "2D Scores Plot" tab at the top of the page to

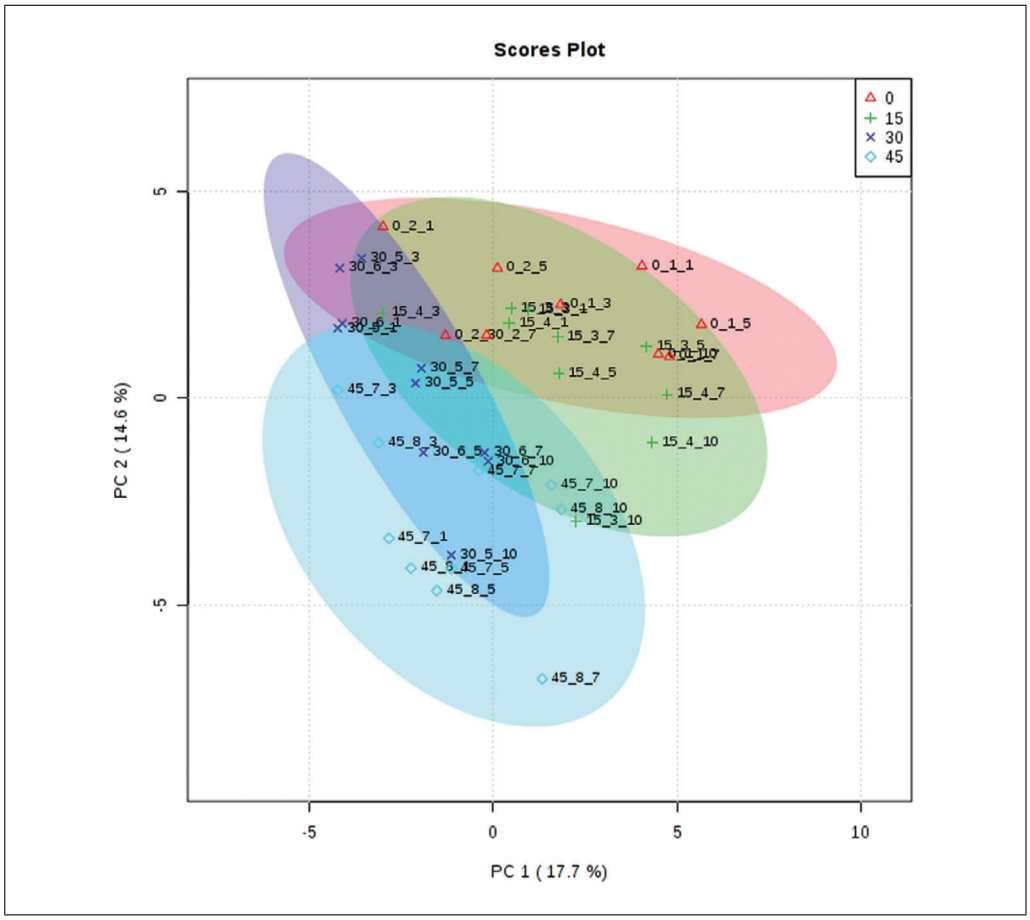


Figure 26 A screenshot showing a PCA score plot of the first two principal components. The shaded areas indicate the 95% confidence ellipse regions based on the data points for individual groups.

see the legend). In Figure 25, users should inspect the PCA plots for any clusters or patterns of interest such as outliers or a clear separation of groups. Using PCA for outlier detection is covered in the Basic Protocol 1 (step 14). In this case, no obvious outlier is found. On the PC1 direction (along the Y-axis of the figures on the first row or along the X-axis of the figures on the first column), there is clearly good separation between green (group 15) and dark blue (group 30) samples. On the PC2 direction (along the Y-axis of the figures on the second row or along the X-axis of the figures on the second column), there is clear separation between red (group 0) and cyan (group 45). In addition, when PC1 and PC2 are combined, there is a good separation between red (group 0) and blue (group 30) samples along the diagonal line. No additional separation is identified along the PC3 direction. The group separation is provided mainly by PC1 and PC2. We can use the "2D Scores Plot" to get a detailed view of the clustering trends.

4. Click the "**2D Score Plot**" tab to get a more detailed score plot. The default is PC1-PC2 (Fig. 26). The main direction of separation among groups 0, 15, 30, and 45 is evident from this plot. Group 0 and 45 are well separated, while group 30 overlaps significantly with both group 15 and group 45. We can further identify the influential compounds that contribute to the separation pattern.

MetaboAnalyst permits flexible exploration or visualization of the score plots between any set of PCs. If, for instance, one discovered in the previous step that PC1 and PC3 offered better separation, a user could enter 3 in the "Specify PC on Y-axis" to obtain the corresponding score plot.

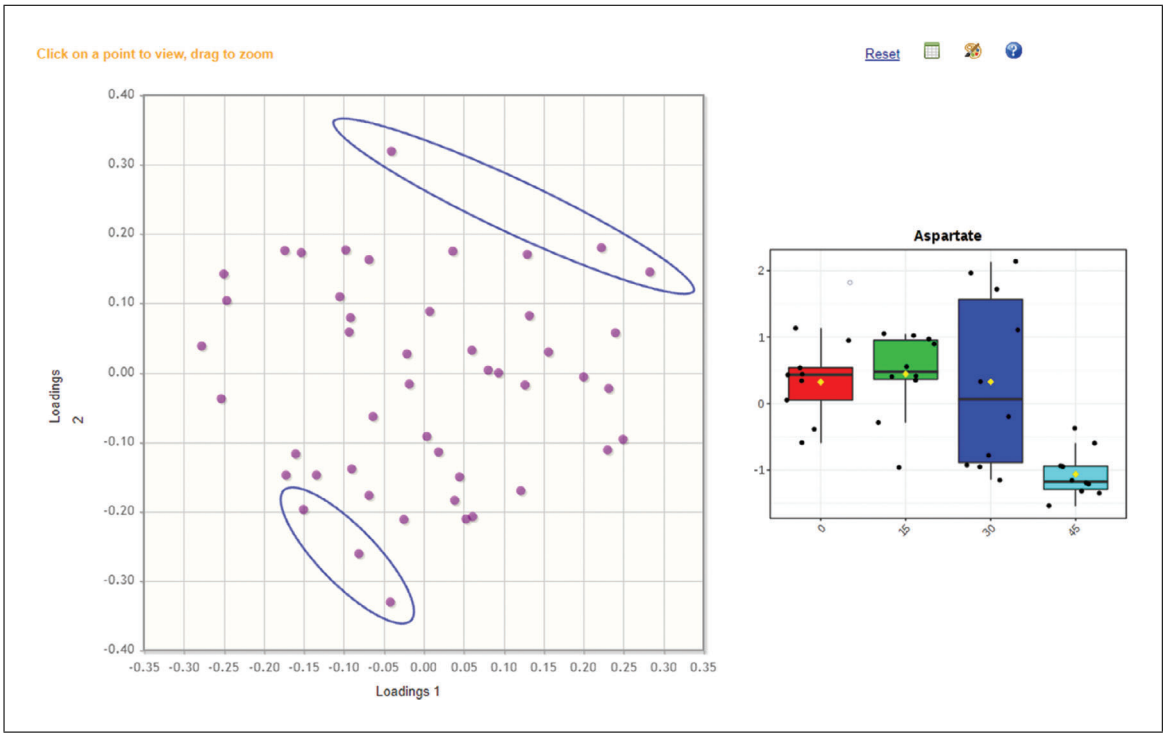


Figure 27 A screenshot showing a PCA loading plot of the first two principal components. Clicking an individual data point will display the corresponding compound as a box plot in to the right. The blue circles on the loading plot indicate the most influential data points located on the outermost areas along the direction of separation as identified in the corresponding score plot.

5. Click the "**Loadings Plot**" tab to view the PCA loadings plot. The default view shows the loadings for PC1 and PC2 (Fig. 27). Users can directly click on each point (compound) to visualize the corresponding box plot to the right of the plot. We can identify the most influential compounds (as shown in the two ellipses) by clicking on those outermost data points along the main direction of separation (top right to bottom left) such as *Aspartate, Isobutyrate*, and *3-PP*, located on the top right, and *Endotoxin, Glucose*, and *Methyalmine*, located on the bottom left.

6. Click the "**Synchronized 3D Plots**" to further explore the PCA results in an interactive 3D score (on the left) and loading plot (on the right) of the first three PCs. Use your mouse/touchpad to rotate and zoom in and out of each plot (Fig. 28). Notice that the rotations are synchronized between the two 3D plots. Click on any feature in the loading plot to view a boxplot comparing the concentrations of the selected feature between the four groups.

Data overview and pattern discovery with heatmap and hierarchical clustering

7. A heatmap provides another visually intuitive overview of large "omic" data sets. It can be visualized with or without hierarchical clustering. Using a heatmap for outlier detection was described in Basic Protocol 1 (step 15). To generate a heatmap, click the "Heatmap" hyperlink on the navigation tree. By default, MetaboAnalyst performs hierarchical clustering on both samples (columns) and variables (rows) to generate a heatmap with dendrograms drawn on both sides. In this case, the goal is to investigate the patterns of compound concentration that change across different groups, not clusters identified or organized by the dendrograms. Check the "Do not organize" check box and make sure that the "Samples" option is selected. Keep all other options as their default values, and click the "Submit" button. The result is shown in Figure 29. Several interesting patterns are evident. For example, as

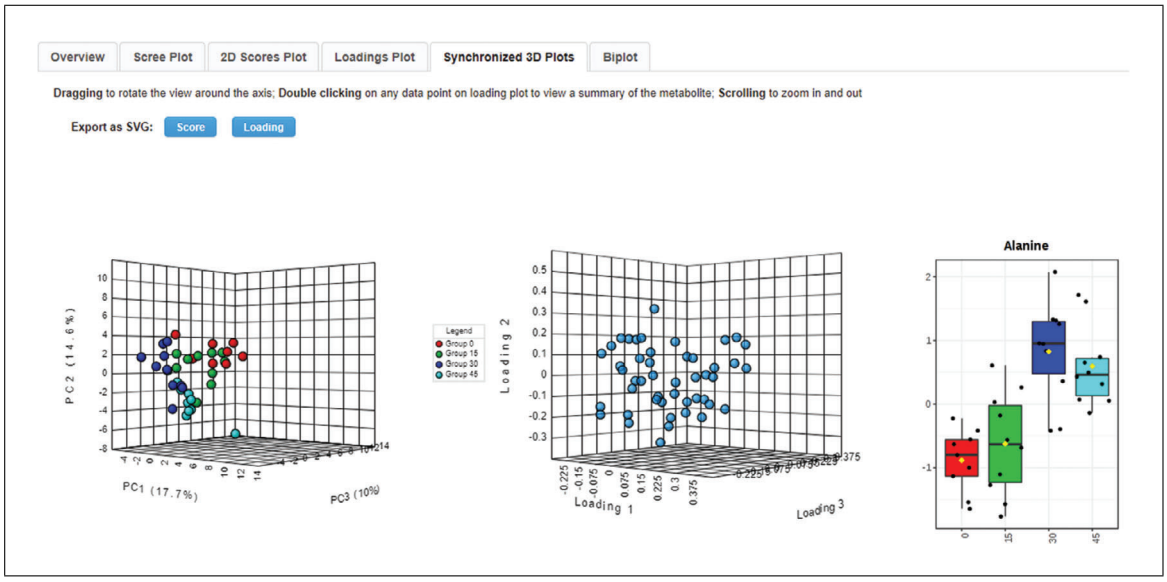


Figure 28 A screenshot of a 3D principal component analysis (PCA) score and loading plot highlighting *Alanine* with a boxplot.

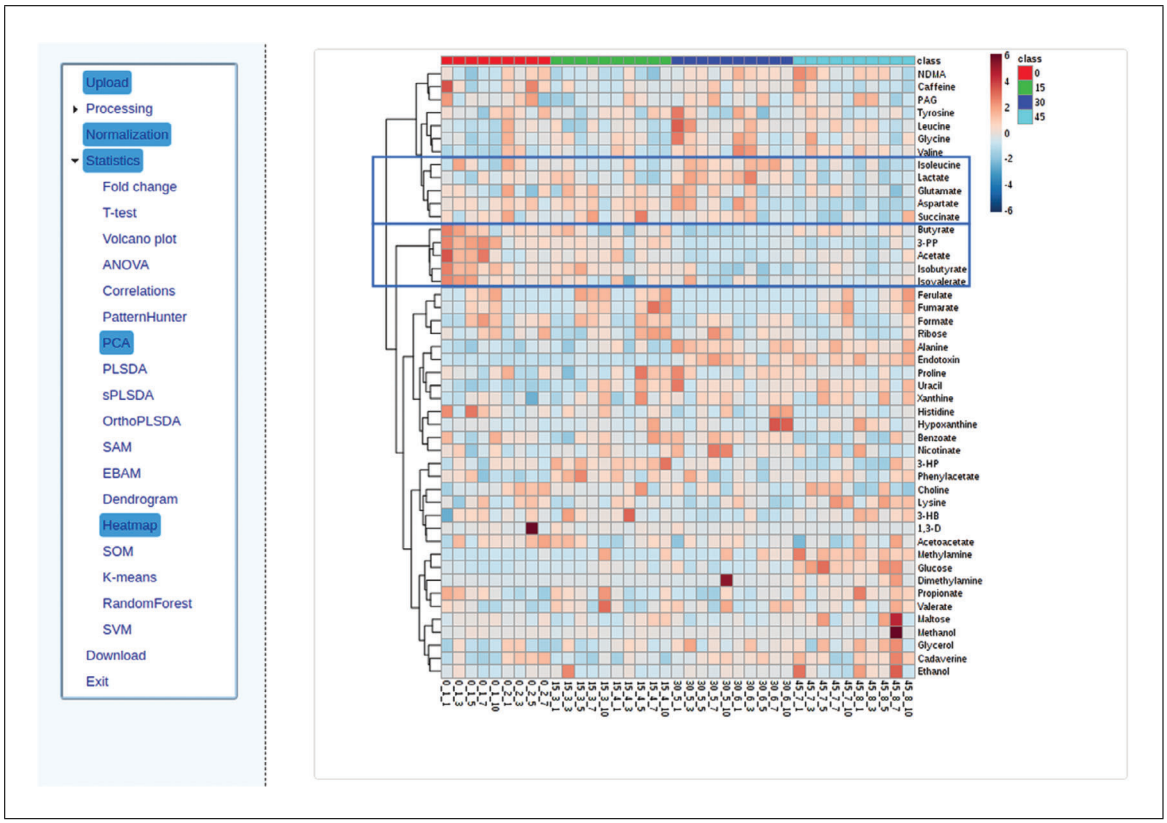


Figure 29 A screenshot showing a heatmap overview with the default color contrast. The hierarchical clustering was performed only on the compounds, while samples are ordered based on their group labels. The two blue boxes indicate the two regions with interesting concentration variations. See text for more details.

shown in the two boxes, there is a group of compounds (*Butyrate*, *3-PP, Acetate*, *Isobutyrate* and *Isovalerate*) that show a decreasing concentration trend from group 0 to group 45, and another group of compounds (*Isoleucine, Lactate*, *Aspartate*, and *Glutamate*) that only show decreased concentration in group 45. Note that many of these compounds are almost identical to the compounds we identified by PCA in the previous step.

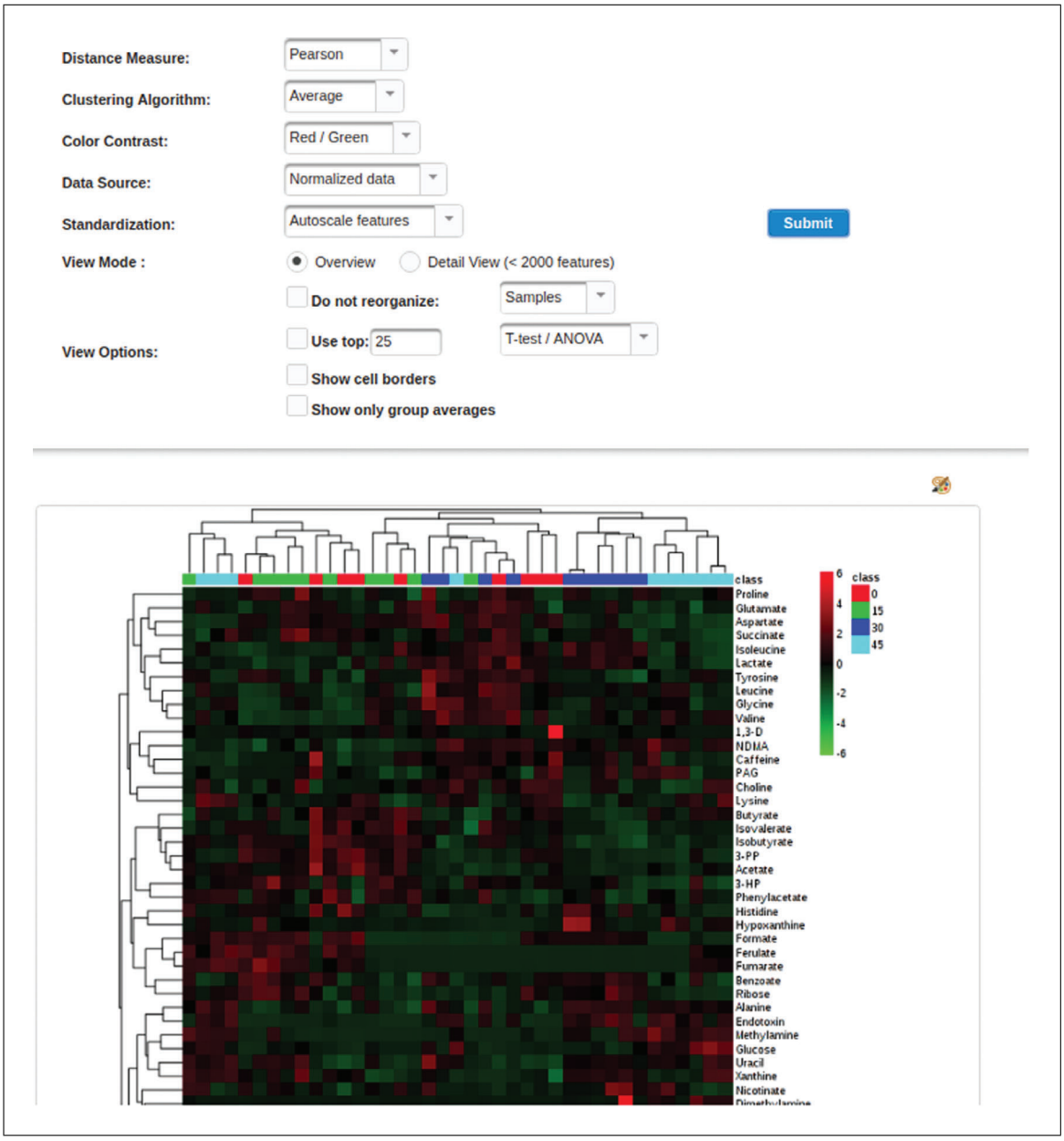


Figure 30 A screenshot showing a heatmap overview with hierarchical clustering. The color contrast was set to "Red/Green," and the clustering algorithm was set to "Average."

Groups are ordered numerically or alphabetically. Users are advised to choose proper group labels to control their order shown in the heatmap. We also suggest that users abbreviate long sample or variable names to avoid them being truncated in the heatmap. For larger datasets, some names may become too crowded to read; users can regenerate the image using the "Detail View" option.

8. MetaboAnalyst provides several options for customizing both the heatmap and the hierarchical clustering. For instance, under "Distance Measure," users can specify different distance measures (Euclidean, Pearson, and Minkowski), different clustering algorithms (Average, Single, Complete and Ward), and different color contrasts (Default, Green/Red, Topo colors and Heat colors). For example, select "Pearson" for the distance measure, select "Average" (linkage) for the clustering algorithm, select "Red/Green" for the color contrast, uncheck the "do not reorganize" check box, uncheck the "Show cell borders," and then click the "Submit" button. The result is shown in Figure 30. It can be seen from this image that the groups formed via hierarchical clustering do not agree well with the true biological groups.

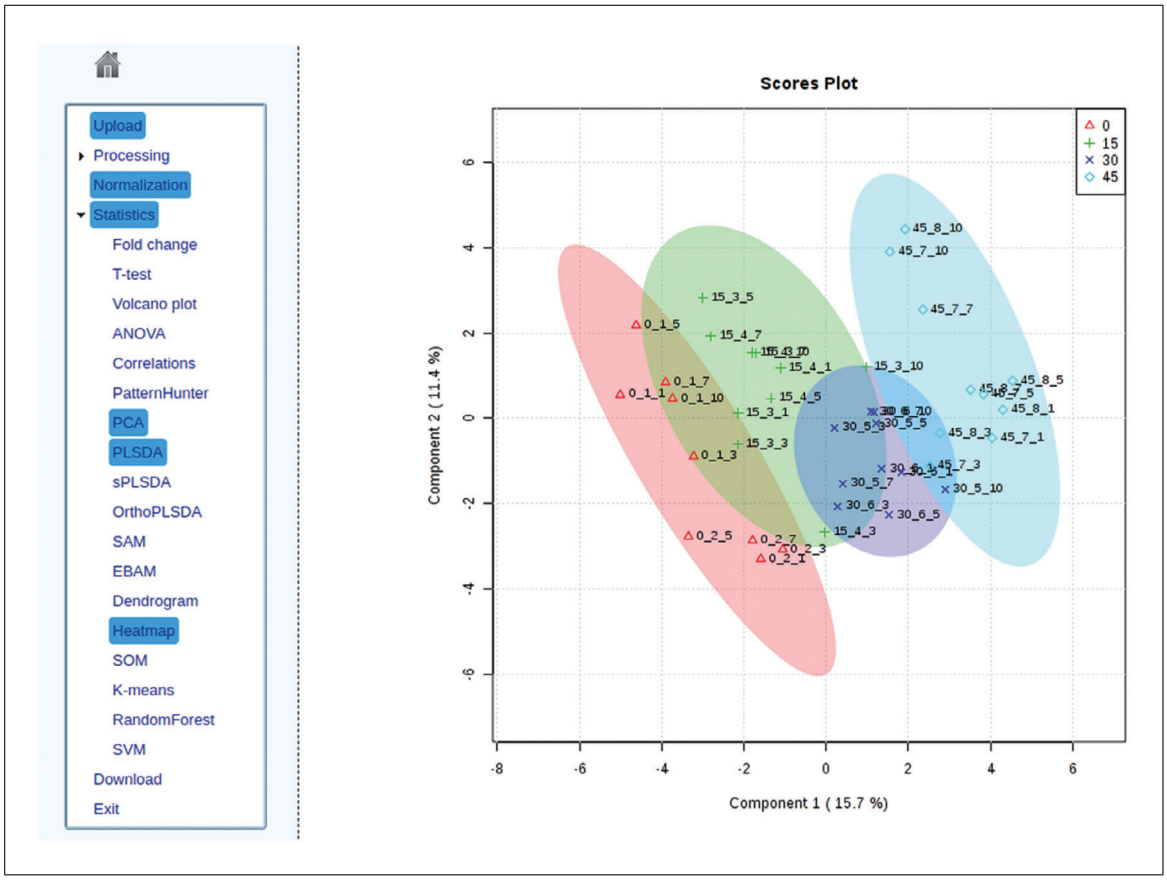


Figure 31 A screenshot showing a PLS-DA score plot of the first two components. The shaded areas indicate the 95% confidence ellipse regions based on the data points for individual groups.

Users should be aware that hierarchical clustering uses some very simple statistical measures to calculate the distance between different biological samples. Only in rare cases will it match the real biological distance under study. In most cases, hierarchical clustering dendrograms tend to distract users from finding patterns across different biological groups. Using a heatmap alone as a visual guide is generally a more useful method as seen in this example. MetaboAnalyst also supports heatmap visualization of a subset of compounds selected based on t-tests/ANOVA, correlation analysis, or PLS-DA VIP scores. An example will be given in step 15 of this protocol.

Data analysis using PLS-DA

9. To perform PLS-DA analysis, click the "PLSDA" hyperlink on the navigation tree. Wait for ~10 sec for MetaboAnalyst to finish its default analysis. Like PCA, the results are presented in a multi-panel page with the pair-wise score plots of the first five components shown as a default. Click the "**2D Scores Plot**" tab at the top of the page to view the scores plot between the first two components (Fig. 31). A much better separation is obtained compared to PCA (Fig. 26).

Users should be aware that the separation in PLS-DA is calculated by maximizing the covariance between the data matrix (X) and the class labels (Y). By default, the program will first convert the class labels into rankings based on their numerical or alphabetical order (i.e., group labels "A, B, C, D" will be 1, 2, 3, 4; while group labels "low, medium, high" will be 2, 3, 1), and then perform PLS regression between the data matrix and numerical Y. For two-group data, this procedure will not affect the visualization pattern, as it will always be between 1 versus 2. For multi-groups, this default approach is meaningful when the group labels correspond to time series, disease severity, or treatment dose (as in this example). However, when group labels do not reflect quantitative differences, users should uncheck the option "Class order matters" (located on the top of the page). In this

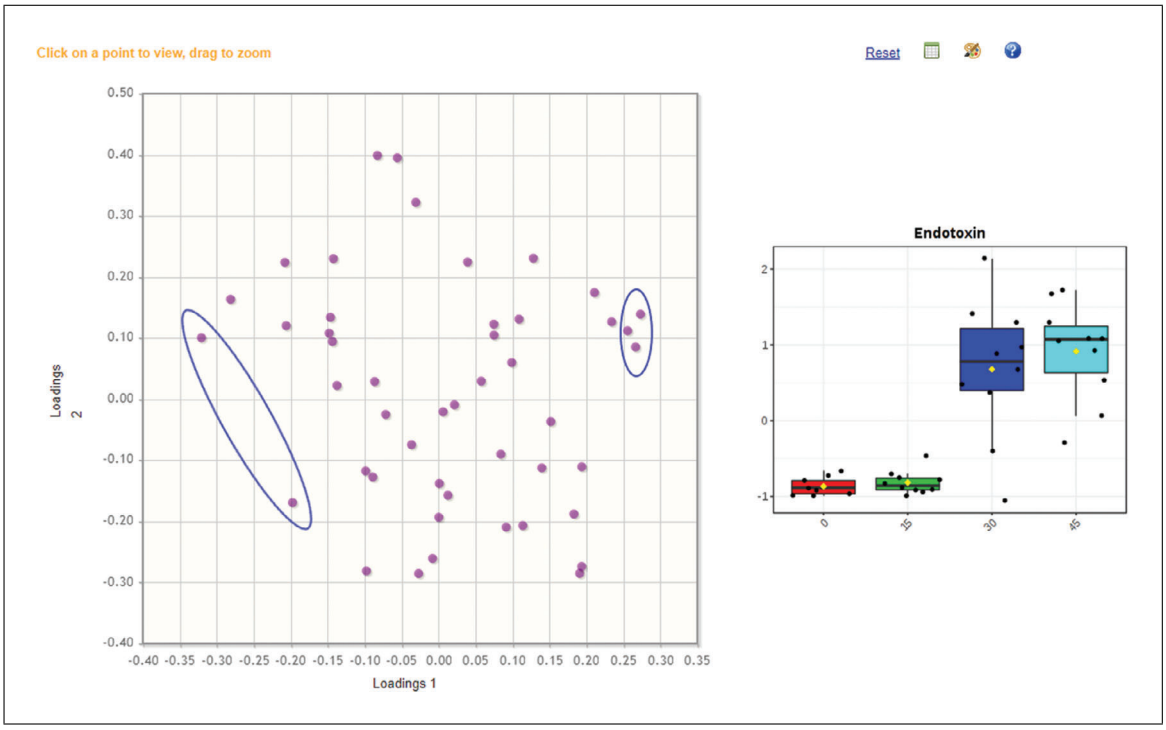


Figure 32 A screenshot showing a PLS-DA loading plot of the first two components. The circles indicate the most influential data points located on the outermost areas along the direction of separation as identified in the corresponding score plot. Clicking on a data point will display the corresponding boxplot to the right.

case, PLS-DA will be performed using a general linear model in which group labels will be coded using the model matrix rather than numerical values.

10. Click the "**Loadings Plot**" tab, and the result is shown (Fig. 32). Click the outermost points (as shown in the ellipses) along the directions of separation to identify the most influential compounds on the right side, *Methylamine*, *Glucose*, and *Endotoxin*, and on the left side, *3-PP* and *Aspartate*. Most of these compounds agree very well with those compounds already identified by both PCA and Heatmap techniques.

11. PLS-DA is a supervised classification method. It uses the group label to maximize the separation between different groups. One important issue associated with PLS-DA is overfitting (Westerhuis et al., 2007). To address this issue, MetaboAnalyst provide two approaches—cross validation and permutation testing. Click the "Cross Validation" tab to view the results from cross validation, which appear as a bar graph (Fig. 33). The purpose of cross validation is to determine the optimal number of components needed to build the PLS-DA model. There are three common performance measures—the sum of squares captured by the model (R^2), the cross-validated R^2 (also known as Q^2), and the prediction accuracy (Accuracy). The default criterion is Q^2, which indicates (marked by a red star) that a three-component model is the best.

Users can click the "View details . . . " link to get the exact numerical values for these performance measures. Please note that due to the nature of random sampling, the results may be slightly different if 10-fold cross validation is used. Using different performance measures may give different answers. For instance, if Accuracy was selected, it shows that using the top five components yields the optimal performance (~80%). In this case, choose the three-component model based on Q^2, which is simpler and therefore less prone to overfitting.

12. Click the "**Permutation**" tab to see the results from MetaboAnalyst's permutation tests. There are two parameters for permutation tests—the number of permutations

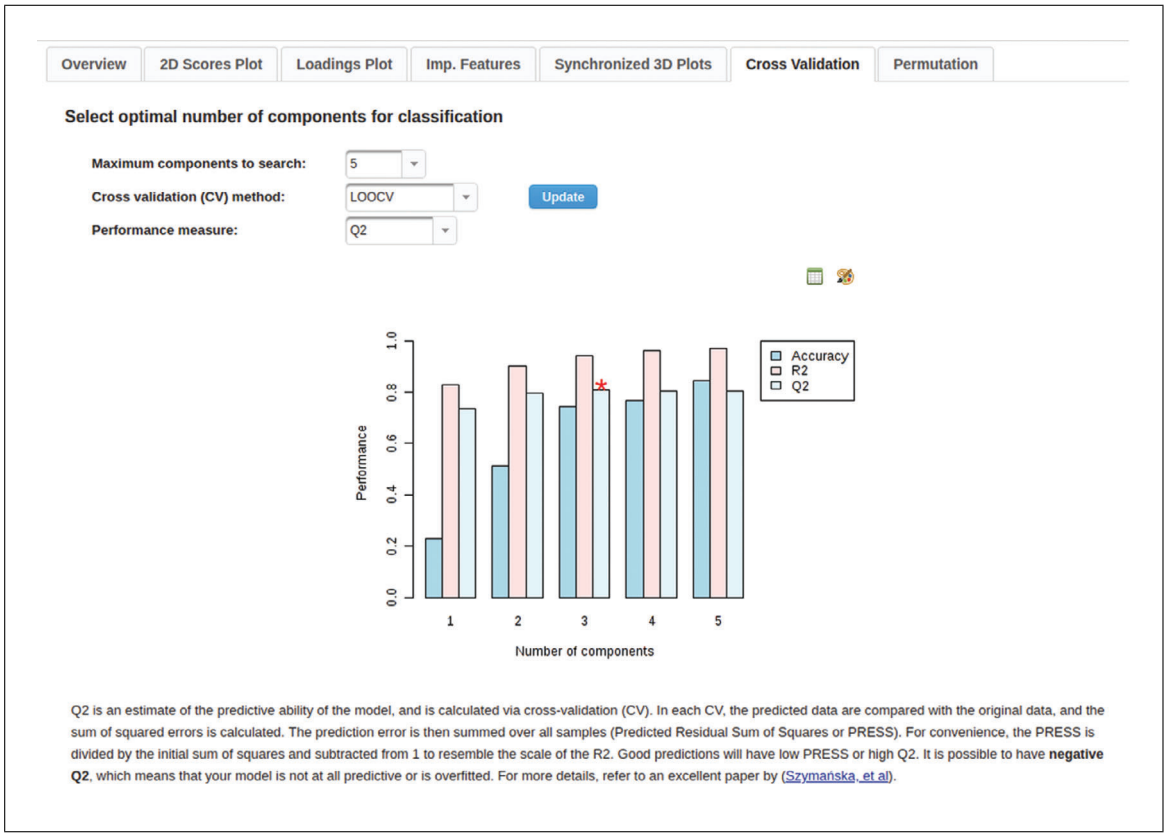


Figure 33 A screenshot showing the PLS-DA cross validation result. The selected performance measure—Q2 shows that the three-component model is best (indicated by a red star).

and the test statistic that will be used as a performance measure. MetaboAnalyst provides two options for its test statistics: (a) the separation distance which is defined as the ratio of the between-group sum of the squares and the within-group sum of squares (B/W-ratio) as suggested by Bijlsma et al. (2006), or (b) the prediction accuracy. For this example, click the radio button for "Separation distance (B/W)" as the test statistic, choose 2000 permutations from the drop-down menu (listed beside "Set permutation numbers") and then click the "Submit" button. After ~30 s, the result should appear (Fig. 34). It is very clear that the B/W distribution based on the original data is very different from the B/W distribution calculated from the permutated data. The *p*-value is significant (<5E-4).

The empirical p-value is calculated as the percentage of times that the B/W calculated from permutated data achieves the same or better results than the B/W calculated from the original data. A p-value of 0.0005 in 2000 permutations means that just once in 2000 permutations (0.0005 × 2000) did the permutated data yield a better performance than the original label.

13. Click the "**Imp. Features**" tab to view the most important or informative compounds that were selected using the three-component model. The default view shows the top 25 compounds ranked based on the variable importance in projection (VIP) score (Fig. 35). The most important compounds include *3-PP*, *Endotoxin*, *Glucose*, *Alanine*, *Methylamine*, *Isobutyrate*, *Uracil*, and *Aspartate.* As might be expected, these compounds agree very well with the previous list generated in Basic Protocol 2.

Two variable importance measures are available in MetaboAnalyst. The default VIP score is a weighted sum of squares of the PLS loadings that considers the amount of explained Y-variance of each component. The other importance measure is based on

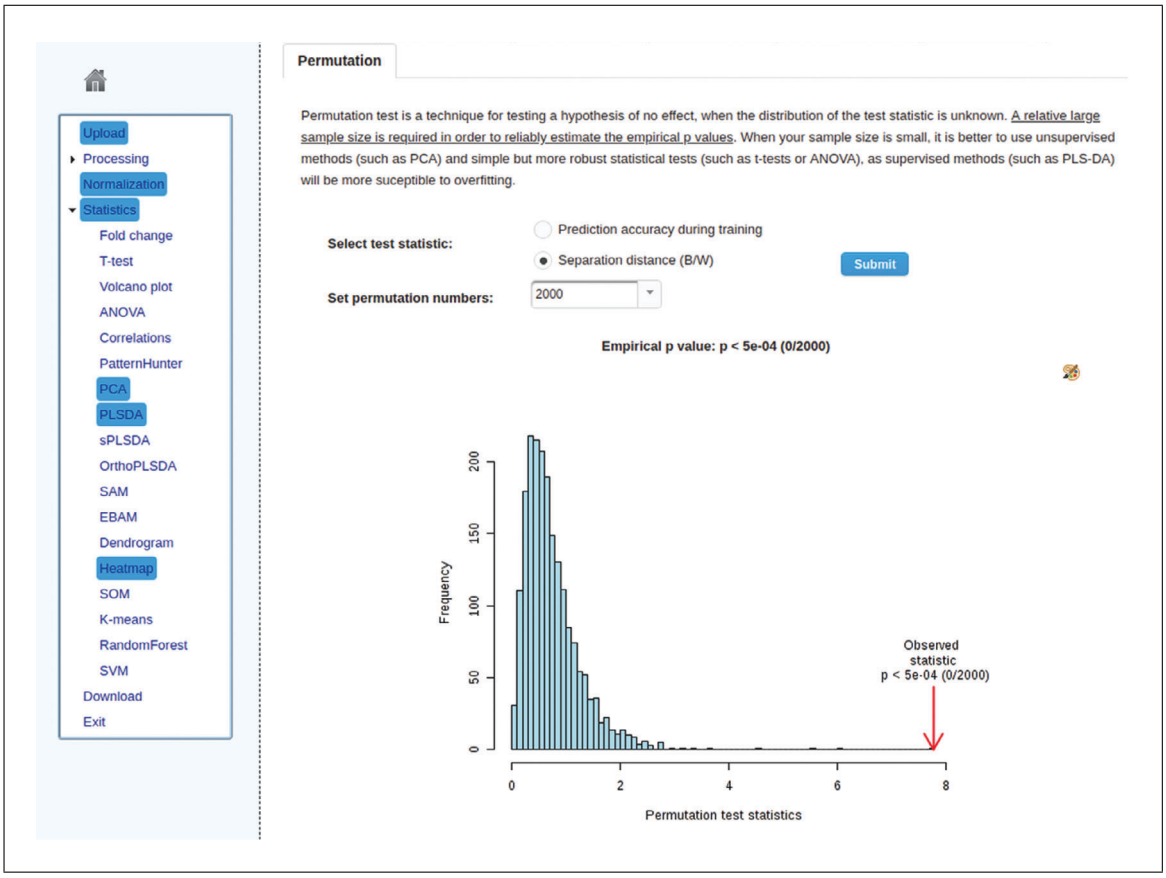


Figure 34 A screenshot showing a PLS-DA permutation result. The histogram indicates that the observed statistic based on the original data is not part of the null distribution formed by those from the permuted data.

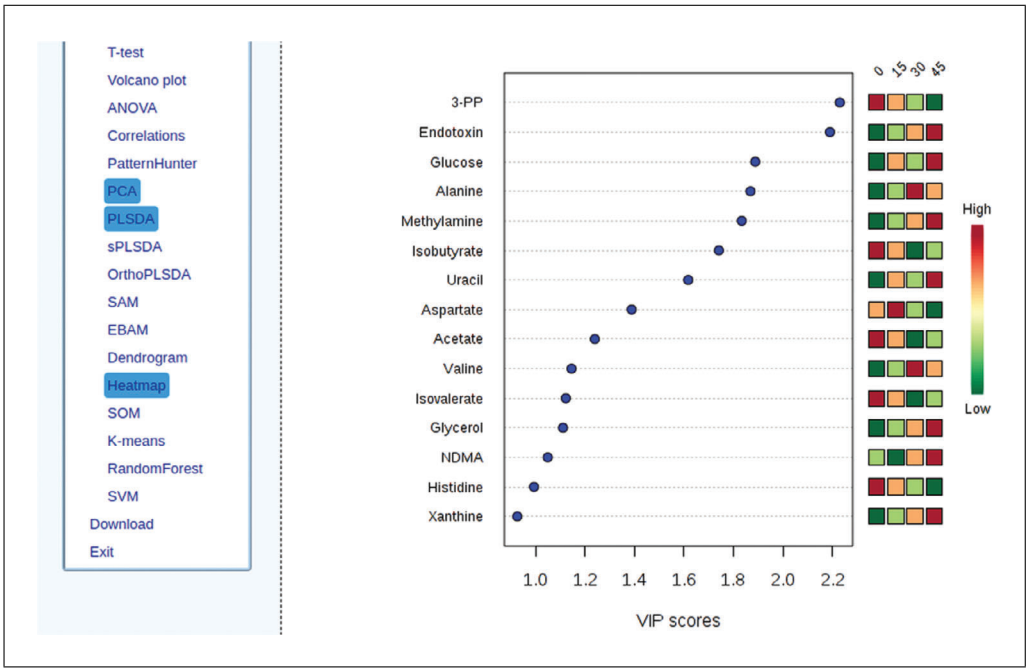
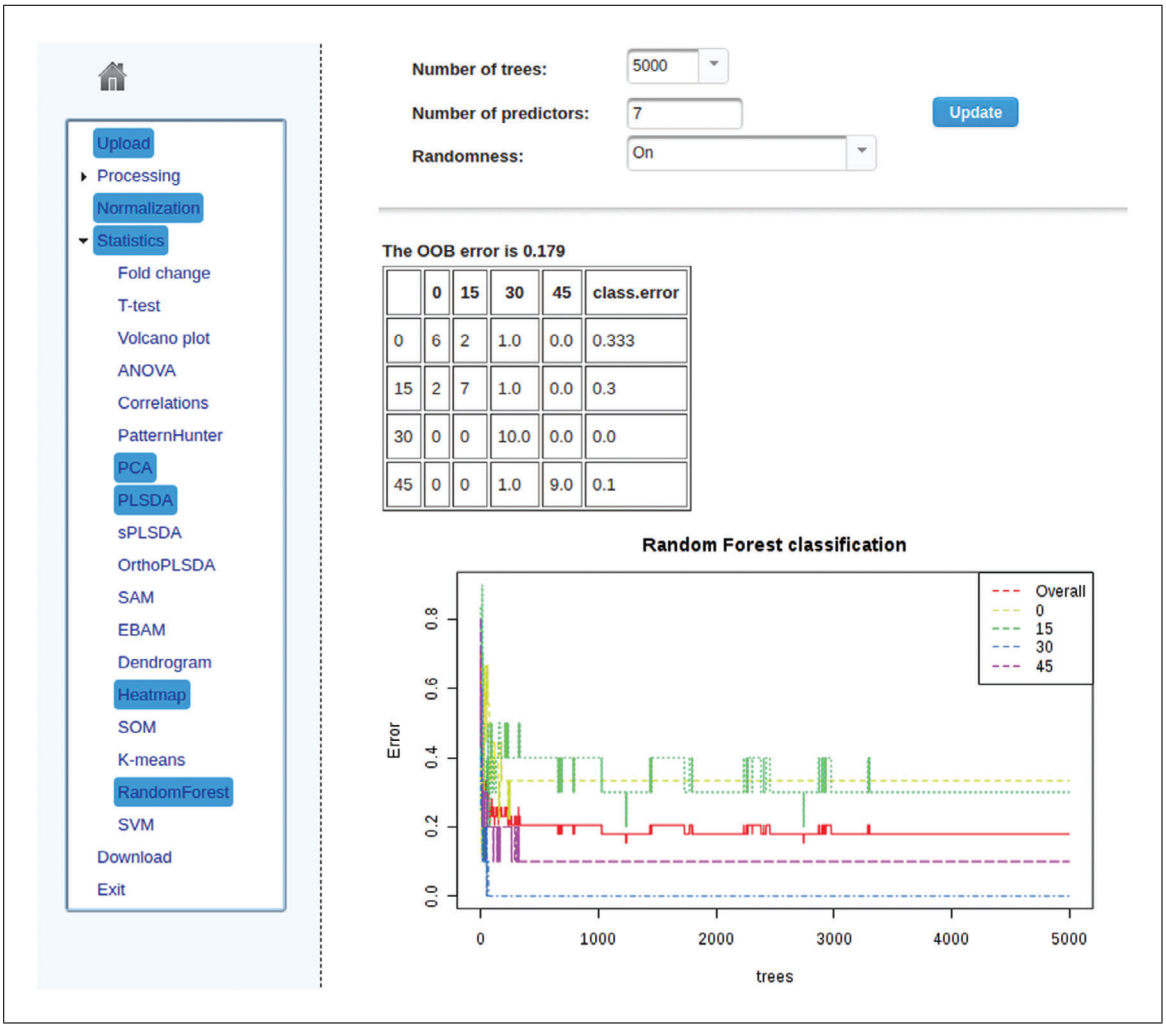


Figure 35 A screenshot showing the most important metabolites ranked by the PLS-DA VIP score. The mini heatmap on the right indicates their concentration variations within different groups.



	0	15	30	45	class.error
0	6	2	1.0	0.0	0.333
15	2	7	1.0	0.0	0.3
30	0	0	10.0	0.0	0.0
45	0	0	1.0	9.0	0.1

Figure 36 A screenshot showing classification results using Random Forest with 5000 trees. The overall (OOB) error is 0.179. The details of the classifications for each group are provided in the table. The bottom image shows the OOB errors versus increasing number of trees. It indicates that using 5000 trees for classification should be enough. as the error rates become stable after ~3000 trees.

a weighted sum of the PLS-regression coefficients. The weights are a function of the reduction of the sums of squares across the number of PLS components.

Data analysis using Random Forests (RF)

14. Click the "RandomForests" hyperlink on MetaboAnalyst's navigation tree on the left. RF is a powerful non-parametric classification method and can be used for both classification and important variable selection (Breiman, 2001). RF uses an ensemble of classification trees, each of which is grown by random feature selection using bootstrap sampling from the original sample set. Class prediction is based on the majority vote of the ensemble. By default, 500 trees are used to build the RF classifier. Use the drop-down menu to adjust this number to increase the number of trees from 500 to 5000 (leaving all other parameters unchanged), and then click the "Submit" button. The result is shown in Figure 36.

During tree construction, about one-third of the instances are left out of the bootstrap sampling process. These "left-out" data are then used as test data to obtain an unbiased estimate of the classification error, known as the 'out-of-bag' (OOB) error. The OOB error using the given parameters is 0.179, as shown in Figure 36. Note that the results can be slightly different due to the random nature of the algorithm.

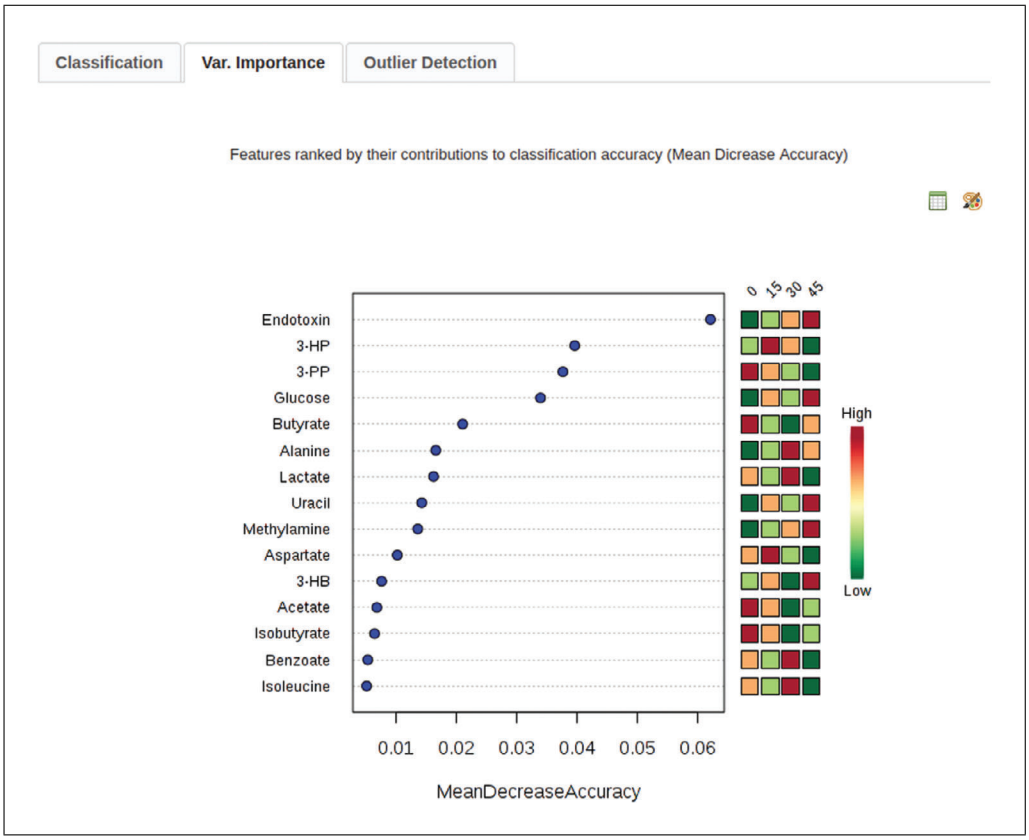


Figure 37 A screenshot showing the top 15 important metabolites ranked by Random Forest classification. They are ranked in order of decreasing prediction accuracy (`MeanDecreaseAccuracy`) when being permuted.

15. Random forests classification also provides a variable importance measure. Variable importance is evaluated by measuring the increase of the OOB error when it is permuted. Click the "**Var. Importance**" tab to view the results (Fig. 37). Compared to the list of important compounds selected by partial least squares VIP score, most of them are seen to be quite similar. The only difference is *3-HP (3-Hydroxyphenylacetate)*, which was not identified using the other approaches.

Further investigation of the selected compounds

16. MetaboAnalyst offers two ways to allow users to further examine the patterns of concentration change for these important or informative compounds identified in PLS-DA or RF. Click the table icon near the top right corner of the PLS-DA VIP plot (Fig. 35) to access the "Feature Details View" of the ranked list of important compounds identified by VIP Analysis. Users can visualize the concentration distribution in the form of box plots and bar graphs by clicking on the corresponding compound name, as shown in Figure 17. The other approach is to use a heatmap. For example, users can easily visualize the top 15 compounds selected by the PLS-DA VIP score. To demonstrate, click the "Heatmap" link on the navigation tree to go to the heatmap page. Leave the first three options as default. Check the "Do not re-organize" check box and make sure that the "Both" option is selected for data visualization without hierarchical clustering. Select the "View Options" checkbox and specify it to display the top 15 features. Make sure the "PLS-DA VIP" is selected from the drop-down menu. Finally, click the "Submit" button. The result is shown in Figure 38. Compared to a box plot, the heatmap allows simultaneous visualization of all of the most important compounds, which greatly facilitates data summarization and pattern discovery.

Figure 38 A screenshot showing a heatmap visualization (without clustering) for the top 15 metabolites selected by the PLS-DA VIP score.

Data download

17. Click the "Download" hyperlink at the bottom of the navigation tree. The download page displays a table showing all figures and result tables, as well as an R Command History file generated during the analysis session. Click the "Generate Report" button to create a PDF report that describes all the steps performed with background information, followed by the results in the form of bulleted lists, tables, and images as appropriate. Users should download the analysis report, the zip file and click "Exit" to complete this session.

BASIC PROTOCOL 4 FUNCTIONAL INTERPRETATION OF METABOLOMIC DATA

The output from a successful statistical analysis in any metabolomics study is usually a long list of features (or metabolites) that have changed significantly or showed interesting patterns of coordinated change under the different conditions. Obtaining this kind of list is usually not the end of one's analysis; rather, it is the starting point for data interpretation and hypothesis generation. Over the past decade, many computer-assisted data interpretation approaches have been developed and tested. Among them, group-based significance tests and pathway analysis methods have gained widespread acceptance among many researchers involved in "omics" data analysis (Draghici et al., 2007; Efron & Tibshirani, 2007; Goeman, van de Geer, de Kort, & van Houwelingen, 2004; Goffard & Weiller, 2007; Subramanian et al., 2005). These two approaches allow for the incorporation of existing biological knowledge into the data analysis process, which greatly facilitates data interpretation. Generally, functional analysis can only be applied when the compound identities and/or their concentrations are known. However,

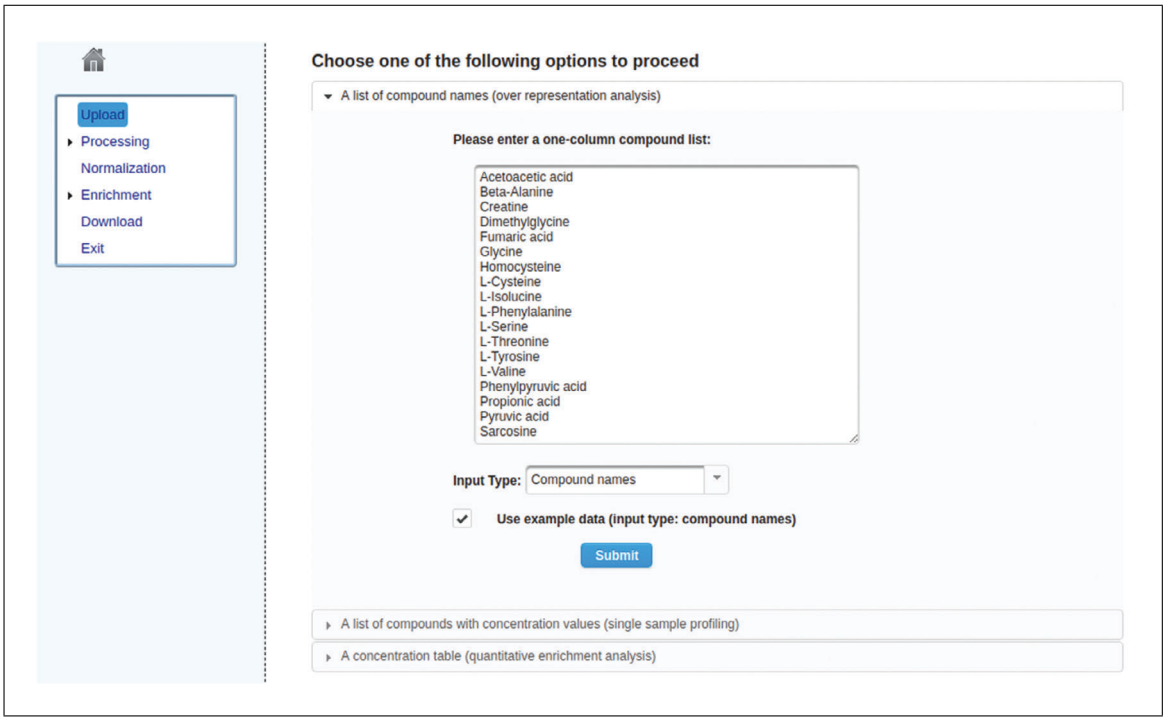


Figure 39 A screenshot showing the data upload view for over-representation analysis (ORA). A list of 18 compound names has been entered. Note the "Compound names" option has been selected to indicate the type of input data.

MetaboAnalyst 4.0 now supports the prediction of pathway-level activity directly from MS peak lists, which will be covered in Basic Protocol 9. This protocol describes how to use data produced from standard quantitative metabolomic approaches or targeted metabolomic approaches (i.e., concentration tables) to perform MSEA and MetPA.

Necessary Resources

Hardware

A computer with internet access

Software

An up-to-date web browser, such as Google Chrome (*http://www.google.com/chrome*), Mozilla Firefox (*http://www.mozilla.com/*), Safari (*http://www.apple.com/safari/*), or Internet Explorer (*http://www.microsoft.com/ie/*). JavaScript must be enabled in the web browser.

Files

None

Over-representation analysis

1. Go to MetaboAnalyst home page (*https://www.metaboanalyst.ca*). Click the "click here to start" to enter the "Module Overview" page. Click on the "Enrichment Analysis" button (upper right of the circle panel) to enter the corresponding "Data Upload" page. There are three drop-down panels corresponding to three different types of enrichment analysis. The over representation analysis (ORA) panel is open by default (Fig. 39).

   MetaboAnalyst's enrichment analysis accepts three data types. The ORA accepts a list of compound names entered in a single-column format. The single sample profiling (SSP) requires a list of compound concentrations entered as two-column table. The quantitative

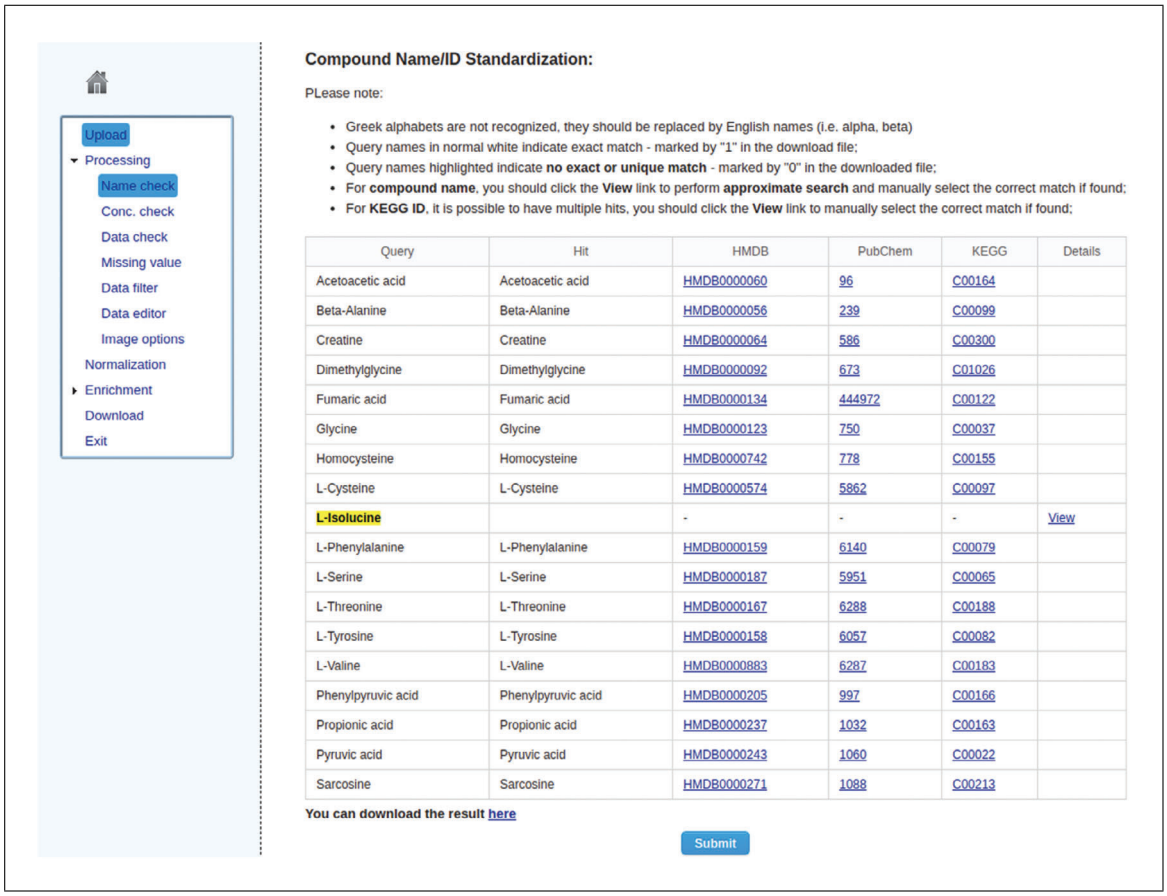


Figure 40 A screenshot showing the result table from compound name standardization. The name highlighted in yellow indicates a compound with an approximate match. Users need to click the "View" link to make the correction manually.

enrichment analysis (QEA) requires a concentration table in comma-separated value format (`.csv`) or tab delimited format (`.txt`).

2. In this protocol, we will use the example compound list. Check "Use the example data (input type: compound names)." A list of 18 compound names should appear in the text box above. Make sure the "Input Type" is set to Compound names, and click the "Submit" button.

   Alternately, users can copy and paste a list of compound names or identifiers used by one of the major metabolomics databases (HMDB, KEGG, PubChem, ChEBI, and METLIN).

3. The "Compound Name Standardization" page is shown (Fig. 40). The purpose of compound name standardization is to match the compound identity from users' data to that used by MetaboAnalyst's knowledgebase. Compounds without exact matches will be flagged using a different color. In this case, *L-Isolucine* is highlighted, which is a typo introduced intentionally to illustrate the function. Users can click the "View" link in the "Details" column to perform an approximate search. All candidate names will be shown in a pop-up window (Fig. 41). Users can manually pick the correct name from this candidate list. The highest-ranked match identified by MetaboAnalyst is *L-Isoleucine*, which is correct. Check the box beside this compound name and click the "OK" button. After this is complete, press the "Submit" button.

   This is a critical step. Compounds without matches in the database will be excluded from the downstream analysis. If the step is completed properly, the table should be entirely filled under the "Hit" column.

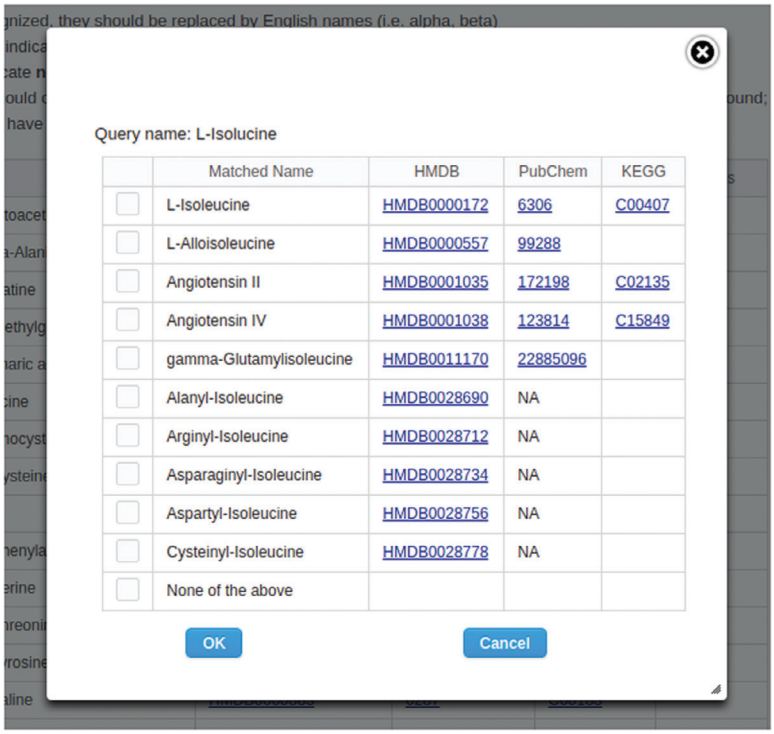


Figure 41 A screenshot showing the available candidate compounds for "L-Isoleucine" identified by the approximate matching algorithm.

4. The next page shows the parameter settings for enrichment analysis (Fig. 42). For demonstration purposes, use the default "pathway-associated metabolite sets" and leave the other options unchanged. Click the "Submit" button to proceed to the next step.

   There are two options that users can specify for enrichment analysis. The first one is to select a metabolite set library. MetaboAnalyst currently contains eight built-in metabolite-set libraries based primarily on studies of human diseases and human metabolism. Users are required to provide their own metabolite set library if they wish to perform MSEA for other species. Click the hyperlink with the text "Click here to upload your own customized metabolite set library." The data upload page is shown, which contains detailed instructions on how to prepare the metabolite set file. An example data set is also available for download. The second option is to specify a reference metabolome. Over-representation analysis uses a reference metabolome to calculate a background distribution to determine if the matched metabolite set is more enriched for certain metabolite sets compared to random chance. The default is to use all the metabolites of the metabolite set library as the reference metabolome. However, this is inappropriate, as currently no single analytical platform can measure all these metabolites with equal probability. Thus, any observed metabolite enrichment may be introduced by the choice of the platform rather than by the experimental conditions. To alleviate this issue, MetaboAnalyst allows users to upload a platform-specific reference metabolome. Clicking on the "Upload a reference metabolome . . . " radio button will produce the reference metabolome upload page (Fig. 43).

5. The default enrichment analysis result is shown as an enrichment network (Fig. 44). In the network, each node represents a metabolite set, with its color corresponding to its *p*-value in the enrichment analysis and its size corresponding to its fold enrichment (hits/expected) to the query. Two metabolite sets are connected by an edge (line) if the number of shared metabolites between the two sets is greater than 20%. From the network, it is apparent that *Homocysteine Degradation* has the largest fold

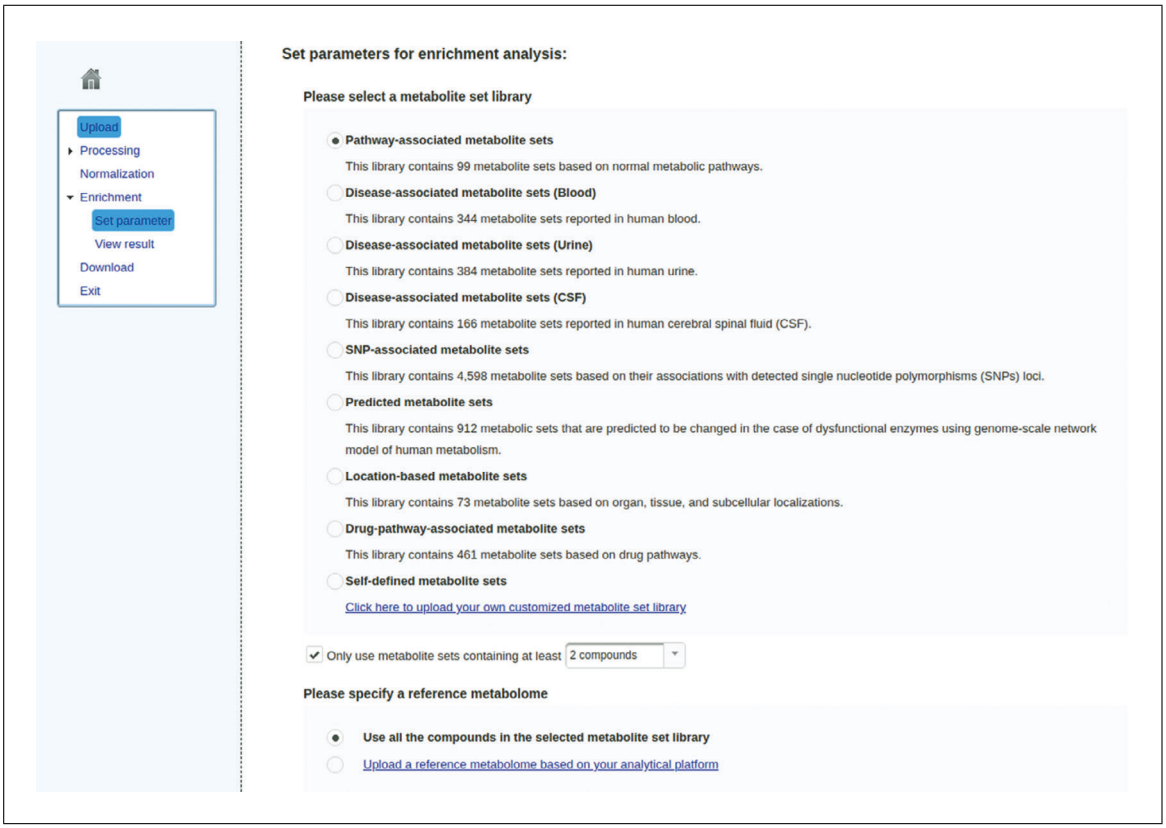


Figure 42 A screenshot showing the parameter settings for enrichment analysis. See the text for further details.

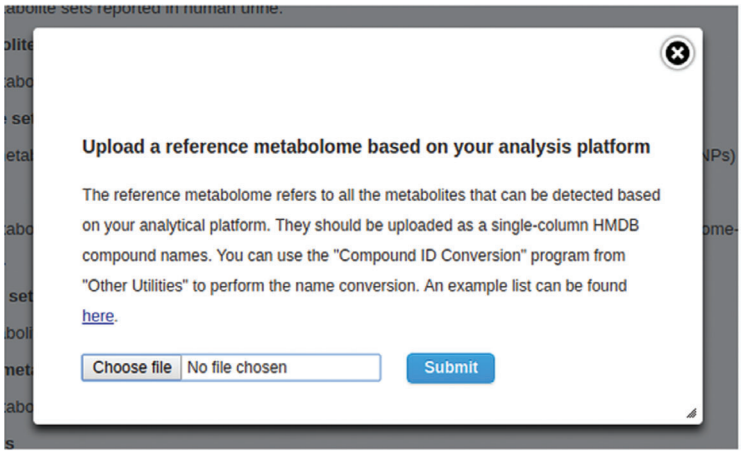


Figure 43 A screenshot showing the dialog for uploading a platform-specific reference metabolome.

enrichment, and that many enriched pathways including *Methionine Metabolism*, *Glycine and Serine Metabolism*, *Phenylalanine and Tyrosine Metabolism*, and *Ammonia Recycling* are highly interconnected. The network is also interactive, allowing users to zoom in/out and drag/drop nodes with their mouse or touch-pad. Double-clicking on a metabolite set will open a dialog box to view more details about that metabolite set. For example, double-click *Homocysteine Degradation*, and all compound members of that pathway with matched metabolites will be highlighted in red (*L-Serine*, *L-Cysteine*, and *Homocysteine*). When available, the corresponding pathway link from SMPDB (Frolkis et al., 2010) is also presented. Users can follow

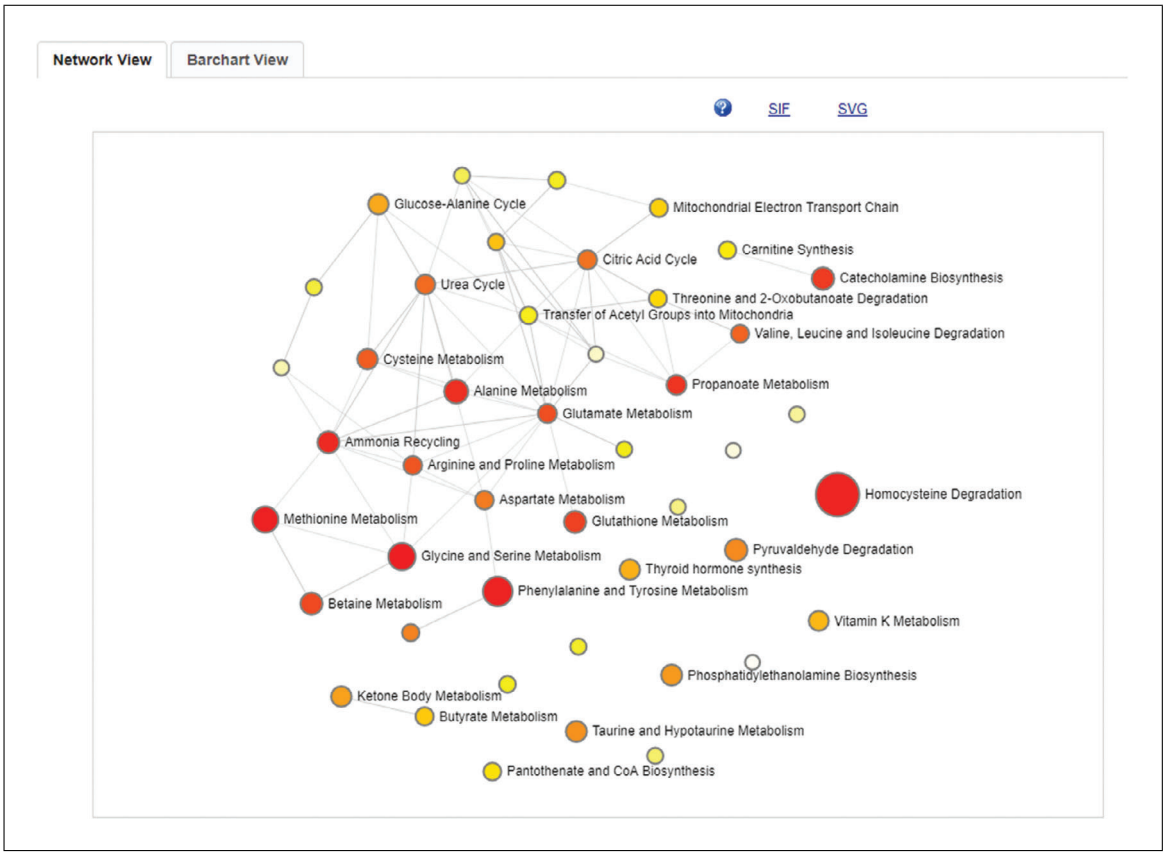


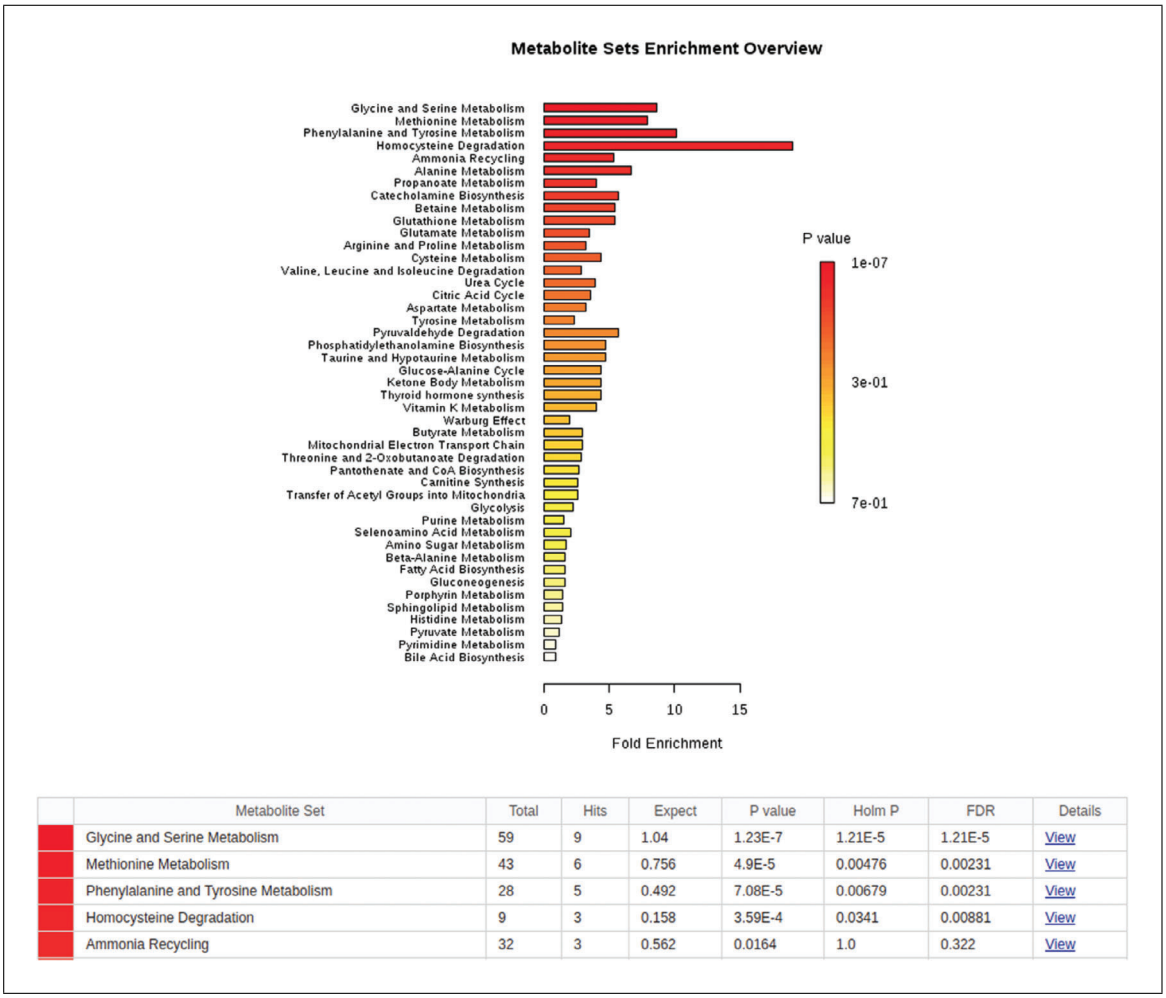
Figure 44 A screenshot showing a network summary from an over-representation analysis (ORA). The nodes represent metabolite sets enriched in the data. Further details are available in the text.

the link to browse the detailed pathway map within SMPDB. The MSEA result can also be visualized as a Barchart. Select the "Barchart View" to visualize this. Now the results are presented both graphically (top) and in a detailed table (bottom) (Fig. 45). The horizontal bar graph summarizes the most significant metabolite sets identified during this analysis. The bars are colored based on their *p*-values (lower *p*-values are redder) and the bar length is based on the fold enrichment. In this case, the top-ranked pathway is *Glycine, Serine, and Threonine metabolism*. Move to the result table below, and users can click the "View" link in the "Details" column of the top pathway to view the detailed information about any metabolite set of interest.

6. When one has finished exploring the MSEA results, click the "Download" hyperlink on the navigation tree (left side) to create and download the analysis report and the accompanying graphical and textual output. Click the "Exit" button on the left side to end this session.

Single sample profiling

7. This approach is only applicable when metabolite concentrations are measured from a common human biofluid such as cerebral spinal fluid (CSF), blood, or urine. Return to the "Module Overview" page and click on the "Enrichment Analysis" button. On the "Data Upload" page, click the second drop-down panel labeled "A list of compounds with concentration values (single sample profiling)" to display its content (Fig. 46). Click on the "Use the example data" check box. A two-column list of about 25 metabolite names and concentrations (in micromolar units) is generated in the text box. Click "Submit."



	Metabolite Set	Total	Hits	Expect	P value	Holm P	FDR	Details
	Glycine and Serine Metabolism	59	9	1.04	1.23E-7	1.21E-5	1.21E-5	View
	Methionine Metabolism	43	6	0.756	4.9E-5	0.00476	0.00231	View
	Phenylalanine and Tyrosine Metabolism	28	5	0.492	7.08E-5	0.00679	0.00231	View
	Homocysteine Degradation	9	3	0.158	3.59E-4	0.0341	0.00881	View
	Ammonia Recycling	32	3	0.562	0.0164	1.0	0.322	View

Figure 45 A screenshot showing a bar chart summary of the results from an over-representation analysis (ORA). A graphical summary is present on the top and the details are given in the table below.

Note that the concentrations must be provided using standard concentration units (µmol for blood and CSF, and µmol/mmol_creatinine for urine).

8. The next page involves compound name standardization as described in **step 3**, above. Manually fix the *L-isoleucine* match and then click the "Submit" button to continue to the next page. A table of concentration comparisons between the uploaded data and the corresponding reference concentrations collected from the published literature will appear (Fig. 47). Compounds are identified as being H (high), M (medium), or L (low). In this example, four compounds (*L-threonine*, *L-tryptophan*, *L-tyrosine*, and *L-alanine*) are all identified as being unusually high.

   *By default, a compound will be selected only if its measured concentration is above or below **all** the reference concentrations. Users can manually select or unselect a compound using the corresponding checkbox in the last column. Click the image icon in the "Details" column to view a graphical summary of the concentration comparison (Fig. 48). The table below displays the original literature reports on these concentrations.*

9. The four compounds identified from **step 8** will be subject to over-representation analysis (ORA). The procedures are identical to steps 4 to 6 described above.

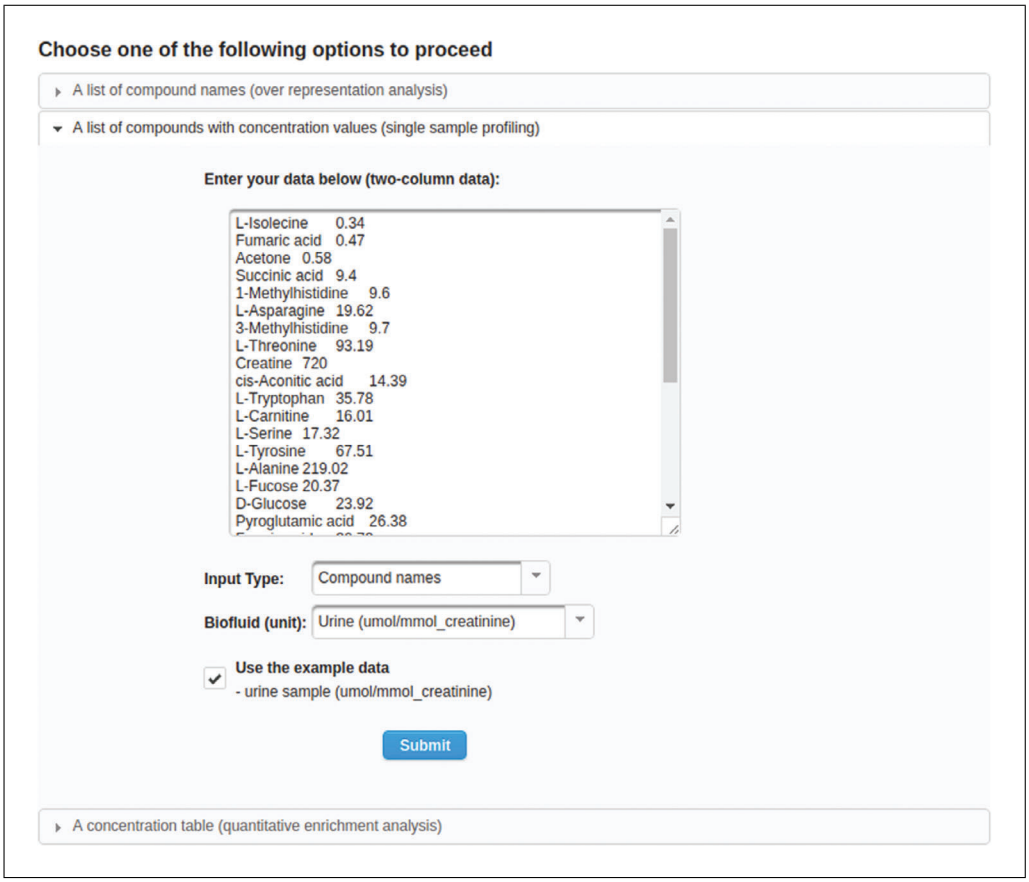


Figure 46 A screenshot showing the data upload view for single sample profiling (SSP). Users must provide the names of the compounds and their concentrations in a two-column format. The compound label and the Biofluid must be properly selected from the drop-down menus.

Quantitative enrichment analysis

10. This approach is used when the data input is a concentration table. Return to the "Module Overview" page and click on the "Enrichment Analysis" button to enter the "Data Upload" page. Click the third drop-down panel labeled with "A concentration table (quantitative enrichment analysis)" to display its content. Locate "Try our test data" and make sure "Data 1" is selected. Then, click the "Next" button. Go through compound name standardization, data integrity checking, data normalization, and parameter specification as described previously.

    This data set contains metabolite concentrations of 77 urine samples from cancer patients, measured by ^{1}H NMR (Eisner et al., 2010). The data contain two groups: a control group and a cachexic group (significant muscle loss). In this case, we choose to use normalization by a pooled sample from the "control" group followed by log transformation. Unlike ORA, users do not need to use a reference metabolome for quantitative metabolomics analysis. The procedure is based on the Globaltest (Goeman et al., 2004), which directly calculates the association between the metabolite sets and the phenotype without referring to a background.

11. Both network and barchart views of the results are available (see Figs. 44 and 45). The Barchart View of the results from QEA is shown in Figure 49. Like Figure 45, the top panel shows a graphical summary of the important metabolite sets (pathways) and the bottom panel shows a detailed table. Click the image icon in the "Details" column, to visualize more information about the matched metabolite set (Fig. 50).

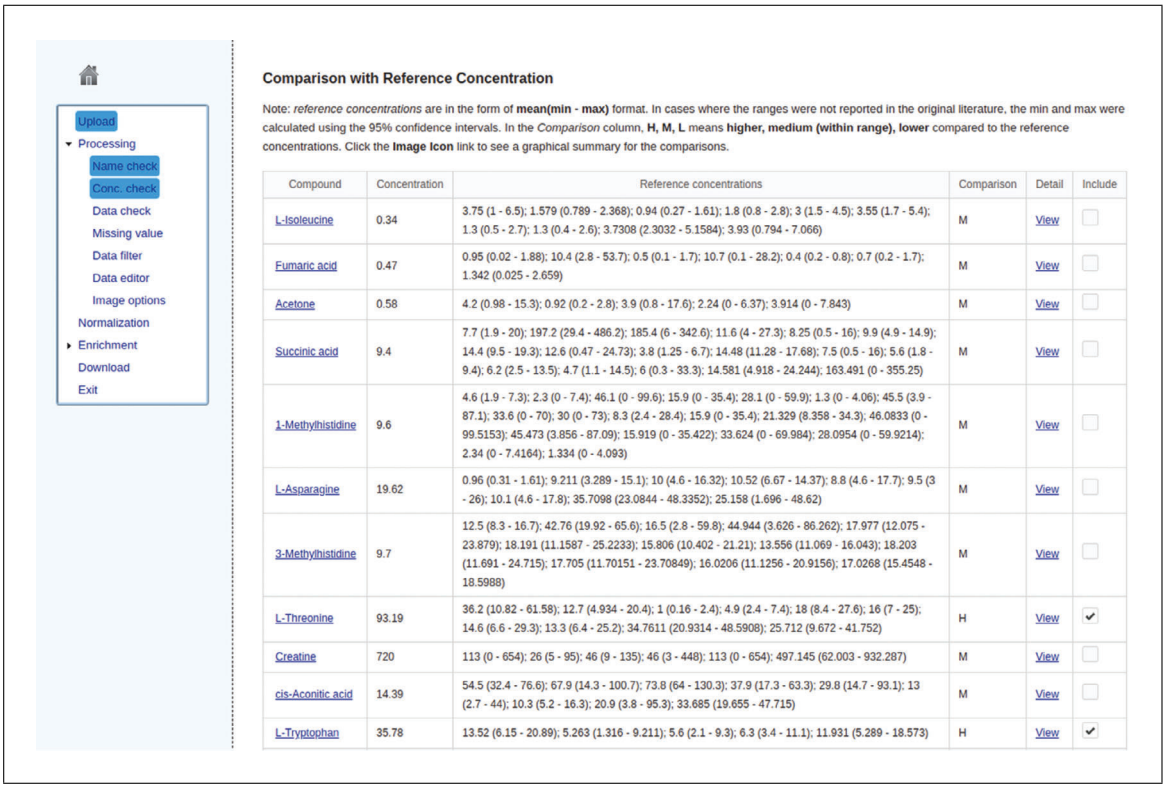


Figure 47 A screenshot showing the result table from the metabolite concentrations in comparison with reference concentrations. A detailed graphical comparison can be obtained by clicking the "View" hyperlink in the "Detail" column.

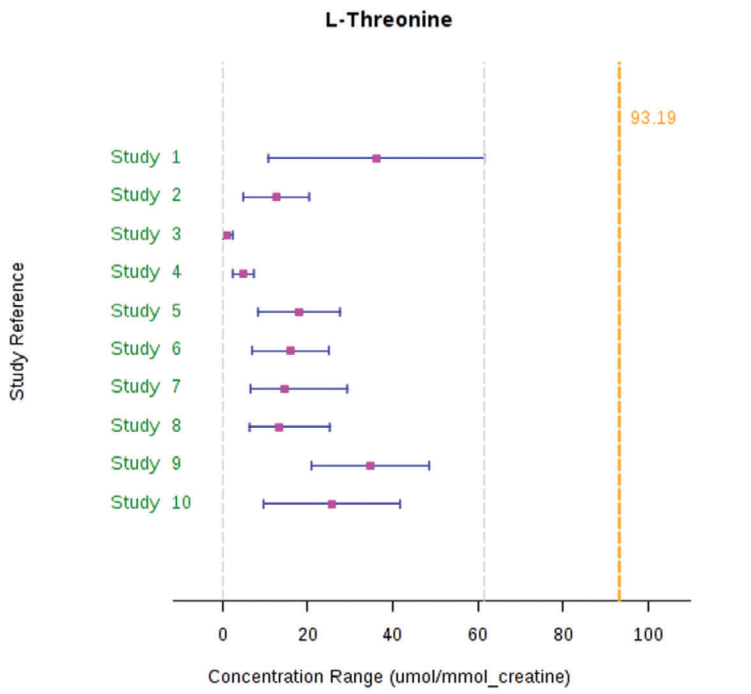
Users can easily tell whether a particular compound is positive or negatively associated with a phenotype of interest based on its concentration distribution across different groups, as indicated by the corresponding box plot. The associated p-values are also given.

Pathway analysis

12. Return to the MetaboAnalyst home page, click on "click here to start" to enter the "Module Overview" page, and then click the "Pathway Analysis" button to enter the corresponding "Data Upload" page (Fig. 51). Here, users can enter either a list of important compound names, KEGG IDs, or HMDB IDs, or upload a compound concentration table. The process shares almost the same steps as described for Enrichment Analysis. Scroll down to the bottom half of the page—"Or upload a concentration table (.csv or .txt)." Check the "Use example data" check box, and then click the "Submit" button at the bottom of the page.

    Alternatively, users can go to the "Data Formats" page (left side of the home page menu) and download the data labeled as "Compound concentration data—human, two groups" in CSV format, and then upload the data from this page. Users need to specify the type of compound label as "common name" and the class label as "discrete" to proceed.

13. Go through compound name standardization, data processing, and data normalization as described in the previous steps. For this demonstration, choose "Auto scaling" and leave other options as defaults. After clicking the "Normalize" button and then clicking the "Proceed" button at the bottom of the Data Normalization Page, users will then be directed to the page for setting the parameters for pathway analysis (Fig. 52). In this case, accept the default selections for all the parameters (including the selection of the *Homo sapiens* KEGG pathway library). Click the "Submit" button at the bottom of the page.



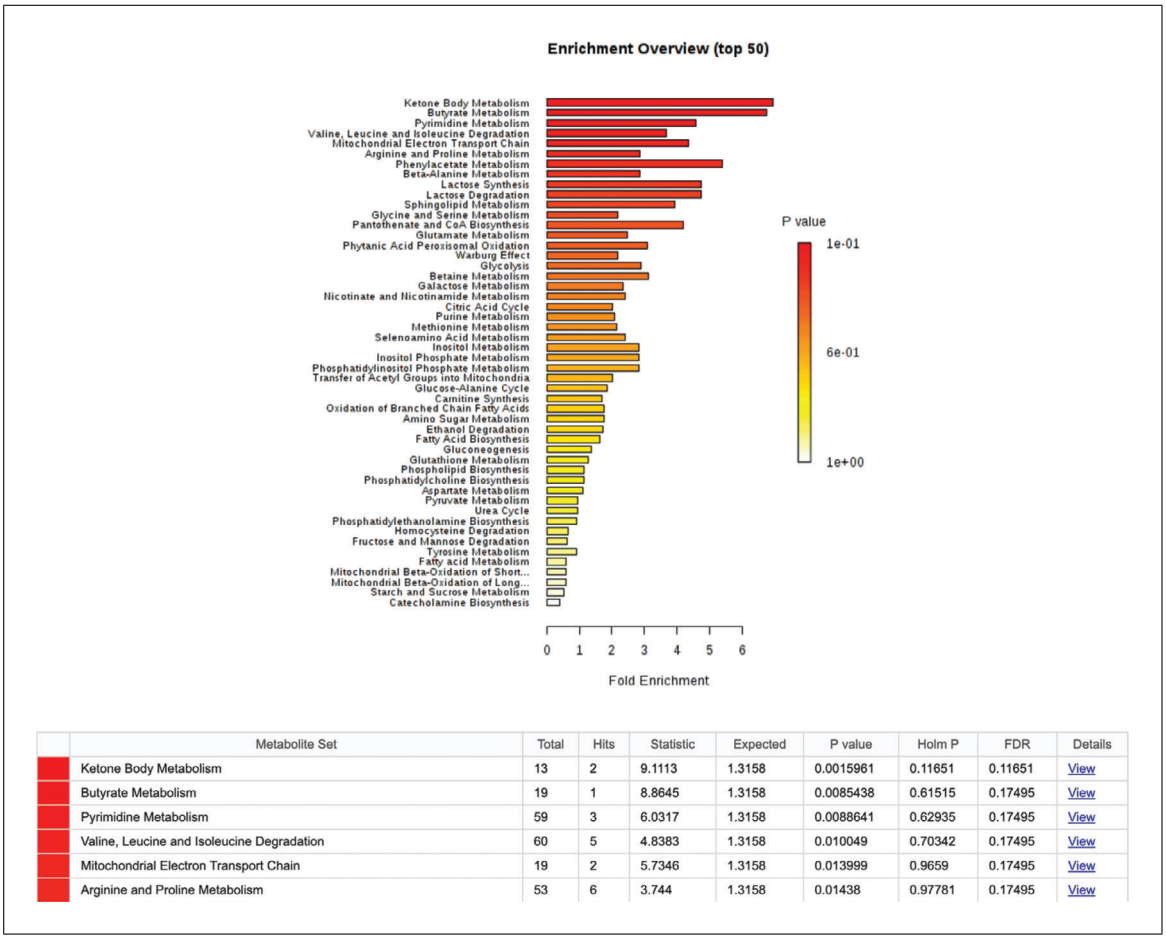
Study	Concentration	Reference	Note
Study 1	36.2 (10.82 - 61.58)	(Pubmed)	Both
Study 2	12.7 (4.934 - 20.4)	(Pubmed)	Both
Study 3	1 (0.16 - 2.4)	Geigy Scientific Tables, 8th Rev edition, pp. 165-177. Edited by Cornelius Lentner.; West Cadwell, N.J. : Medical education Div., Ciba-Geigy Corp.; Basel, Switzerland c1981-1992. (Pubmed)	Both
Study 4	4.9 (2.4 - 7.4)	Geigy Scientific Tables, 8th Rev edition, pp. 165-177. Edited by Cornelius Lentner.; West Cadwell, N.J. : Medical education Div., Ciba-Geigy Corp.; Basel, Switzerland c1981-1992. (Pubmed)	Female

Figure 48 A screenshot showing the concentration comparisons for *L-Threonine*. The original literature citations for these reference concentrations are listed below.

Three parameters need to be specified for pathway analysis—the pathway library, the algorithm for pathway enrichment analysis, and the algorithm for topological analysis. MetaboAnalyst currently contains KEGG pathway libraries for 21 model organisms and SMPDB pathway libraries for two model organisms. Pathway enrichment analysis supports Globaltest (Goeman et al., 2004) and GlobalANCOVA (Hummel, Meister, & Mansmann, 2008), which are two similar algorithms designed for testing differentially expressed genes or metabolites in functionally related groups. Degree centrality and betweenness centrality are two measures to estimate the importance of a compound within a given metabolic pathway. The former measures the number of connections the node of interest has to other nodes, while the latter measures the number of shortest paths going through the node of interest.

14. The results from pathway analysis are presented in two parts—a graphical output on the top panel and a table giving all the numerical details in the bottom panel (Fig. 53). Users can intuitively explore the results by pointing and clicking on various graphical elements.

There are three types of viewing options. The upper left panel is the metabolome view, which displays all the matched pathways as circles. The color and size of each circle are based on its p-value and pathway impact value, respectively. Users should pay attention to those pathways in the top right diagonal region, which indicates that metabolites involved in those pathways are significantly changed, and that they are more likely to



	Metabolite Set	Total	Hits	Statistic	Expected	P value	Holm P	FDR	Details
	Ketone Body Metabolism	13	2	9.1113	1.3158	0.0015961	0.11651	0.11651	View
	Butyrate Metabolism	19	1	8.8645	1.3158	0.0085438	0.61515	0.17495	View
	Pyrimidine Metabolism	59	3	6.0317	1.3158	0.0088641	0.62935	0.17495	View
	Valine, Leucine and Isoleucine Degradation	60	5	4.8383	1.3158	0.010049	0.70342	0.17495	View
	Mitochondrial Electron Transport Chain	19	2	5.7346	1.3158	0.013999	0.9659	0.17495	View
	Arginine and Proline Metabolism	53	6	3.744	1.3158	0.01438	0.97781	0.17495	View

Figure 49 A screenshot showing the bar chart summary of results from a quantitative enrichment analysis (QEA). A graphical summary is presented on the top and the details are given in the table below.

have significant impacts on the pathways based on their positions. Hovering the pointer over different nodes will show the corresponding pathway names. Clicking the node of interest will launch the corresponding pathway view on the right panel (if nothing is clicked, the right panel will remain blue). Users can zoom or drag to focus on a section of the pathway by clicking on the navigation/manipulation tools at the bottom of the panel. Clicking on any matched compound node (colored yellow, orange or red) will generate a pop-up diagram showing the corresponding compound which contains a detailed summary of the compound concentrations, importance measures, as well as the p-value. Clicking the pathway name hyperlink at the top of the right panel will direct users to the corresponding pathway (either KEGG or SMPDB, depending on the initial choice of the pathway library). The lower table lists the pathway names (clicking a pathway name will update the pathway diagram on the upper right), the matched metabolites (which will be displayed in a pop-up window as a highlighted list), the statistical values (p, log-p, FDR, impact), and links to two different pathway databases (SMPDB and KEGG).

15. When finished exploring the results, click the "Download" hyperlink on the navigation tree to create and download the analysis report as well as all the graphical and textual results.

    The pathway images will only be generated when users click the corresponding pathway nodes on the pathway view. Users should spend some time exploring various pathways of interest before going to the download page.

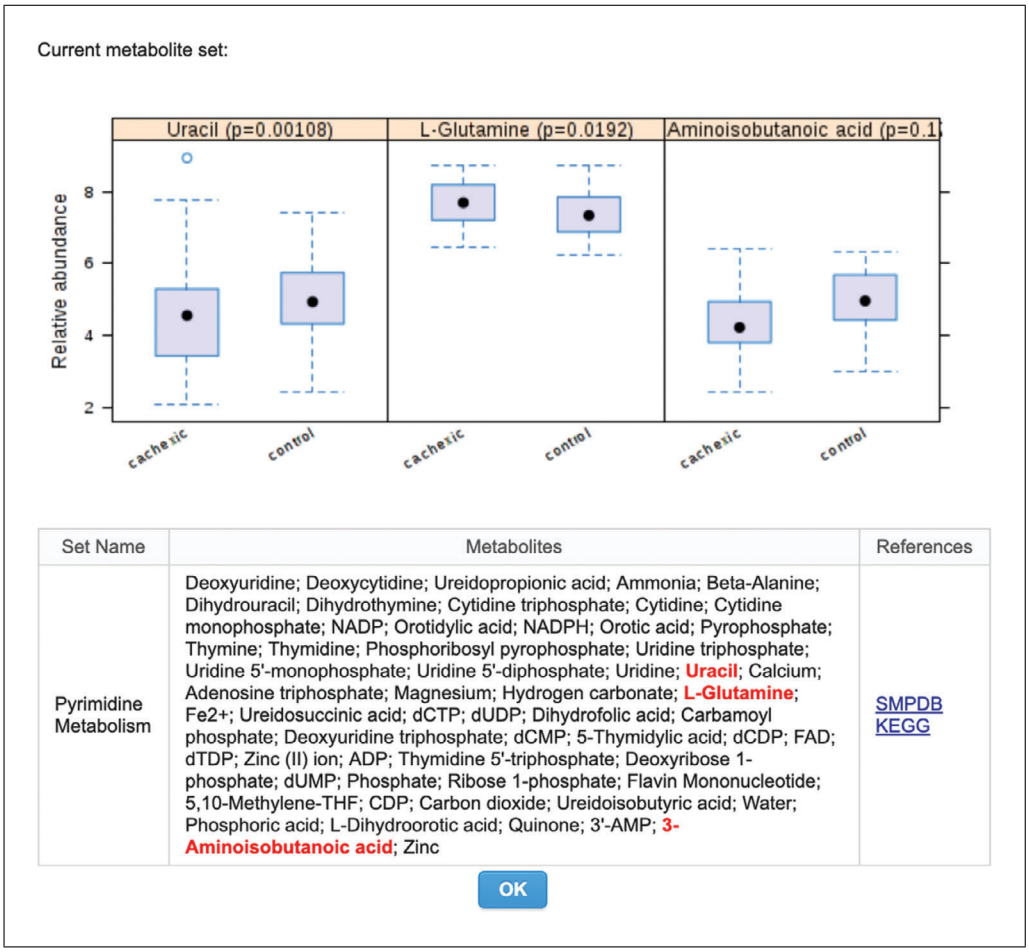


Figure 50 A screenshot showing the details of a matched metabolite set with the metabolite set plot on the top. Please refer to the text for further details.

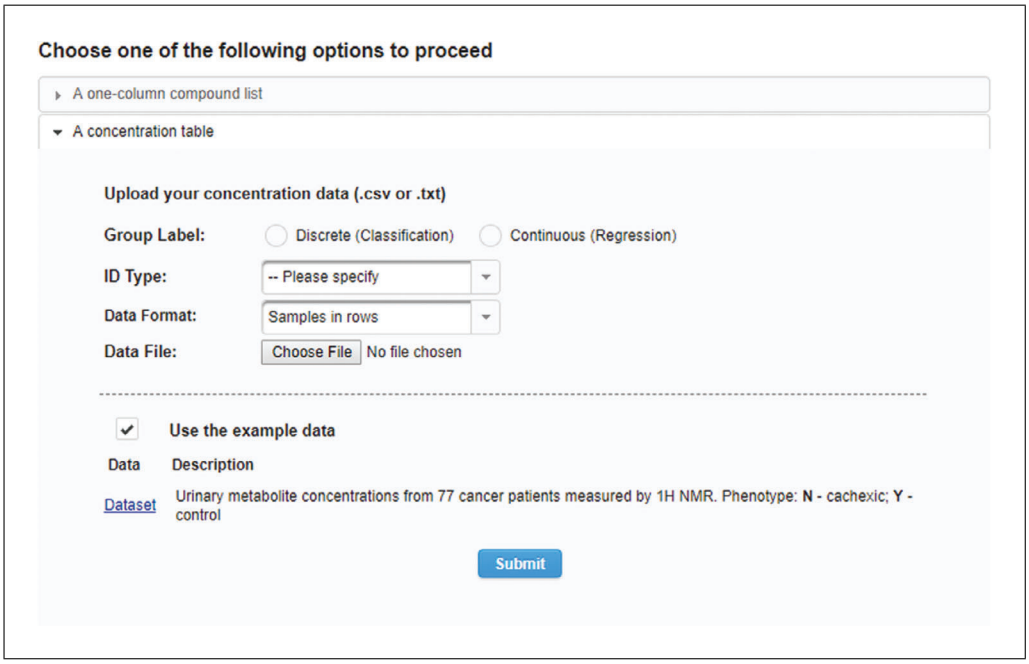


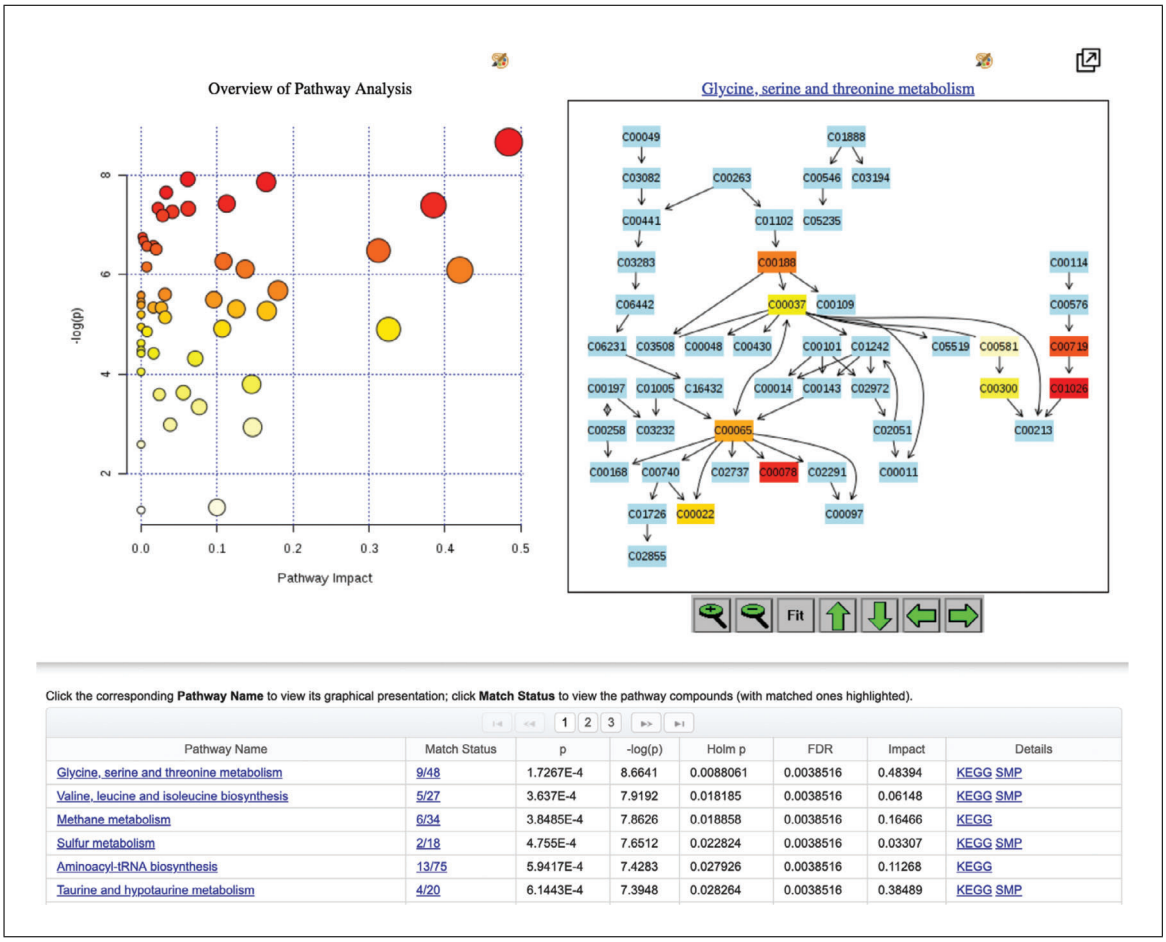
Figure 51 The “Data Upload” page for the Pathway Analysis module. Users can upload either a list of metabolites or a metabolite concentration table.

Figure 52 A screenshot showing the parameter settings for pathway analysis. Please refer to the text for further details.

BASIC PROTOCOL 5

BIOMARKER ANALYSIS BASED ON RECEIVER OPERATING CHARACTERISTIC (ROC) CURVES

Current metabolomics studies can be put into two general categories: (1) those that aim to understand biological processes, and (2) those that aim to identify biomarkers. Studies in the first group focus primarily on gaining improved biological understanding through the analysis of metabolite profiles using various univariate and multivariate statistical methods. Methods such as MSEA or pathway analysis are usually employed to understand the biological processes involved in the conditions of interest. These tasks are well supported in MetaboAnalyst and have been described in great detail in the previous four basic protocols. Studies in the second group (i.e., biomarker studies) have been



Pathway Name	Match Status	p	-log(p)	Holm p	FDR	Impact	Details
Glycine, serine and threonine metabolism	9/48	1.7267E-4	8.6641	0.0088061	0.0038516	0.48394	KEGG SMP
Valine, leucine and isoleucine biosynthesis	5/27	3.637E-4	7.9192	0.018185	0.0038516	0.06148	KEGG SMP
Methane metabolism	6/34	3.8485E-4	7.8626	0.018858	0.0038516	0.16466	KEGG
Sulfur metabolism	2/18	4.755E-4	7.6512	0.022824	0.0038516	0.03307	KEGG SMP
Aminoacyl-tRNA biosynthesis	13/75	5.9417E-4	7.4283	0.027926	0.0038516	0.11268	KEGG
Taurine and hypotaurine metabolism	4/20	6.1443E-4	7.3948	0.028264	0.0038516	0.38489	KEGG SMP

Figure 53 A screenshot showing the results from a pathway analysis, with the left panel showing the "metabolome view" and the right the pathway view. Clicking any matched compound node on the pathway will launch the corresponding compound view. The numerical details are given by the table below.

focused primarily on identifying metabolite biomarkers for disease diagnosis, predictive markers, prognostic markers, and markers of exposure. To address this critical need, an ROC curve based approach for biomarker analysis has been developed (Xia, Broadhurst, Wilson, & Wishart, 2013), and is now part of MetaboAnalyst 4.0

A promising biomarker or biomarkers must have high sensitivity (i.e., to give a positive test result when the disease is present) and high specificity (i.e., to give a negative test result when the disease is absent). ROC curve analysis is widely used to describe the trade-off between sensitivity and specificity with regard to biomarker performance. ROC curves allow one to see how the diagnostic performance varies along all possible threshold values. The area under the ROC curve (AUROC) value is a robust measure for comparing the performance of different biomarker models. ROC-curve biomarker analysis is generally considered the best method for developing and assessing the performance of medical diagnostic tests (Xia et al., 2013). The MetaboAnalyst biomarker module supports three common ROC-based analysis modes: (1) classical univariate ROC curve analysis, (2) multivariate ROC curve exploration, and (3) manual biomarker model creation and evaluation. This protocol will provide a detailed description of these three analytical modes.

Necessary Resources

Hardware

A computer with internet access

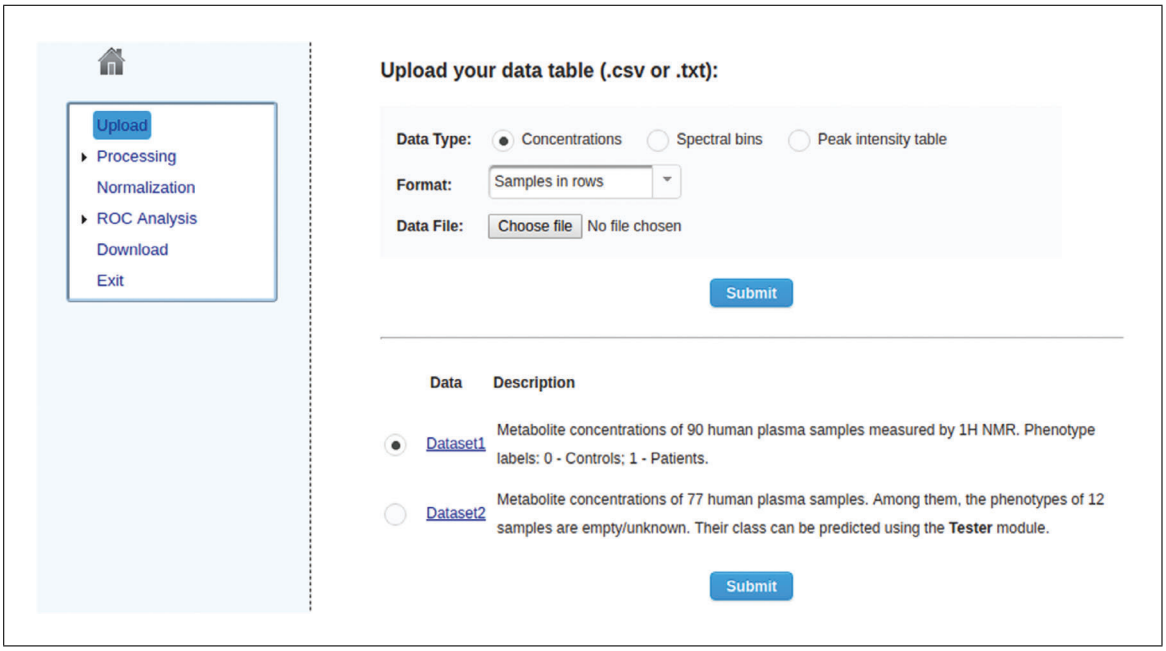


Figure 54 The "Data Upload" page for the Biomarker Analysis module. Users need to upload a data table containing two groups (for binary comparison). Note that the navigation tree or menu is located on the left panel with the current step highlighted (dark blue).

Software

An up-to-date web browser, such as Google Chrome (*http://www.google.com/chrome*), Mozilla Firefox (*http://www.mozilla.com/*), Safari (*http://www.apple.com/safari/*), or Internet Explorer (*http://www.microsoft.com/ie/*). JavaScript must be enabled in the web browser.

Files

None

Data upload and processing

1. Go to the MetaboAnalyst home page (*https://www.metaboanalyst.ca*). Click on "click here to start" at the top of the page to enter the "Module Overview" page, and click "Biomarker Analysis" button on the top left side of the circle. On the "Data Upload" page, click on the *Dataset1* radio button on the bottom half of the page and click the "Submit" button (Fig. 54).

   This test data set contains metabolite concentrations of 90 human plasma samples (60 healthy pregnant controls and 30 pregnant patients with pre-eclampsia) measured by ^{1}H NMR. Alternatively, users can download this data set and then upload it to MetaboAnalyst. To do this, right click the "Dataset1" and choose "Save Link As .." to save the CSV file into your computer. To re-upload the data, specify the data type as "Concentrations." Make sure "Samples in rows" is selected for this data format. Browse through one's file manager to select the downloaded data, and then click "Submit" to upload the data. Note that ROC curve analysis is only applicable for two group comparisons. Users need to edit or re-label their data when there are multiple groups.

2. The "Data Integrity Check" page indicates that this data passed all the integrity checks with a total of five missing values detected. Click the "Skip" button to go to the "Data Normalization" page.

   By default, missing values will be replaced by a very small value, assuming that the corresponding concentrations are below the detection limit. MetaboAnalyst replaces them with half the lowest concentration present in the dataset. When there is a large

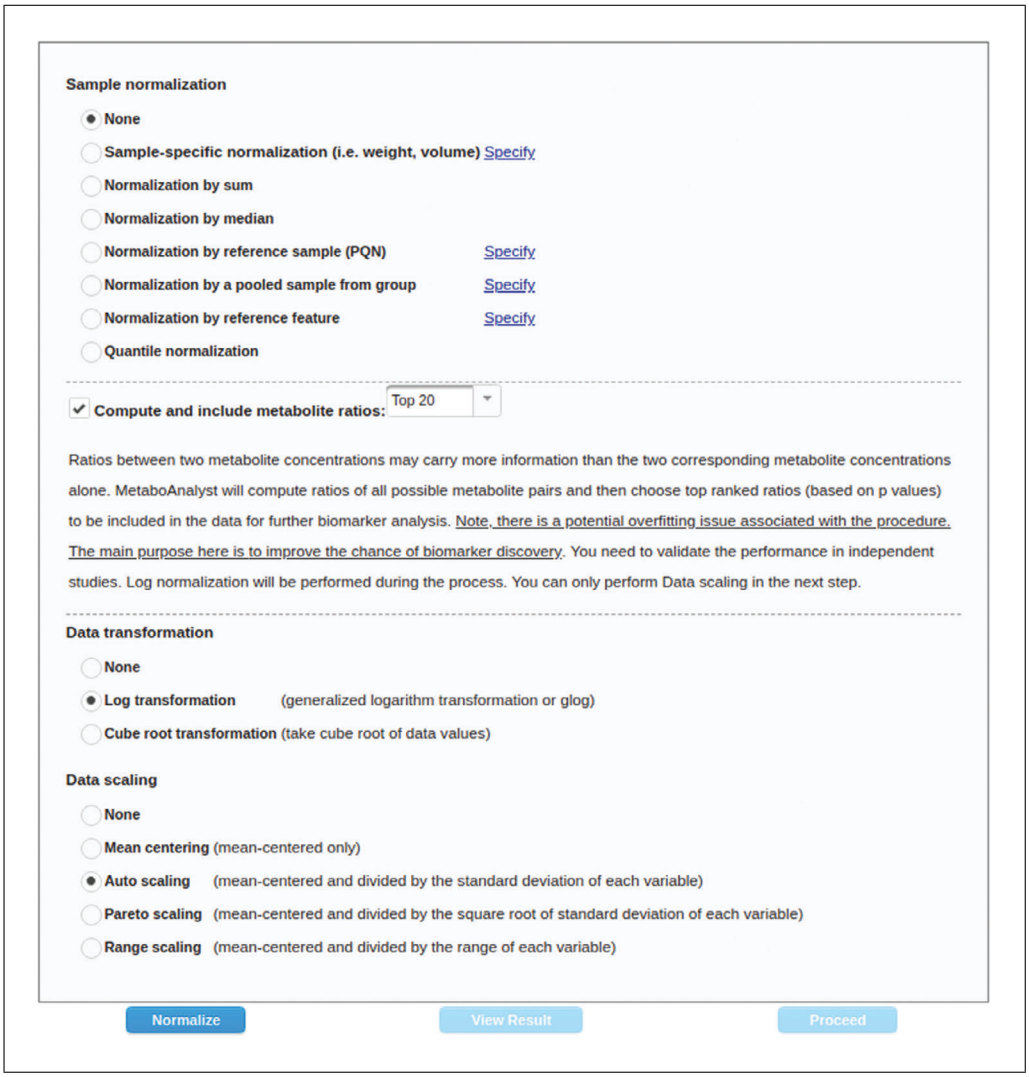


Figure 55 A screenshot of the "Data Normalization" page for biomarker analysis. Note the "Compute and include metabolite ratios" option is selected for biomarker discovery rather than for performance evaluation.

fraction (>25%) of missing values present, users are advised to apply "Low quality data filtering" and "Missing value estimation." The corresponding instructions have been described in steps 6 to 8 of Basic Protocol 1.

3. On the "Data Normalization" page, keep "None" for the Sample normalization, check "Compute and include metabolite ratios" with the default set to Top 20, select "Log transformation" for Data transformation, and select "Auto scaling" for Data scaling. Click the "Submit" button at the bottom of the page to proceed. A screenshot is shown in Figure 55.

   Some studies have suggested that metabolite ratios can be more useful as biomarkers. MetaboAnalyst can compute pair-wise ratios between all possible metabolite concentrations and then incorporate those top ranked ratios (i.e., based on p-values from t-tests) into a concentration data table for subsequent analysis. Note that due to the potential issue of over fitting, this procedure is mainly designed for biomarker discovery rather than performance evaluation.

4. The graphical summary displayed on the "Data Normalization Results" page shows that the data look reasonably "bell-shaped" after the normalization procedure. After clearing the pop-up window by clicking the "X" on its top corner, click the "Proceed" button at the bottom of the "Data Normalization" page. The "ROC Analysis Options"

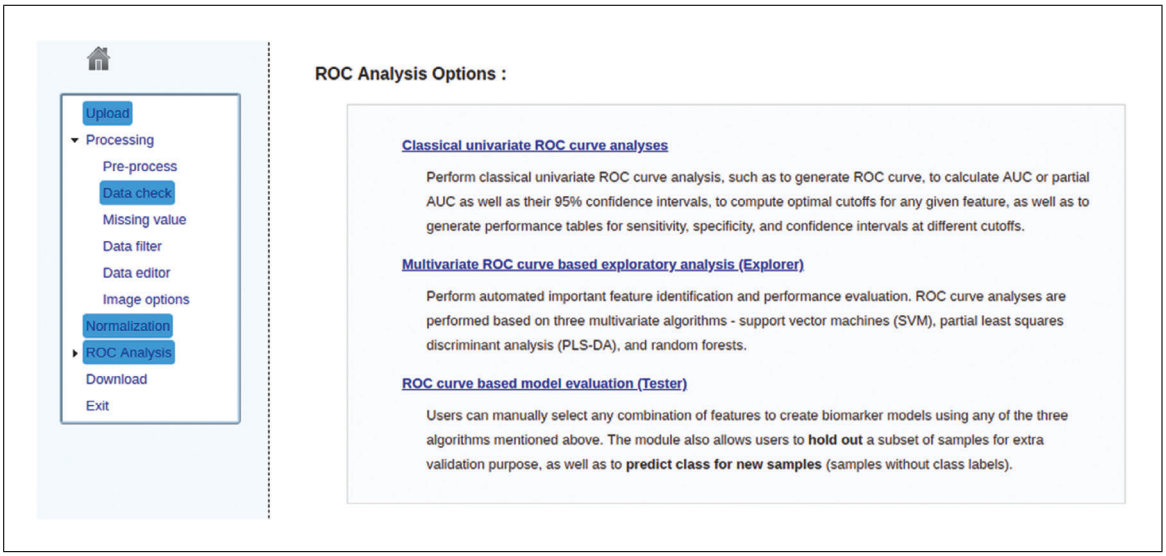


Figure 56 The "ROC Analysis Options" page showing the three main ROC analysis paths.

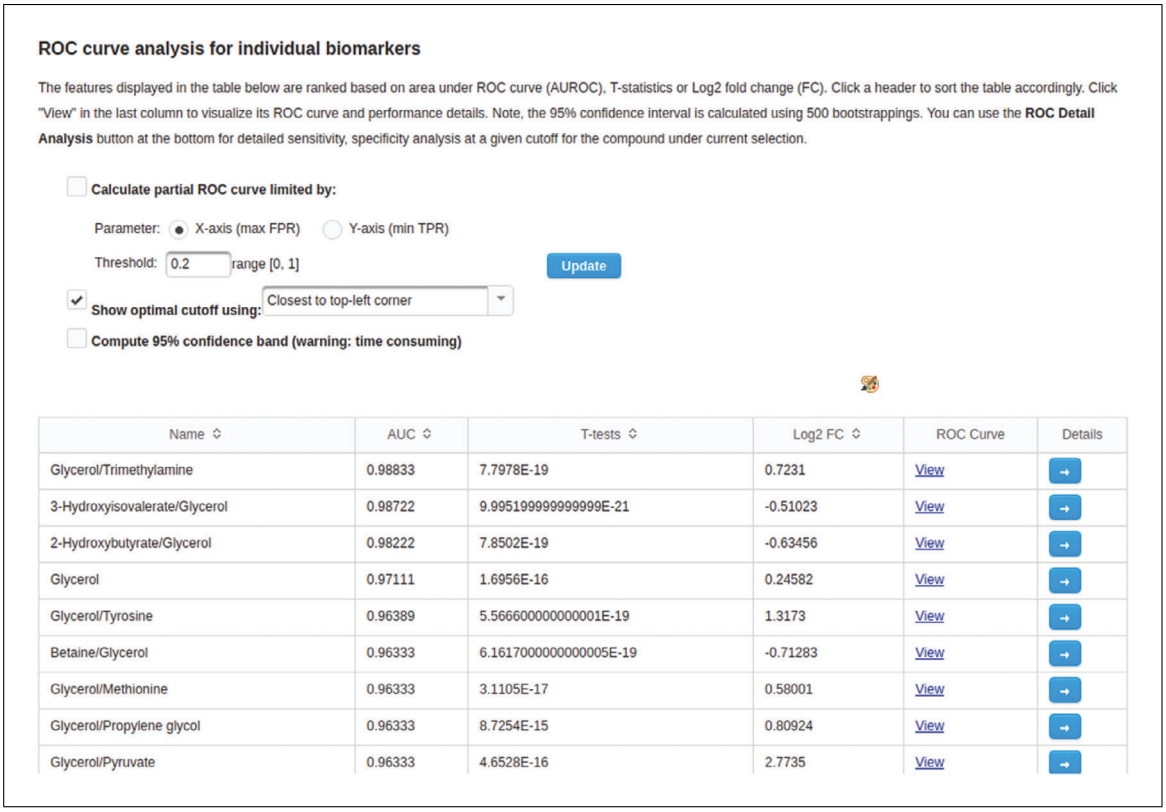


Figure 57 A screenshot showing an ROC curve analysis result for individual biomarkers. Users can visualize the ROC curve and perform a more detailed analysis for each potential biomarker.

page should be displayed (Fig. 56), showing the three ROC analysis modes: (1) classical univariate ROC curve analysis, (2) multivariate ROC curve explorer, and (3) ROC curve-based model evaluation.

Classical univariate ROC curve analysis

5. Click the "classical univariate ROC curve analysis" link on the page. MetaboAnalyst will perform ROC-curve analysis for each feature in the data set. The results will be displayed as a table containing all features ranked by their AUROC values. A screenshot is shown in Figure 57.

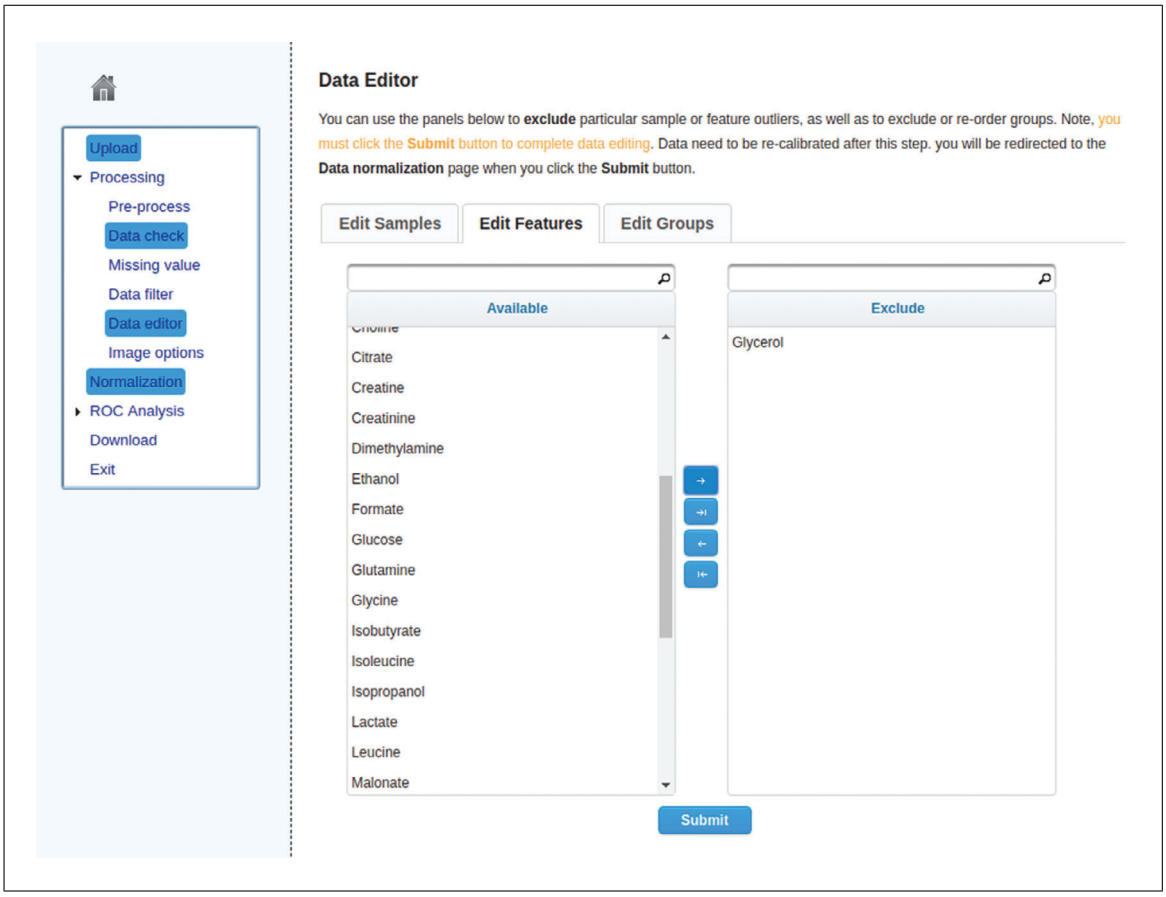


Figure 58 A screenshot of the "Data Editor" page. This illustrates how to exclude a metabolite (Glycerol) from further analysis. Make sure to click the "Submit" button at the bottom to complete the procedure.

In this case, glycerol and metabolite ratios involving glycerol dominate the first page of biomarkers, with AUROC values all close to 1. Glycerol is a well-known common source of contamination (from lubricants or from ultrafiltration devices). For demonstration purposes, exclude it from further analysis for the remainder of this tutorial.

6. Click to expand the "Processing" hyperlink on the MetaboAnalyst navigation tree (left side panel), then locate and click the "Data editor" hyperlink to enter the corresponding page. The "Data editor" page contains three tabs to allow users to exclude samples, features, or groups. In this case, click the "**Edit Features**" tab. In the list box, select the *glycerol* from the "Available" panel, and click the right-hand arrow to move it to the "Exclude" panel. The result is shown in Figure 58. Click the "Submit" button to perform the actual feature exclusion. Users will then be re-directed to the Normalization page. Please repeat steps 3 to 5 to get back to the "ROC curve analysis for individual biomarkers" page.

   After editing the data, it is always advisable to re-do the data processing and normalization steps to ensure that the updates are propagated to each stage of the analysis.

7. In the updated table, the metabolite *acetate* seems to give the best performance, with an AUC ~0.83. Using metabolite ratios further improves the performance, with the top one being *3-hydroxyisovalerate/acetate*, with an AUC ~0.89. To view the corresponding ROC curve together with the 95% confidence intervals, select the "Compute 95% confidence band" checkbox and click the corresponding "View" button in the first row. The confidence interval is computed using 500 bootstrap replications. The result is shown in Figure 59.

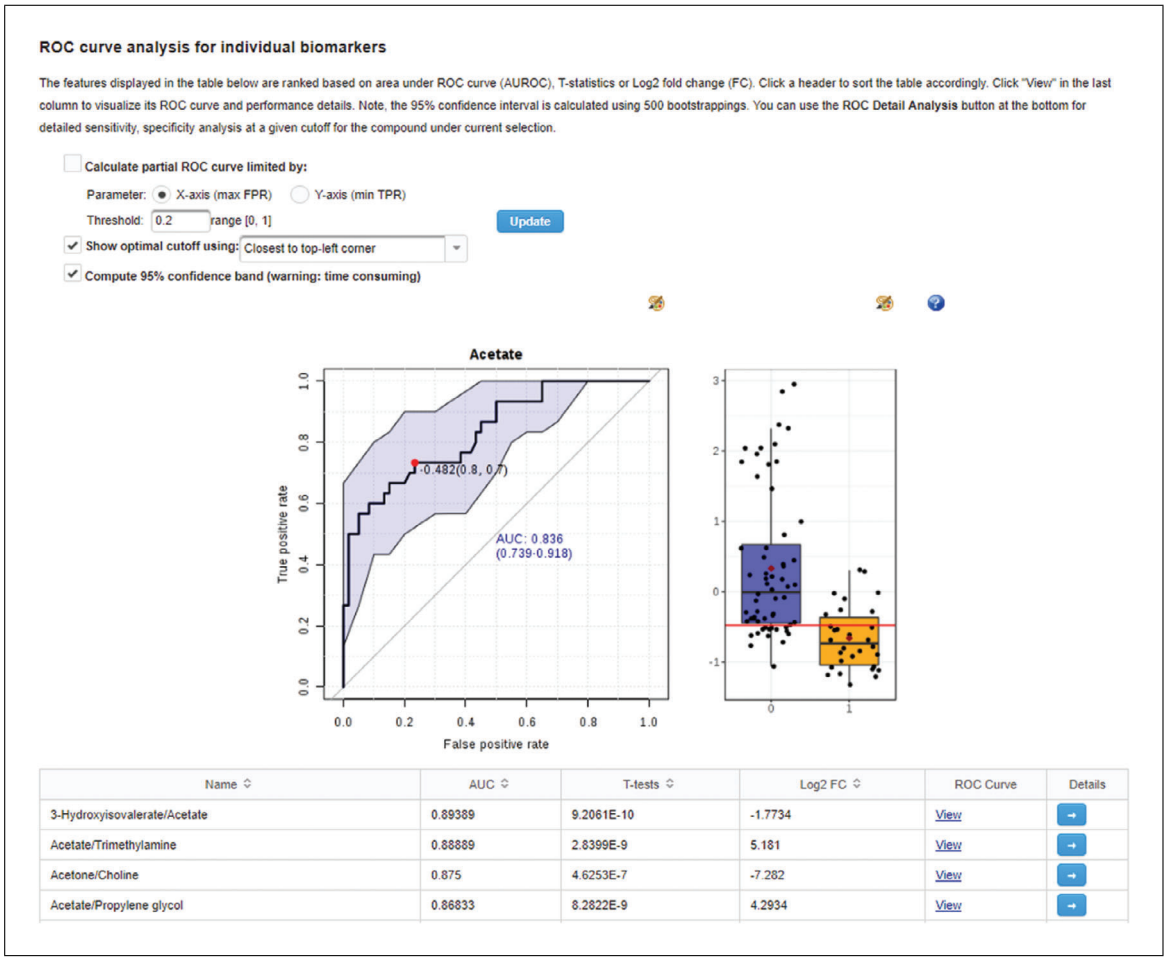


Figure 59 A screenshot of the ROC curve and associated boxplot for a potential biomarker based on compound ratios.

The left-hand plot shows the ROC curve with a 95% confidence interval in shadow. The solid red dot indicates the mathematically optimal cutoff with the associated sensitivity and specificity values. The optimal cutoff can be computed using either the point closest to the top-left corner (default) or the one farthest from the diagonal line (Youden). The right-hand-side plot is the box-and-whisker plot showing the distribution of abundance values between the two groups. The optimal cutoff is indicated as a horizontal dotted red line.

8. The default mathematically optimal cutoff may not be appropriate in some cases, and a user may want to further explore the sensitivities and specificities at different cutoffs. To do this, click the right-hand arrow in the last column ("Details") for the first feature. On the new page, the default ROC image is shown at the top of the page. The "Query ROC" panel below allows users to easily explore the cutoff-sensitivity/specificity relationship. For instance, to get the minimal cutoff value to reach 100% sensitivity, enter 1.0 in the input area for Sensitivity and click the "Submit" button. The corresponding result is shown in Figure 60. The table at the bottom indicates the sensitivity and specificity values as well as the positive and negative likelihood ratios at different cutoff values. Users can download this table using the corresponding link at the bottom of the table, or click the "Download" hyperlink on the navigation tree (left side panel) to download all the results.

It is important to be aware that MetaboAnalyst only performs its analyses based on users' requests. This is particularly true for image generation. Therefore, users are advised to perform ROC curve analysis and visualization for metabolites of interest (and their ratios) before they download their data. All the images and analysis results, as well as

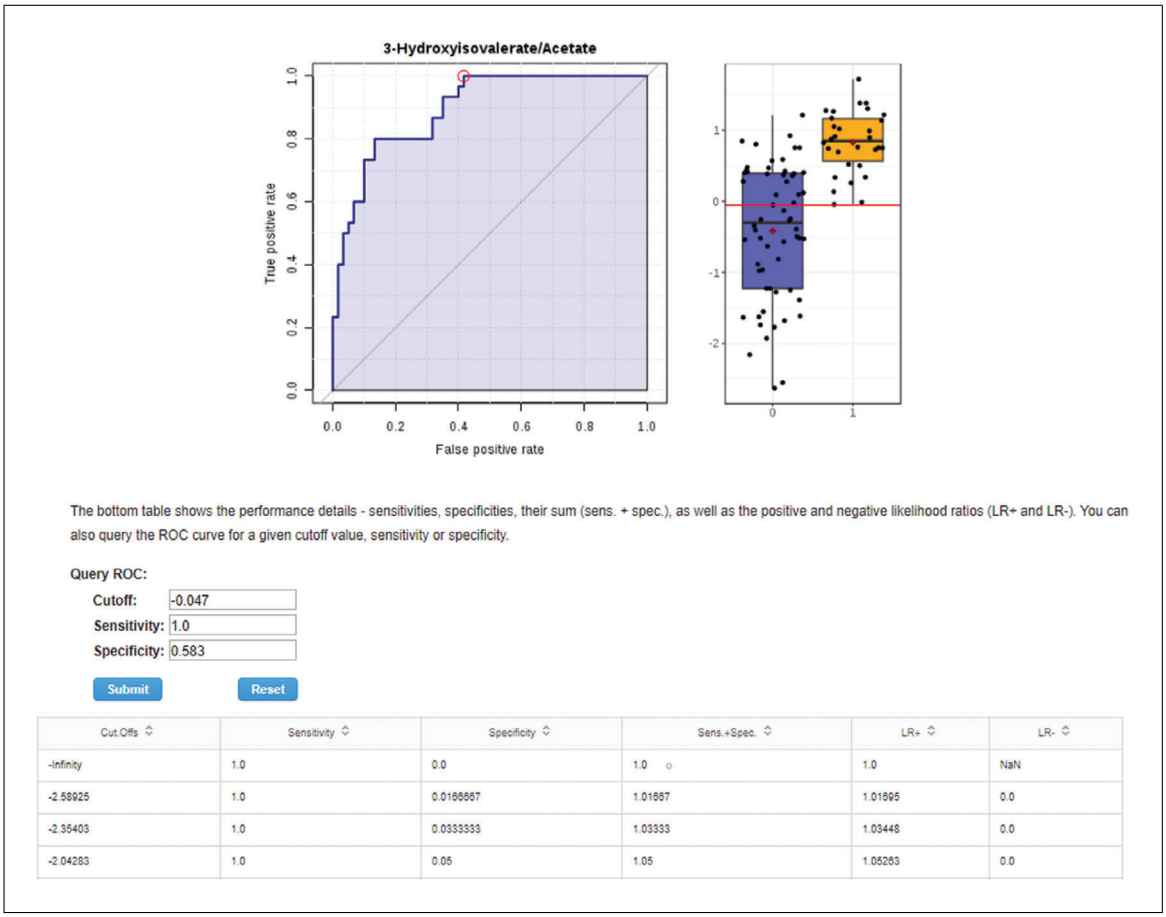


Figure 60 A screenshot of the "Detailed ROC Curve Analysis" page. Users can enter the specific values to query a ROC curve about a corresponding cutoff as well as sensitivity and specificity values. The results will be computed and highlighted on the graph.

the normalized data containing metabolite ratios, can be downloaded as a single zip file from the Download page.

Multivariate biomarker identification and performance evaluation

9. Click the "Normalization" hyperlink on the navigation tree (the menu on the left side) to return to the "Data Normalization" page. In this case, uncheck the "Compute and include metabolite ratios" option and re-perform normalization using "Log transformation" and "Auto scaling."

   The previous procedure of computing and including the top metabolite ratios is designed to increase the chance for identification of promising individual biomarkers. However, the procedure is based on information using the complete dataset, which will slightly increase the risk of over-fitting. This can lead to over-optimistic performance estimations based on cross validations. For demonstration purposes, we will choose to focus on the metabolite concentration data only.

10. On the "ROC Analysis Options" page, click the "Multivariate ROC curve based exploratory analysis (Explorer)" link. The next page shows different options for various multivariate or machine-learning algorithms to perform biomarker identification/prediction (Fig. 61). In this case, leave the default "Linear SVM" for the classification method and "SVM built-in" for the feature ranking method. Click the "Submit" button to proceed.

    In general, simple models based on a small number of biomarkers are preferred over complex models using many biomarkers. Simple biomarker models are more robust and cost-effective and less prone to over-fitting. To identify such models with good

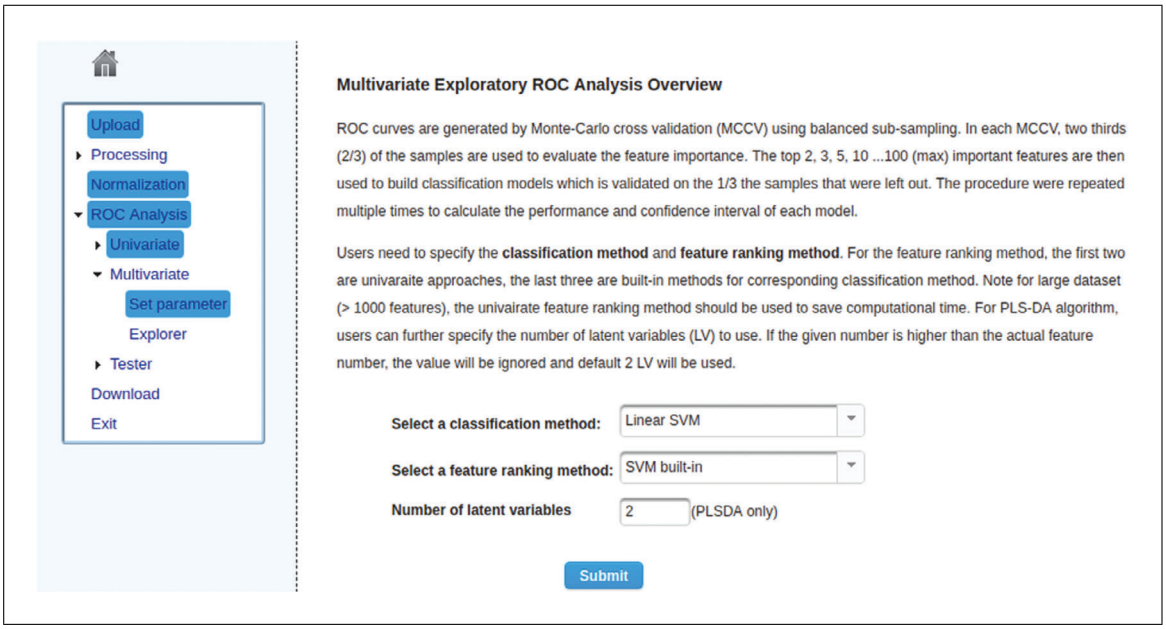


Figure 61 A screenshot of the "Multivariate Exploratory ROC Analysis Overview" page. Users can select different classification algorithms and feature selection algorithms to build specific biomarker models.

performance, MetaboAnalyst uses an algorithm based on Monte-Carlo cross validation (MCCV) through balanced subsampling. The framework is coupled with three well-established algorithms—partial least squares discriminant analysis (PLS-DA), Support Vector Machine (SVM), and Random Forest (RF). In each MCCV, two-thirds of the samples are used to evaluate feature importance. The top 2, 3, 5, 10,... 100 (max) important features are then used to build appropriate classification models that are validated on the one-third of the samples that were left out. The procedures are repeated multiple times to calculate the performance and confidence interval for each model.

11. After a few seconds, the "Results" page will be displayed containing multiple views organized under different tabs (Fig. 62). The first tab, "**ROC View**," provides an overview comparing the ROC curves for all models created by MetaboAnalyst. On this page, users can also choose to visualize the ROC curve for a model. For instance, model #3 (five features) gives relatively good performance with just five features. Select this model and click "Update" to visualize the corresponding ROC curve.

12. Click the "**Prob. View**" tab to see detailed predictions on the class probabilities with regards to different samples in the cross-validation assessments for Model #3. Note that, due to balanced subsampling, the optimal cutoff threshold will always be at the center (0.5). Users can check the option "Label samples classified to the wrong groups" to show those sample names that tend to be misclassified. The confusion matrix summarizes the predictions in each category based on the cross-validation. A screenshot is shown in Figure 63.

    MetaboAnalyst does not provide sensitivity and specificity values, as users can define positive and negative cases using different labels (0 versus 1, control versus case, normal versus abnormal, etc.). However, users can easily calculate sensitivity and specificity values from the confusion matrix. In a confusion matrix, the column names are the actual class labels, while the row names are the predicted class labels. In this case, assuming 1 is for a positive case and 0 is for a negative case, the sensitivity is 0.97 (29/30) and the specificity is 0.82 (49/60).

13. Click the "**Sig. Features**" tab to see a list of features that are important for constructing the biomarker model. Figure 64 shows the top 15 metabolites ranked by their frequencies of being selected for the 2-feature panel of Model #1.

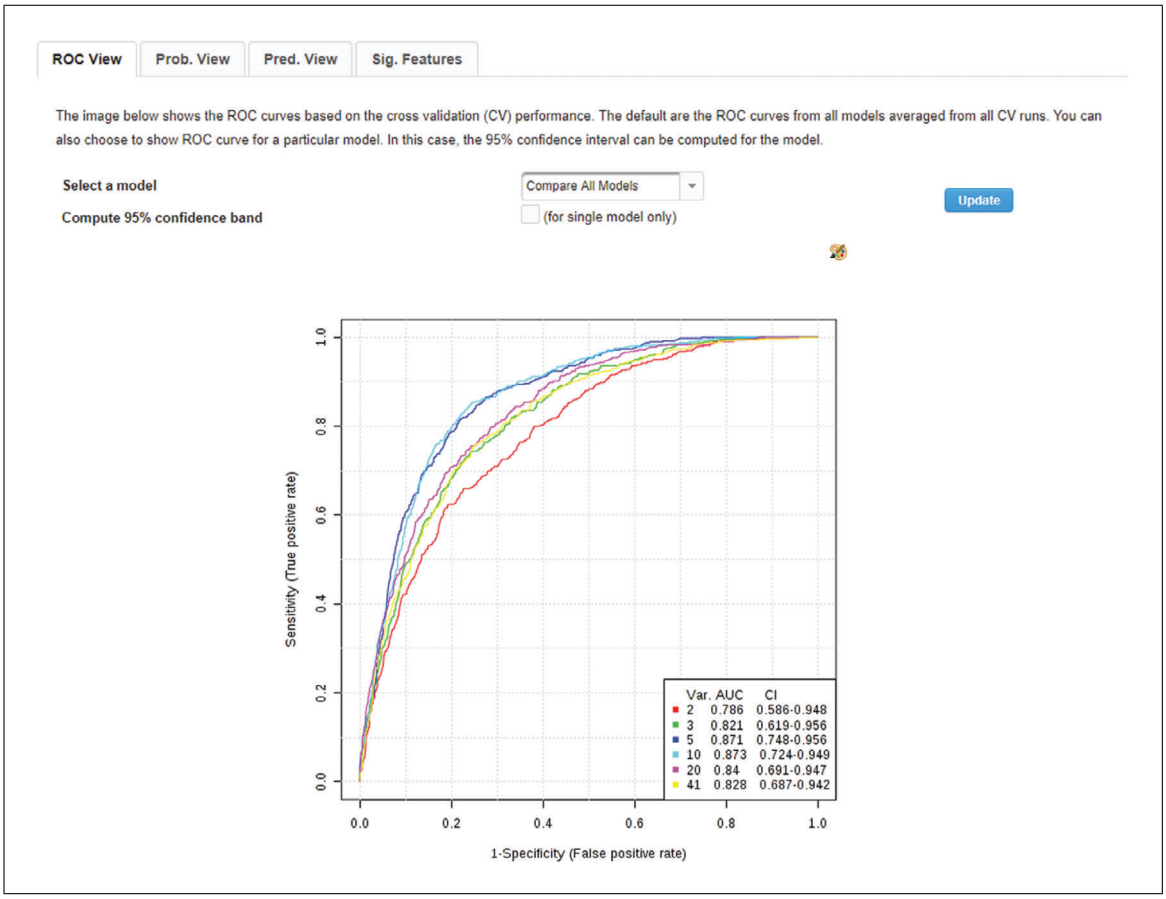


Figure 62 A screenshot showing an overview of ROC curves from different biomarker models using different numbers of features. Users can choose to visualize the ROC curve for an individual model using the drop-down menu beside "Select a model."

The processes of feature selection, model building, and performance evaluation are performed multiple times through MCCV. The most important features can be slightly different using different subsets of data. Figure 64 shows that Choline is always selected in the model, while 3-hydroxyisovalerate is selected 90% of the time based on the SVM feature selection algorithm.

14. After completing the exploratory multivariate biomarker analysis, click the "Download" hyperlink on the navigation tree to download all the results, the analysis report, and images.

Manual biomarker model creation and performance evaluation

15. Click the "Upload" hyperlink on the navigation tree (left side panel) to return to the "Module Overview" page, and then click the "Biomarker Analysis" button to re-enter the "Data Upload" page. In this case, select the "Using the example data" checkbox and make sure that *Dataset2* is chosen (Fig. 65). Click the "Submit" button.

This analysis mode allows users to manually create a biomarker model to predict class labels for new samples (indicated by leaving class labels empty). Dataset2 contains 77 human urine samples obtained from cancer patients with and without cachexia, and 12 unlabeled samples. Including samples with known labels together with new (unlabeled) samples will allow them to be processed together to significantly reduce potential problems due to batch effects. Batch effects tend to have a strong negative impact in biomarker analysis, especially for studies with small sample sizes.

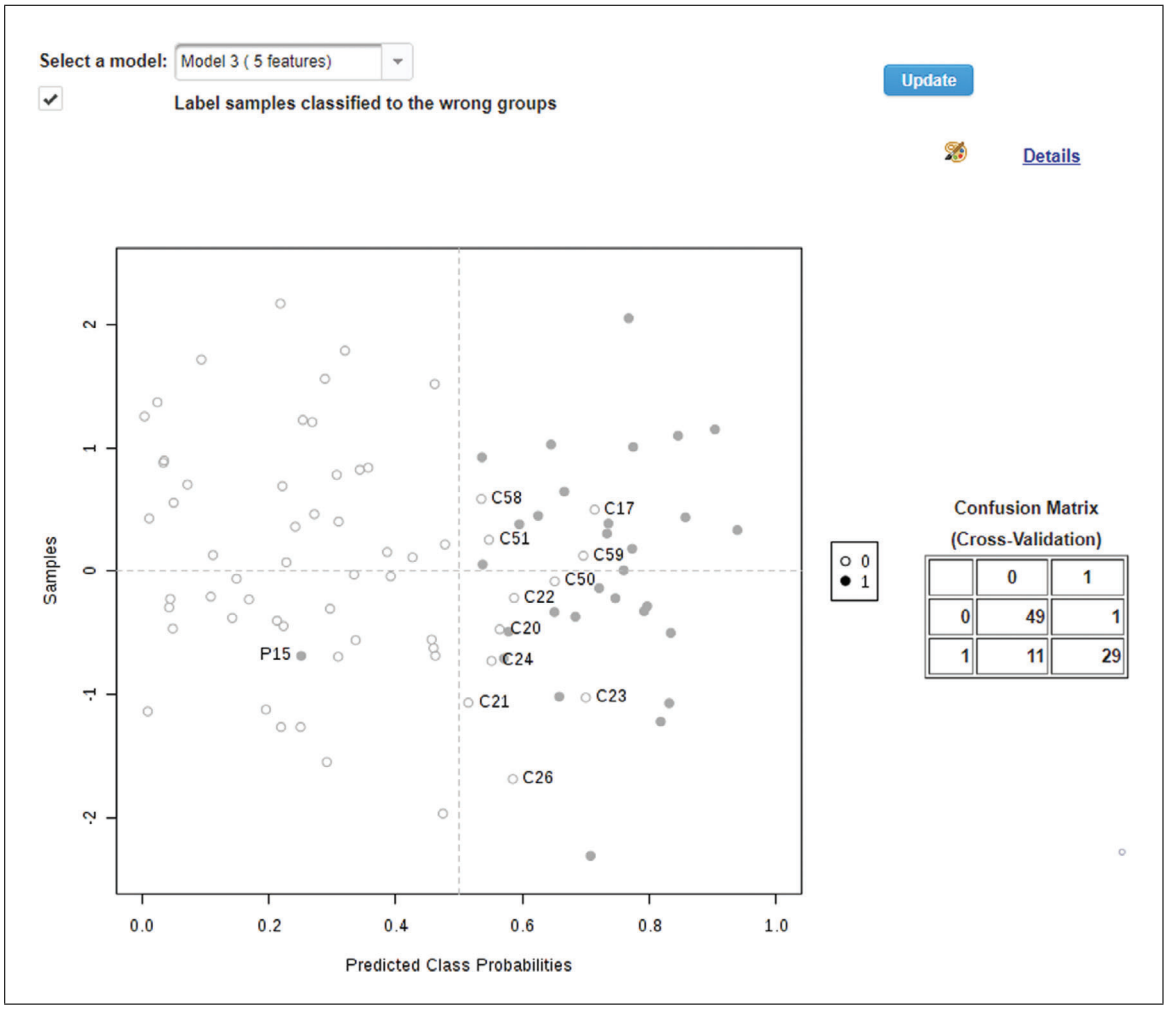


Figure 63 A screenshot showing the predicted class probability for each sample and the performance summary of a selected model (confusion matrix). Samples predicted in the wrong groups are labeled. Users can compute the sensitivity and specificity based on the summary table and their own definition of positive and negative cases.

16. The message from the "Data Integrity Check" page confirms that 12 new samples have been detected. Click the "Skip" button to go to the "Data Normalization Page." Choose "Log transformation" and leave all other options as their default values. Click the "Normalize" button at the bottom of the page and then press the "View Result" button to visualize the results on a pop-up window. The graph on the upper right side should look reasonably bell-shaped. After clearing the pop-up window by clicking the "X" on its top corner, click the "Proceed" button at the bottom of the "Data Normalization" page to move to the "ROC Analysis Options" page. In this case, select the "ROC curve-based model evaluation (Tester)" to enter the "ROC Builder" page. This multi-tab page allows users to manually select variables/features for model creation, to hold out certain samples for validation, and to display new samples (if detected). On the "**Variable Selection**" tab, a table will be displayed showing the available metabolites, their AUCs, T-statistics, Fold changes, and KM cluster sizes. Using the check boxes on the left side of the table, choose *betaine*, *N,N-dimethylglycine*, *glucose*, and *adipate*, which all show good AUCs and relatively high fold changes, and belong to different KM Clusters (Fig. 66). Click the "Next" button at the bottom of the page to enter the "ROC Evaluator" page.

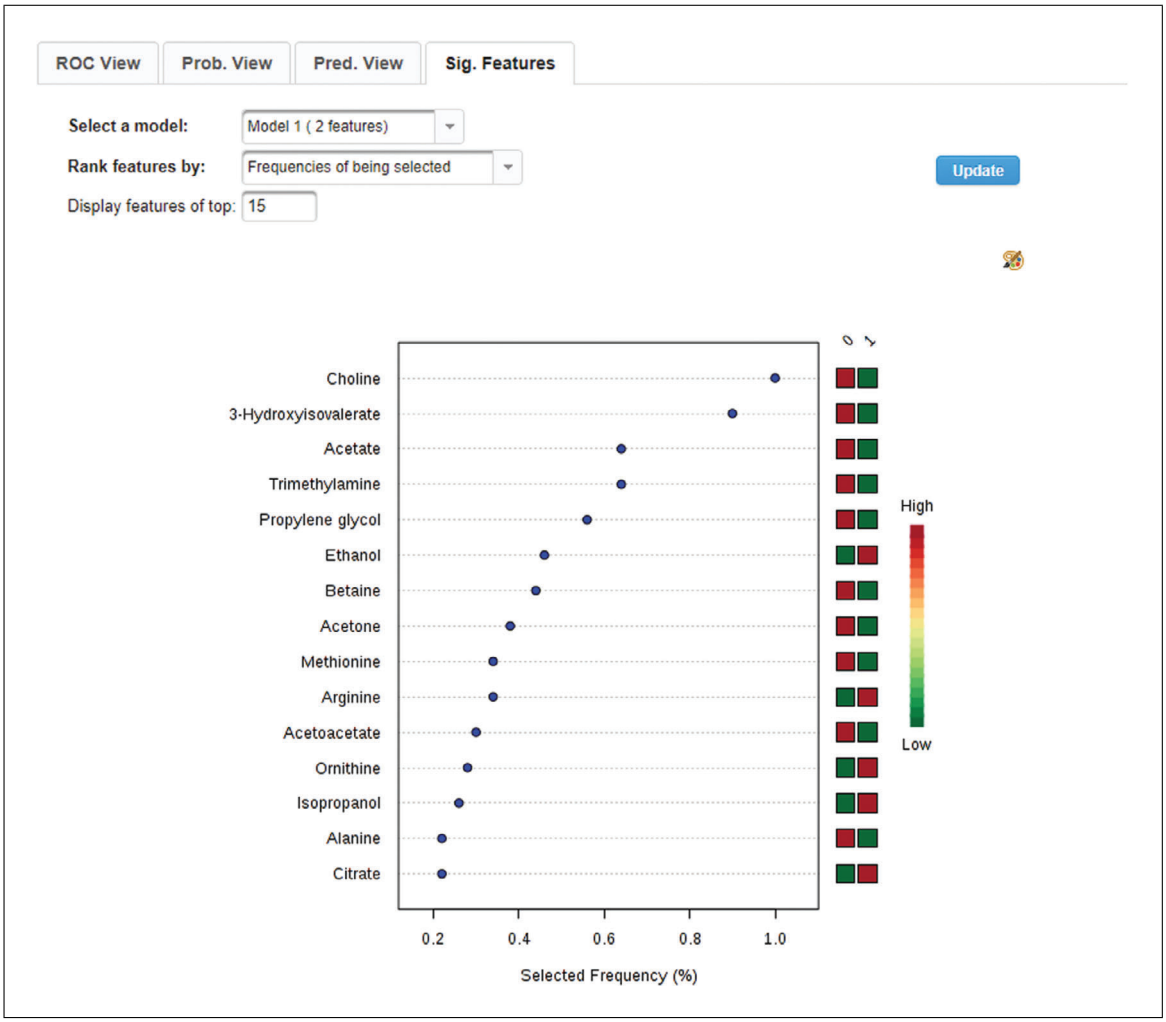


Figure 64 The top 15 significant features ranked based on their frequencies of being selected during cross validation.

K-means (KM) clustering is designed to help reduce the redundancy in biomarker selection, as features in the same cluster tend to behave more similarly. Combining features in different clusters may provide complementary information to improve the performance of a multivariate biomarker model.

17. This page shows the performance of the selected four-biomarker model created using linear SVM (the default algorithm) with an AUC of 0.779. Use the top drop-down menu to select the PLS-DA algorithm, and press the "Submit" button. This selection gives a slightly better AUC value of 0.79. This page also provides other information, including cross validation results and prediction probabilities, to help users get a better idea of the biomarker model. These have been described in more detail in steps 11 and 12. Click the "**New Sample Prediction**" tab (on the far right) to see the predicted class labels and the associated probability scores for the 12 new samples (which have labels such as PIF_141 and NETCR_016_V1) (Fig. 67).

18. Upon finishing this data analysis and visualization protocol, click the "Download" hyperlink on the navigation tree to create analysis report and download all the results and images.

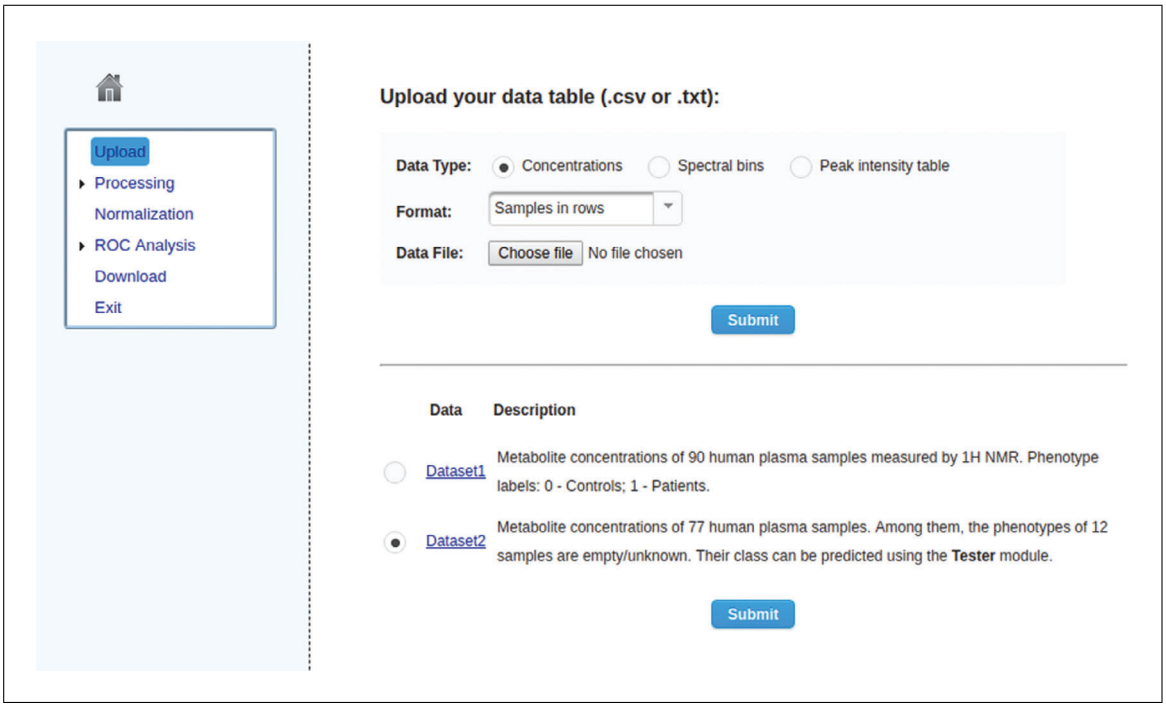
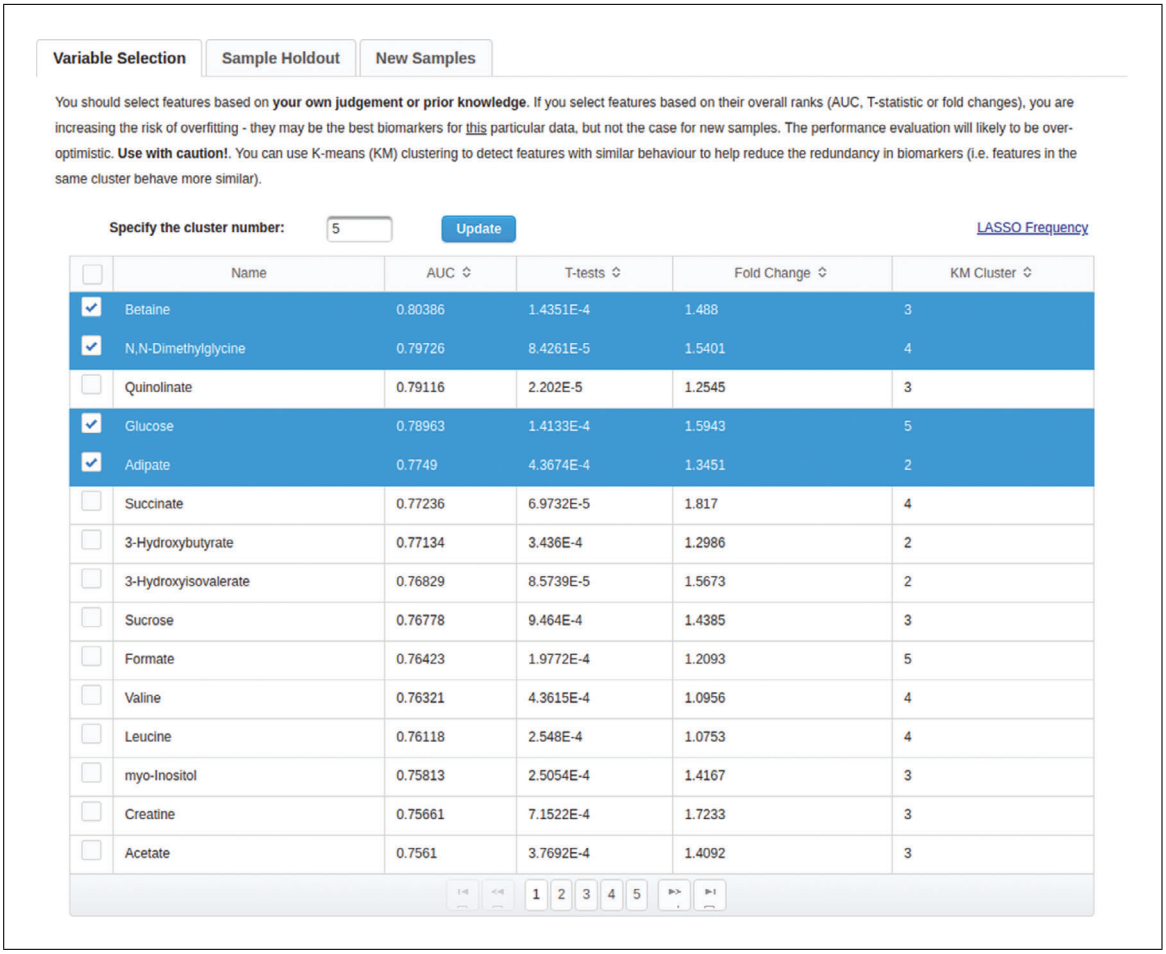


Figure 65 A screenshot showing the "Data Upload" page for the ROC Tester in the Biomarker Module. Note the new samples are indicated in the uploaded dataset by having their class labels empty.



Name	AUC	T-tests	Fold Change	KM Cluster
Betaine	0.80386	1.4351E-4	1.488	3
N,N-Dimethylglycine	0.79726	8.4261E-5	1.5401	4
Quinolinate	0.79116	2.202E-5	1.2545	3
Glucose	0.78963	1.4133E-4	1.5943	5
Adipate	0.7749	4.3674E-4	1.3451	2
Succinate	0.77236	6.9732E-5	1.817	4
3-Hydroxybutyrate	0.77134	3.436E-4	1.2986	2
3-Hydroxyisovalerate	0.76829	8.5739E-5	1.5673	2
Sucrose	0.76778	9.464E-4	1.4385	3
Formate	0.76423	1.9772E-4	1.2093	5
Valine	0.76321	4.3615E-4	1.0956	4
Leucine	0.76118	2.548E-4	1.0753	4
myo-Inositol	0.75813	2.5054E-4	1.4167	3
Creatine	0.75661	7.1522E-4	1.7233	3
Acetate	0.7561	3.7692E-4	1.4092	3

Figure 66 A screenshot showing the "Variable Selection" tab of the ROC Builder page. *Betaine*, *N,N-Dimethylglycine*, *Glucose*, and *Adipate* were selected based on their AUC, fold-change values, and KM Cluster.

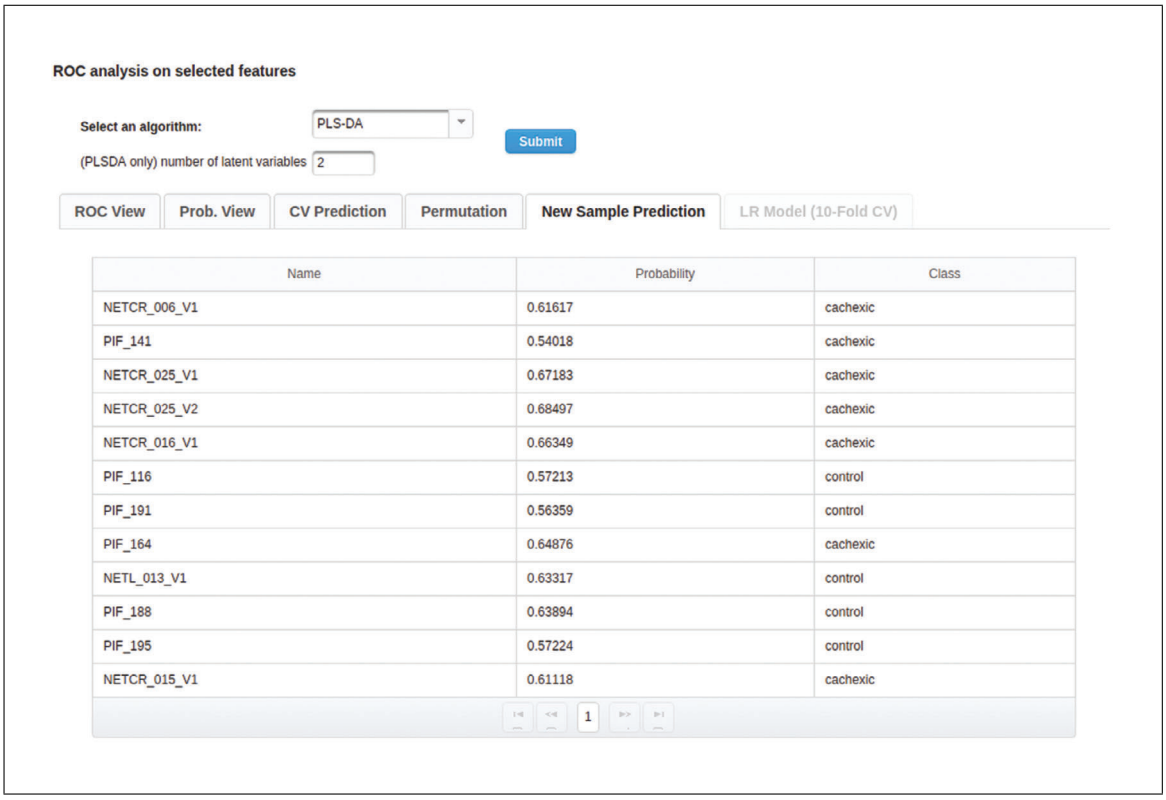


Figure 67 A screenshot showing the predicted class labels and associated probabilities for the 12 new samples.

TIME-SERIES AND TWO-FACTOR DATA ANALYSIS

BASIC PROTOCOL 6

A less common but increasingly important approach to metabolomics is to study metabolic profiles across different time points. These longitudinal metabolomic studies are often used to study disease progression or diet/drug/chemical treatment over extended periods of time. To analyze these types of data, one needs to consider more than just a single factor, using more advanced statistical approaches. These advanced methods include such techniques as multi-way ANOVA, ANOVA-simultaneous component analysis (ASCA; Smilde et al., 2005) and multivariate empirical Bayes time-series analysis (MEBA; Tai & Speed, 2006). These three methods have been implemented in MetaboAnalyst to support the identification of major patterns associated with different experimental factors. They can also be used for the comparison of temporal profiles across biological conditions, as well as the detection of interactions (Xia, Sinelnikov, & Wishart, 2011). This protocol will introduce these methods designed for time-series and two-factor metabolomic data analysis.

Necessary Resources

Hardware

A computer with internet access

Software

An up-to-date web browser, such as Google Chrome (*http://www.google.com/chrome*), Mozilla Firefox (*http://www.mozilla.com/*), Safari (*http://www.apple.com/safari/*), or Internet Explorer (*http://www.microsoft.com/ie/*). JavaScript must be enabled in the web browser.

Files

None

Sample	S1T0	S1T1	S1T2	S1T3	S2T0	S2T1	S2T2	S2T3	S3T0	S3T1	S3T2	S3T3	S4T0	S4T1	S4T2
Phenotype	WT	WT	WT	WT	WT	WT	WT	WT	WT	WT	WT	WT	WT	WT	WT
Time	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2
0.4425/385	126.39	234.11	162.68	115.29	172.44	171.53	180.31	114.94	224.14	195.09	189.45	122.79	212.02	243.48	188.42
0.4498/625	108.9	133.53	128.66	205.38	121.48	112.23	142.97	199.35	127.44	102.45	173.31	252.14	123.96	96.93	118.7
0.4711/463	124.1	166.84	156.18	82.57	158.4	190.09	152.89	100.76	155.75	174.85	153.14	112.76	171.54	150.37	163.49
0.4715/659	688.13	739.41	529.62	354.46	828.48	660.76	545.32	318.51	798.21	714.6	617.95	400.94	815.2	696.85	611.31
0.4766/757	117.8	110.31	114.56	99.99	144.39	155.33	114.47	122	147.16	132.82	129.78	94.86	128.26	161.38	119.12
0.4772/675	182.72	209.42	144.61	92.6	255.9	199.4	179.54	74.58	231.99	196.13	159.57	91.78	188.64	195.08	197.85
0.4781/591	7025.79	6967.26	5000.87	3179.67	7029.79	7022.22	5101.51	3296.96	6685.19	6592.03	5559.12	3692.11	6841.99	6461.22	5273.43

Figure 68 A screenshot showing the data format required for time-series data. Note the term "Time" needs to be used to indicate the time group.

Data upload and processing

1. Go to the MetaboAnalyst home page (*https://www.metaboanalyst.ca*). Click "click here to start" to enter the "Module Overview" page, and then click the "Time Series Analysis" module from the available options. On the "Data Upload" page, check the "Use the example data" to use the first example dataset (*Time series + one experimental factor*). Note that time-series analysis is a special case of two-factor experimental design. MetaboAnalyst requires the term "Time" to be used as the group label to perform time-series analysis (Fig. 68). Otherwise, only regular two-factor analysis will be applied.

   The dataset is a peak list intensity table from a LC-MS metabolomic study of the wound healing process in Arabidopsis thaliana using a time-course experiment that compared wild-type (WT) to a dde2-2 mutant (MT). The two factors are phenotype (WT and MT) and Time (four time points: 0, 1, 2, 3). Please refer to original paper for more information (Meinicke et al., 2008).

2. Click the "Submit" button to perform the data integrity check. Click the "Skip" button to enter "Data Filtering" page. Use the default "inter-quantile range (IQR)" option, click "Submit," and then click "Proceed" to move to the Data Normalization page.

   Data filtering is recommended for processing untargeted metabolomics data so as to remove less informative features and to improve the overall statistical power. In this case, IQR is used to measure the feature variance. The top 5% with the lowest variance will be removed from these data.

3. Choose "Log transformation" and "Pareto scaling" for data normalization. After reviewing the graphical output from the normalization steps, click the "Proceed" button to enter the main "Analysis Overview" page (Fig. 69).

Data overview and pattern discovery with PCA and heatmaps

4. On the "Analysis Overview" page, click the "Interactive PCA Visualization (iPCA)" link to visualize the 3D scatter plots for both the scores and loading values for the top three principal components (Fig. 70). By default, the two experimental factors are indicated using different colors for the primary experimental factor (Factor A), and different shapes for the secondary factor (Factor B). Users can choose to switch this default behavior.

   The interactive scores and loadings plots act identically to the 3D PCA plots from Basic Protocol 3, step 6. Briefly, users can use their mouse/touchpad to zoom in and out or drag to rotate the view around the axis. The current view can be saved as an SVG image using the "Export" button. Users can click on any point in the loading plot to view a summary of the corresponding feature.

5. PCA provides a high-level summary of the main patterns of data variance. The detailed feature (metabolite or peak) abundance profile can be obtained from a heatmap-based visualization of the metabolomic data. Click the "Heatmap2"

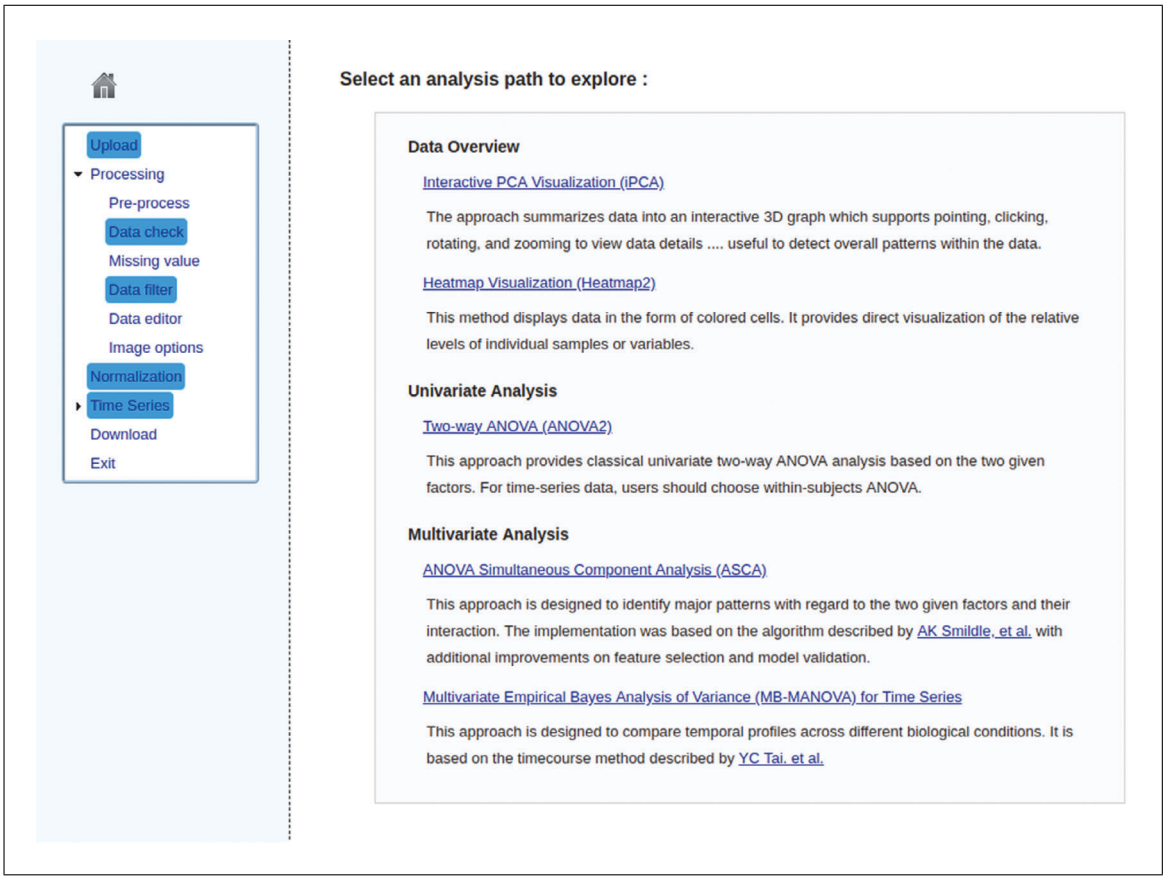


Figure 69 A screenshot showing an overview of the available methods for time-series and two-factor analysis. Note that the last method is only for time-series data, while others can also be used for general two-factor data analysis.

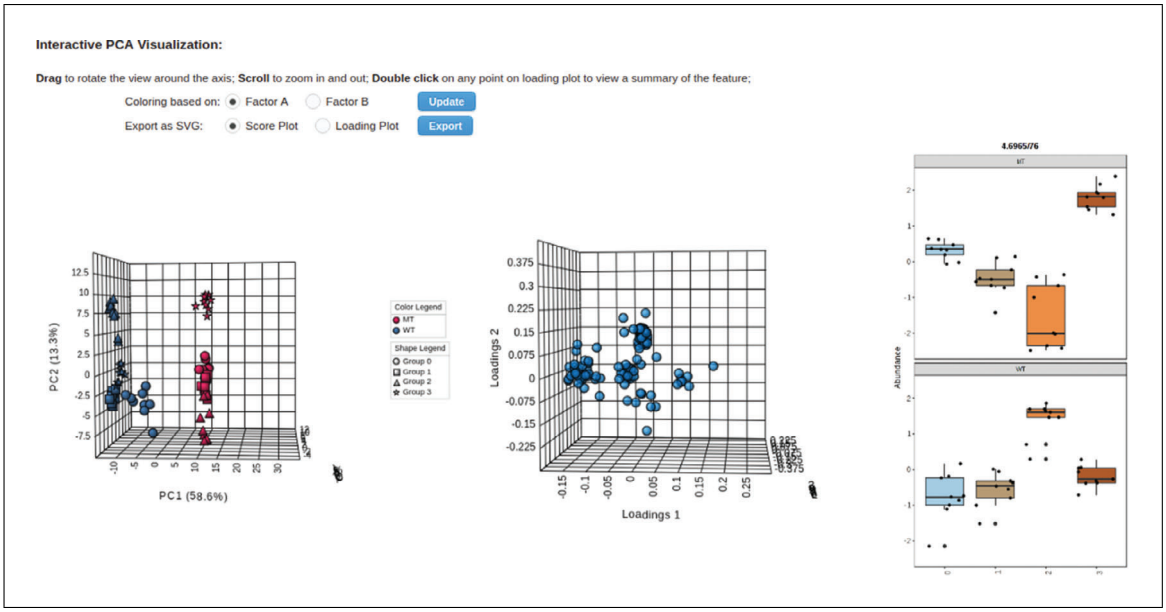


Figure 70 A screenshot showing the interactive 3D PCA score and loading plots from the Time Series module of MetaboAnalyst.

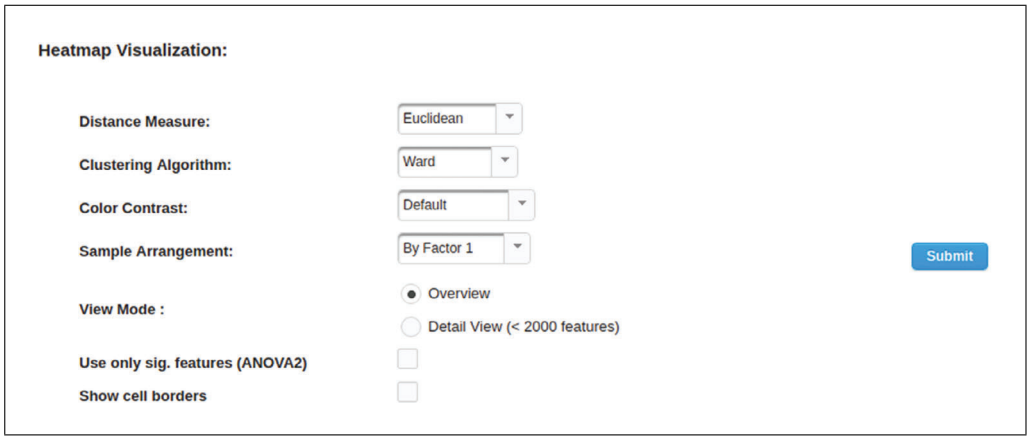


Figure 71 A screenshot showing a comprehensive list of parameters available for users to control the heatmap visualization effects.

link in the left-side navigation tree. As shown in Figure 71, MetaboAnalyst provides a comprehensive list of parameters for cluster generation, color adjustment, sample arrangement, selection of important features, and switching between a high-level overview and a detailed view. Users are encouraged to test different parameters to see how they can improve the plots to meet their specific needs.

6. Set the Color Contrast from "Default" to "Red/Green" and click the "Submit" button. The result is shown in Figure 72. By comparing the peak intensity variations with respect to the class labels, we can clearly see several interesting patterns. For instance, the top cluster of peaks varies significantly with time points. Note that these peaks exhibit opposite trends over the last two time points with respect to the two phenotypes. The peaks in the middle section show an overall higher abundance for the MT phenotype compared to the WT phenotype. However, they also appear to be stable across different time points. Users can switch the View Mode from "Overview" to "Details view," and then scroll to the corresponding regions to view the peak identities. For data sets without clear patterns from the overview, users can choose to visualize only those significant features based on ANOVA analysis, as described below.

Univariate and multivariate statistical analysis

7. Expand the "Time Series" hyperlink on the navigation panel and click the "ANOVA2" hyperlink to enter the page for Two-way ANOVA analysis. As the time series data are collected from repeated measurements of the same subjects over different time points, the module will automatically perform "Within-subjects ANOVA." The result is shown in Figure 73. The Venn diagram summarizes the number of significant features associated with each factor, as well as their interactions. Click the table icon (on the top right corner of the graph) to see the corresponding ANOVA table.

8. By default, the ANOVA table is sorted according to *p*-values based on the primary factor (phenotype). Here, the objective is to identify those features that respond differently between the two phenotypes (i.e., interaction effects). To obtain these features, click the "Interaction" column header to have these interactions sorted by their *p*-values. Click the top feature to view a graphical summary of the result. A screenshot is shown in Figure 74.

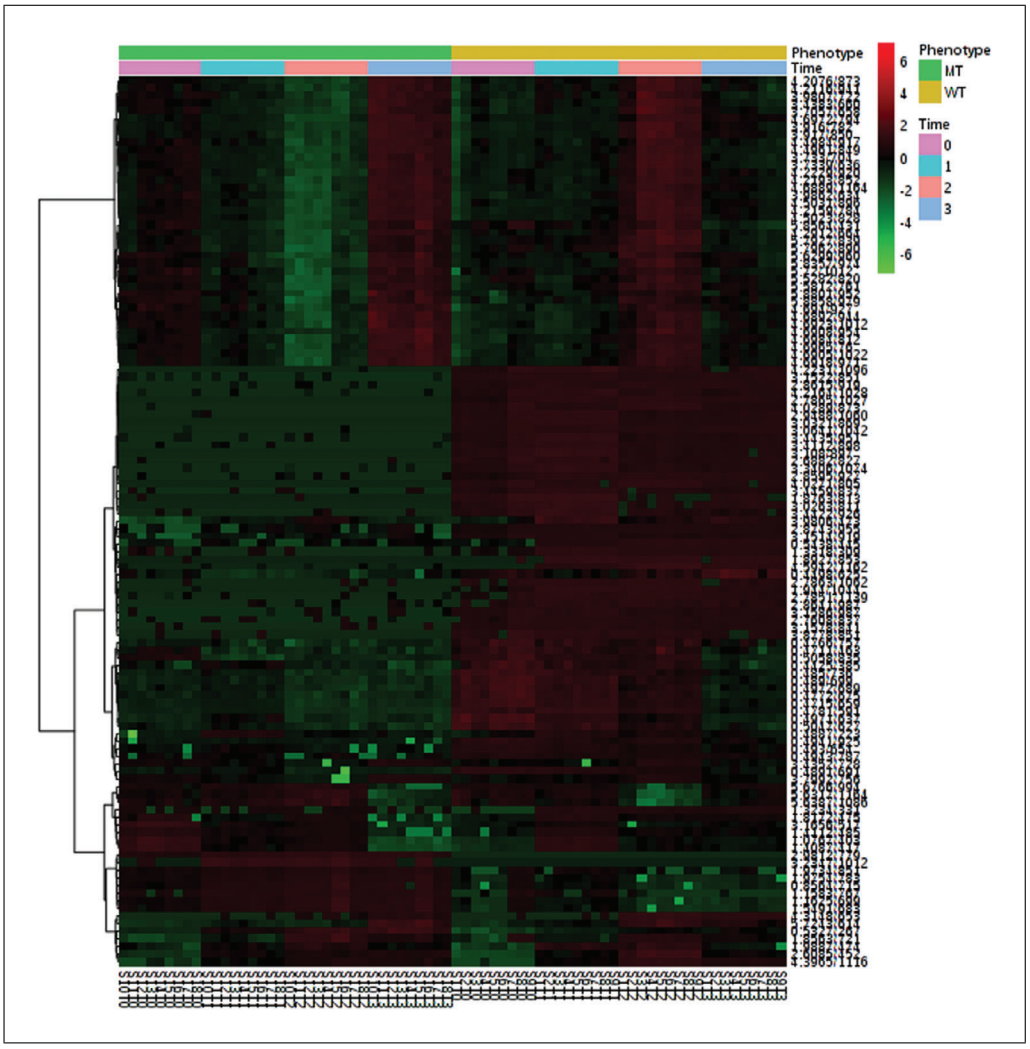


Figure 72 An example output from the heatmap visualization tool using a red/green color gradient. The samples are sorted first by Phenotype (Factor 1) then by the Time (Factor 2).

In this case, the presence of an interaction means the feature (metabolite) responded differently between the two MT and WT phenotypes over the observed period. This is clear based on the boxplots shown in Figure 74. Note that there are also significant changes at each time point within each phenotype. The overall metabolite abundance levels are similar between the two phenotypes. These patterns are consistent with the corresponding p-values.

9. The next step is to further analyze the data using a multivariate version of the two-way ANOVA called ANOVA-simultaneous component analysis (ASCA). This technique is very useful for the detection of major temporal trends. Click the "ASCA" hyperlink on the MetaboAnalyst navigation tree (left menu). An initial analysis is available with the default parameters. Click the "**Major Patterns**" tab to view the major trends associated with each experimental factor as well as their interaction effects. Note that there are two ways of graphically presenting these interactions—*Time* versus *Phenotype* or *Phenotype* versus *Time*. In this context, the second case is more meaningful, as it shows the main patterns of change for different phenotypes through time (Fig. 75). The component1 of interaction effect (left plot) clearly shows the opposite trends occurring over the last two time points between the MT and WT phenotypes. This is consistent with the heatmap visualization shown in Figure 72. Click the "**Sig. Features**" tab to view those variables that follow these trends (i.e., those that are well modeled) as well as those that clearly deviate (outliers).

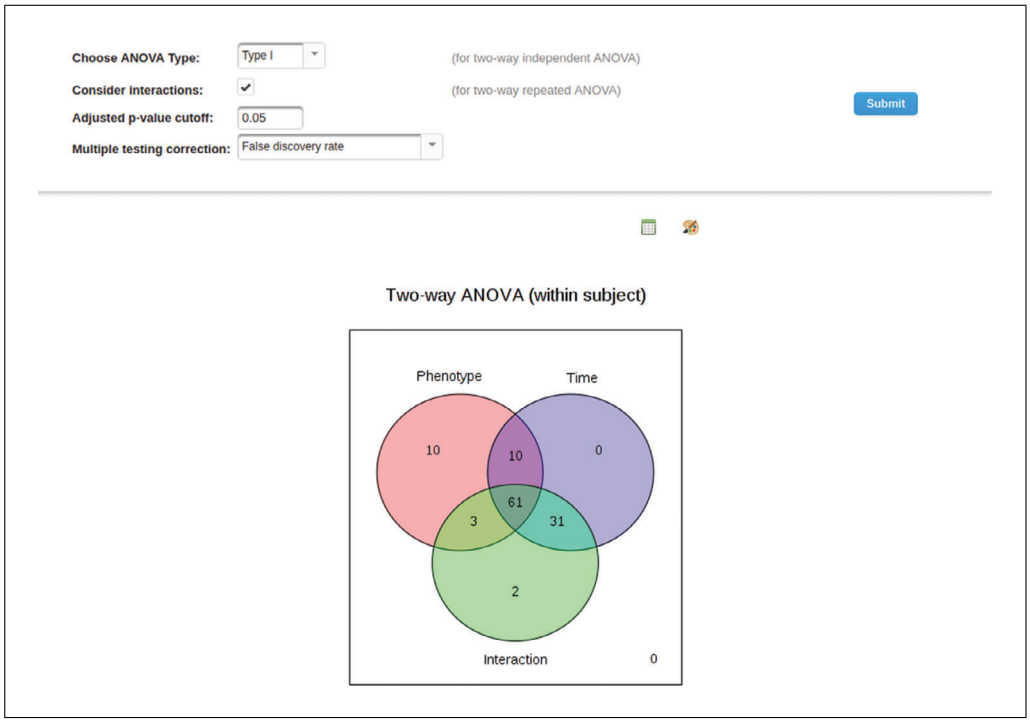
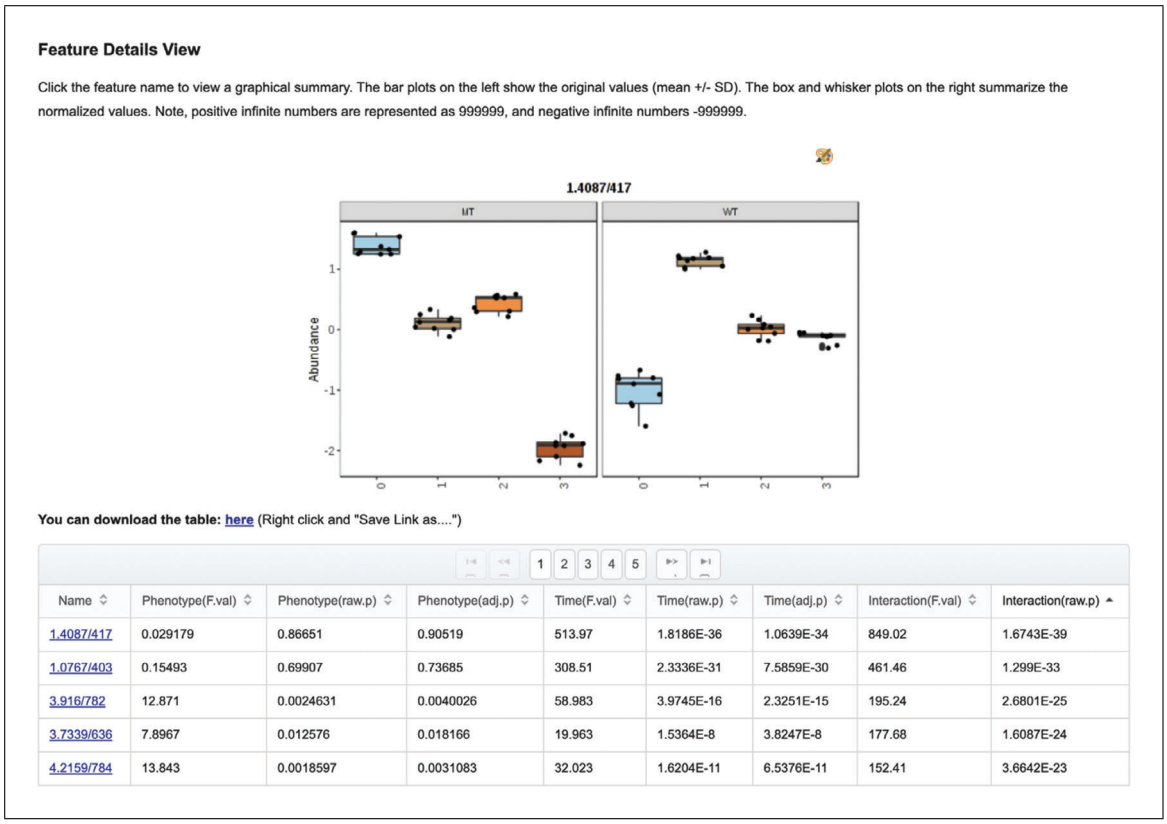


Figure 73 A screenshot showing the two-way ANOVA page. The parameters are shown on the top, and the bottom shows a Venn diagram summary of the significant features.



Name	Phenotype(F.val)	Phenotype(raw.p)	Phenotype(adj.p)	Time(F.val)	Time(raw.p)	Time(adj.p)	Interaction(F.val)	Interaction(raw.p)
1.4087/417	0.029179	0.86651	0.90519	513.97	1.8186E-36	1.0639E-34	849.02	1.6743E-39
1.0767/403	0.15493	0.69907	0.73685	308.51	2.3336E-31	7.5859E-30	461.46	1.299E-33
3.916/782	12.871	0.0024631	0.0040026	58.983	3.9745E-16	2.3251E-15	195.24	2.6801E-25
3.7339/636	7.8967	0.012576	0.018166	19.963	1.5364E-8	3.8247E-8	177.68	1.6087E-24
4.2159/784	13.843	0.0018597	0.0031083	32.023	1.6204E-11	6.5376E-11	152.41	3.6642E-23

Figure 74 A screenshot showing the detailed ANOVA results table. Users can click the feature name to view a corresponding boxplot summary of the abundance profiles across different factors.

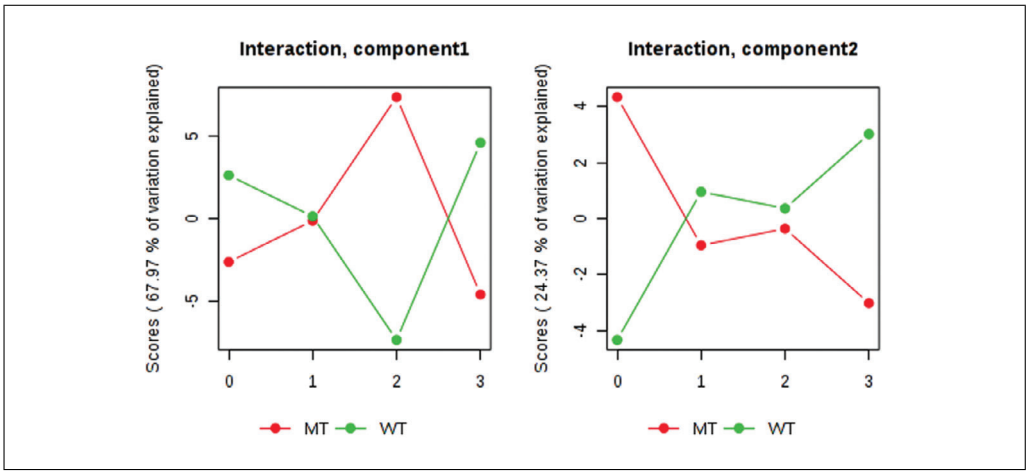


Figure 75 A screenshot showing the major interaction patterns detected by the ASCA method. In general, only the first component is considered in this kind of analysis.

10. The previous analysis clearly shows that there is a group of features with distinctive response patterns between the two phenotypes over the time course being studied. Multivariate Empirical Bayes Analysis (MEBA) is particularly well suited to identify these kinds of differential temporal profiles. Click the "MEBA" hyperlink on the MetaboAnalyst menu tree to perform the analysis. The result is a list of all features ranked by their Hotelling's T^2 values. Clicking any feature name will display its corresponding temporal profiles, as shown in Figure 76.

    The MEBA is among very few methods that can be used for short time series data. The limitation is that it focuses on ranking based on the profile difference and does not provide p-values at the moment. We intend to address this limitation by computing p-values using bootstrap methods.

11. Upon finishing the data analysis and visualization for this protocol, click the "Download" hyperlink on the navigation tree (left side panel) to create an analysis report and download all the results and images.

SAMPLE SIZE ESTIMATION AND POWER ANALYSIS

BASIC PROTOCOL 7

In statistics, power is defined as the probability of detecting an effect, when the effect exists. For instance, if a study comparing the two groups (control versus disease) has a power of 0.8, and assuming the study can be conducted many times, then 80% of the time, a statistically significant difference between the two groups would be achieved. Power analysis requires three parameters: (1) the magnitude of the change or effect size, which is usually defined as the difference of two group means divided by the pooled standard deviation (a larger the effect size will lead to more power), (2) the degree of confidence such as a significance cutoff based on *p*-values or false-positive rates (a more stringent cutoff will lead to reduced power), and (3) the sample size—more samples will generally increase statistical power. In most cases, researchers are interested in predicting sample size for a given power (i.e., 0.8) and a given degree of confidence (i.e., adjusted *p*-values of 0.05). In many cases, the main challenge is how to make a reasonable estimate of the effect size they expect to see in their data. One common approach is to use data sets from a pilot study or from closely related samples obtained from public data repositories. If the pilot data sets have similar characteristics as the experiment to be conducted, the predictions should work well. This approach has been implemented in MetaboAnalyst. For a given pilot data, the method estimates the average power for a given FDR. Users can refer to the original publication for more technical details (van Iterson et al., 2009).

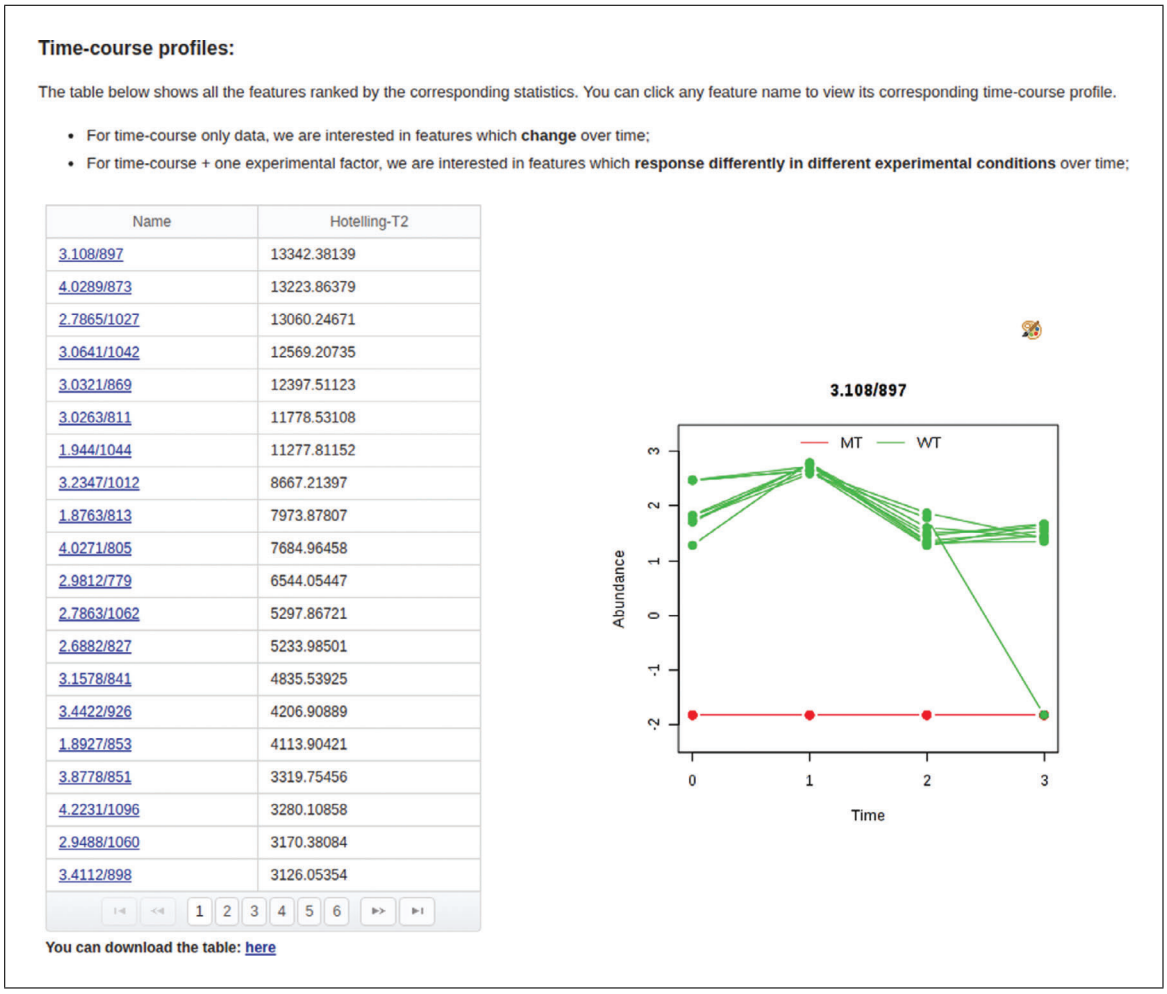


Figure 76 A screenshot showing a ranked list of features with differential temporal profiles based on MEBA analysis. Clicking a feature name will show the detailed time course profile.

This protocol describes the basic steps involved in sample size estimation using data obtained from a pilot metabolomic study.

Necessary Resources

Hardware

A computer with internet access

Software

An up-to-date web browser, such as Google Chrome (*http://www.google.com/chrome*), Mozilla Firefox (*http://www.mozilla.com/*), Safari (*http://www.apple.com/safari/*), or Internet Explorer (*http://www.microsoft.com/ie/*). JavaScript must be enabled in the web browser.

Files

None

1. Go to the MetaboAnalyst home page (*https://www.metaboanalyst.ca*). Click the "click here to start" link at the top of the page to enter the "Module Overview" page, and then click the "Power Analysis" button (lower right side). On the "Data Upload" page, click "Use the example data" check box, and click "Submit." A screenshot of the "Data Upload" page is shown in Figure 77.

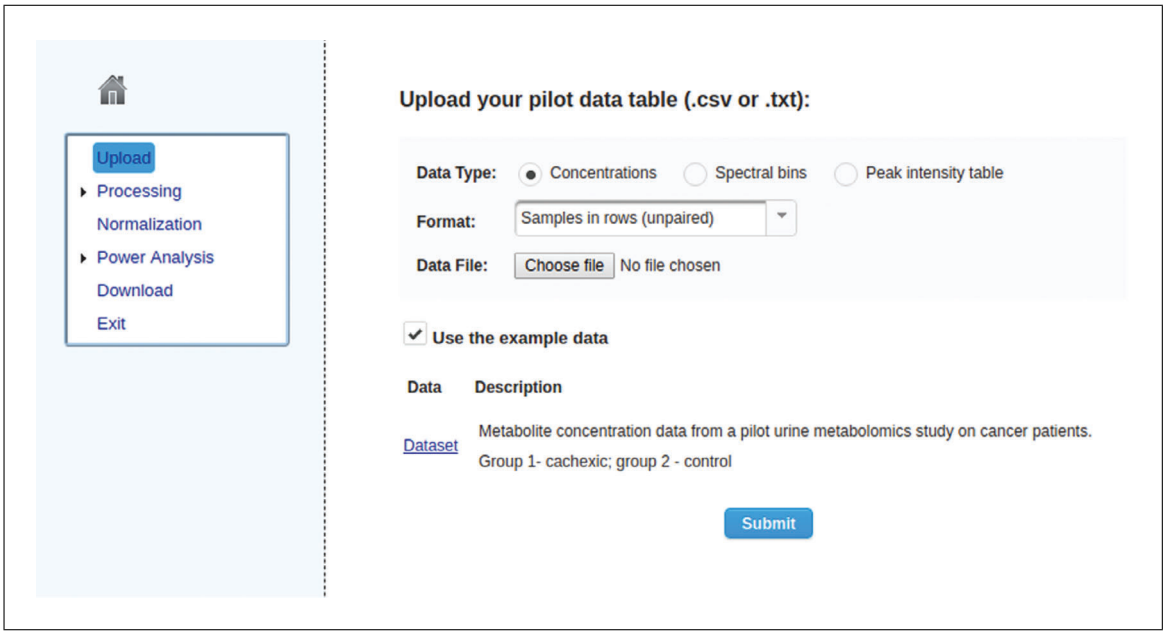


Figure 77 A screenshot showing the "Data Upload" page for the Power Analysis module.

These metabolomics example data are from a pilot study of cancer cachexia (muscle wasting) based on urine metabolic profiles from late-stage cancer patients divided into two groups: cachectic and non-cachectic controls. More details can be obtained from the original publication (Eisner et al., 2011).

2. Perform the data integrity check and then click the "Skip" button to move directly to the "Data Normalization" page. As these are urine data, choose "Normalization by sum" to adjust for potential dilution effects, select "Log transformation," and leave all other options to their default values. Click the "Normalize" button to perform the data normalization and then the "Proceed" button at the bottom of the page to move to the next step.

3. After the data normalization step, the next page shows four diagnostic plots (Fig. 78) to allow users to assess whether the data fit the underlying model. It is expected that the test statistics should be normally distributed while the *p-value*s should follow an extreme-value distribution with a large portion of the values close to zero (i.e., skewed to the left side). In this case, both distributions largely conform to the expectations. Note that the close-to-zero *p*-values are only moderately enriched based on the histogram.

   If the distributions deviate significantly from expectation, users can try either to apply different normalization procedures or to increase the sample size of the pilot dataset.

4. Click the "Submit" button to view the power analysis result. By default, the result shows the predicted powers for up to 200 samples (per group). In this case, we can see that the curve has not reached a plateau yet. Enter 800 as the maximum sample size per group, leave the FDR cutoff as 0.1, and click the "Submit" button. The result is shown in Figure 79. From the graph, it is evident that with around 320 samples per group, the study will have enough power (0.8) based on this pilot data set.

   Note that users can zoom into the graph to obtain more accurate coordinates. To do this, use the mouse to draw a box around the region of interest.

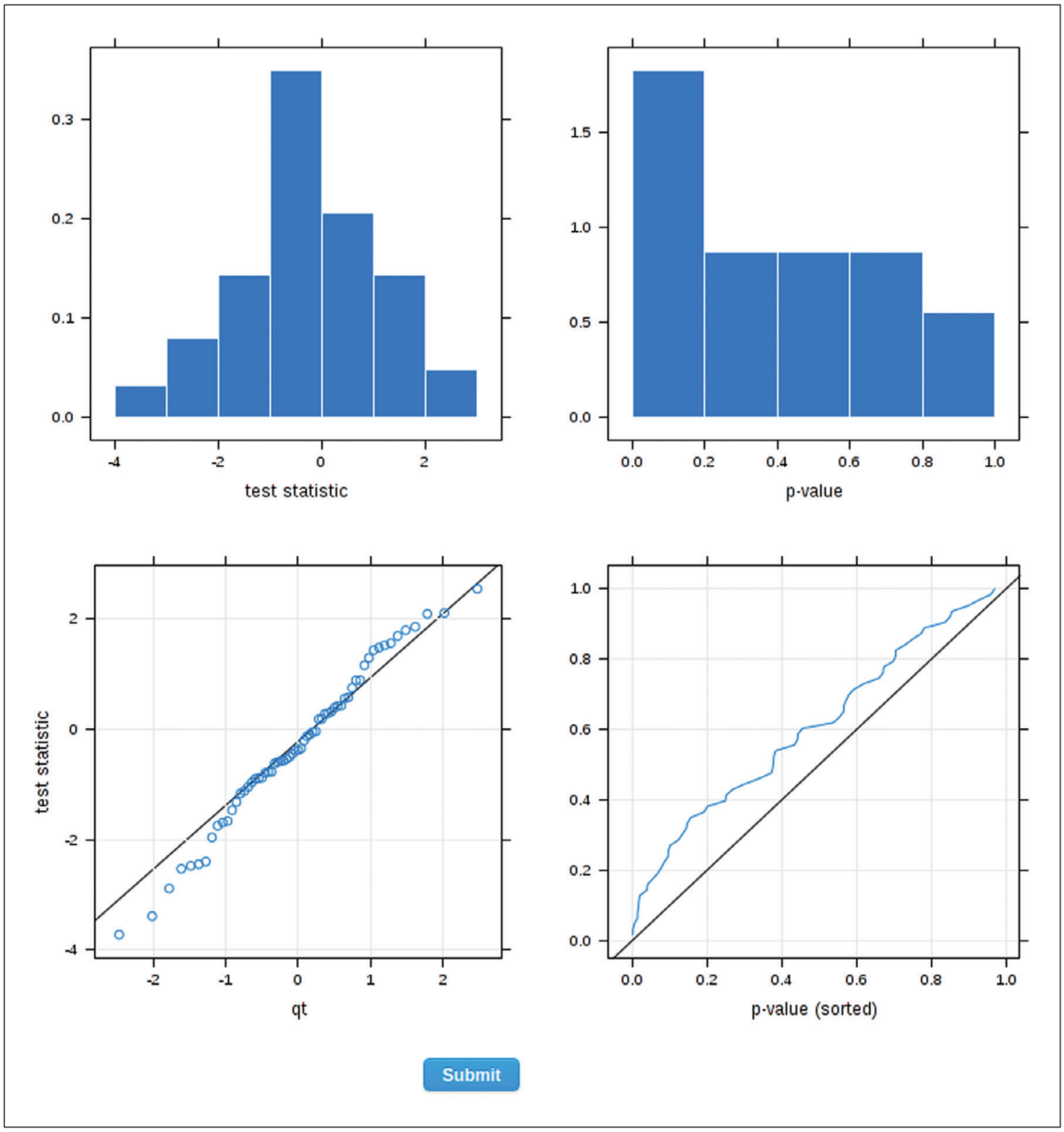


Figure 78 A screenshot showing the diagnostic plots for statistical power prediction. The plots on the left side show whether the statistics follow a normal distribution, and the plot on the right side shows whether *p*-values follow an extreme-value distribution skewed to the left.

5. The graphical display for the Power Analysis module is interactive and is optimized for facile web presentation. MetaboAnalyst can also generate high-resolution static images for most images displayed. To do this, click the palette icon (the one with the paintbrush) on the top right corner of plot to show the "Graphics Center" dialog box. Users can then specify both the size and desired resolution of the image. In this case, use the default parameters and click "Submit." A screenshot is shown in Figure 80.

6. Upon finishing this data analysis, click the "Download" hyperlink on the navigation tree (left side panel) to create an analysis report and download all the results and images.

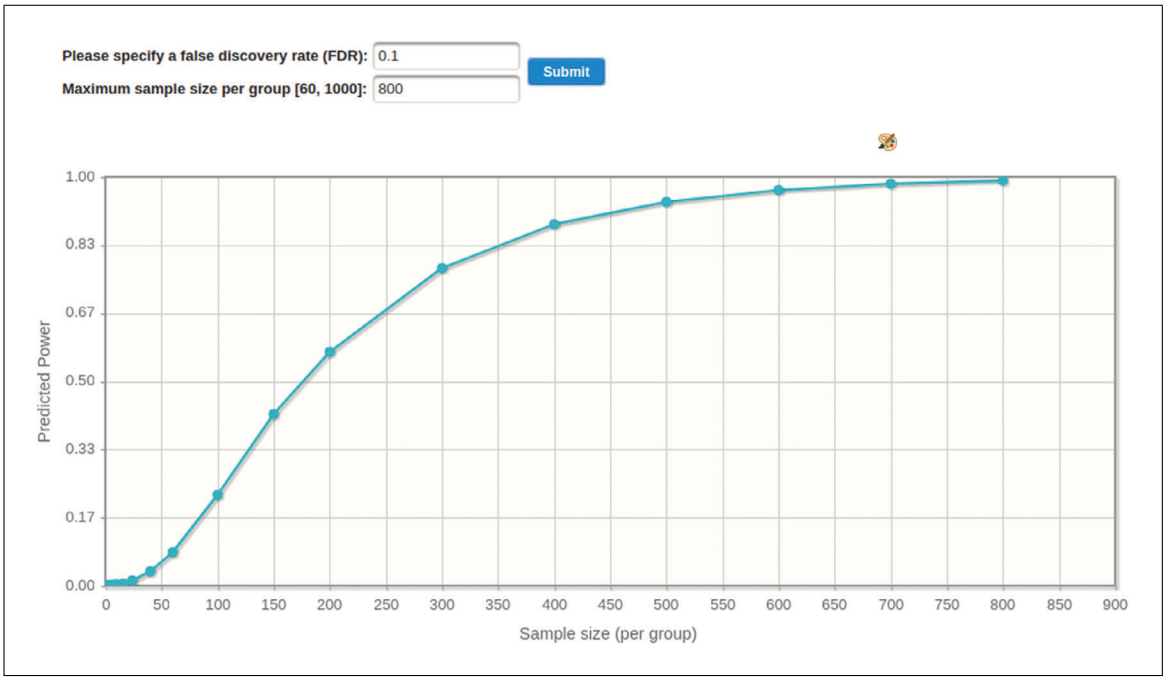


Figure 79 The plot shows the sample size versus the predicted power. Users can specify the maximum sample size to make sure the desired power can be reached.

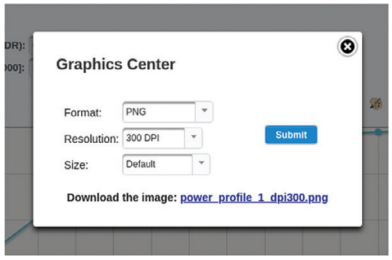


Figure 80 A screenshot showing the "Graphics Center" page. Most images can be re-created here with a specified size or resolution.

JOINT PATHWAY ANALYSIS

BASIC PROTOCOL 8

The "Joint Pathway Analysis" module has been designed to support multi-omics analysis and interpretation (i.e., metabolomics, transcriptomics or proteomics). In particular, it allows users to map significant genes/proteins (identified from gene expression or proteomics studies) together with significant metabolites (identified from metabolomics studies) to metabolic pathways for functional enrichment analysis and pathway topology analysis. The working assumption behind this module is that by integrating evidence based on changes in both gene/protein expression and metabolite concentrations, one is more likely to pinpoint the pathways involved in the underlying biological processes.

Necessary Resources

Hardware

A computer with internet access

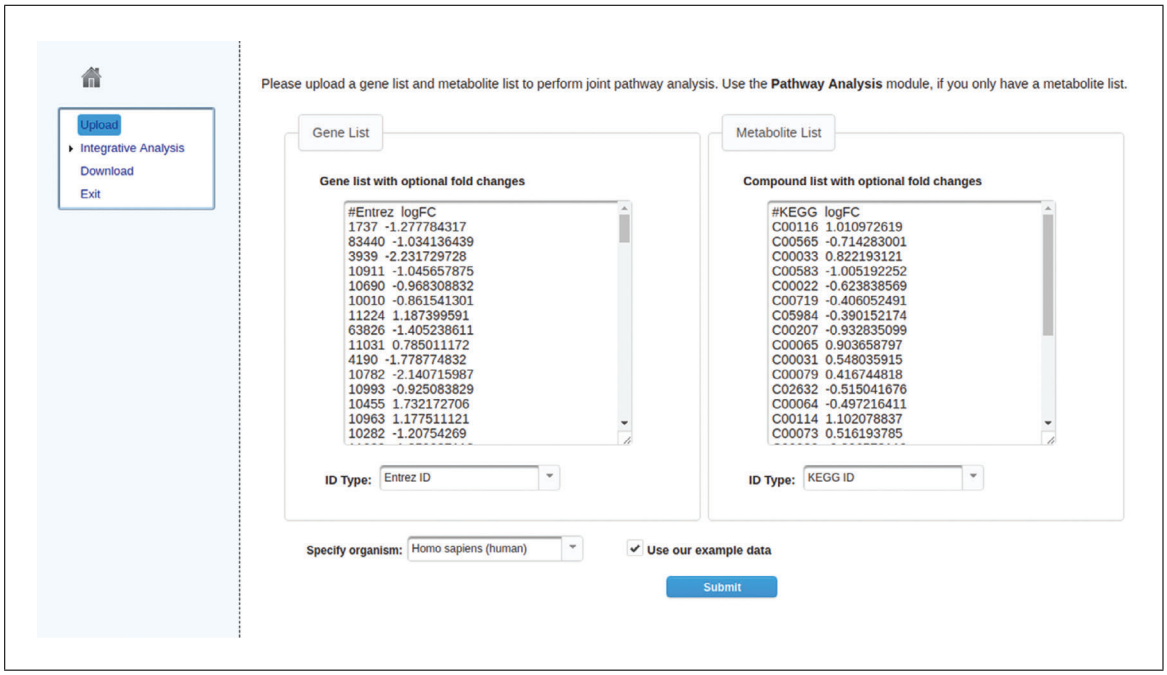


Figure 81 A screenshot showing the data input page for the Joint Pathway Analysis module.

Software

An up-to-date web browser, such as Google Chrome (*http://www.google.com/chrome*), Mozilla Firefox (*http://www.mozilla.com/*), Safari (*http://www.apple.com/safari/*), or Internet Explorer (*http://www.microsoft.com/ie/*). JavaScript must be enabled in the web browser.

Files

None

1. Go to the MetaboAnalyst home page (*https://www.metaboanalyst.ca*). Click the "click here to start" link to enter the "Module Overview" page. Click the "Joint Pathway Analysis" button to enter the "Data Upload" page. Users need to enter a list of genes and a list of metabolites, with optional abundance values such as log fold changes (FC), that have been identified from the same samples or obtained under similar conditions. In this case, use the example data provided by MetaboAnalyst. This particular example comes from an integrative transcriptome and metabolome analysis that aimed to identify potential biomarkers of intrahepatic cholangiocarcinoma (ICC) in 16 individuals (Murakami et al., 2015). The example data consist of 51 genes (official gene symbols) and 13 metabolites (KEGG identifiers). Note that their log-FCs were fabricated for demonstration purposes. Select the "Use our example data" option by clicking on the check box. Click the "Submit" button to perform ID mapping for the specified gene and metabolite identifiers. A screenshot is shown in Figure 81.

   MetaboAnalyst 4.0 currently supports three common organisms—human, mouse, and rat. More organisms will be added based on user requests. Five common gene ID types are supported, including official gene symbols, Entrez IDs, RefSeq IDs, Genbank Accession numbers, and Ensemble Gene IDs. Three types of metabolite IDs are supported, including common compound name, HMDB ID, and KEGG compound ID.

2. The next page shows the result table of the genes and metabolites mapped to the underlying database. Users can delete incorrect mappings. After reviewing the table, click the "Submit" button at the bottom of the page to enter the "Analysis Parameters"

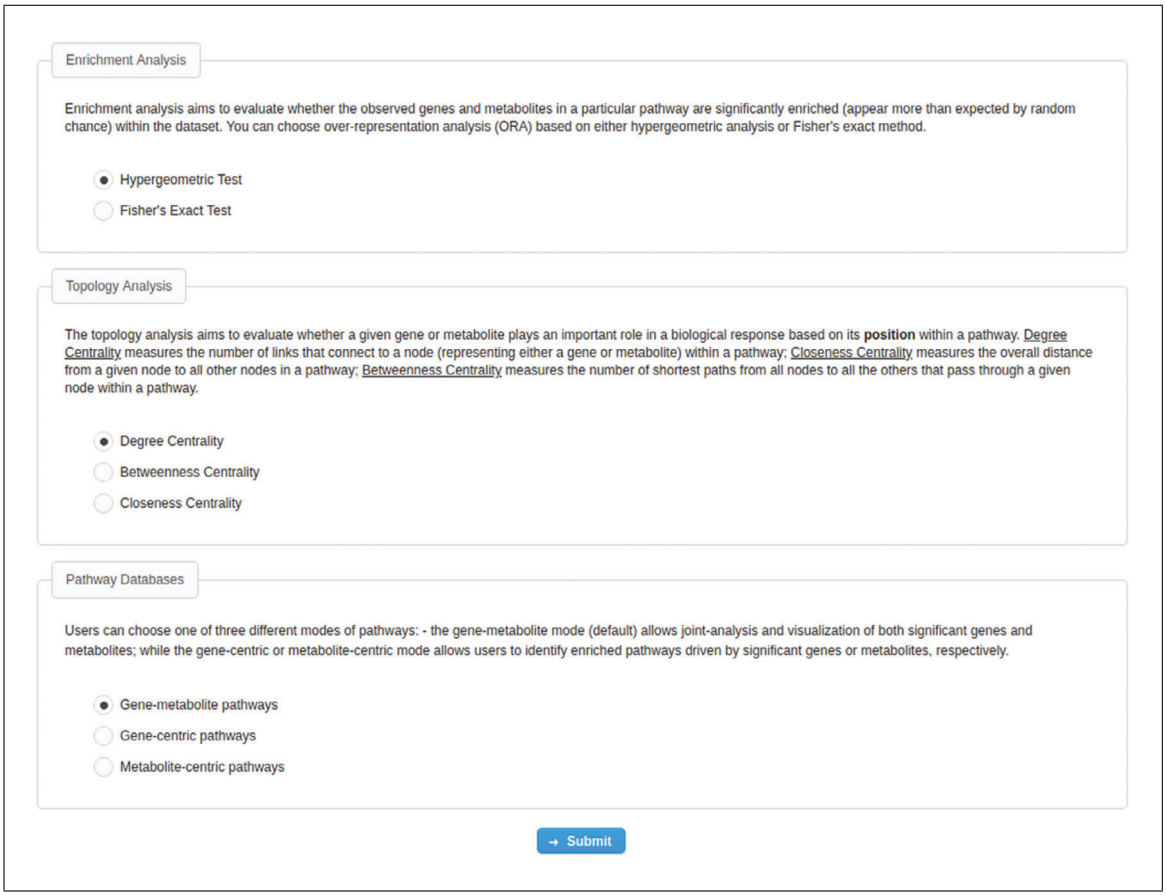
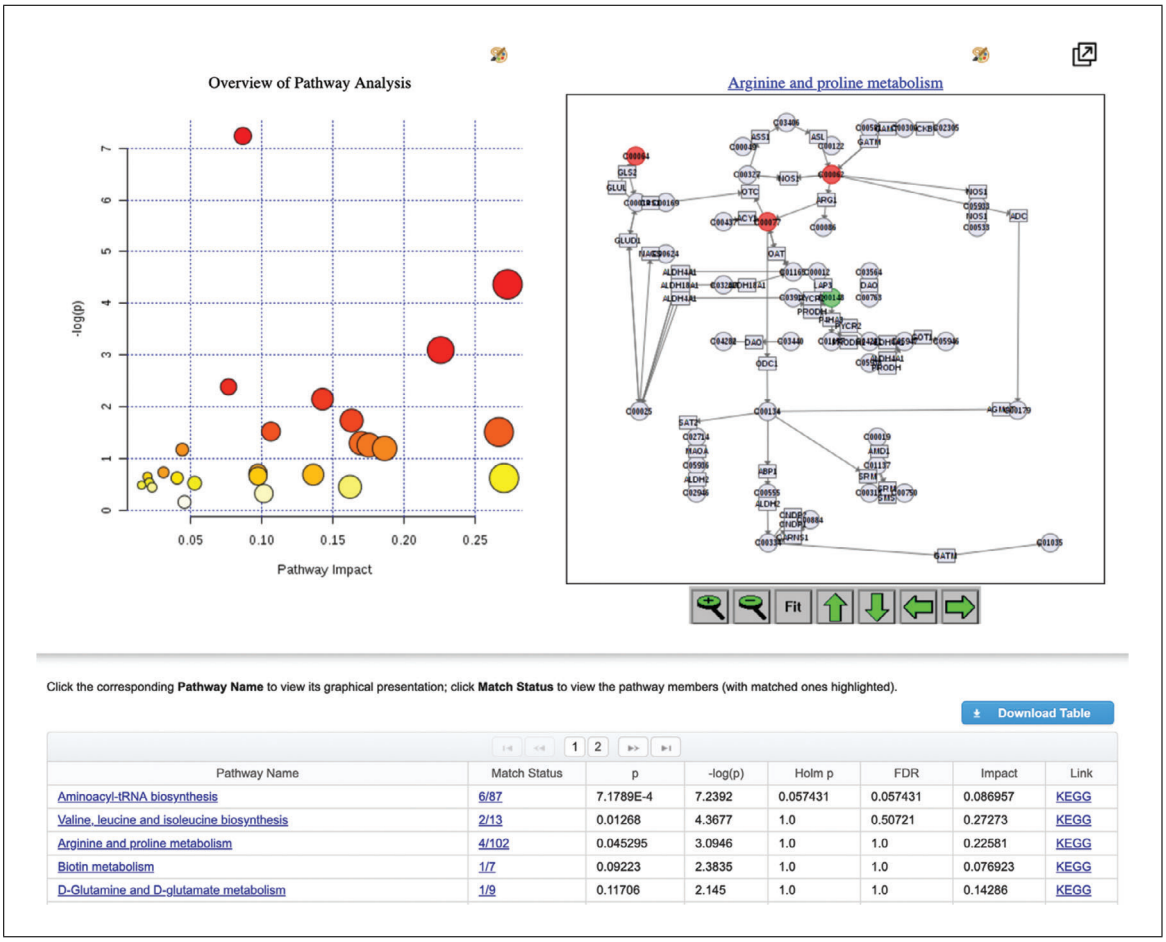


Figure 82 A screenshot showing the available parameters for the Joint Pathway Analysis module. See text for more details.

page (Fig. 82). Users can choose among several different methods for enrichment analysis, topology analysis, and pathway type (gene-metabolite, gene-centric or metabolite-centric). Keep the parameters selection as default and click "Submit."

Enrichment analysis is used to evaluate whether the observed genes or metabolites in a pathway appear more frequently than expected by random chance within a given dataset. This can be tested using either hypergeometric analysis or Fisher's exact test. Topology analysis evaluates the potential importance of a given gene or metabolite based on its position within a pathway. For instance, "degree centrality" measures the number of links that connect to a node within a pathway; "closeness centrality" measures the overall distance from a given node to all other nodes in a pathway; and "betweenness centrality" measures the number of shortest paths from all nodes to all the others that pass through a given node within a pathway. Users can choose one of three different pathway modes: the "gene-metabolite pathways" mode allows joint-analysis and visualization of both genes and metabolites (note pathways containing only genes or metabolites will be excluded in this type of analysis), while the "gene-centric pathways" or "metabolite-centric pathways" modes allow users to identify enriched pathways driven by significant genes (alone) or significant metabolites (alone) for comparative analysis.

3. The results page is identical in design to that from Pathway Analysis (Fig. 53, Basic Protocol 4). The top shows the pathway visualization, and a detailed table is at the bottom. The "Overview of Pathway Analysis" plot displays all matched pathways as circles, with the color and size of each circle corresponding to its *p*-value and pathway impact value, respectively. Click on any circle for its corresponding pathway plot to appear on the right in the "Pathway Viewer" (Fig. 83). The matched nodes are highlighted (red for upregulated and green for downregulated). If one clicks on



Pathway Name	Match Status	p	-log(p)	Holm p	FDR	Impact	Link
Aminoacyl-tRNA biosynthesis	6/87	7.1789E-4	7.2392	0.057431	0.057431	0.086957	KEGG
Valine, leucine and isoleucine biosynthesis	2/13	0.01268	4.3677	1.0	0.50721	0.27273	KEGG
Arginine and proline metabolism	4/102	0.045295	3.0946	1.0	1.0	0.22581	KEGG
Biotin metabolism	1/7	0.09223	2.3835	1.0	1.0	0.076923	KEGG
D-Glutamine and D-glutamate metabolism	1/9	0.11706	2.145	1.0	1.0	0.14286	KEGG

Figure 83 A screenshot showing the pathway visualization and results table for the Joint Pathway Analysis module. In the pathway viewer, matched nodes are highlighted in different colors. Users can click a given node to get more details.

a node, a dialog box will appear, to provide more details about the corresponding gene or compound. Users can also use the buttons below to zoom in and out.

Keep in mind that unlike transcriptomics where the entire transcriptome is routinely measured, current metabolomic technologies capture only a small portion of the metabolome. This difference can lead to potentially biased results. To address this issue, the current implementation in MetaboAnalyst 4.0 allows users to explore the enriched pathways based either on the joint evidence or on the evidence obtained from one particular 'omic platform by selecting different pathway analysis modes in step 2.

4. Upon finishing this data analysis and visualization, click the Download hyperlink on the navigation tree (left side panel) to enter the "Download" page. Users can create an analysis report and download the result table together with all the graphical outputs.

BASIC PROTOCOL 9

MS PEAKS TO PATHWAY ACTIVITIES

Liquid chromatography mass spectrometry (LC-MS) has now become a major workhorse in untargeted metabolomics. Many bioinformatics tools have been developed to support spectral processing, such as XCMS (Smith et al., 2006), MZmine (Pluskal et al., 2010), OpenMS (Röst et al., 2016), and MS-DIAL (Tsugawa et al., 2015). However, high-throughput downstream functional analysis has continued to be a major bottleneck. After peak picking and alignment, the conventional procedures typically require annotation of these peak lists prior to downstream analysis. This process is very time consuming

and error prone. MS peak lists are defined by a combination of a mass-to-charge ratio (*m*/*z*) and retention time. Despite the high mass accuracy of modern instruments, it is not unusual for several metabolites to match a single MS peak. It is also possible that completely unknown metabolites may have *m*/*z* values identical to known metabolites. The use of tandem MS (or MS/MS) experiments, which provide additional information beyond *m*/*z* values or retention time, can be used to distinguish or identify some compounds, but obtaining MS/MS adds additional time and cost, and so most untargeted metabolomics studies still rely on high-resolution MS data. One promising approach is to leverage the collective power of metabolic pathways to help resolve the ambiguities in metabolite annotations. In particular, the *mummichog* algorithm (Li et al., 2013) bypasses the bottleneck of metabolite identification prior to pathway analysis by employing a priori pathway/network knowledge to directly infer biological activity based on MS peaks.

The underlying assumption in *mummichog* is that if a list of significantly enriched peaks truly reflects biological activity, the representation of these true metabolites would be enriched on localized structures such as pathways, while false matches would be distributed at random across the metabolic networks. This concept can also be tested using the popular Gene Set Enrichment Analysis (GSEA) algorithm (Subramanian et al., 2005) and has been implemented in MetaboAnalystR 2.0 and MetaboAnalyst 4.0. The original *mummichog* is based on ORA and tests if certain pathways are enriched in the significant peaks as compared to null models based on peak lists of the same size that are randomly drawn from the reference peak list. In comparison, GSEA is a cut-off-free method that evaluates the overall differences of two distributions based on Kolmogorov-Smirnov tests. This protocol describes the basic steps involved in pathway activity prediction using an example pre-ranked list of MS peaks obtained from a pediatric cohort of inflammatory bowel disease (IBD) [Integrative HMP (iHMP) Research Network Consortium, 2014].

Necessary Resources

Hardware

A computer with internet access

Software

An up-to-date web browser, such as Google Chrome (*http://www.google.com/chrome*), Mozilla Firefox (*http://www.mozilla.com/*), Safari (*http://www.apple.com/safari/*), or Internet Explorer (*http://www.microsoft.com/ie/*). JavaScript must be enabled in the web browser.

Files

None

MS peaks upload and parameter selection

1. Go to the MetaboAnalyst home page (*https://www.metaboanalyst.ca*). Click the "click here to start" link to enter the "Module Overview" page, and then click the "MS Peaks to Pathways" button to enter the "Peak Uploads" page. Users must first specify the "Mass Accuracy" and the "Analytical Mode" based on their MS instruments used in generating the data, as well as the format of their data. Users can then upload their list of *m*/*z* features by selecting the "Choose File" button. For the purpose of this protocol, use the second example dataset, which consists of a subset of fecal samples from 12 pediatric IBD patients and 12 controls collected from the Integrative Human Microbiome Project [Integrative HMP (iHMP) Research Network Consortium, 2014]. Select the checkbox next to "IBD1," and then click "Submit" to continue (Fig. 84).

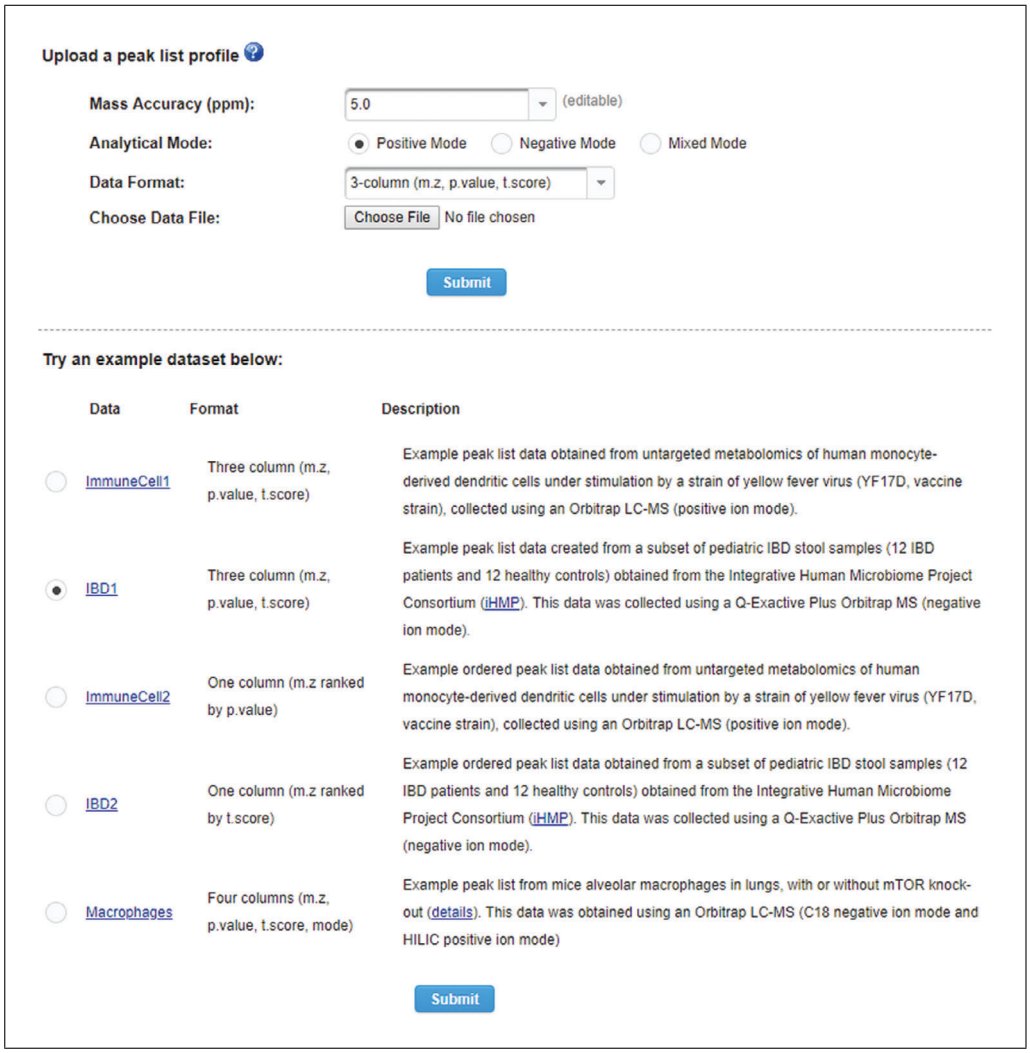


Figure 84 The Data Upload page for the MS Peaks to Pathways module. The module accepts various data formats containing the pre-ranked list of MS peaks. Several example datasets are available for users to explore.

The input peak list data should be saved as a tab-delimited (`.txt`) file containing either: (1) a three-column table with m/z features, p-values, and statistical scores; (2) a two-column table containing m/z features and either p-values or t-scores; or (3) a one-column table ranked by either p-values or t-scores. If these values have not yet been calculated, users can upload their m/z peak list files or peak tables to the MetaboAnalyst Statistical Analysis module to perform their statistical analysis of choice (t-test or fold-change analysis), and then upload these results into the "MS Peaks to Pathways" module. If the input is a three-column table, both the mummichog and GSEA algorithms (and their combination) can be applied. If only p-values (or ranked by p-values) are provided, then only the mummichog algorithm will be applied. If only t-scores (or ranked by t-scores) are provided, then only the GSEA algorithm will be applied. Users should note that their selection of p-value cut-off could affect the results of this module when using the mummichog algorithm. Changing this value may alter which compounds are considered significant, and thereby modify the output of the mummichog algorithm. We therefore recommend that users explore different p-value cut-off values during their analysis, although the algorithm was shown to be robust in the original publication (Li et al., 2013).

2. After uploading the data, a data integrity check is performed to ensure that the uploaded data are of suitable quality for further analysis. MetaboAnalyst checks the format of the uploaded data, the total number of *m/z* features, and the user specified

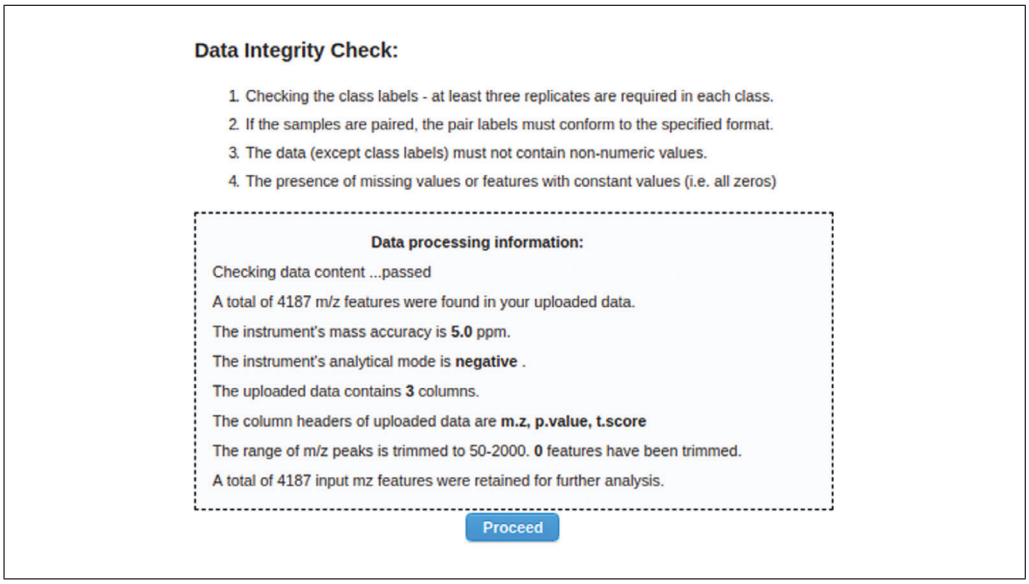


Figure 85 The Data Integrity Check page for the MS Peaks to Pathways module. The dataset passed the check and was successfully parsed, with 4187 *m*/*z* features.

parameters. In this case, the example data passed all the integrity checks (Fig. 85). Click "Proceed" to continue.

A typical input peak list data may contain ~5000 peaks, with 5% to 10% of them significant. Only peaks with mass range [50-2000] will be used for downstream analysis. Note the pathways analysis using the mummichog algorithm will be unreliable when there are very few significant peaks (i.e., <30). In this case, users should adjust the p-value cutoff for more robust results.

3. The next page shows the "Library View," where users can select which algorithm to be used for pathway enrichment (*mummichog*, GSEA, or their integration) as well as the pathway library that best fits their organism (Fig. 86). For the *mummichog* algorithm, users can use either the "Default cutoff," which considers the top 10% or top 500 peaks (whichever is smaller based on the user's uploaded MS peaks) as significant, or specify a *p*-value cutoff themselves.

   The original mummichog algorithm requires users to specify a pre-defined cutoff based on either t-statistics or fold changes. In comparison, the GSEA method considers the overall ranks of all features involved. It can therefore detect subtle and consistent changes that could be missed using the ORA method.

4. The pathway knowledgebase consists of five genome-scale metabolic models obtained from the original *mummichog* package (version 1.0.10), a manually curated *Danio rerio* genome-scale metabolic model, and an expanded pathway library of 21 organisms derived from KEGG and soon SMPDB. For demonstration purposes, we will use the *mummichog* algorithm and then briefly demonstrate the third approach—combining *mummichog* and GSEA. First, select the *mummichog* algorithm and keep the "Default *p*-value cutoff" at 0.2. Keep the pathway library selected to *Homo sapiens* (human) [MFN] library, which was manually curated by the original authors from a number of sources (KEGG, BiGG, and Edinburgh Model; Li et al., 2013). Click "Submit" to proceed to the next step.

   The genome-scale metabolic model of Danio rerio was manually curated from the KEGG zebrafish model, as well as the human BiGG and Edinburgh Models, designated with [MFN] at the end of its name. The remaining four genome-scale metabolic models were directly derived from BioCyc, denoted with [BioCyc] at the end of their names. The KEGG

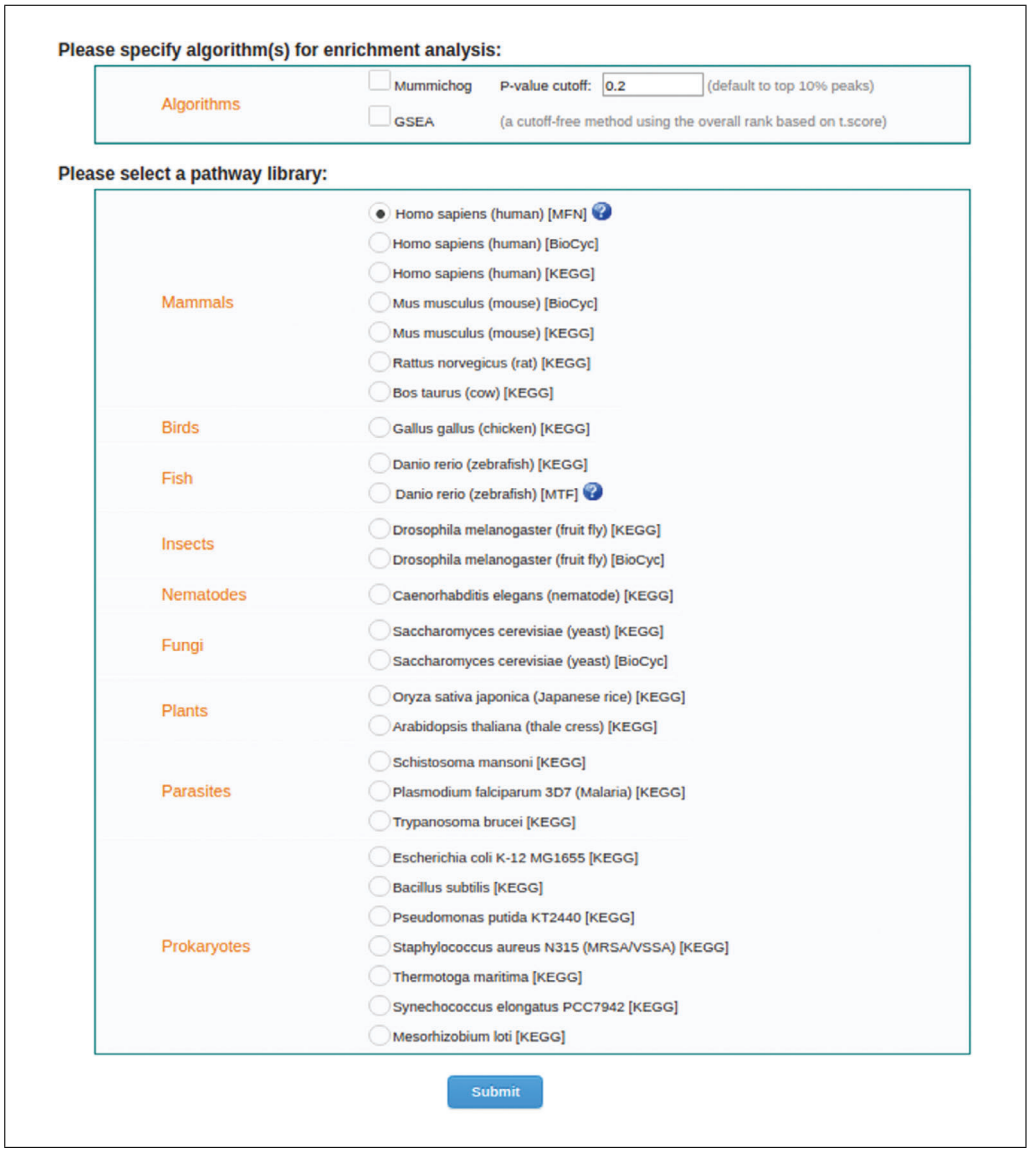


Figure 86 The parameters page for the MS Peaks to Pathways module. At the top, users must specify which algorithm/s to apply. Next, users must specify which pathway library to select. By default, the selected pathway library is *Homo sapiens* (human) [MFN].

pathway libraries are designated with [KEGG], and the SMPDB pathway libraries will be designated with [SMPDB].

Prediction of enriched metabolic pathways

5. The results of the predicted pathway activity using the *mummichog* algorithm are shown in a pathway summary plot at the top of the page and summarized in a detailed table below (Fig. 87). The summary plot displays all matched pathways as circles with colors and sizes corresponding to their *p*-values and enrichment factors, respectively. The enrichment factor of a pathway is calculated as the ratio between the number of hits found in user-uploaded data and the expected number of hits for that pathway.

Hover over the circles to display the pathway names and double-click any circle to view a list of candidate compound hits from the uploaded MS peak list (Fig. 88). In the dialog, metabolites in red represent significant hits and metabolites listed in blue represent non-significant yet present hits. Meanwhile, the results table consists of the total number of compounds, total number of hits, significant hits, hits expected, raw

Pathway Name	Total	Hits (all)	Hits (sig.)	Expected	P-value	Gamma P	Details
Vitamin E metabolism	54	33	19	6.2557	3.2091E-4	0.0010809	View
De novo fatty acid biosynthesis	106	15	7	12.28	0.098383	0.22401	View
Bile acid biosynthesis	82	52	19	9.4993	0.11503	0.18117	View
Vitamin D3 (cholecalciferol) metabolism	16	10	5	1.8535	0.12025	0.30053	View
Fatty Acid Metabolism	63	11	5	7.2983	0.1715	0.37519	View

Figure 87 A screenshot of the pathway activity profile plot in MetaboAnalyst summarizing the results of the *mummichog* algorithm in the MS Peaks to Pathways module.

p-values, and Gamma-adjusted p-values (based on permutations) per pathway. In this case, vitamin E metabolism, De novo fatty acid biosynthesis, and Bile acid biosynthesis are the top-ranked enriched pathways.

6. There are three buttons at the top of the result table. Users can click the "Pathway Hits" to download the pathway results or click the "Compound Hits" to download the matched metabolites including the query mass, matched compound, matched form, and mass difference between the query and the matched compound. The "Explore Results in Network" button allows users to visualize their results in the context of the KEGG global metabolic network. Click "Explore Results in Network" to continue to the next step.

Network visualization of pathway prediction results

7. On this page, users can view the global metabolic network to visually assess the peak matching patterns of their data. The metabolic network visualization is based on the manually curated KEGG global metabolic network. The page consists of three sections: the top toolbar, the left panel containing the pathway analysis results, and the central area displaying the metabolic network. To demonstrate the utility of this page, select the checkbox next to *de novo fatty acid biosynthesis* from the left section of the page to highlight all matched compounds (Fig. 89). Here, the metabolites of significantly enriched pathways are represented as nodes on the network. Solid nodes represent significantly enriched features detected in the data, while empty nodes represent compounds detected in the data but are not significant.

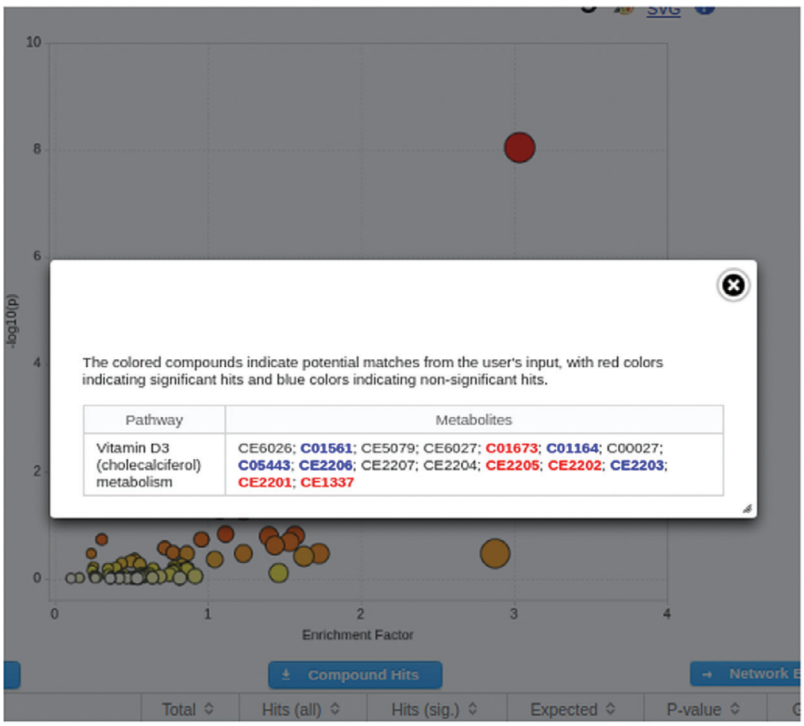


Figure 88 A screenshot of the dialog containing a list of potential compound matches in the selected pathway from the user-uploaded list of MS peaks.

The bottom left corner of the network visualization page (Fig. 89) shows all matched compounds in the current pathway. For instance, after highlighting the de novo fatty acid biosynthesis pathway, the box lists all the pathway's compounds that match with the user's data, with significant hits highlighted in red. Clicking on a compound name from this box will link users directly to the KEGG page for that compound. Users can double click any highlighted compound node to examine all matched isotopic or adduct forms. An example is shown in Figure 90.

8. The toolbar at the top of the page provides several interactive options to customize the network such as changing the background color (black or white), changing the view style (KEGG, expression or plain), and highlighting user-specified pathways in any color. Mouse scrolling for zooming in and out of the network is also enabled. The customized maps can be downloaded as PNG or SVG files for publication/report purposes.

   To use the highlight feature, click the colored box at the top of the toolbar next to "Highlight." A color palette will appear, where you can use the color bar on the right-hand side of the palette to select any color. Alternatively, users can also type in the hex code. In this case, replace `#ffff00` (yellow) with `#5cb9f2` (blue), then select "choose" to use this color (Fig. 91). Click on the "Bile acid biosynthesis" pathway on the left-hand side to highlight the matched compounds in the selected color (Fig. 92).

Integrated prediction of enriched metabolic pathways

9. After exploring the results in the metabolic network, click the "Library" page to return to the "Library View." For the pathway algorithms, keep the *mummichog* algorithm checked but now also select GSEA. Keep the pathway library selection as "Homo Sapiens [MFN]" and press "Submit" to perform the integrated *mummichog* and GSEA analysis.

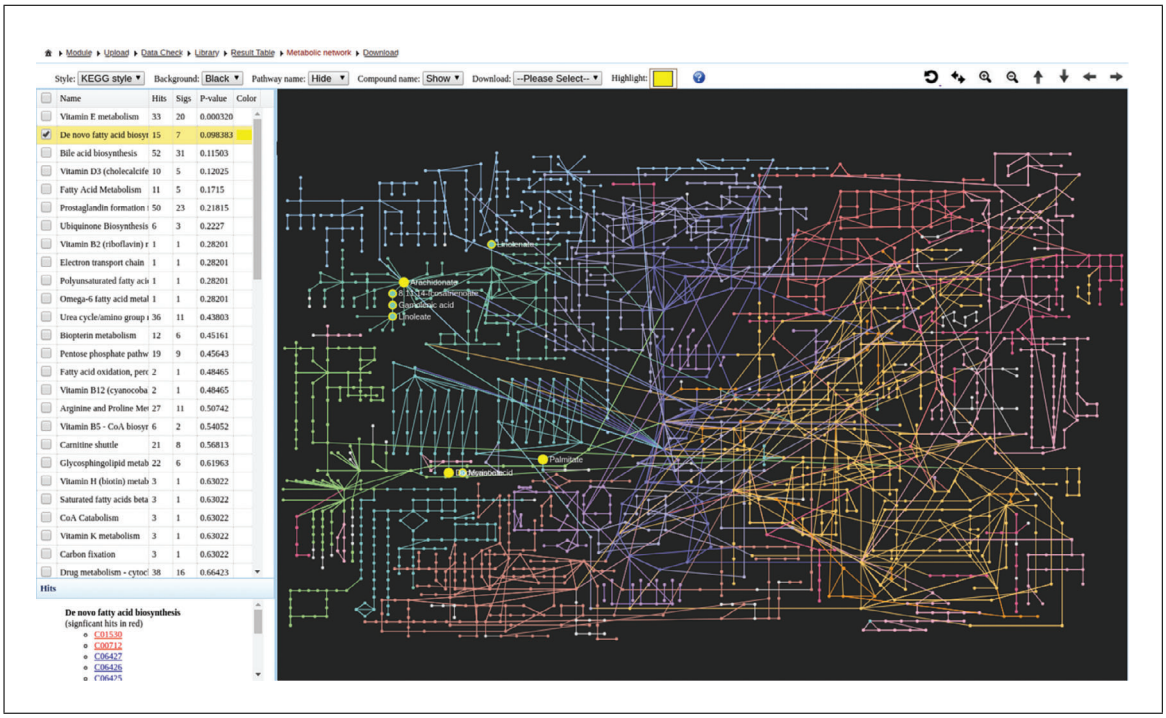


Figure 89 A screenshot of the KEGG global metabolic network. The compounds involved in the *De novo fatty acid biosynthesis* pathway are highlighted.

Figure 90 A screenshot of the dialog box for *Palmitate*, showing corresponding isotopic/adduct matches.

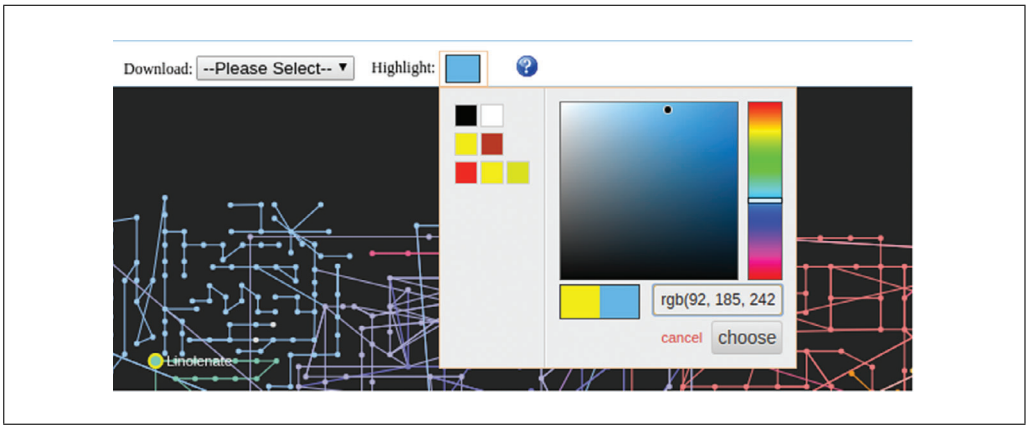


Figure 91 A screenshot of the color palette for customization of node colors in the KEGG global metabolic network. Here, the default code for yellow (`#ffff00`) was replaced with one for blue (`#5cb9f2`).

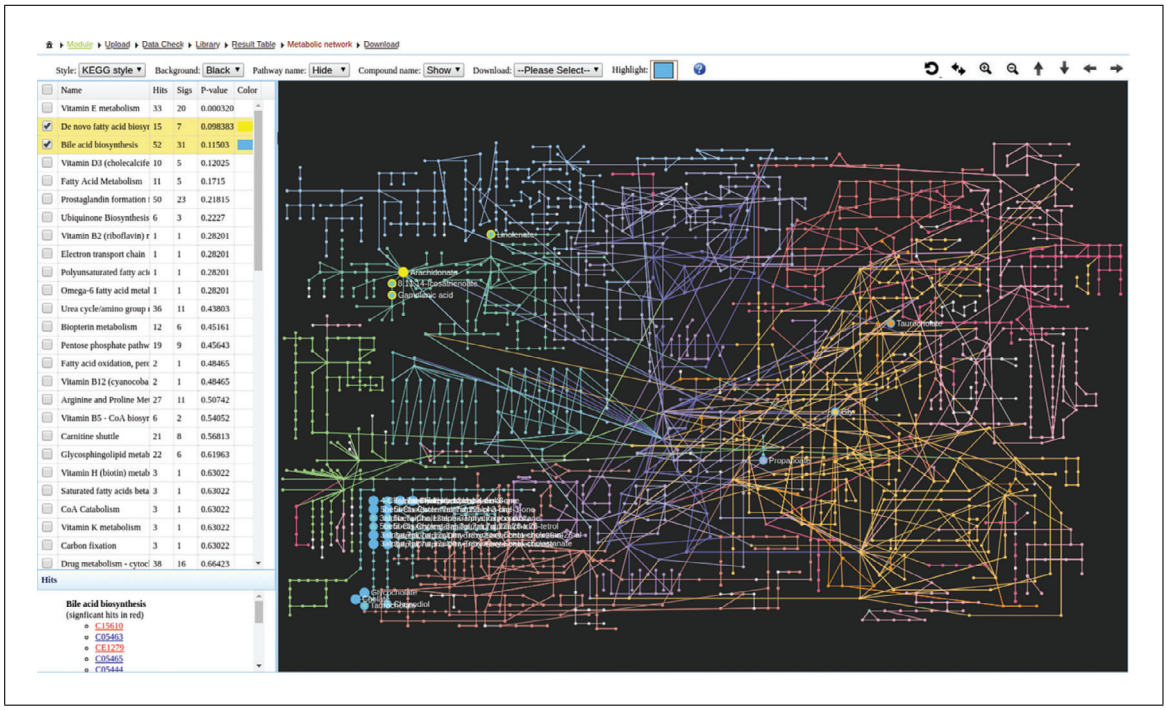
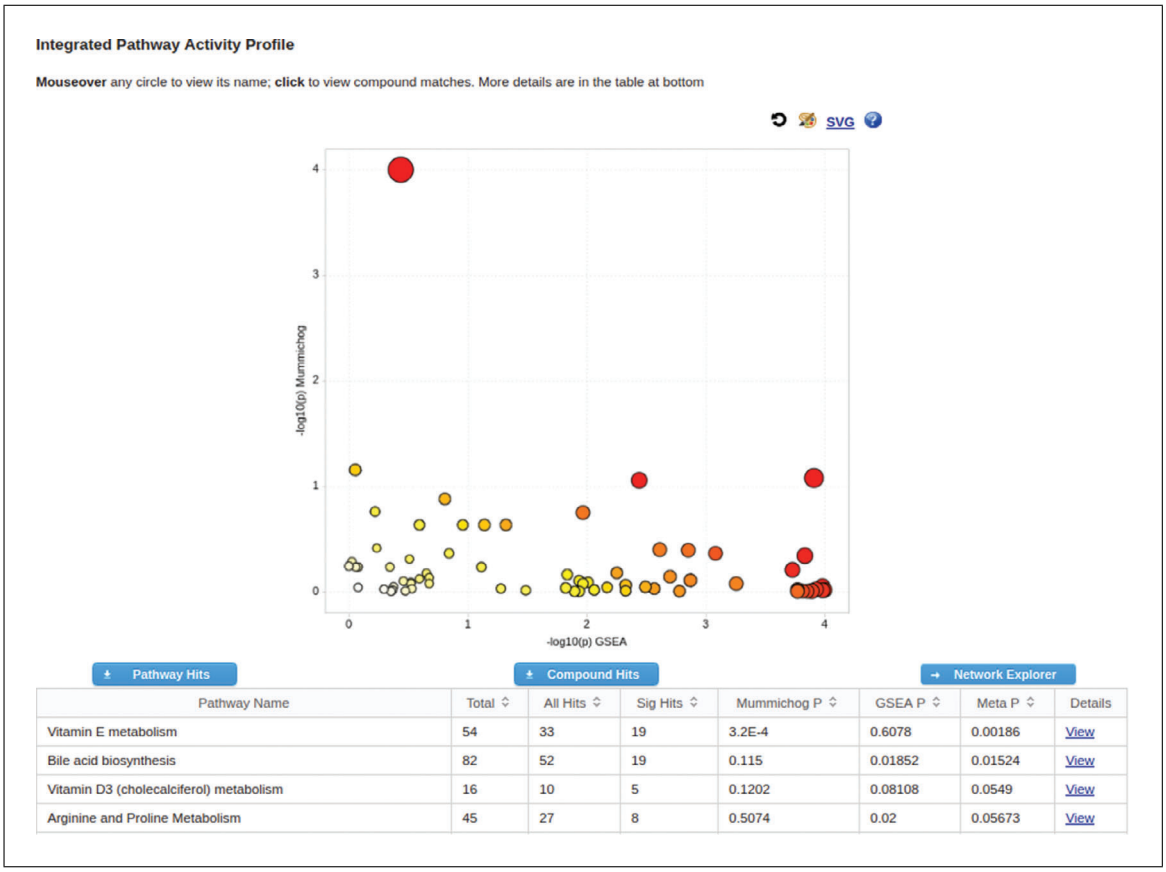


Figure 92 A screenshot of the KEGG global metabolic network highlighting all matched compounds between the example data and *De novo fatty acid biosynthesis* (highlighted in yellow) and *Bile acid biosynthesis* (highlighted in blue).

10. The next page displays the results of the combined *mummichog* and GSEA *p*-values (Fig. 93). The layout is identical to the *mummichog* results page (Fig. 87). The scatter plot at the top graphically summarizes the *p*-values from *mummichog* (y-axis) and GSEA (x-axis). The circles represent matched pathways, and the size and color of the circles correspond to their transformed combined *p*-values. *Bile acid biosynthesis* was consistently identified as one of the most enriched pathways in pediatric IBD patients as compared to healthy controls using each algorithm individually as well as combined.

11. Click the Download hyperlink of the navigation tree (left side of the page) to enter the "Download" page. Create the analysis report and download the R Command History,



Pathway Name	Total	All Hits	Sig Hits	Mummichog P	GSEA P	Meta P	Details
Vitamin E metabolism	54	33	19	3.2E-4	0.6078	0.00186	View
Bile acid biosynthesis	82	52	19	0.115	0.01852	0.01524	View
Vitamin D3 (cholecalciferol) metabolism	16	10	5	0.1202	0.08108	0.0549	View
Arginine and Proline Metabolism	45	27	8	0.5074	0.02	0.05673	View

Figure 93 A screenshot of the pathway summary plot integrating *mummichog* (y-axis) and GSEA (x-axis) *p*-values. The size and color of the circles correspond to the transformed combined *p*-values.

images, and result tables. Click "Logout" in the center of the page to complete this session.

BIOMARKER META-ANALYSIS

BASIC PROTOCOL 10

Biomarker identification is an important area of research in metabolomics, and biomarker validation is challenging due to inconsistencies in identified biomarkers among similar studies. With the growing applications of metabolomics in biomarker discoveries and the establishment of several metabolomics data repositories, there is a growing interest in performing secondary analysis of published metabolomics datasets collected under similar conditions. This is a practice often referred to as "meta-analysis." When executed properly, meta-analysis can leverage the collective power of multiple studies to help overcome study biases and small effect sizes to improve the precision in identifying true patterns and robust biomarkers. A key concept in meta-analysis is that it is generally not advisable to directly combine multiple independent datasets into a single large dataset and to analyze them as a single unit due to significant batch effects. Instead, it is suggested that meta-analysis should be performed based on summary statistics (such as *p*-values, effect sizes, etc.). This protocol will describe how to perform biomarker meta-analysis that allows users to integrate individual metabolomics studies to help identify robust metabolic biomarkers.

Necessary Resources

Hardware

A computer with internet access

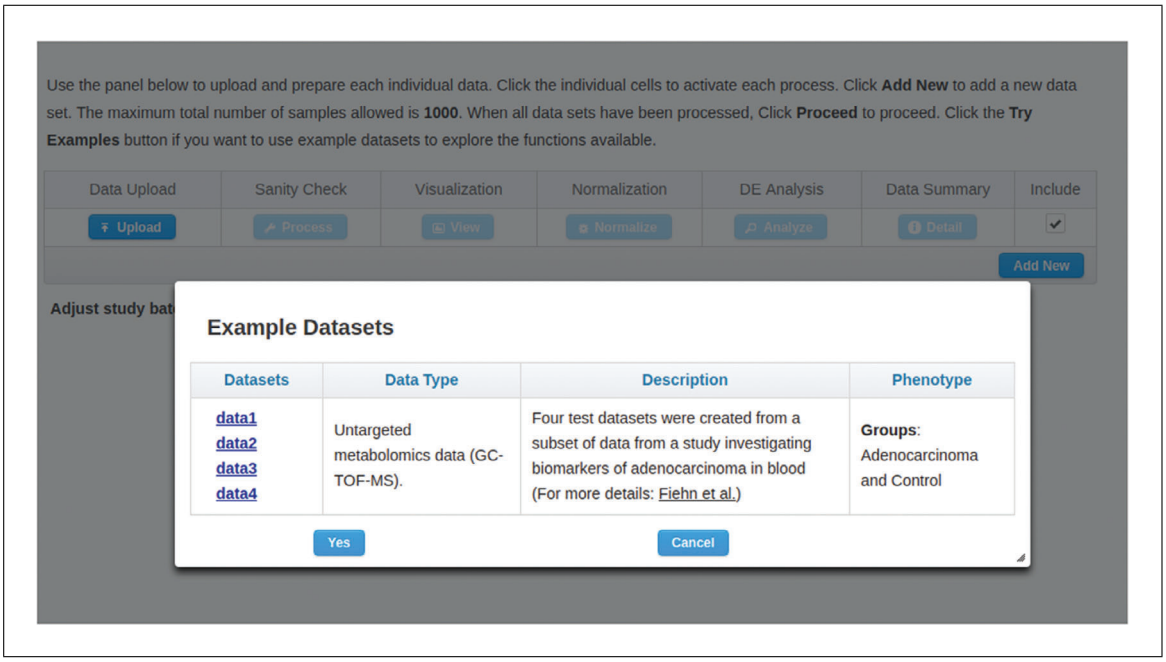


Figure 94 A screenshot of the Example Datasets dialog in the Biomarker Meta-Analysis module, which consists of a table with the example datasets available for download (right click each dataset and Save link as . . .), the data type, a brief description of the datasets, and the phenotypes of the datasets.

Software

An up-to-date web browser, such as Google Chrome (*http://www.google.com/chrome*), Mozilla Firefox (*http://www.mozilla.com/*), Safari (*http://www.apple.com/safari/*), or Internet Explorer (*http://www.microsoft.com/ie/*). JavaScript must be enabled in the web browser.

Files

None

Data upload and processing

1. Go to the MetaboAnalyst home page (*https://www.metaboanalyst.ca*). Click the "click here to start" link to enter the "Module Overview" page, and then click the "Biomarker Meta-Analysis" button to enter the data preparation page. In the center of the page is a table-based navigation panel containing six columns corresponding to the six steps in data processing—Data Upload, Sanity Check, Visualization, Normalization, DE Analysis, and Data Summary. One table row allows users to upload and process one dataset. To upload a second dataset, users need to first click the "Add New" button at the bottom right corner of the navigation panel to add a new row to the table. In this protocol, we will use the example data, which consist of four datasets created from a subset of data from a study investigating biomarkers of adenocarcinoma (Fahrmann et al., 2015). Click the "Try Examples" button in the bottom left of the page. A pop-up window will appear containing a table with the datasets (data 1-4) as well as further details about the example data (Fig. 94). Right click on each dataset (1-4) and select "Save link as . . . " to save the files onto one's computer. Click "Cancel" to close the Example Datasets dialog.

Before uploading individual datasets to the Biomarker Meta-Analysis module, ensure that all features names (compound names, spectral bins, peaks) are consistent between the individual studies. At least 25% of the features must match between the studies, or else the analysis will fail. Also make sure that the group labels are also consistent between the studies. Finally, all sample identifiers must be unique.

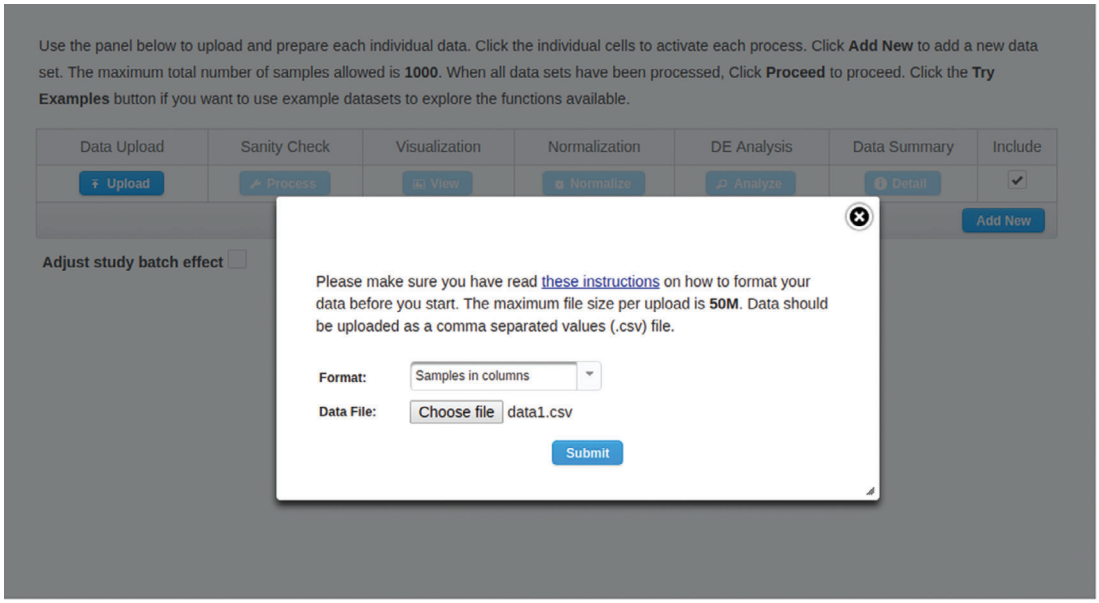


Figure 95 A screenshot of the Data Upload dialog. The format of the dataset is specified as "Samples in columns," and the data file selected is `data1.csv`.

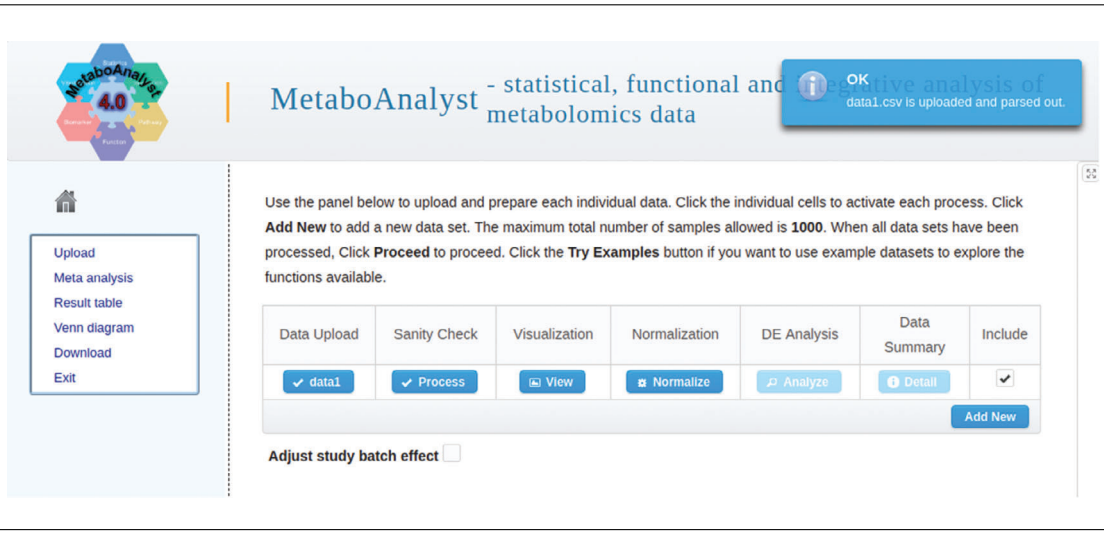


Figure 96 A screenshot of the successful upload message in the top-right corner for `data1.csv`.

2. To upload an individual dataset, click the "Upload" cell under the Data Upload column of the panel. First, specify the data format (samples in rows or columns). In this case, the example data are formatted as "Samples in columns." Next select "Choose file" and locate the first example file `data1.csv`, which was downloaded in the previous step (Fig. 95). Click "Submit" to upload the data. A message will pop up at the top-right corner of the screen summarizing the data upload attempt. If successful, the data name (`data1`) will appear in the "Data Upload" panel on the right (Fig. 96). Following the uploading of a dataset, the remaining buttons in the panel will now be activated.

3. Click "Process" to perform a sanity check of the uploaded data, which will verify that all sample names are unique, identify group labels, and check the number of samples and features uploaded (Fig. 97). Click "Done" to close the Sanity Check.

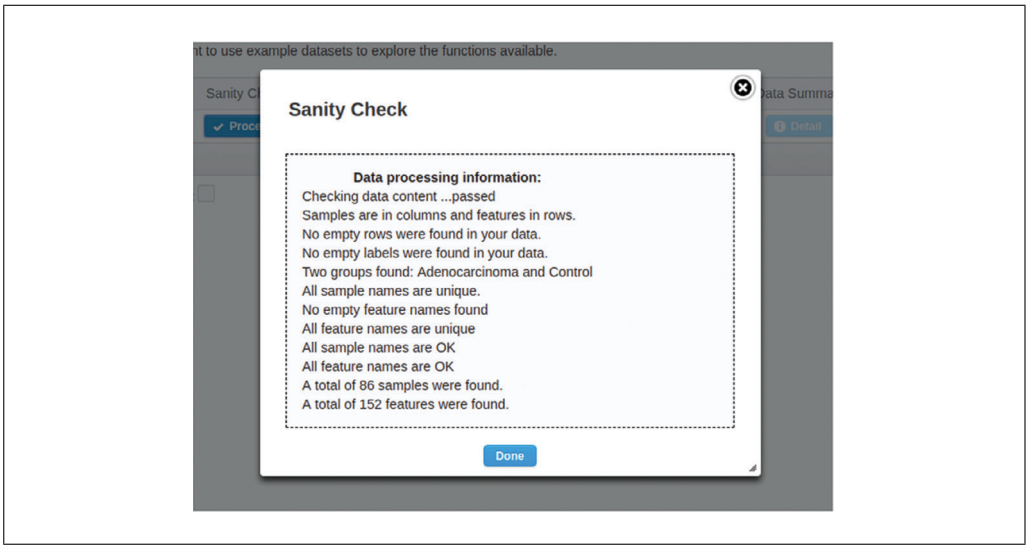


Figure 97 A screenshot of the Sanity Check dialog. This summarizes the results of the sanity check, including verifying that sample names are unique, identifying group labels, and checking the number of samples and features uploaded.

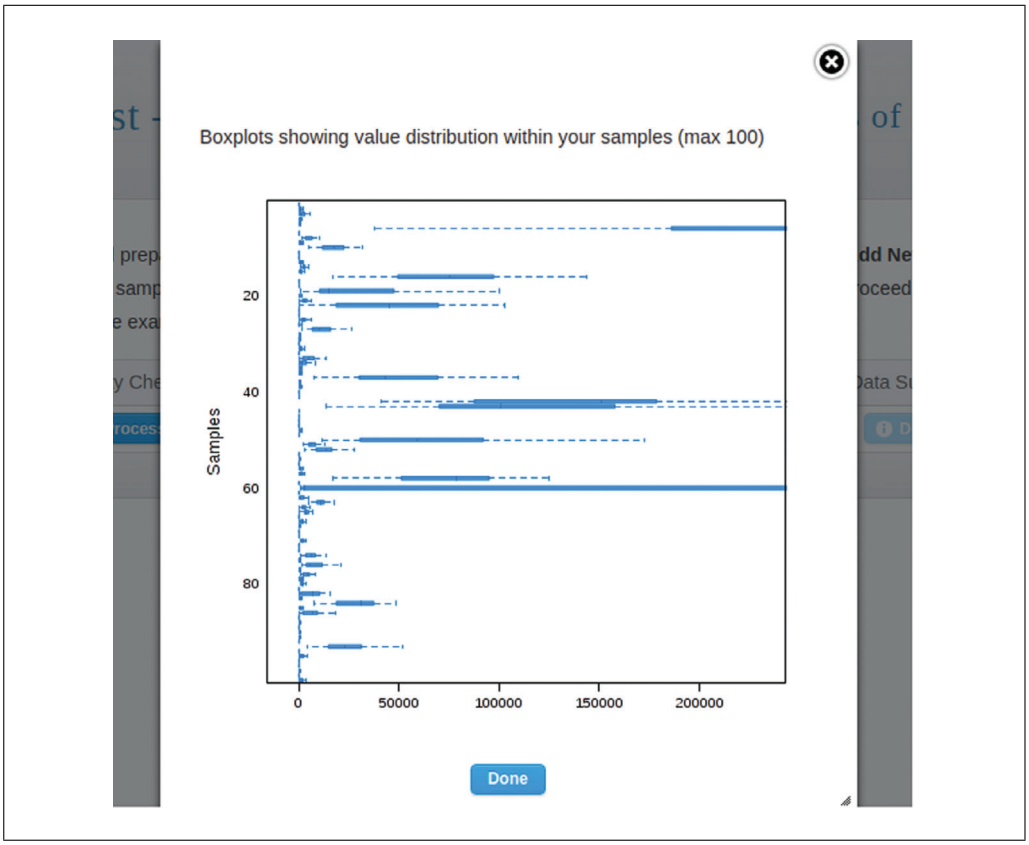


Figure 98 A screenshot of the box plot showing the value distributions of the first 100 samples of the uploaded example data. It is clear from the boxplots that the distributions are very uneven among the samples.

4. Click "View" to see the distribution of the first 100 samples in the uploaded data (Fig. 98). It is evident from the boxplots that the distributions are very uneven among the samples, indicating that further data normalization is necessary. Click "Done" to close the pop-up window.

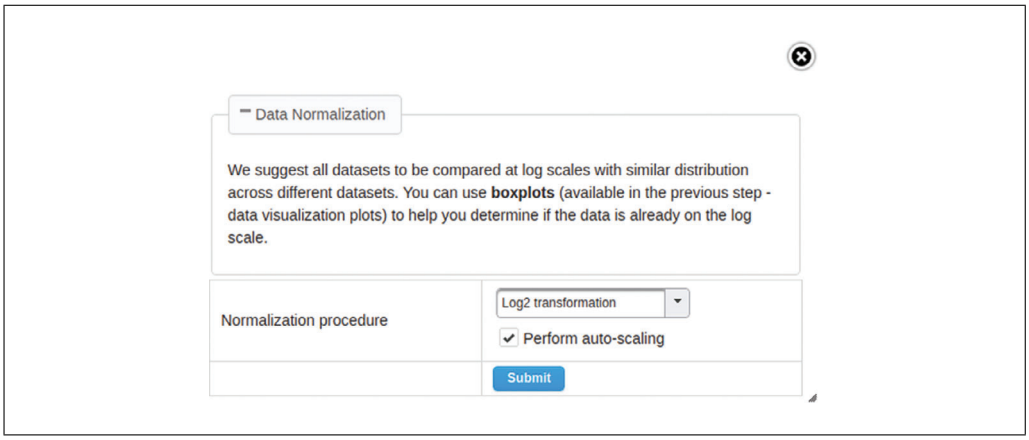


Figure 99 A screenshot of the Normalization dialog showing the two options for data normalization in the Biomarker Meta-Analysis module: Log2 transformation and auto-scaling.

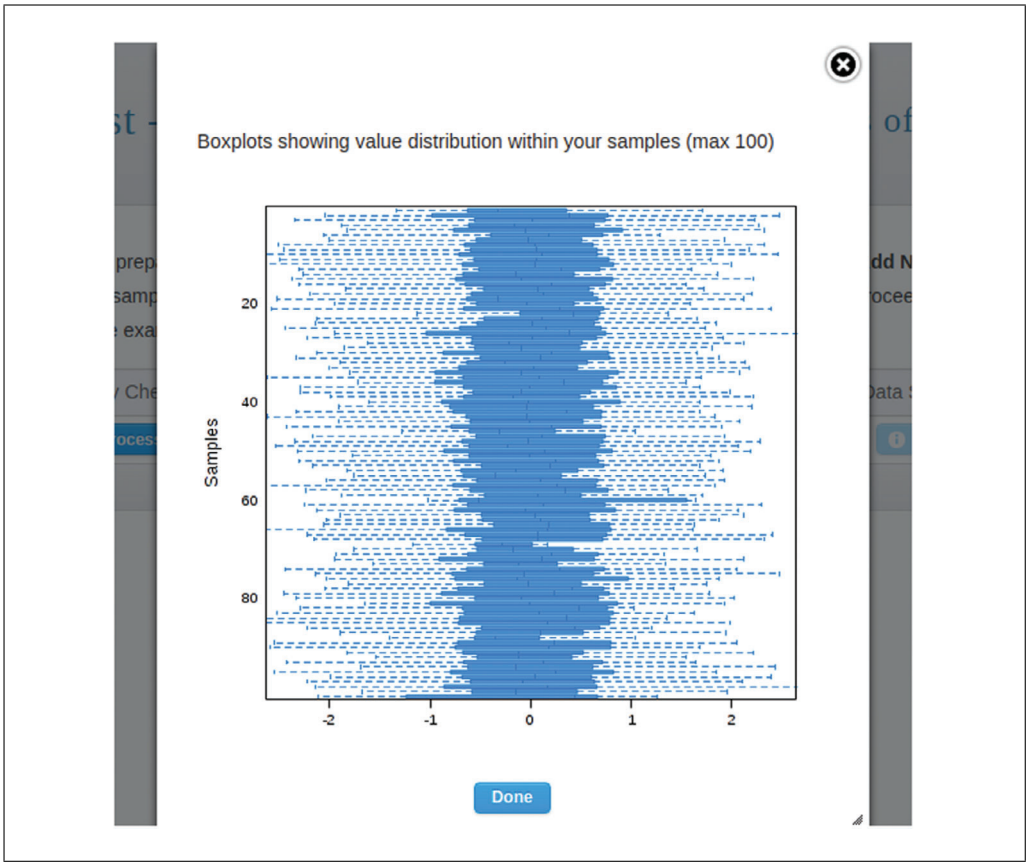


Figure 100 A screenshot of the box plot showing the value distributions of the first 100 samples of the uploaded example data following data normalization.

5. Click the "Normalize" cell under the Normalization header. Here, MetaboAnalyst provides two options—Log2 transformation and (optional) auto-scaling. In this case, select both methods to first transform the data into log scale and then scale to unit variance (Fig. 99). A pop-up window will appear in the top-right of the screen to indicate that the normalization was completed.

   Click the "View" button again. The boxplot shows that the data normalization was successful, as the data are now evenly distributed around 0 (Fig. 100). Click "Done" to close the boxplots.

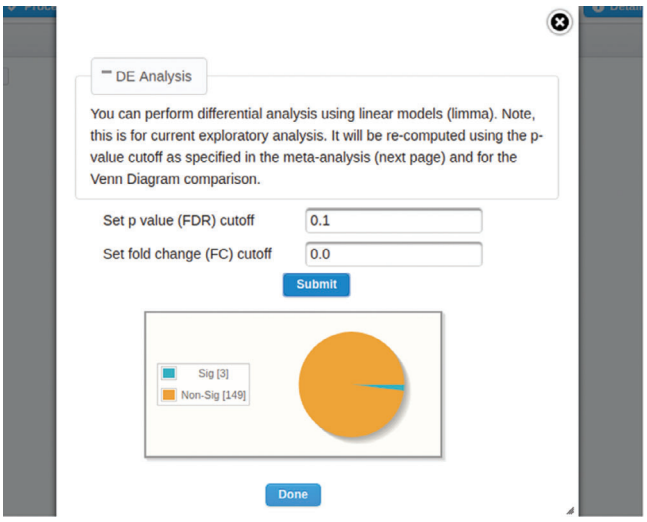


Figure 101 A screenshot of the Differential Expression (DE) Analysis dialog. This shows the results of the DE analysis in a pie chart, with a total of three significant metabolites (blue) and 149 non-significant metabolites (orange).

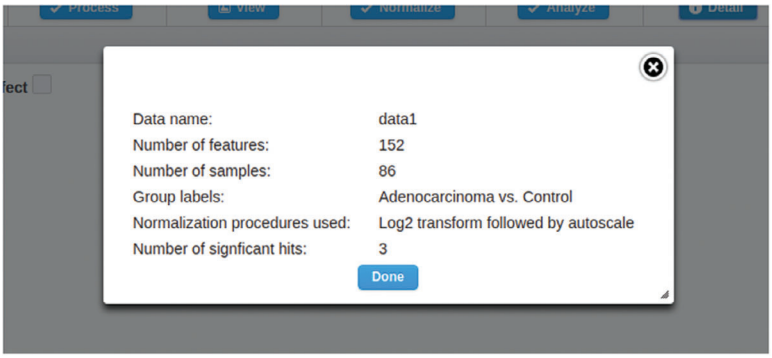


Figure 102 A screenshot of the Data Summary dialog, which shows an overview of the data processing steps performed on the uploaded dataset.

6. Following data normalization, click "Analyze" under the DE Analysis header. A dialog box appears to allow users to perform differential analysis using linear models (limma) on their data. In this case, the default *p*-value and fold change cut-offs will be used. Press "Submit" to continue. A pie chart will appear in the bottom of the dialog box indicating the number of significantly different features (Fig. 101). The result shows 3 significantly different features and 149 non-significantly different features at an adjusted *p*-value cut-off of 0.1. Click "Done" to close the dialog box.

7. Click the "Details" under the Data Summary header to obtain an overview of the data processing steps performed on the uploaded data. The dialog (Fig. 102) shows the number of uploaded features and samples, group labels, normalization procedures applied, and the number of significant features detected. Click "Done" to close the dialog box. The first dataset is now ready.

8. Repeat steps 2 to 7 for the remaining datasets (`data2`, `data3`, `data4`). However, instead of repeating these steps, we will use the preloaded example datasets for the remainder of the tutorial. Click the "Try Examples" button to bring up the dialog and click "Yes" to perform the default data uploading and processing procedures

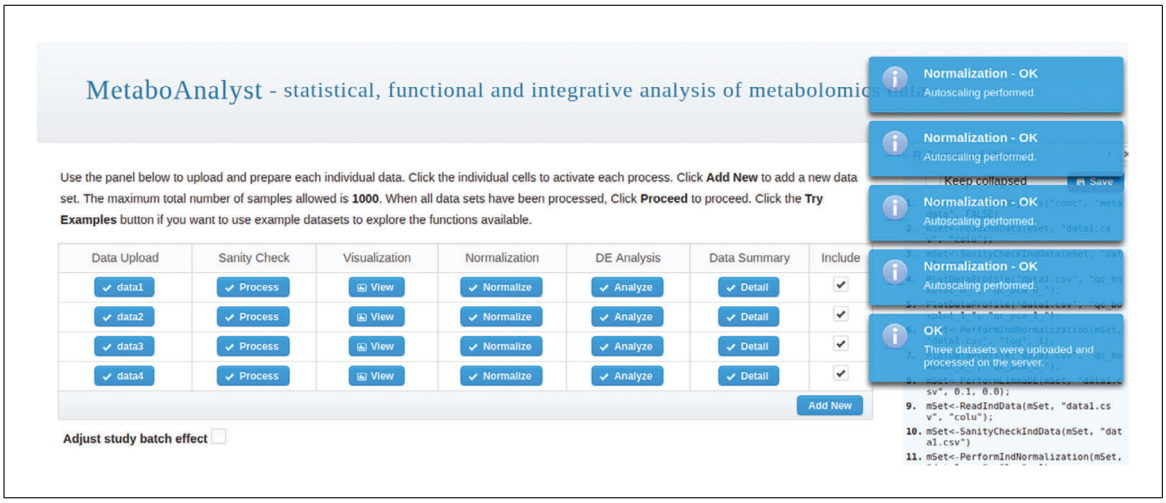


Figure 103 A screenshot of the "Try Examples" message indicating that the example datasets were successfully uploaded and processed.

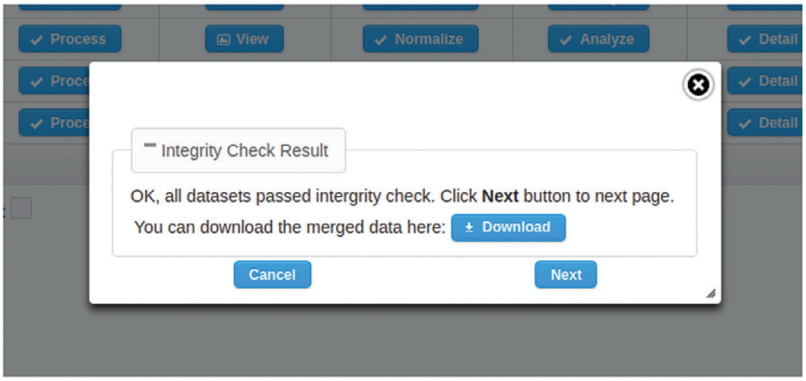


Figure 104 A screenshot of the Integrity Check Result dialog. This indicates that the uploaded datasets passed the necessary integrity check. The check will verify if group labels are consistent and if at least 25% of features are common, and determine the total number of uploaded samples.

for all four example datasets (Fig. 103). A message will appear indicating that the example datasets were successfully uploaded and processed.

9. Click "Proceed" on the bottom right of the page. An integrity check will be performed to determine if group labels are consistent and if at least 25% of features are common, as well as the total number of uploaded samples, etc. A dialog box will appear confirming that the uploaded datasets passed the integrity check (Fig. 104). Click "Next" to continue.

   Note that a maximum of 1000 samples can be uploaded. In addition, after uploading all intended datasets, users can uncheck the boxes in the "Include" column of the data-upload navigation panel to exclude a study from the meta-analysis. Select the "Adjust study batch effect" checkbox under the data upload panel (Fig. 103) if you would like batch effect to be accounted for in the meta-analysis. The batch effect adjustment is based on the well-established ComBat procedure (Johnson, Li, & Rabinovic, 2007). It is very effective in adjusting for batch effects in high-dimensional omics data with small sample sizes. The procedure consists of three steps: (1) features absent in more than 80% of samples are eliminated to minimize noise, and the remaining values are then standardized to have similar overall mean and variance; (2) information is pooled across features from a batch to determine batch effect estimates that could affect multiple features, such as increased abundance level and higher variability; and (3), the estimated

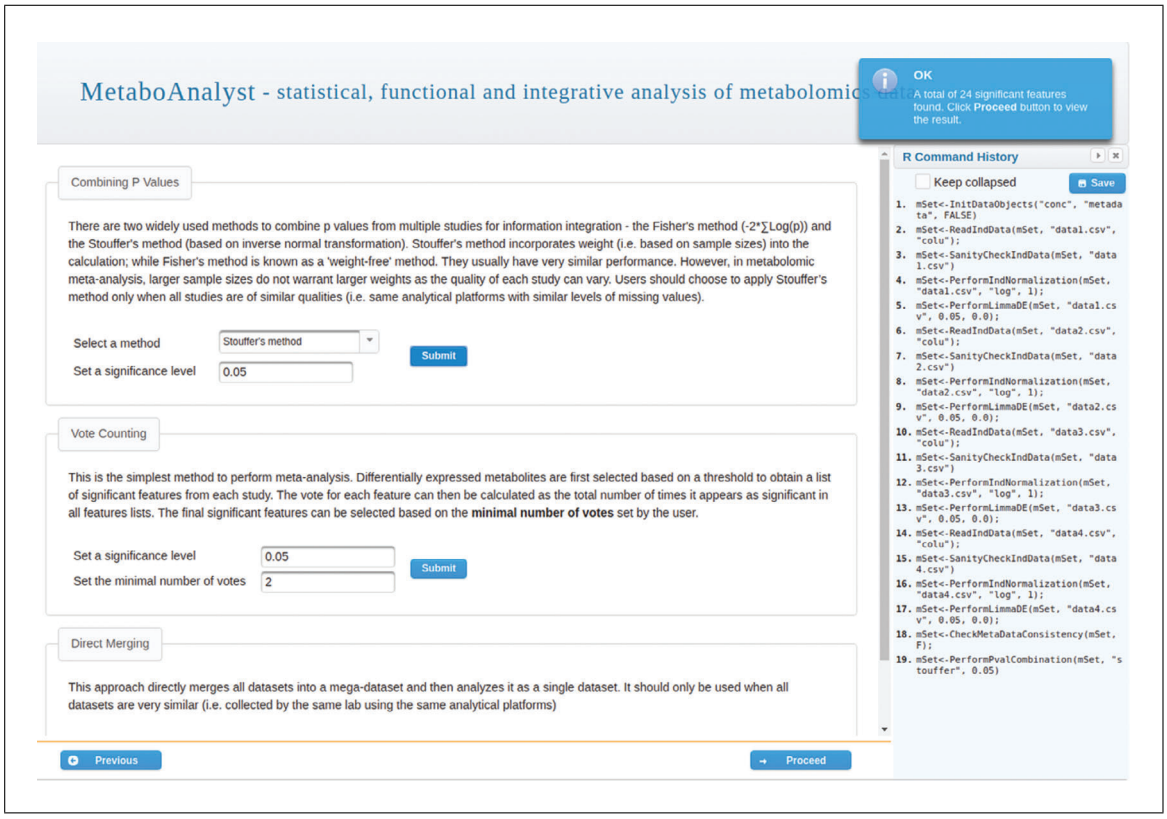


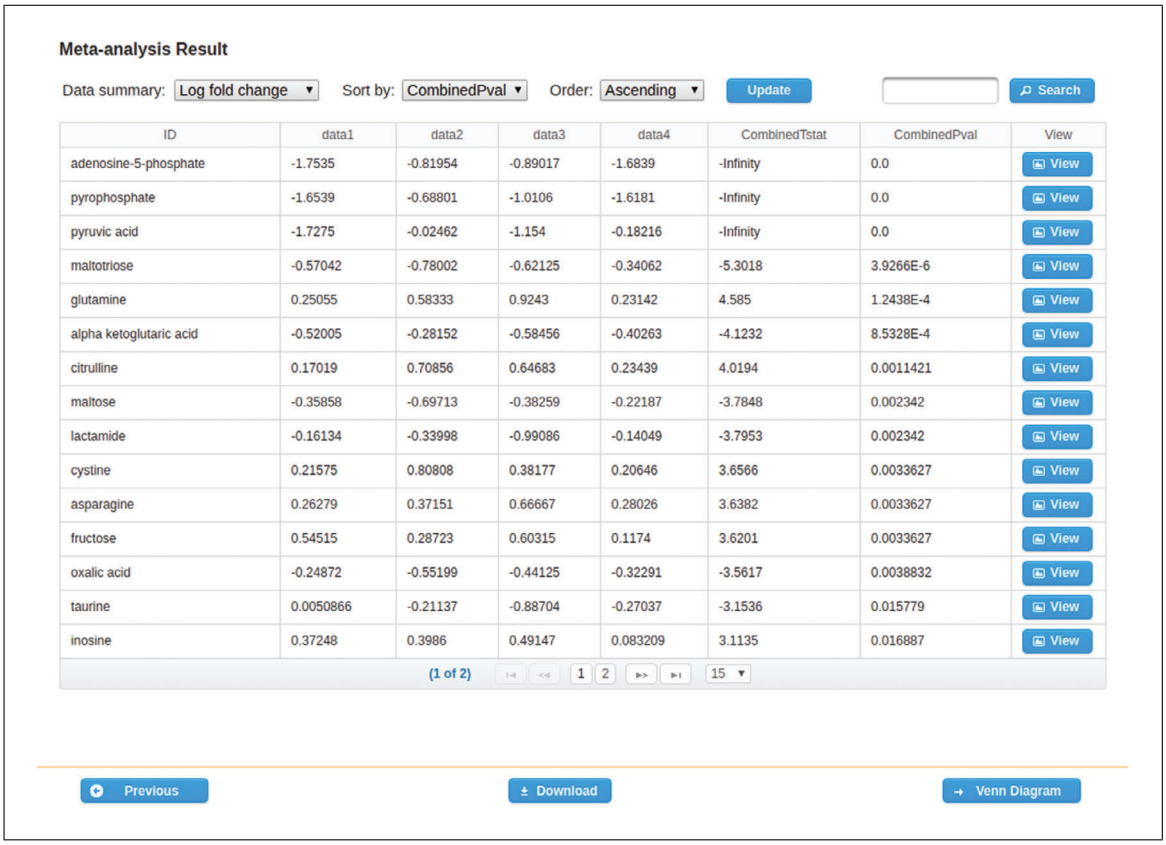
Figure 105 A screenshot showing the total number of significant features identified in the meta-analysis using the "Combining P Values" method. Stouffer's method and a *p*-value cutoff of 0.05 was used.

batch effects are then adjusted to obtain normalized data, which are more suitable for statistical comparisons and meta-analysis.

Statistical meta-analysis

10. The page shows three commonly used statistical methods to perform meta-analysis, "Combining P Values," "Vote Counting," and "Direct Merging." For this demonstration, use the popular *p*-value combination method. In the "Combining P Values" box, click the "Select a method" box to choose either Fisher's method or Stouffer's method for combining *p*-values. Stouffer's method incorporates weight based on sample sizes into the calculation, while Fisher's method is known as a "weight-free" method. As the four example datasets are of similar quality (generated by the same GC-TOF-MS instrumentation and all studies have no missing values), select Stouffer's method to give more weight to larger studies. By default, the significance level in "Set a significance level" is 0.05, which will be kept for the meta-analysis. Click "Submit" to perform the *p*-value combination. A dialog box will appear in the top-right corner of the screen, informing users of the total number of significant features identified in the meta-analysis (Fig. 105). In this case, 24 significant features were identified. Click "Proceed" at the bottom right corner of the page to view the detailed results of the meta-analysis.

Combining p-values** is a well-established approach to perform meta-analysis. MetaboAnalyst supports the two most popular methods, Fisher's method and Stouffer's method, which have similar levels of performance and can be easily interpreted as larger scores reflecting greater differential abundance. The main difference is that weights (i.e., based on sample sizes) are incorporated into the calculation in Stouffer's method, whereas Fisher's method is known as a weight-free method. Note that a larger sample size does not warrant larger weights, as the quality of each study can be variable. Users should choose to apply Stouffer's method only when all studies are of similar quality. **Vote



ID	data1	data2	data3	data4	CombinedTstat	CombinedPval	View
adenosine-5-phosphate	-1.7535	-0.81954	-0.89017	-1.6839	-Infinity	0.0	View
pyrophosphate	-1.6539	-0.68801	-1.0106	-1.6181	-Infinity	0.0	View
pyruvic acid	-1.7275	-0.02462	-1.154	-0.18216	-Infinity	0.0	View
maltotriose	-0.57042	-0.78002	-0.62125	-0.34062	-5.3018	3.9266E-6	View
glutamine	0.25055	0.58333	0.9243	0.23142	4.585	1.2438E-4	View
alpha ketoglutaric acid	-0.52005	-0.28152	-0.58456	-0.40263	-4.1232	8.5328E-4	View
citrulline	0.17019	0.70856	0.64683	0.23439	4.0194	0.0011421	View
maltose	-0.35858	-0.69713	-0.38259	-0.22187	-3.7848	0.002342	View
lactamide	-0.16134	-0.33998	-0.99086	-0.14049	-3.7953	0.002342	View
cystine	0.21575	0.80808	0.38177	0.20646	3.6566	0.0033627	View
asparagine	0.26279	0.37151	0.66667	0.28026	3.6382	0.0033627	View
fructose	0.54515	0.28723	0.60315	0.1174	3.6201	0.0033627	View
oxalic acid	-0.24872	-0.55199	-0.44125	-0.32291	-3.5617	0.0038832	View
taurine	0.0050866	-0.21137	-0.88704	-0.27037	-3.1536	0.015779	View
inosine	0.37248	0.3986	0.49147	0.083209	3.1135	0.016887	View

(1 of 2) 1 2 15

Figure 106 A screenshot showing the detailed results table of the meta-analysis. The enriched pathways are ranked by their combined *p*-value across the datasets. Click "View" to visualize their differences across the datasets.

> ***counting*** *is the most primitive yet simplest and most intuitive method of meta-analysis. In this approach, significant features are first selected based on some criteria (e.g., adjusted* $p < .05$ *and same direction of fold change) for each data set. The vote for each feature can then be calculated by counting the total number of times it occurs as being statistically significant across all data sets. This method is statistically inefficient and should be considered as a last resort in situations when other meta-analysis methods cannot be applied.* ***Direct merging*** *of different data sets into a single mega-data set ignores the inherent bias and heterogeneity of data sets that originate from different sources, as many factors (experimental protocols, technical platforms, raw data processing procedures and so forth) can potentially contribute to observed differences. Therefore, this approach should only be used when data sets are very similar (i.e., from the same lab and same platform without batch effects).*

11. The results of the "Combining P Values" meta-analysis are listed in a table on the new page (Fig. 106). In this table, the features are ranked by their combined *p*-values. It is evident that *adenosine-5-phosphate*, *pyrophosphate*, and *pyruvic acid* are the top three biomarkers identified, all with a combined *p*-value of 0.

12. Click "View" to view a box-plot summary of the expression pattern of *adenosine-5-phosphate (A5P)*. As shown in Figure 107, the concentration of *A5P* is consistently higher in patients with adenocarcinoma than in healthy controls across all uploaded datasets.

Result exploration using an interactive Venn diagram

13. Click "Venn Diagram" at the bottom of the page to visualize the meta-analysis results in a Venn diagram. A dialog box will appear summarizing results from analyzing individual datasets as well as the result of the meta-analysis (named `meta_dat`). Note that the Venn diagram implemented in MetaboAnalyst supports a maximum

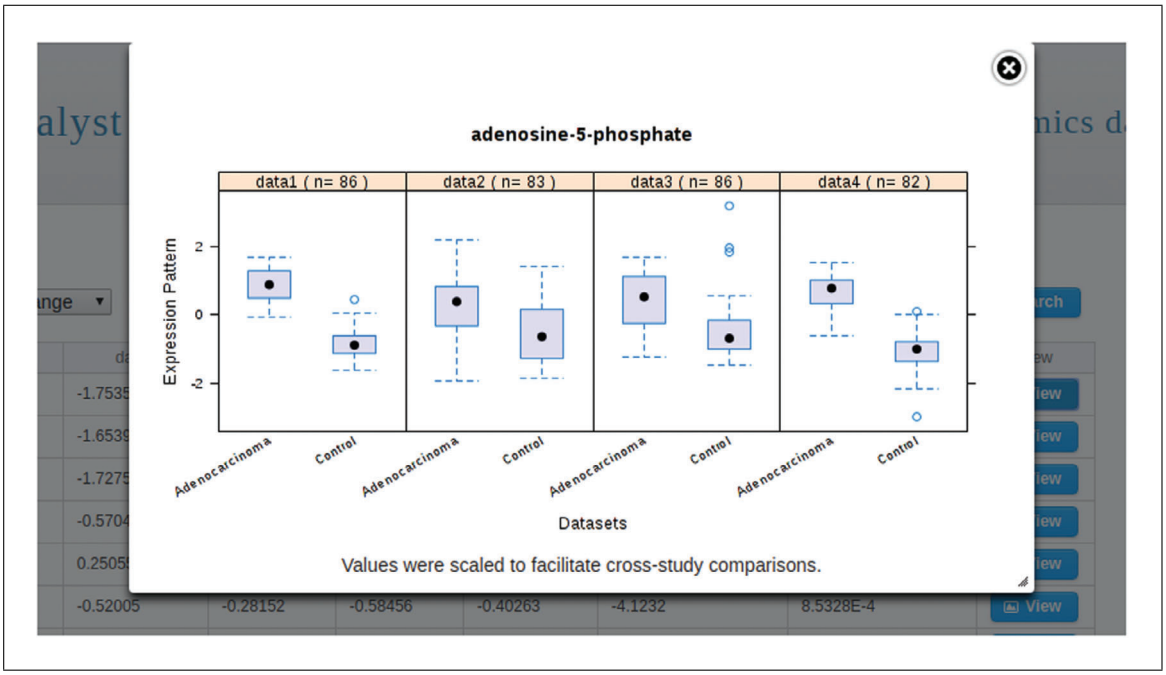


Figure 107 A screenshot showing a box plot of the different concentrations of Adenosine-5-Phosphate across the different uploaded datasets between Adenocarcinoma and Healthy Control patients.

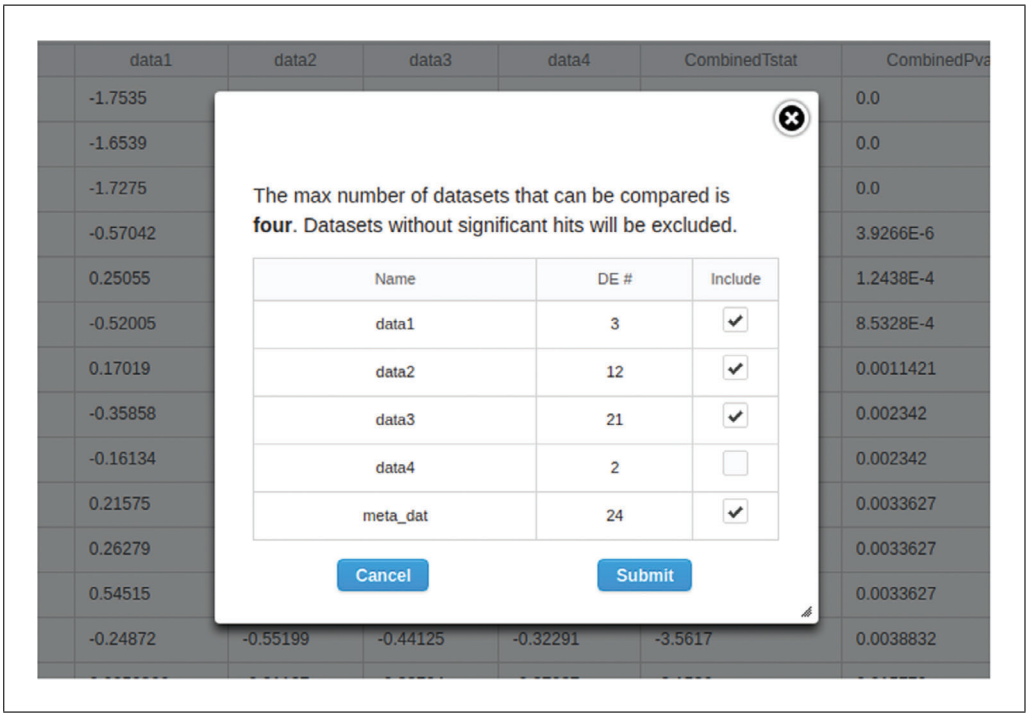


Figure 108 A screenshot of the Venn diagram dialog for the Biomarker Meta-Analysis module. The table shows the names of all datasets included in the meta-analysis, the number of differently expressed features in each dataset, and whether the dataset will be included in the construction of the Venn diagram. In this case, data4 will not be included in the Venn diagram.

of four datasets. In this case, we exclude dataset #4, as it has the lowest number of DE features (Fig. 108). Click "Submit" to continue.

14. The Venn diagram is shown on the new page. Users can explore different combinations of the four datasets and identify common or unique features. Click directly on the center of the Venn Diagram: the area becomes shaded and a list of

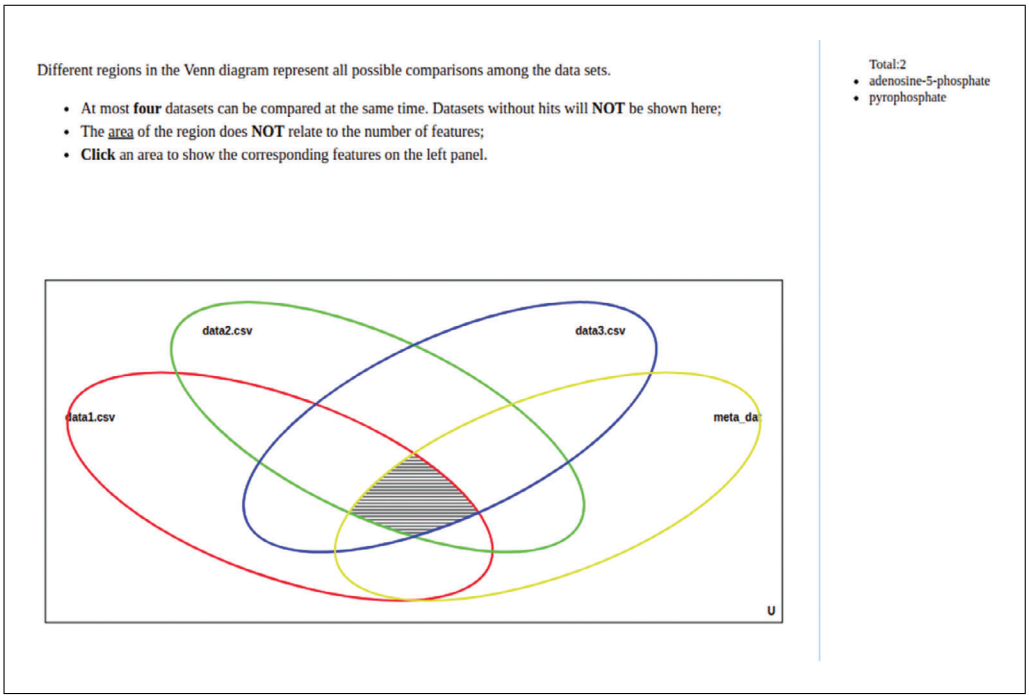


Figure 109 A screenshot of the Venn diagram that represents all possible combinations of overlapping significant features between the example data (1-3) and the meta-analysis result. A total of two significant features (Adenosine-5-Phosphate and Pyrophosphate) are consistently identified across all the datasets.

corresponding features appears in the right-side panel (Fig. 109). A total of two features (*A5P* and *pyrophosphate*) are significant in all three datasets, 1, 2, and 3, and the meta-analysis. Click on any areas of interest to view the possible combinations of features.

Data download

15. Click the "Download" hyperlink on the navigation tree to enter the "Download" page. Users can view all results generated during the analysis, as well as create a PDF analysis report. Download the results and the analysis report. Click "Logout" at the bottom of the results table to exit the session.

KNOWLEDGE-BASED NETWORK EXPLORATION OF MULTI-OMICS DATA

BASIC PROTOCOL 11

Metabolomics is increasingly applied together with other omics technologies such as transcriptomics, proteomics, and metagenomics to gain functional insight into complex diseases/conditions. However, interpretation of multi-omics data at a systems level remains a significant challenge. A common strategy is to analyze each set of omics data individually, and then piece together the "big picture" using the significant features (genes, proteins, metabolites, etc.) identified from individual omics analysis. Biological networks are useful and flexible means to represent our knowledge at a systems level. By harnessing the power of networks and a priori biological knowledge, these lists of significant features can be co-projected onto knowledge-based networks to reveal important links between the molecules of interest. The networks can also be used to identify their associations with diseases or other interesting phenotypes. Network visualization of users' data can be employed to gain novel insights or assist users with the development of new hypotheses. The Network Explorer module in MetaboAnalyst 4.0 was developed to support a network-based approach for multi-omics data integration and interpretation. The aim of this module is to provide an easy-to-use tool that permits the mapping of

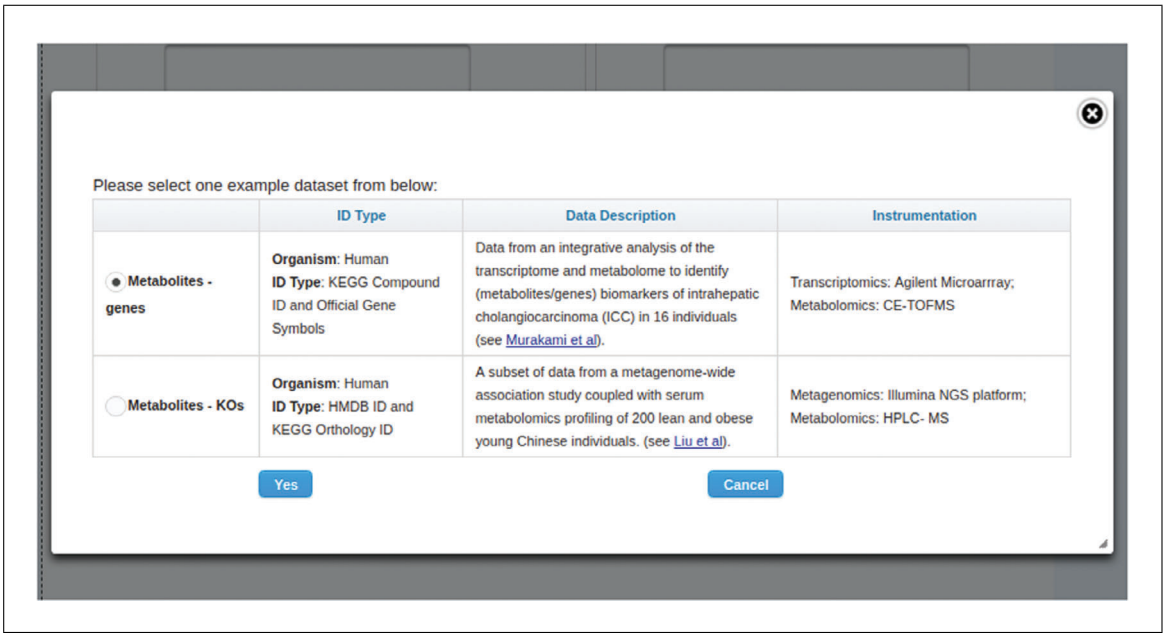


Figure 110 A screenshot of the Example Data Upload dialog for the Network Explorer module. The metabolites-genes data are selected, which were derived from humans and consist of KEGG compound IDs and Entrez Gene IDs.

metabolites and/or genes (including KEGG orthologs or KOs) onto different types of molecular interaction networks, as well as molecule-phenotype association networks. MetaboAnalyst 4.0 supports the integration of transcriptomics and metabolomics data, as well as metagenomics and metabolomics data. This protocol will give detailed instructions on how to use the Network Explorer module using a list of metabolites and genes.

Necessary Resources

Hardware

A computer with internet access

Software

An up-to-date web browser, such as Google Chrome (*http://www.google.com/chrome*), Mozilla Firefox (*http://www.mozilla.com/*), Safari (*http://www.apple.com/safari/*), or Internet Explorer (*http://www.microsoft.com/ie/*). JavaScript must be enabled in the web browser.

Files

None

Data upload and processing

1. Go to the MetaboAnalyst home page (*https://www.metaboanalyst.ca*). Click the "click here to start" link to enter the "Module Overview" page. Click the "Network Explorer" button to enter the corresponding "Data Upload" page. Users must enter a list of genes and a list of metabolites, with optional abundance values (i.e., log fold changes). These genes and metabolites should be obtained from the same biological samples or obtained under similar conditions. For this demonstration, use the example lists provided in MetaboAnalyst's example data set. Click the "try our example data" hyperlink at the top-right corner of the screen. A dialog box will appear showing the two example datasets with corresponding details such as Organism, ID type, study design, and instrument used (Fig. 110). In this case, select

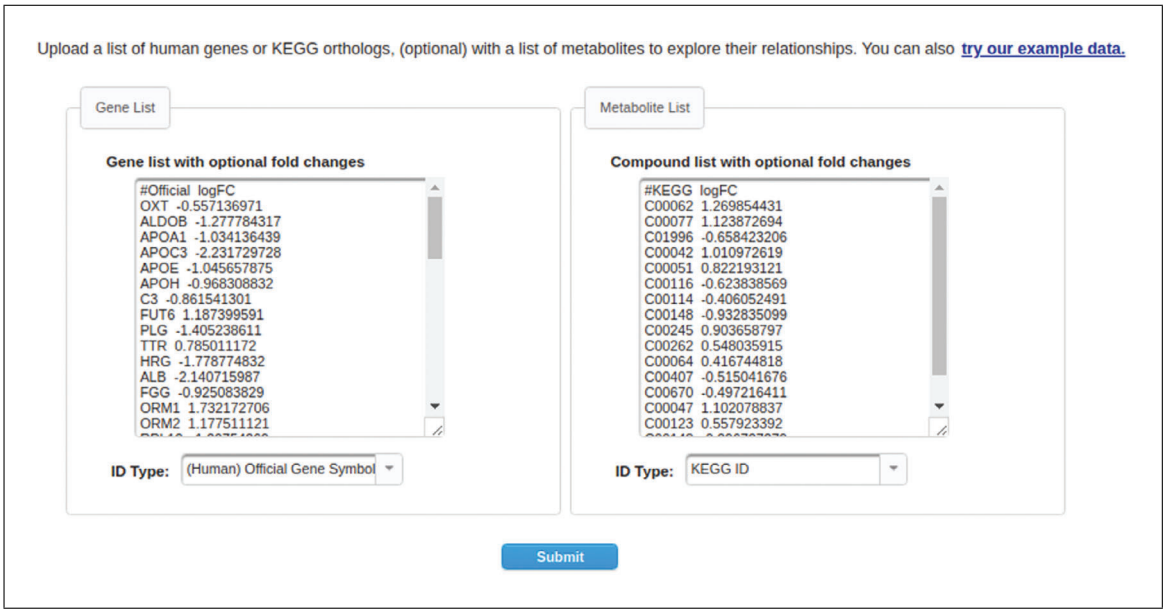


Figure 111 A screenshot of the Data Upload page for the Network Explorer module using the example data. The ID types for the gene list was set to (Human) Entrez ID and set to KEGG ID for the metabolite list.

the default "Metabolites-genes" example data, which consist of data taken from an integrative transcriptome-metabolome study on potential biomarkers of intrahepatic cholangiocarcinoma (ICC) in 16 individuals, using gene expression microarray and capillary electrophoresis time-of-flight mass spectrometry (CE-TOF-MS) analysis of tumor and surrounding non-tumor tissues (Murakami et al., 2015). For demonstration purposes, the example data were slightly modified (all log fold-changes were fabricated, and three metabolites were added). Click "Yes" to upload these data (Fig. 111). Click the "Submit" button to perform gene and metabolite ID mapping.

MetaboAnalyst 4.0 currently supports four types of gene identifiers—Entrez, Ensembl gene ID, official gene symbol, and KEGG Ortholog (KO), as well as three types of metabolite identifiers (compound name, HMDB and KEGG ID).

2. The next page shows the results of the gene/metabolite mapping to the underlying MetaboAnalyst databases. There are two tabs on this page—Compound Name Mapping and Gene Name Mapping. For the gene name mapping result table, genes highlighted in red indicate cases where there are no matches to the database, and genes in gray indicate their KO are not available (Fig. 112). Users can delete any matches from further analysis if they believe the matches are erroneous. The name mapping results can be downloaded using the hyperlinks at the bottom of the tables. After reviewing the results, click "Submit" at the bottom of the page.

Selection of network analysis parameters

3. The Network Explorer module currently provides five types of networks including KEGG global metabolic network, gene-metabolite interaction network, metabolite-disease interaction network, metabolite-metabolite interaction network, and metabolite-gene-disease interaction network (Fig. 113). Detailed descriptions are provided under each network option. For this tutorial, click on the "Metabolite-Gene-Disease Interaction Network" hyperlink at the bottom of the page.

MetaboAnalyst uses a comprehensive database on associations between genes, metabolites, and diseases based on various experimental and computational evidence. It currently contains 11,502 genes, 1900 metabolites, and 132 diseases, making for a total of

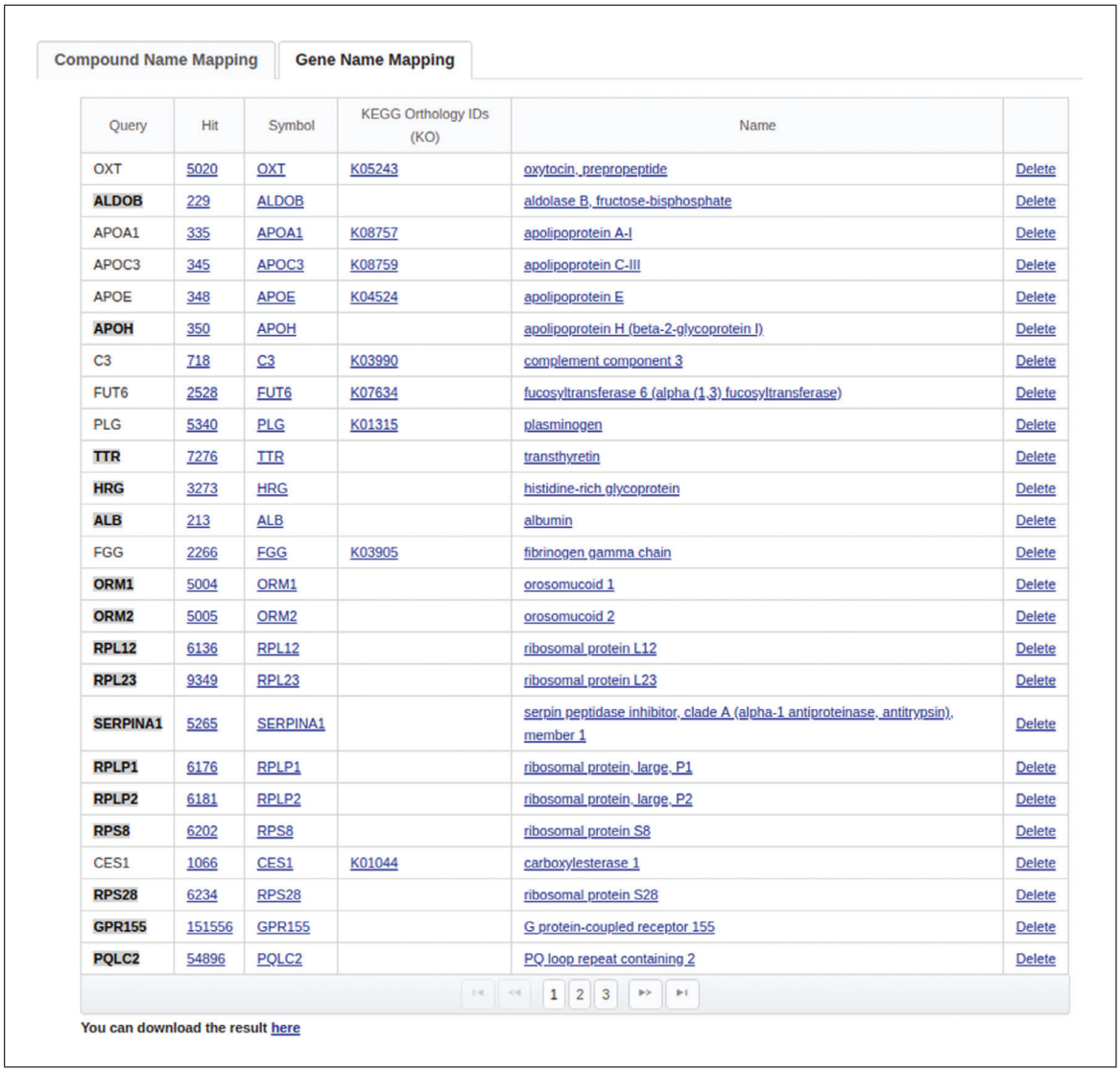


Figure 112 A screenshot of the metabolite and gene mapping results for the Network Explorer module. Genes highlighted in gray represent queries with incomplete matches to the underlying MetaboAnalyst database.

139,570 gene-metabolite associations, 78,200 metabolite-metabolite associations, and 519 metabolite-disease associations. The metabolite-disease associations were based primarily on published literature obtained from HMDB. The gene-chemical and chemical-chemical associations for creating the gene-metabolite and metabolite-metabolite networks, respectively, were extracted from STITCH (Kuhn, von Mering, Campillos, Jensen, & Bork, 2007), using only highly confident interactions. Finally, the metabolite-gene-disease network is an integration of the gene-metabolite, metabolite-disease, and gene-disease associations (Yao et al., 2015).

4. The next page provides an overview of the mapping of the input data to the "Metabolite-Gene-Disease Interaction Network" (Fig. 114). The metabolites and/or genes (called seeds) are mapped onto the selected network to create subnetworks containing these seeds and their direct neighbors (i.e., first-order subnetworks). This often produces one large subnetwork with several smaller ones. Subnetworks with at least three nodes are listed in the table, all of which can be visually explored in the next step. These subnetworks can be downloaded as SIF (Simple Interaction Format) files to be explored in other tools such as Cytoscape. Click "Proceed" to continue to the next step.

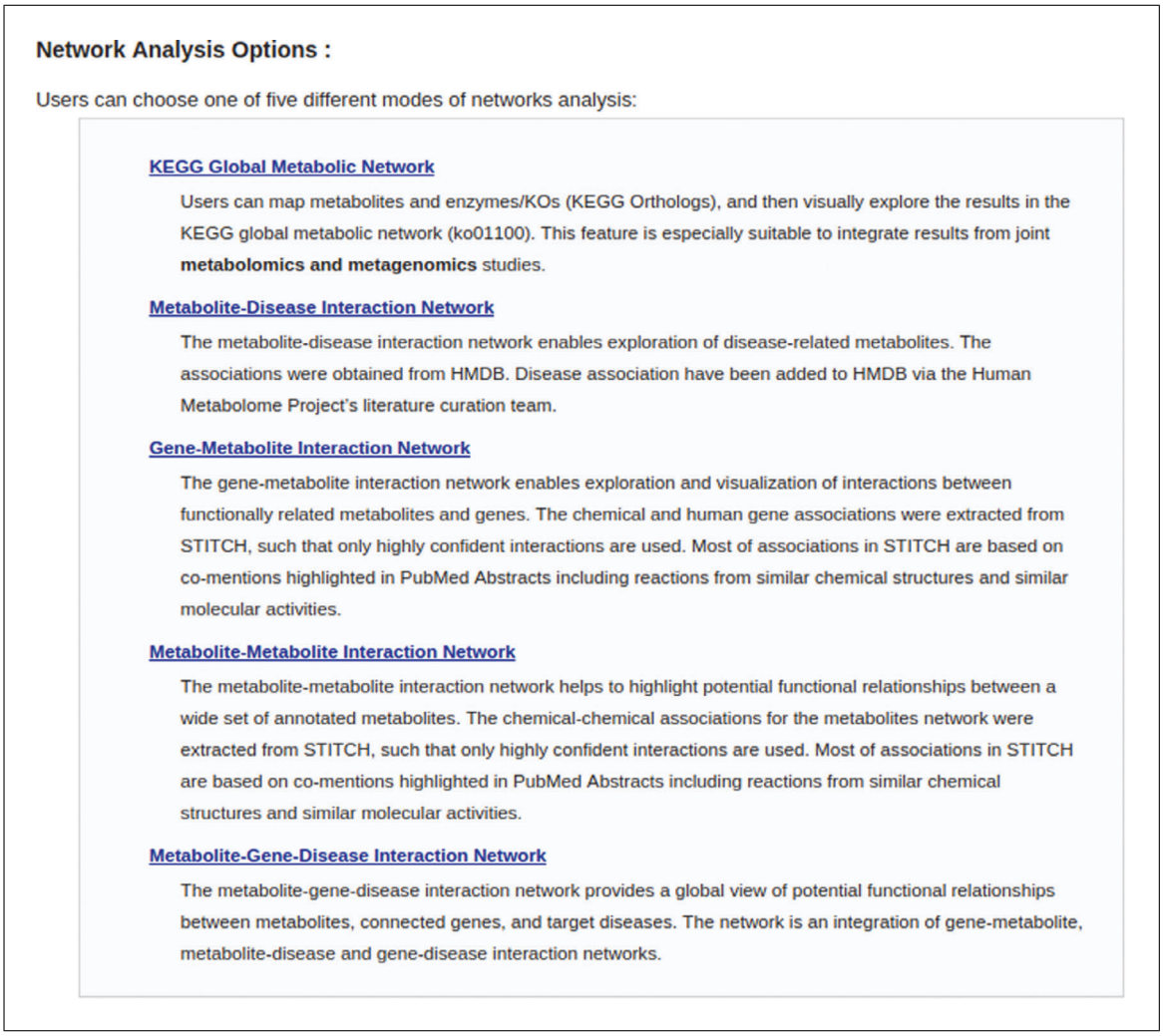


Figure 113 A screenshot showing the five options for network analysis: the KEGG global metabolic network, a gene-metabolite interaction network, a metabolite-disease interaction network, a metabolite-metabolite interaction network, and a metabolite-gene-disease interaction network.

Basic network visualization

5. The next page shows the default view of the first subnetwork ("subnetwork1"; Fig. 115). The page consists of four sections—the top toolbar, the Node Explorer on the left, the Function Explorer on the right, and the Network Viewer at the center. In this network, metabolites are represented as diamonds, genes are represented as circles, and diseases are represented as squares. The size of a node corresponds to its node degree (further discussed in step 6). Users can use the mouse scroll wheel to zoom in and out of the network, directly drag and drop nodes, or click on a node/edge to view more details.

   To change the color of the nodes, click the colored box on the top left side of the network viewer (Fig. 116). A color palette will appear; select a color and then click "choose." The new color will be applied when a user double clicks a node or highlights nodes involved in a function, as described below.

Exploration of important nodes in the network

6. Potentially important nodes can be identified based on their positions in the network. The assumption is that nodes in key positions are more likely to play important roles than marginal or relatively isolated nodes. The Node Explorer

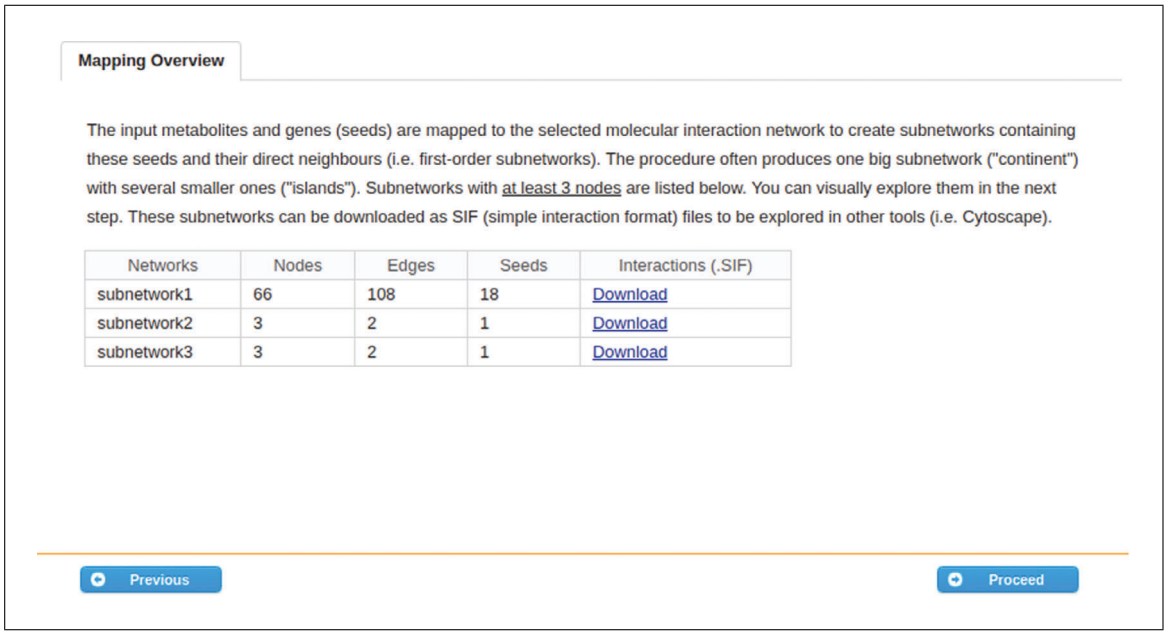


Figure 114 A screenshot showing the mapping overview results using the example gene-metabolite data mapped to the Metabolite-Gene-Disease Interaction network. Three subnetworks were created from the mapping.

table provides two popular topological measures—node degree and betweenness centralities. For instance, *L-Arginine* may be an important node, as it has the highest node degree and highest betweenness centrality. If a user clicks on the checkbox next to *L-Arginine*, the network will automatically zoom into this node (Fig. 116). It is evident that *L-Arginine* is involved in many congenital metabolic diseases including Hyperlysinemia, Argininemia, and Tyrosinemia. Remember that these links are metabolite-disease annotations from HMDB. The "Current Selections" box at the bottom left corner of the page provides further details of a selected feature, including a database link (KEGG, GenBank, or OMIM). For *L-Arginine*, a link to the KEGG compound database (C00062) appears in the box. Clicking the link will open a new page showing the KEGG compound. Returning to the MetaboAnalyst page, relationships between the different nodes can be further explored by clicking on the corresponding edges.

Node degree refers to the number of links a node has to other nodes, while betweenness centrality measures the number of shortest paths that pass through that node, taking the global network structure into consideration. Note that users can sort the Network Explorer table based on ID, Name, Degree, Betweenness, or Expression values by clicking the corresponding column header. To delete nodes (with their associated edges) from the current network, first select the nodes from the Node Explorer in the left-hand section. Then, click the "Delete" button at the top of the Node Explorer table. A confirmation dialog box will appear asking if the user really wants to delete these nodes. Click "OK" to delete the nodes. Deleting nodes will trigger network re-arrangement, especially if hub nodes are removed. In addition, "orphan" nodes, or nodes that are no longer connected to any other nodes, may be produced due to removal. These nodes will also be excluded during re-arrangement. Finally, the Node Explorer table can be saved as a `.csv` *file by clicking on the blue disk button on the top right corner of the table.*

Exploration of functional insights into the interaction network

7. To gain further functional insights into any uploaded data, pathway and gene ontology (GO) analysis can be performed. For demonstration purposes, we will perform enrichment analysis on all the nodes (Fig. 117). In the Function Explorer panel,

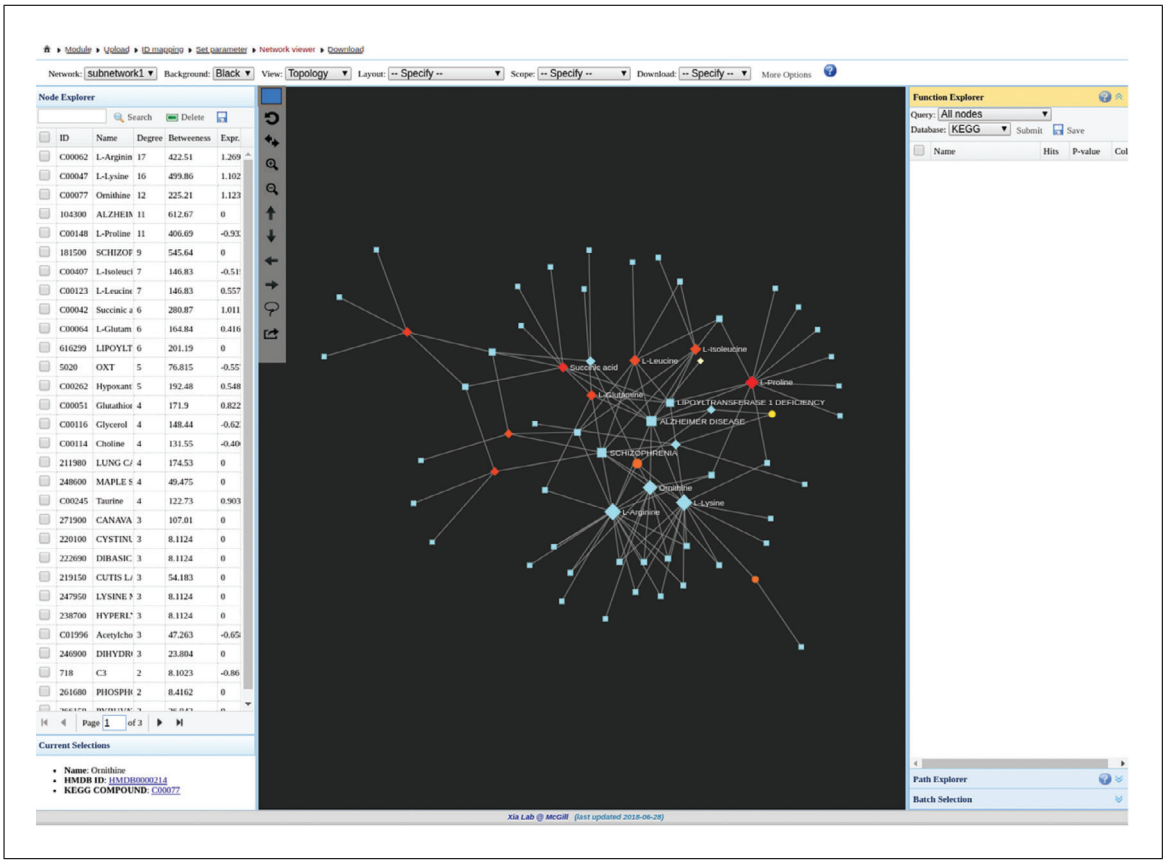


Figure 115 The Network View page of the Metabolite-Gene-Disease Interaction network (subnetwork1).

click the "Query" drop-down menu, where four query options will appear: (1) all nodes, (2) upregulated nodes, (3) downregulated nodes, and (4) highlighted nodes. Up- and downregulated nodes are based on their expression levels, if provided, and highlighted nodes are those selected in the Node Explorer menu or highlighted in the network. In this case, select "All nodes," set the database to "GO:BP," and click "Submit" to perform GO enrichment analysis. From the results, the top three GO:BP terms are "positive regulation of multicellular organismal process," "negative regulation of multicellular organismal process," and "regulation of response to external stimulus." To highlight the nodes involved in these GOs, click the colored box on the vertical toolbar in the network and select a green color (i.e., `#adff00`). Next, select the checkboxes adjacent to the top three terms to highlight them in the network (Fig. 117).

The Function Explorer supports enrichment analysis to KEGG pathways for metabolites, and enrichment analysis to gene ontologies (Biological Process [GO:BP], Cellular Component [GO:CC], and Molecular Function [GO:MF]) and pathways (KEGG/Reactome). It tests whether any defined pathways or functional groups based on the selected database are significantly enriched amongst the selected queries within the network. Hypergeometric tests are used to compute the enrichment p-values.

Path explorer

8. The path explorer (right-side panel) can be used to find the shortest path between any two nodes in the network. For instance, to find the shortest path between *L-Arginine* and the gene *C3* (complement component 3), type `L-Arginine` in the "From" box, and `C3` in the "To" box, and then click "Submit." There are three shortest paths between these two nodes, all crossing four nodes. The first path will be highlighted

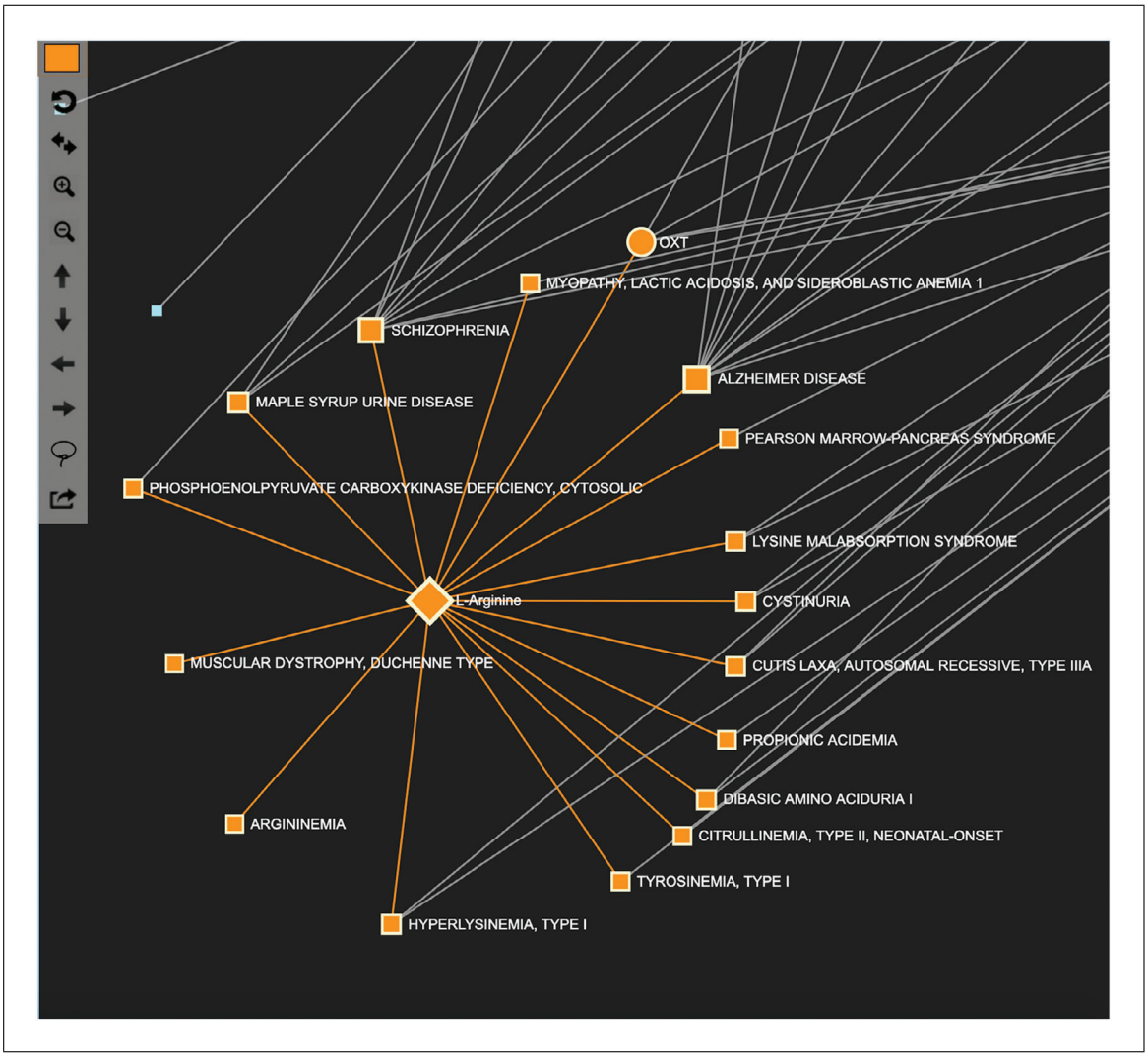


Figure 116 A screenshot of a zoomed-in view of the Metabolite-Gene-Disease Interaction network, highlighting connections between L-Arginine and several congenital metabolic diseases. Users need to set the Scope option at the top menu bar to "Node-neighbors," then double click the node to highlight.

in the network (Fig. 118). Users can choose different highlight colors and then click to view other paths.

Batch selection of nodes

9. Instead of manually selecting nodes by searching the table in Node Explorer or clicking nodes in the network, users can directly perform batch selection of nodes in the network. Here we show how to highlight five metabolites (*L-Proline*, *Succinic acid*, *Glycerol*, *Choline*, and *Taurine*). First, click the colored box on the vertical toolbar to select the node highlight color. In this case, pink is selected. Second, click the "Batch Selection" menu from the bottom right side of the page (under the Function Explorer and Path Explorer). Enter the list of node IDs or names (one entry per row) and click "Submit." The selected nodes are now highlighted in the current network (Fig. 119).

Network customization

10. MetaboAnalyst provides a suite of tools for users to customize their network using the toolbar at the top of the page. For instance, from the drop-down menu next to "Network," users can explore all of the different subnetworks that were created

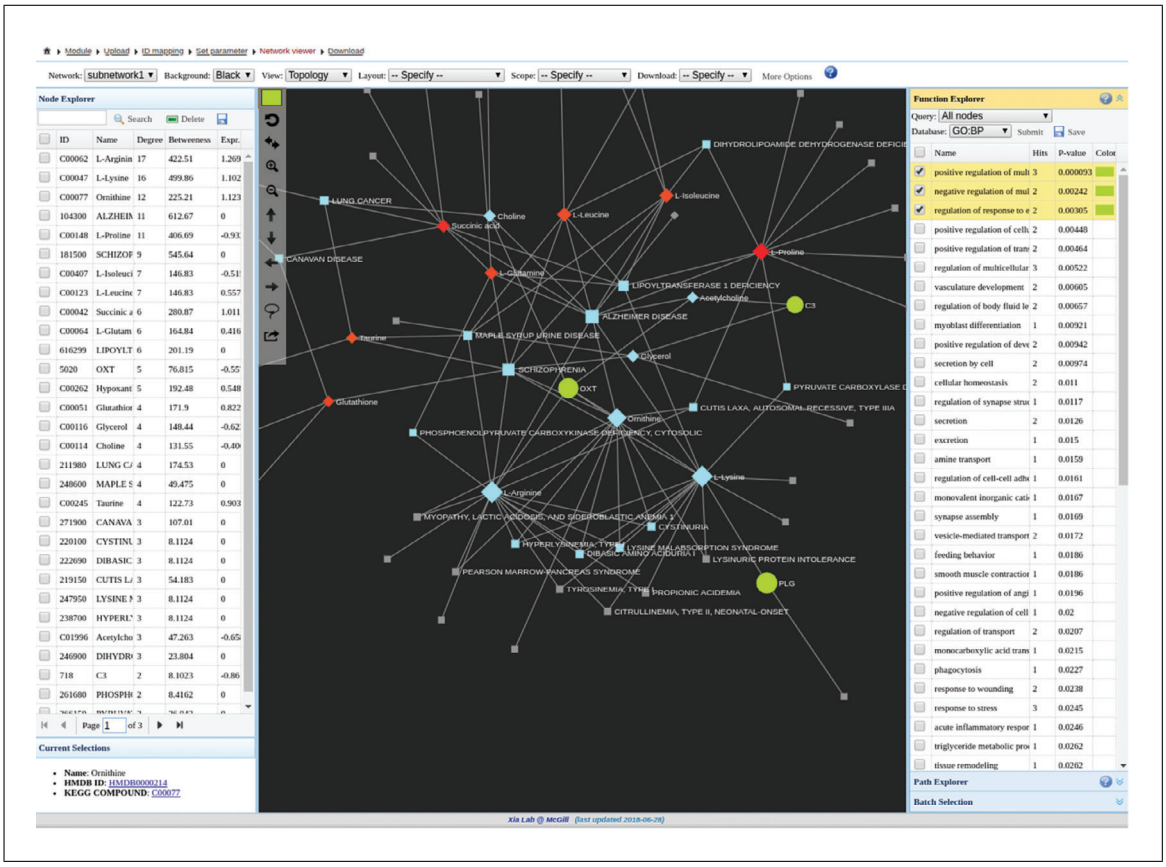


Figure 117 A screenshot of the top three enriched GO:BP terms highlighted in green on the Metabolite-Gene-Disease Interaction network.

in step 4. As the remaining subnetworks have very few nodes, keep working with subnetwork1. Click the "Background" drop-down menu and select white. Following this, from the drop-down menu next to "View," change the coloring of the network from "Topology" (default) to "Expression"; nodes will then be colored by their expression levels (if provided). Next, change the composition of the network by clicking the drop-down menu next to "Layout," which will reveal six options available for automatic arrangement of network nodes. Selecting one of these options will automatically re-organize the features. For demonstration purposes, select "Fruchterman-Reingold." Finally, click the "Download" drop-down menu and select PNG. A download dialog will appear—right-click the PNG image and select "Save link as" to save it to a local computer (Fig. 120). Users can also choose to download the current network in SVG format.

The "Scope" drop-down menu contains several options for highlighting and dragging nodes: (1) single node, to highlight/move only the node being clicked (default), (2) node-neighbors, to select a node and its direct neighbors, (3) all-highlights, to select all highlighted nodes and their direct neighbors, and (4) current function, to select all nodes from a pathway in the Function Explorer. To extract nodes from the network, first highlight the nodes (double-click in the network) or select them in the Node Explorer menu. Next, click the "Extract" icon on the left tool bar of the network view window. This will prompt an Extract Confirmation dialog box to appear, asking users to confirm the creation of a new module. Click "OK" to extract the nodes. The network view will automatically switch to the new module, which will be named "module," and will now be available in the Network drop-down menu.

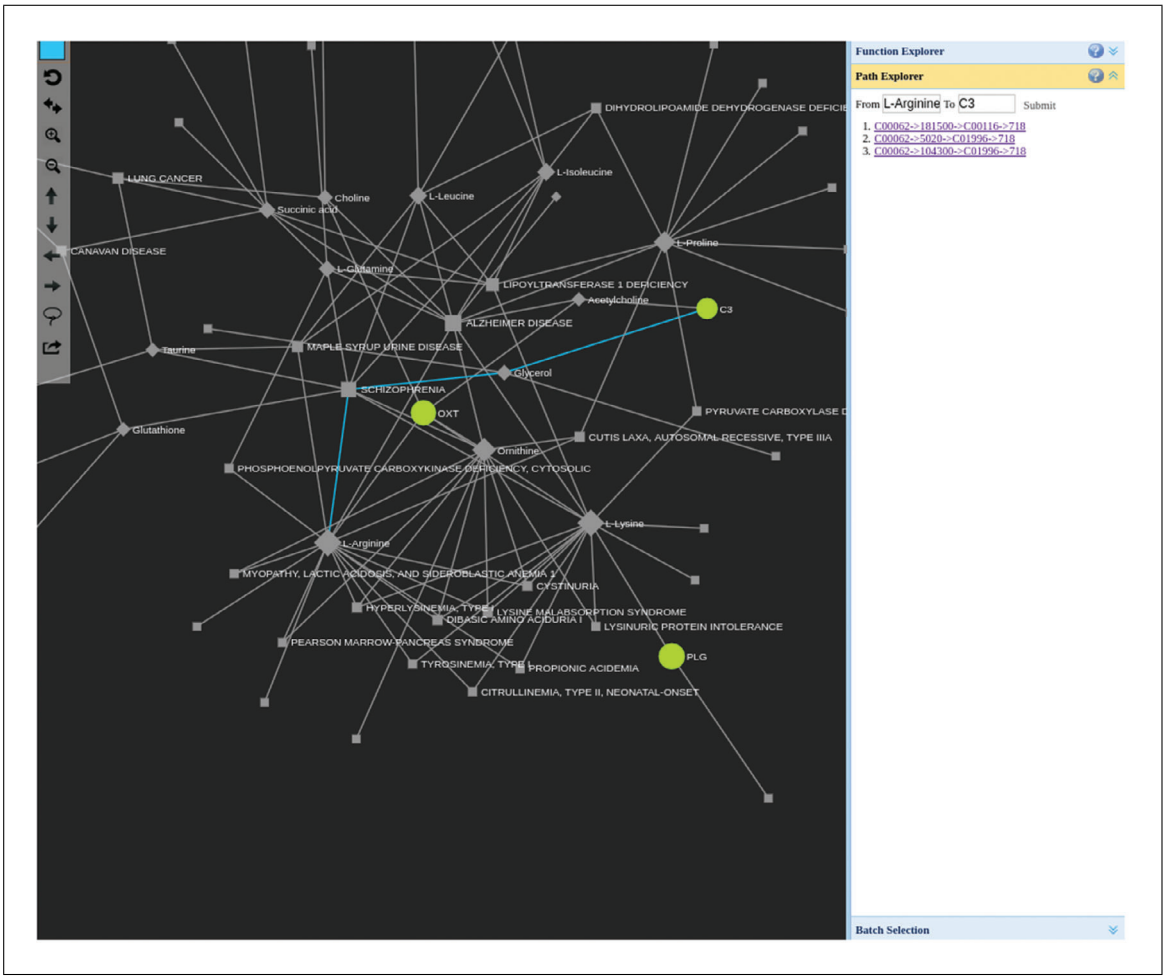


Figure 118 A screenshot of the shortest path between L-Arginine to complement component 3 (C3).

BASIC PROTOCOL 12

MetaboAnalysR INTRODUCTION

Despite its ease of use and wide accessibility, MetaboAnalyst presents inherent limitations in terms of flexibility of data analysis, batch processing, and handling of big data. To address these issues, we have developed MetaboAnalystR (Chong et al., 2019; Chong & Xia, 2018)—a companion R package based on the R codebase underlying the MetaboAnalyst web server. The R package has been thoroughly tested to ensure that the same commands will produce identical results from both interfaces. MetaboAnalystR therefore provides a flexible solution for more advanced users to tailor the analysis to their data, as well as to extend package capabilities by develop customized metabolomics workflows. It complements the web server by enabling more transparent, flexible, and reproducible analysis of big metabolomics data. The R package was recently upgraded to support raw spectral preprocessing and functional interpretation for LC-MS-based global metabolomics data. The MetaboAnalystR package is available from *https://github.com/xia-lab/MetaboAnalystR*. This protocol will first provide detailed instructions for installing the MetaboAnalystR package and then demonstrate how to (i) reproduce and customize analysis from the MetaboAnalyst web server, (ii) perform batch processing, and (iii) perform raw data pre-processing.

Necessary Resources

Hardware

A computer with internet access

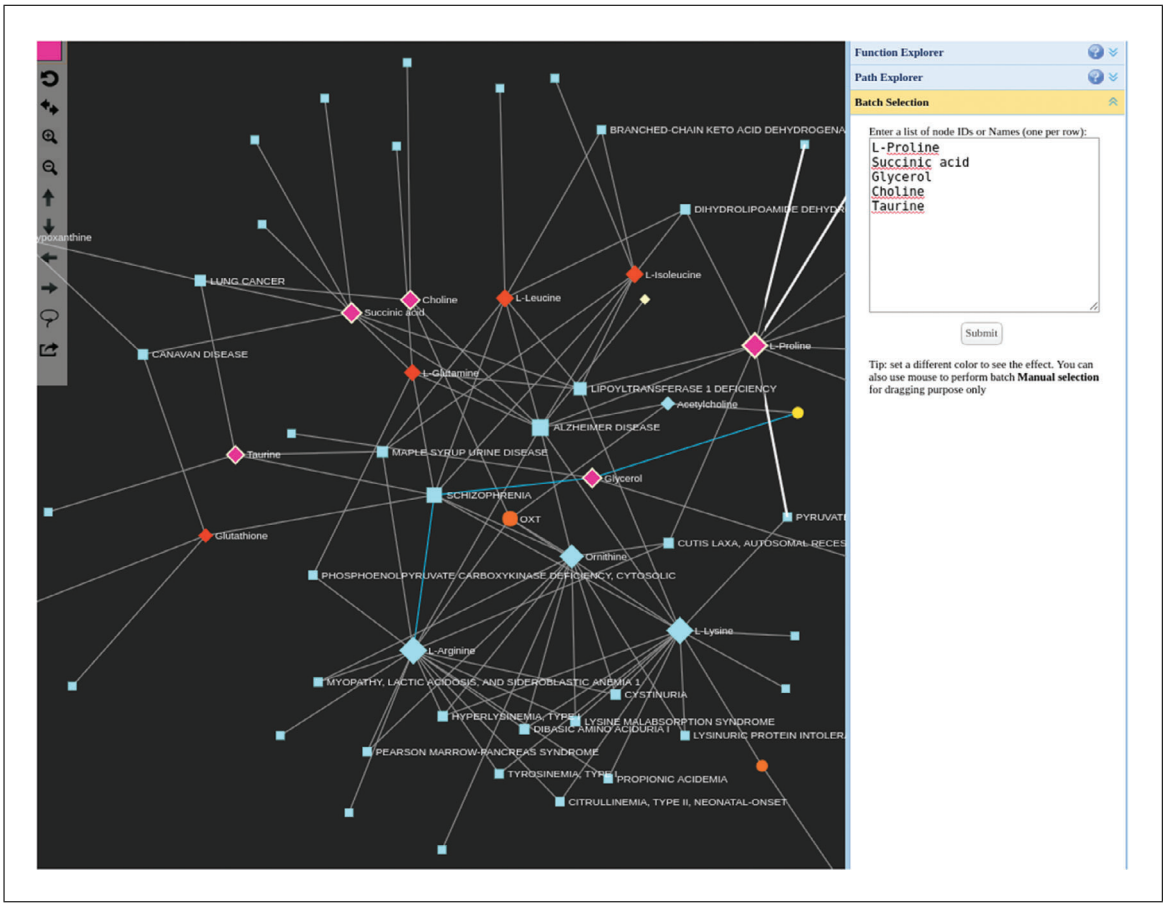


Figure 119 A screenshot of a batch-selection of five nodes, highlighted in pink, in the Metabolite-Gene-Disease Interaction network.

Software

An up-to-date web browser, such as Google Chrome (*http://www.google.com/chrome*), Mozilla Firefox (*http://www.mozilla.com/*), Safari (*http://www.apple.com/safari/*), or Internet Explorer (*http://www.microsoft.com/ie/*). JavaScript must be enabled in the web browser.
Required: R ≥ Version 3.4.4 installed on the computer
Optional: R Studio ≥ 1.1.383 installed on the computer

Files

Batch processing files: zipped file containing 12 NetCDF spectra of 12 mouse spinal cord samples (`goo.gl/qLMXkc`)

Raw Data pre-processing files: zipped file containing 12 mzML files (fecal samples from 6 pediatric IBD patients and 6 healthy controls—*https://www.dropbox.com/s/1v0a9b3y6pe0cbl/iHMP.zip?dl=0*)

Installation of the MetaboAnalystR package

1. Go to the MetaboAnalystR GitHub (*https://github.com/xia-lab/MetaboAnalystR*). Under the "Getting Started" heading, go to "Step 1: Install package dependencies." Next, open a new R or R Studio session and copy and paste the "metanr_packages" function from Step 1 into the R console. Using this function will automatically

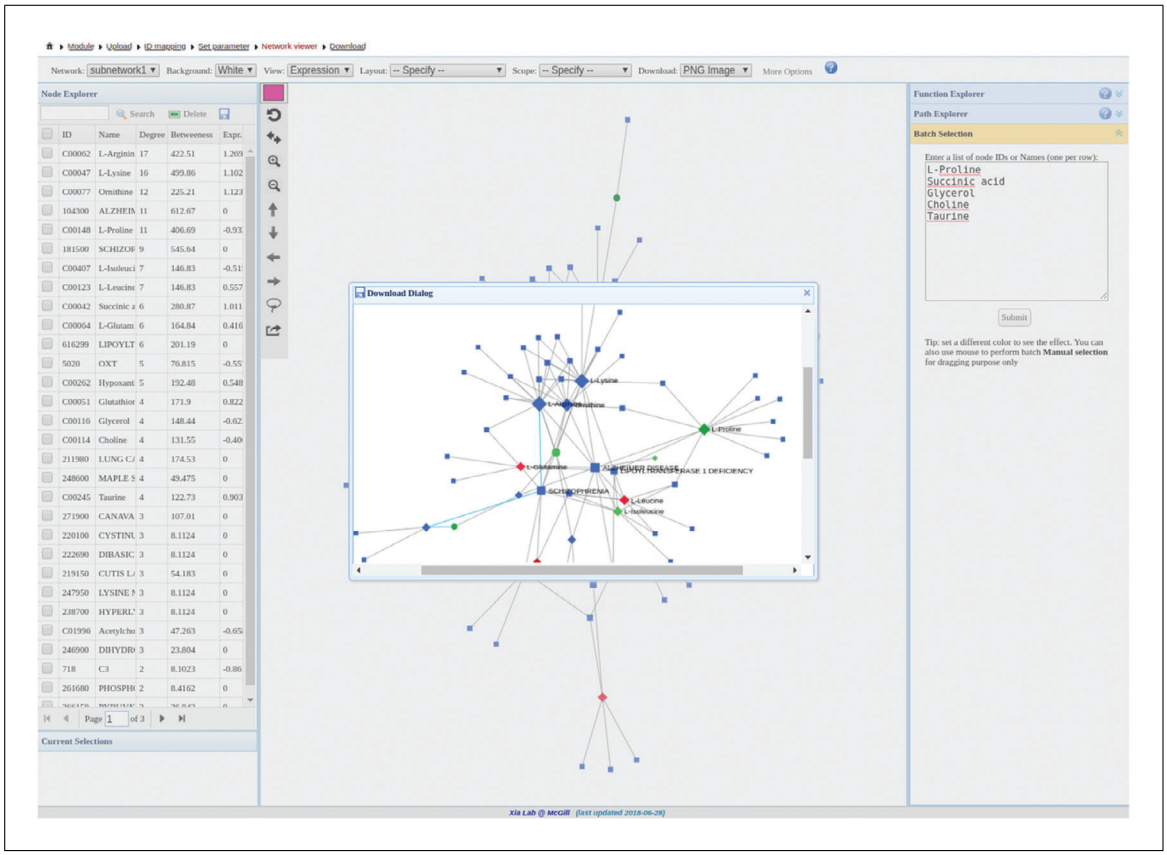


Figure 120 A screenshot of the download dialog for the customized Metabolite-Gene-Disease Interaction network.

install any/all missing dependencies, and a printed message will appear informing users whether any R packages were installed (Fig. 121).

These dependencies are other R packages that MetaboAnalystR uses to perform various functionalities and must be a minimum version number as specified by MetaboAnalystR. Note that if installation of MetaboAnalystR fails, a good place to start is to verify that all package dependencies have been correctly installed.

2. The second step is to install the MetaboAnalystR package itself. On the GitHub, three options are available with detailed instructions (Fig. 122): (a) using the R package **devtools** (Wickham & Chang, 2016), (b) cloning the GitHub, and (c) manually downloading the `.tar.gz` file. This tutorial will go over option A, which will be executed within the current R session. Install the R development tools package "devtools" and then install MetaboAnalystR using the `install_github()` function (Fig. 123).

   If installation of MetaboAnalystR using the devtools method is unsuccessful, follow the steps on GitHub for options b and c. As with the MetaboAnalyst web version, the MetaboAnalystR package is under active development.

3. Detailed tutorials illustrating typical workflows with example data are available as vignettes within the MetaboAnalystR package. To view these vignettes in a web browser (Fig. 124), use the command:

```
browseVignettes("MetaboAnalystR")
```

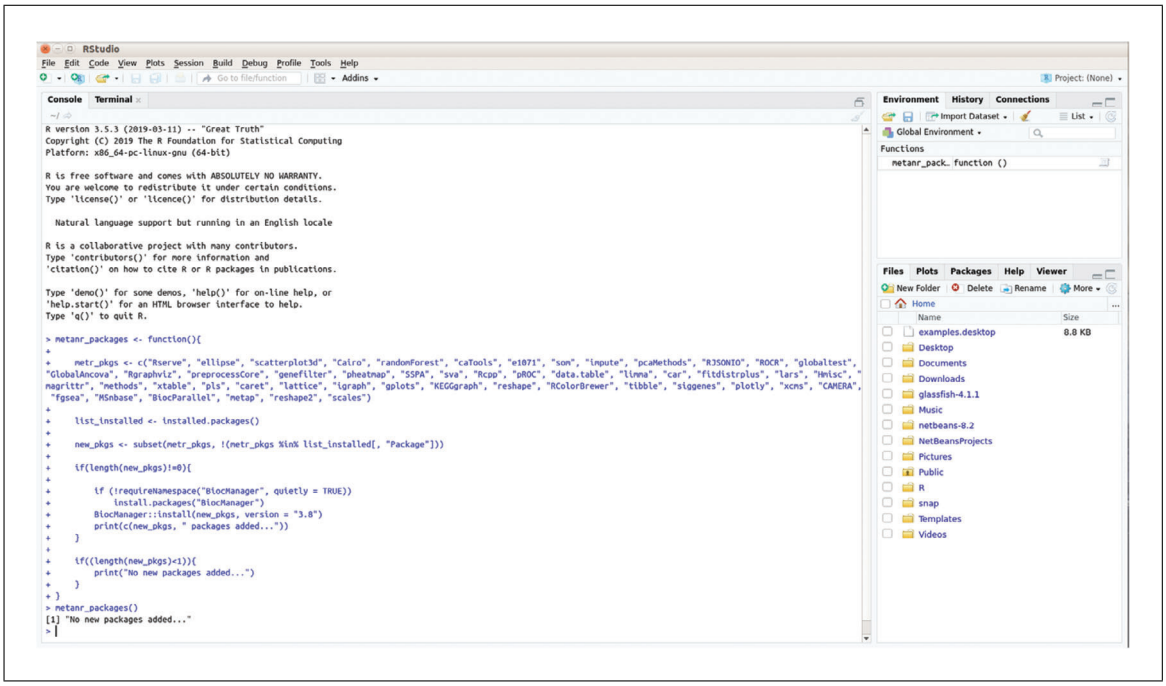


Figure 121 A screenshot of a R session showing the outcome of the `metanr_packages` function to download any missing package dependencies. Here, no new package dependencies were installed.

or to view them directly in R, use the command:

```
vignette(package="MetaboAnalystR")
```

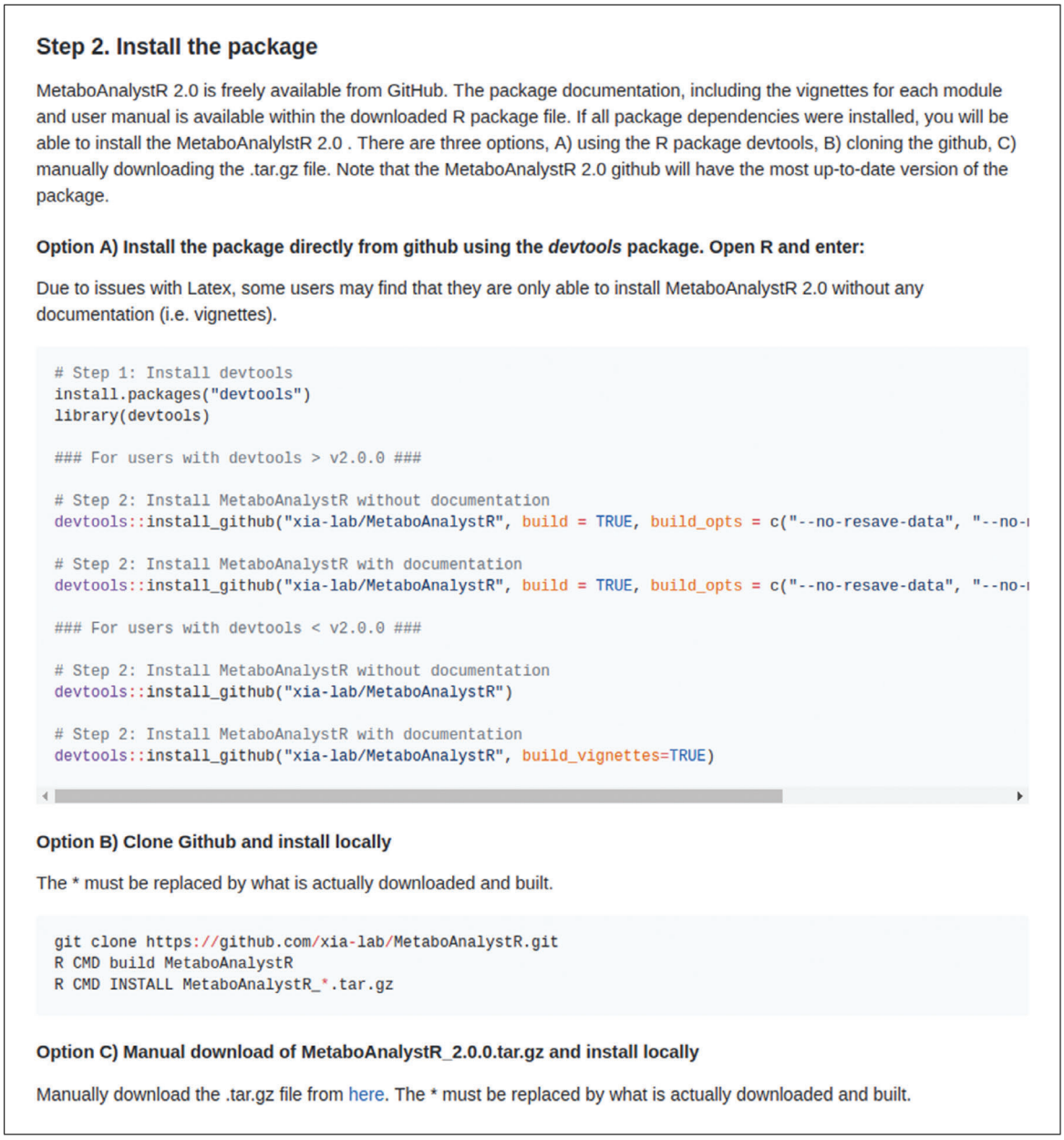
It is highly recommended that users read through/follow the relevant vignettes prior to performing their analyses to get familiar with the R package. These vignettes will give step-by-step explanations as well as tips for using the R package (Fig. 125).

4. MetaboAnalystR also comes with comprehensive function-level as well as package-level manuals embedded in the package. To view the manual for a function, in this case the `InitDataObjects` function (Fig. 126), use the command:

```
??MetaboAnalystR::InitDataObjects
```

Reproducibility and flexibility

5. The next steps will demonstrate how the R command history from the MetaboAnalyst web server can be input into R to obtain identical results using the MetaboAnalystR package. Here, the MS Peaks to Pathways module will be used to infer pathway activity directly from a list of LC/MS peaks. For further details, please refer to Basic Protocol 9. To begin, go to the MetaboAnalyst homepage (*https://www.metaboanalyst.ca*) and press "Click here to start" to enter the "Module Overview" page. Select the "MS Peaks to Pathways" button to enter the "Peak Uploads" page. Click the circle next to "ImmuneCell1" and then press "Submit." Following the "Data Integrity Check," click "skip." Next, on the "Library View" page, click the *mummichog* algorithm and ensure that "Homo sapiens (human) [MFN]" is selected. Press "Submit" to continue. Briefly, from the results, *Aminosugars metabolism*, *Tryptophan metabolism*, and *Sialic acid metabolism* and are the top enriched pathways from the example data. Next click the "Download" hyperlink on the left-side navigation tree, where the `Rhistory.R` file will be located. This file contains the entire set of R



Step 2. Install the package

MetaboAnalystR 2.0 is freely available from GitHub. The package documentation, including the vignettes for each module and user manual is available within the downloaded R package file. If all package dependencies were installed, you will be able to install the MetaboAnalylstR 2.0 . There are three options, A) using the R package devtools, B) cloning the github, C) manually downloading the .tar.gz file. Note that the MetaboAnalystR 2.0 github will have the most up-to-date version of the package.

Option A) Install the package directly from github using the *devtools* package. Open R and enter:

Due to issues with Latex, some users may find that they are only able to install MetaboAnalystR 2.0 without any documentation (i.e. vignettes).

```
# Step 1: Install devtools
install.packages("devtools")
library(devtools)

### For users with devtools > v2.0.0 ###

# Step 2: Install MetaboAnalystR without documentation
devtools::install_github("xia-lab/MetaboAnalystR", build = TRUE, build_opts = c("--no-resave-data", "--no-

# Step 2: Install MetaboAnalystR with documentation
devtools::install_github("xia-lab/MetaboAnalystR", build = TRUE, build_opts = c("--no-resave-data", "--no-

### For users with devtools < v2.0.0 ###

# Step 2: Install MetaboAnalystR without documentation
devtools::install_github("xia-lab/MetaboAnalystR")

# Step 2: Install MetaboAnalystR with documentation
devtools::install_github("xia-lab/MetaboAnalystR", build_vignettes=TRUE)
```

Option B) Clone Github and install locally

The * must be replaced by what is actually downloaded and built.

```
git clone https://github.com/xia-lab/MetaboAnalystR.git
R CMD build MetaboAnalystR
R CMD INSTALL MetaboAnalystR_*.tar.gz
```

Option C) Manual download of MetaboAnalystR_2.0.0.tar.gz and install locally

Manually download the .tar.gz file from here. The * must be replaced by what is actually downloaded and built.

Figure 122 A screenshot of the MetaboAnalystR GitHub showing three options available to install the package: (1) using devtools, (2) cloning the GitHub, or (3) manual downloading of the `.tar.gz` file.

commands in the order in which they were invoked in a user's session. Right-click the `Rhistory.R` hyperlink and select "Save link as..." to save the file in a specified working directory.

6. Return to the R session and ensure that the working directory is set to the directory that contains the `Rhistory.R` file. In the R console, enter the following command to open the file (Fig. 127):

```
file.edit("Rhistory.R")
```

7. In the Rhistory file, locate the `Read.PeakListData` function. Note that the path to the input data is `Replacing_with_your_file_path` and replace it with the dataset from the web server *https://www.dropbox.com/s/cbakdss04sglh4h/mummichog_mzs.txt* (Fig. 127).

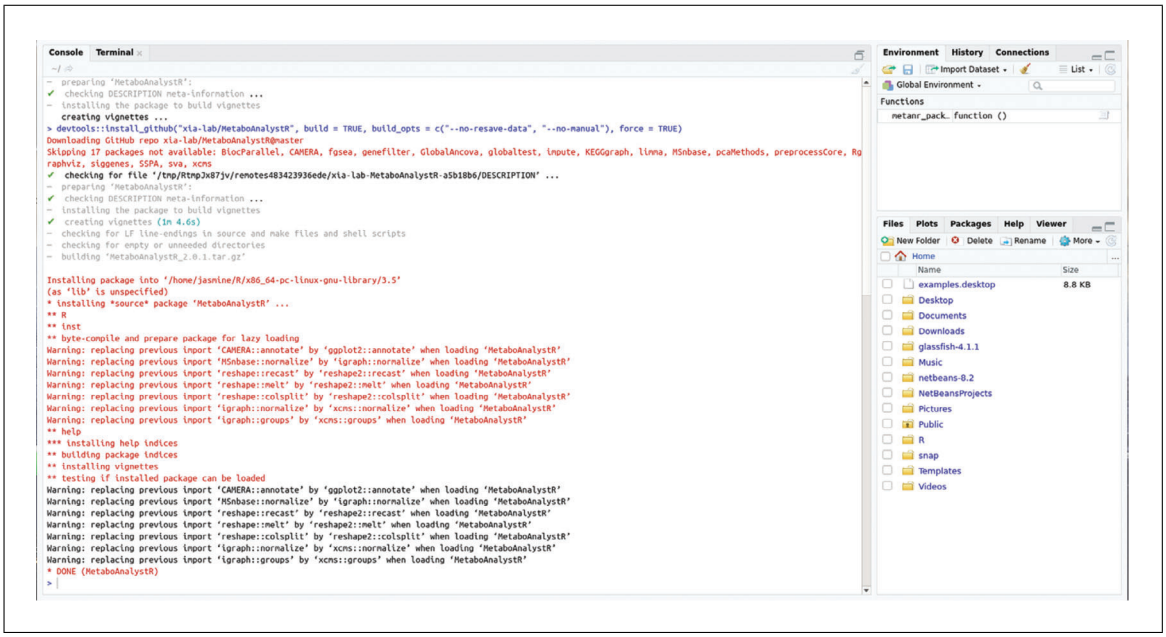


Figure 123 A screenshot of an R session showing how to use the devtools R package to directly install MetaboAnalystR from the GitHub server.

Figure 124 A screenshot of the vignettes browser in a web browser.

8. The first step is to load the MetaboAnalystR package:

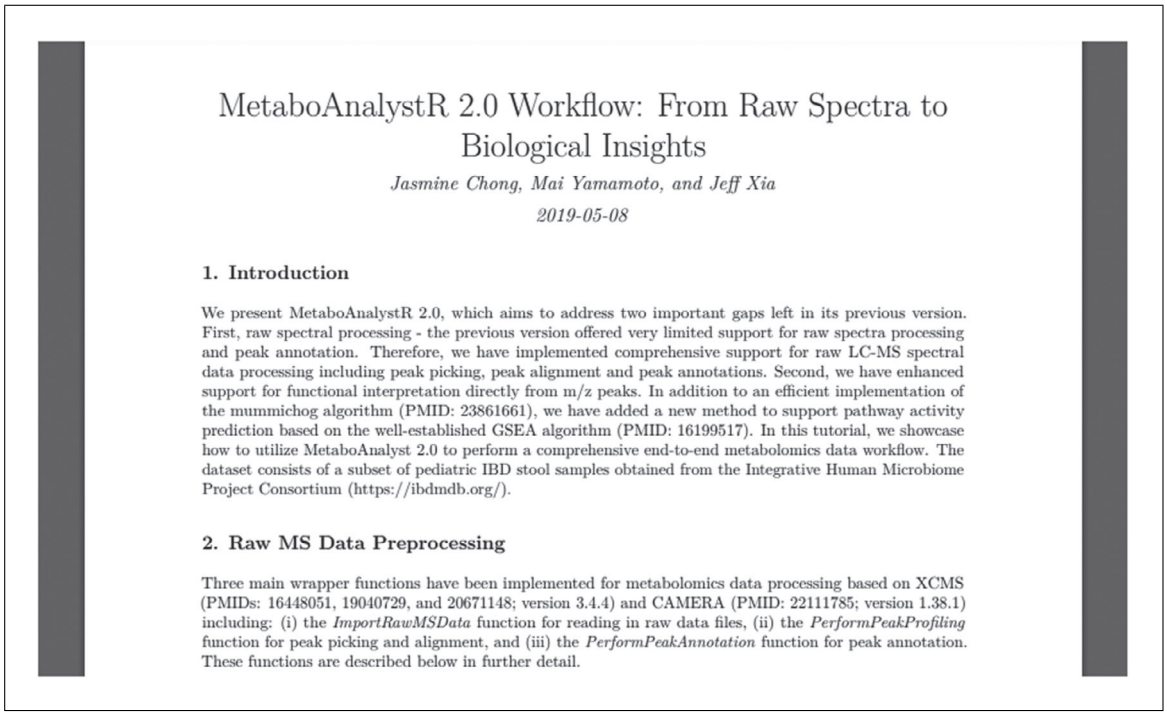
```
library(MetaboAnalystR)
```

9. The first function called in every module is the *InitDataObjects* function, which constructs the `mSetObj` object:

```
mSet <- InitDataObjects("mass_all", "mummichog", FALSE)
```

The `mSetObj` is a fundamental component of all analyses performed by the package, and stores user's data/results for further processing and analysis. In this function, users must specify the data type and type of analysis to be performed. For this module, the data type is `mass_all` and the type of analysis is `mummichog`. In the R console, copy and paste the following command from the Rhistory file. A message will appear informing the user that the "`R objects initialized ...`."

For those using the R code from the MetaboAnalyst web server, the suggested name of the `mSetObj` is `mSet`, which must be called consistently throughout the workflow. Furthermore, the MetaboAnalystR package directly creates plots/tables/analysis outputs in the current working directory. It is not necessary to call any plotting functions onto



MetaboAnalystR 2.0 Workflow: From Raw Spectra to Biological Insights

Jasmine Chong, Mai Yamamoto, and Jeff Xia

2019-05-08

1. Introduction

We present MetaboAnalystR 2.0, which aims to address two important gaps left in its previous version. First, raw spectral processing - the previous version offered very limited support for raw spectra processing and peak annotation. Therefore, we have implemented comprehensive support for raw LC-MS spectral data processing including peak picking, peak alignment and peak annotations. Second, we have enhanced support for functional interpretation directly from m/z peaks. In addition to an efficient implementation of the mummichog algorithm (PMID: 23861661), we have added a new method to support pathway activity prediction based on the well-established GSEA algorithm (PMID: 16199517). In this tutorial, we showcase how to utilize MetaboAnalyst 2.0 to perform a comprehensive end-to-end metabolomics data workflow. The dataset consists of a subset of pediatric IBD stool samples obtained from the Integrative Human Microbiome Project Consortium (https://ibdmdb.org/).

2. Raw MS Data Preprocessing

Three main wrapper functions have been implemented for metabolomics data processing based on XCMS (PMIDs: 16448051, 19040729, and 20671148; version 3.4.4) and CAMERA (PMID: 22111785; version 1.38.1) including: (i) the *ImportRawMSData* function for reading in raw data files, (ii) the *PerformPeakProfiling* function for peak picking and alignment, and (iii) the *PerformPeakAnnotation* function for peak annotation. These functions are described below in further detail.

Figure 125 A screenshot of the MetaboAnalystR 2.0 vignette.

the created `mSetObj`*. Note that every command must be run in sequence; please do not skip any commands, as this may lead to errors downstream.*

10. The second function in the Rhistory file is `SetPeakFormat`, which tells MetaboAnalystR what format your data are in. The third function is `UpdateInstrumentParameters`, where users can set the mass accuracy of their instrument (`instrumentOpt`) and the mode of the MS instrument (`msModeOpt`). The fourth function is `Read.PeakListData`, which reads in the list of *m/z* features. Following this, the fifth function, `SanityCheckMummichogData`, verifies that the *m/z* features list is in the correct format, removes duplicate *m/z* features, and trims *m/z* features to those within a range of 50 to 2000 *m/z*. The sixth function is `SetPeakEnrichMethod`, where users specify which algorithm(s) they wish to perform. Since the *mummichog* algorithm is selected, the sixth function is `SetMummichogPval`, where users set the *p*-value cutoff. Next is the `PerformPSEA` function, which performs the actual enrichment analysis. Here, users must indicate the pathway library that best fits their organism (`lib`) and the type of *p*-value (`enrichOpt`). Last is the `PlotPeaks2Paths` function, which creates a graphical summary of the enrichment analysis. Copy and paste these functions from the Rhistory to the R console (Fig. 128).

11. The results of the *mummichog* algorithm and the potential matched compounds are saved in the working directory as CSV files, as shown in the bottom right panel (Fig. 128). The results can be directly viewed in the R terminal:

```
mSet$mummi.resmat
```

From this matrix (Fig. 128), it is evident that results between the web server and the R package are identical, with *Aminosugars, Tryptophan*, *Sialic acid* metabolism as the top three enriched pathways, with the same *p*-values.

12. To demonstrate the flexibility of the R package, the downloaded Rhistory file will be modified. First, set the *p*-value cut-off in `UpdateMummichogParameters`

Figure 126 A screenshot of the help page for the `InitDataObjects` function.

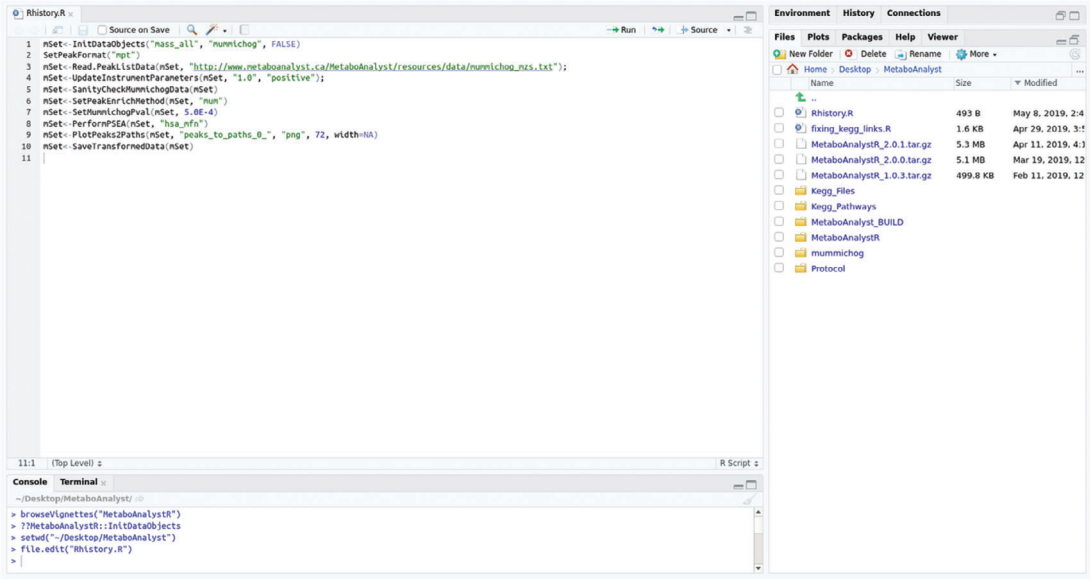


Figure 127 A screenshot of the downloaded "Rhistory" file for the MS Peaks to Pathways module using example data from the MetaboAnalyst web server.

to a far more stringent cutoff of 1.0E-6 (F1 below). Second, on the web, the number of permutations executed in `PerformMummichog` is 100, and cannot be changed by users online. However, this number can be modified in MetaboAnalystR. To the `PerformMummichog` function, add a "1000" after "gamma" to increase the

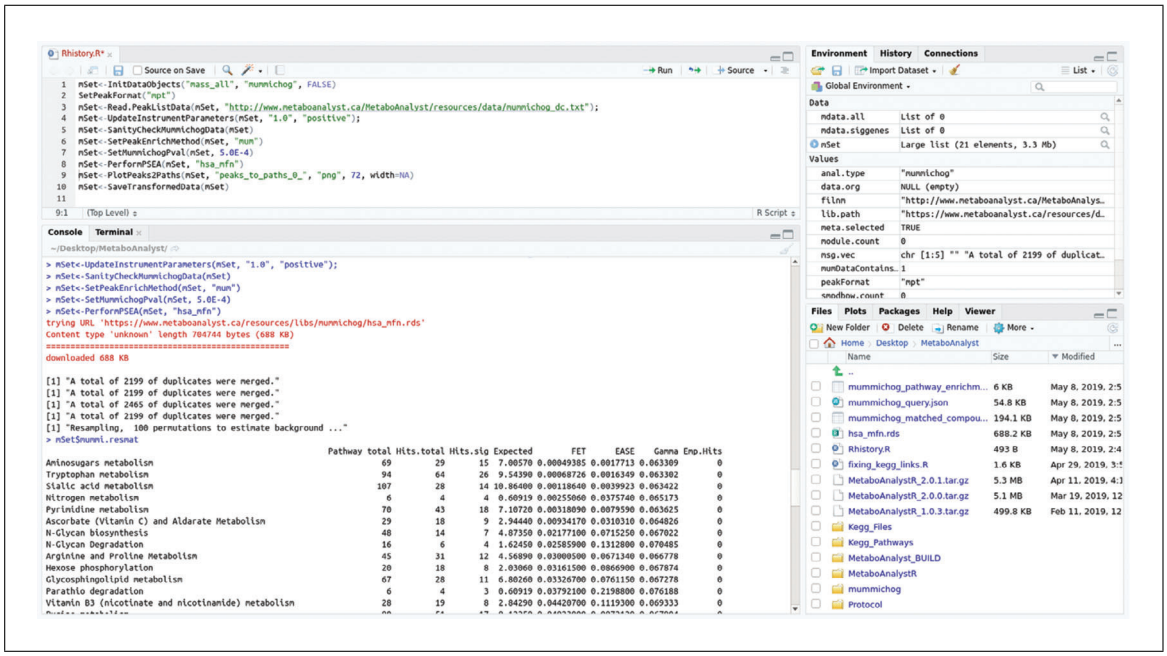


Figure 128 A screenshot of the altered "Rhistory" file for the MS Peaks to Pathways module. The entire "Rhistory" was run in the console. The printed messages from the *mummichog* analysis are visible in the console, and the resulting files from the analysis are saved in the current working directory.

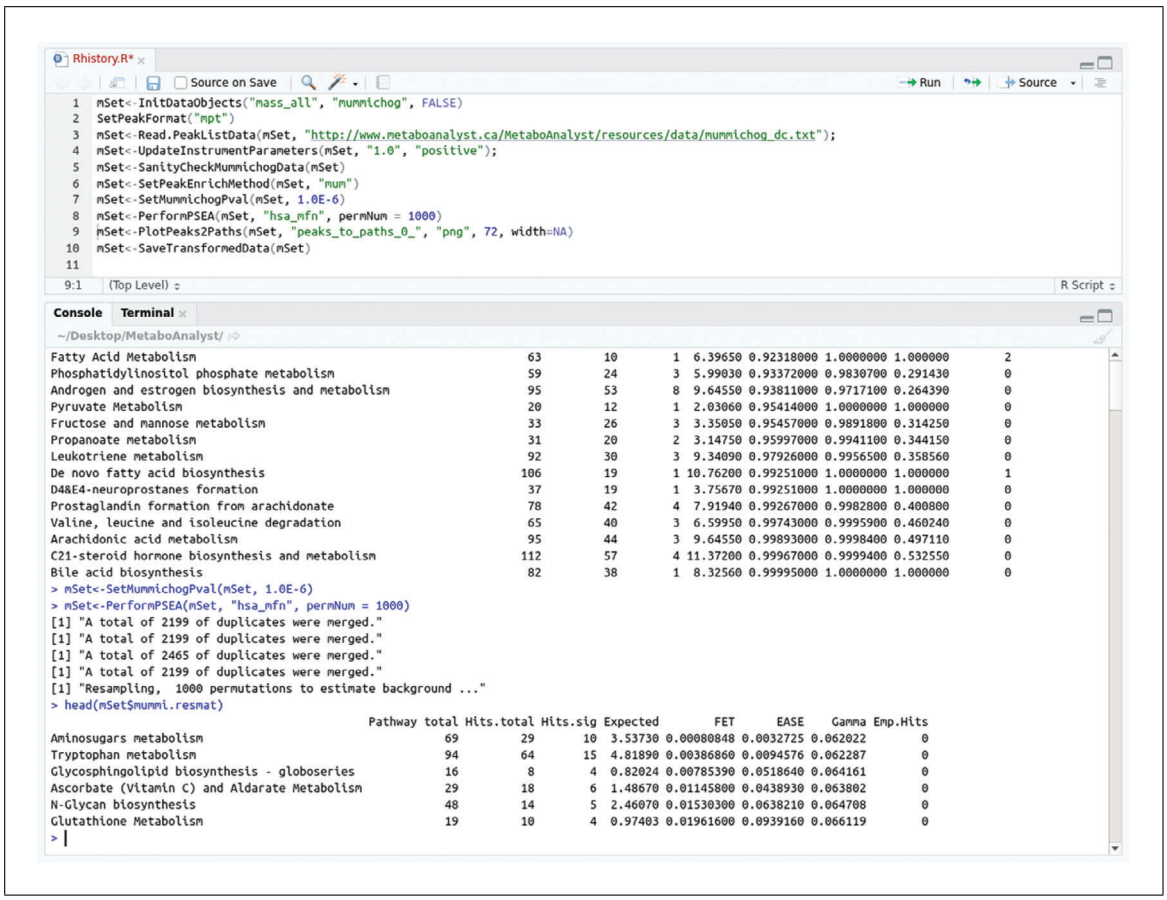


Figure 129 A screenshot of an R session with the altered "Rhistory" file run in the R console (*p*-value cutoff = 1.0E-06 and permutations = 1000). The top 6 enriched pathways are also visible at the bottom of the image.

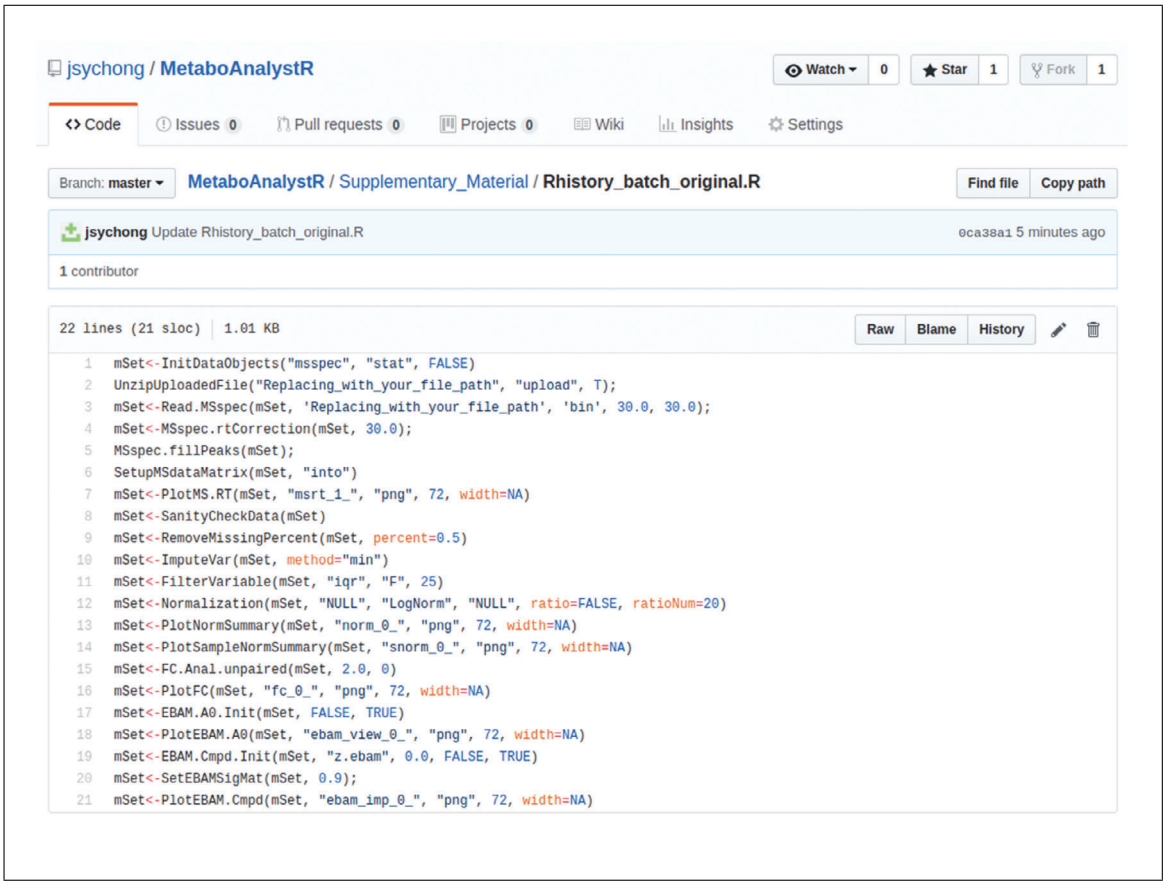


Figure 130 A screenshot of the "Rhistory" file of the pilot data downloaded from the MetaboAnalyst web server to be used in later steps for batch processing.

number of permutations (F2 below). The modified functions are shown in the gray box below. Next, execute the modified Rhistory file and explore the results (Fig. 129):

```
F1. mSet <- UpdateMummichogParameters(mSet, "0.1",
"positive", 1.0E-6)
F2. mSet <- PerformMummichog(mSet, "hsa_mfn", "fisher",
"gamma", 1000)
```

Batch processing

13. The MetaboAnalyst web server is unaccommodating for the analysis of big metabolomics data, with a file size limit of 50 MB on the public server. 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). To start, download the necessary files in the link provided at the beginning of the MetaboAnalystR protocol in a new folder named `batch`. This zipped file contains raw 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). Previously, a zipped file containing a subset of the mice samples, three wild-type and three enzyme-inactivated, was uploaded to the MetaboAnalyst web-server. On these data, missing value imputation and data normalization were performed, followed by fold-change (FC) analysis

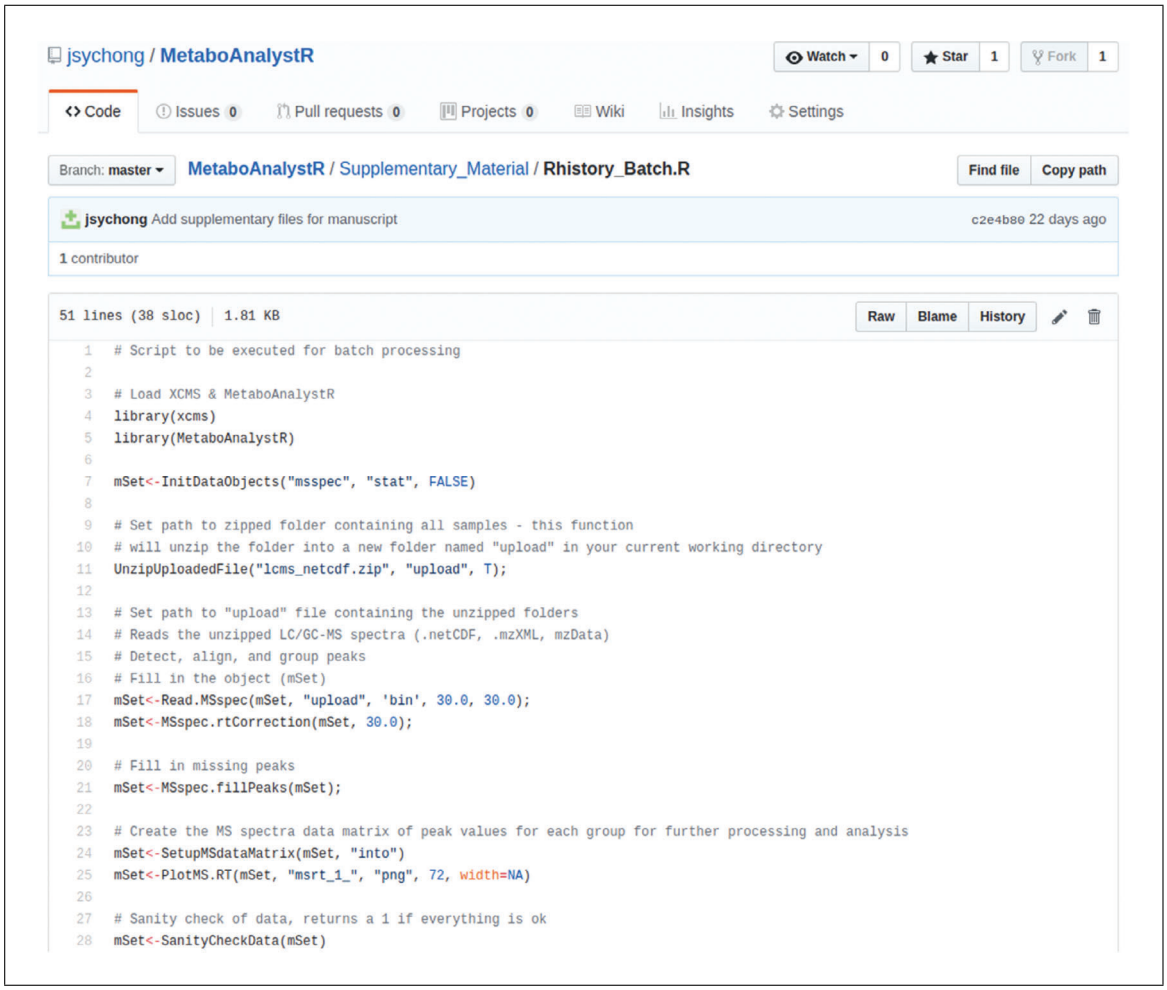


Figure 131 A screenshot of an example "Rhistory" file that has been modified for batch processing of big metabolomics data.

and Empirical Bayesian Analysis of Microarray (and Metabolites) (EBAM). The Rhistory file was then saved (`Rhistory_batch_original.R`) and is available from the MetaboAnalystR GitHub (`goo.gl/j1bXNF`) (Fig. 130). Follow the link and copy and paste the R code into an empty R Script in the current R session, then save it to the `batch` folder.

14. Within the R Script, minor edits must be made prior to performing batch processing. The MetaboAnalystR and the XCMS packages must be loaded, and the file path in the function `UnzipUploadedFile` must be modified to the zip file containing all 12-mouse samples. The file path of the `Read.MSspec` function must also be adjusted to the "upload" folder, which will contain the unzipped files in the current working directory. An example of the modified R Script with detailed step-by-step explanations per function is also available on the MetaboAnalystR GitHub (`Rhistory_Batch.R`, `goo.gl/zkR13D`) (Fig. 131).

15. Ensure that the downloaded zipped files and the modified R Script are in the "batch" folder. Then, open a BASH command line and set the current directory to the "batch" folder. Next, enter the code in the gray box below to execute the batch processing of the LC-MS data. Using this single step, the data will be processed, filtered, and normalized, and FC and EBAM analyses will be performed (Fig. 132). The outputs (plots/tables) of the batch processing will be found in the `batch` folder.

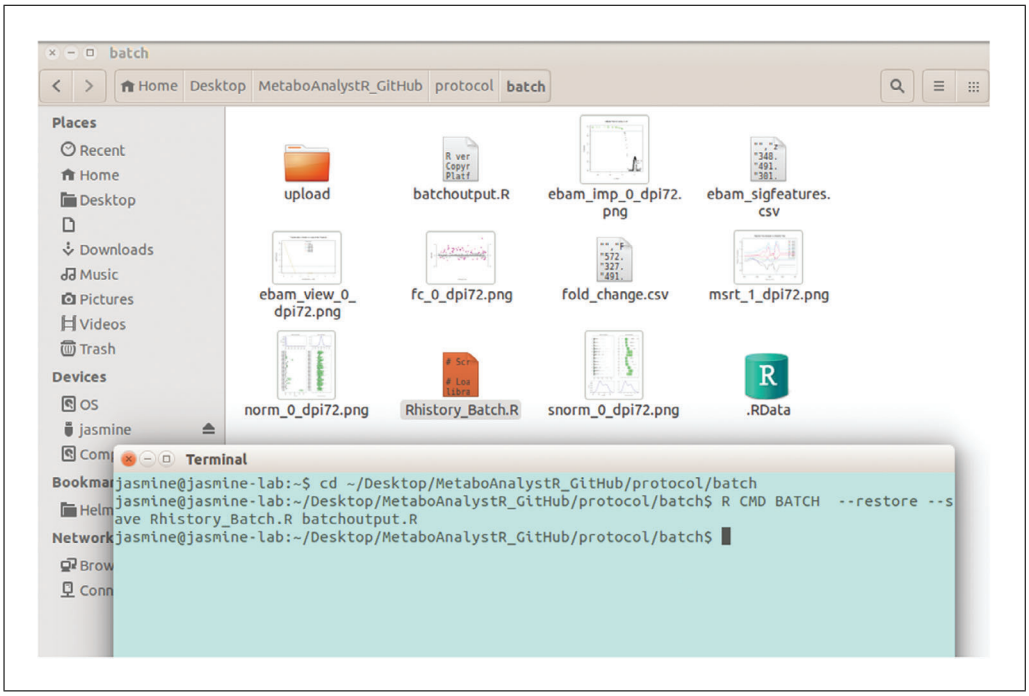


Figure 132 A screenshot of a terminal showing the BASH command to perform batch processing, as well as the results of the command saved into the current working directory.

```
R CMD BATCH --restore --save Rhistory_Batch.R
batchoutput.R
```

In the above command, the `Rhistory_Batch.R` file contains the modified R code to be executed, and `batchoutput.R` is the name of the output file. The results at the end of the session are saved as an `.RData file using --save`.

Raw spectra preprocessing of high-resolution LC-MS data with MetaboAnalystR

16. While local installation of the MetaboAnalyst web server supports raw metabolomics data pre-processing, the public version has entirely disabled this option. The MetaboAnalystR 2.0 package was therefore created for users to directly process their raw spectral data and then directly perform a wide range of metabolomics data analyses. These next steps will demonstrate how MetaboAnalystR 2.0 seamlessly integrates the XCMS (Smith et al., 2006) and CAMERA (Kuhl, Tautenhahn, Bottcher, Larson, & Neumann, 2012) R packages to create a complete metabolomics data analysis workflow. Here, users will perform peak identification, retention time correction, and peak alignment using an example dataset to create a usable peak table, and then perform OPLS-DA. To start, download the necessary files in the link provided at the beginning of the MetaboAnalystR protocol in a new folder named `iHMP`.

17. The R commands used in the next steps can be found in the "MetaboAnalystR 2.0 Workflow: From Raw Spectra to Biological Insights" vignette, available in the R package. To open this vignette, follow the R commands in step 3 of this protocol. From the vignette, enter the first library command, which will load the MetaboAnalystR packages.

    Three main wrapper functions have been implemented for metabolomics data processing including: (i) the `ImportRawMSData` function for reading in raw data files, (ii)

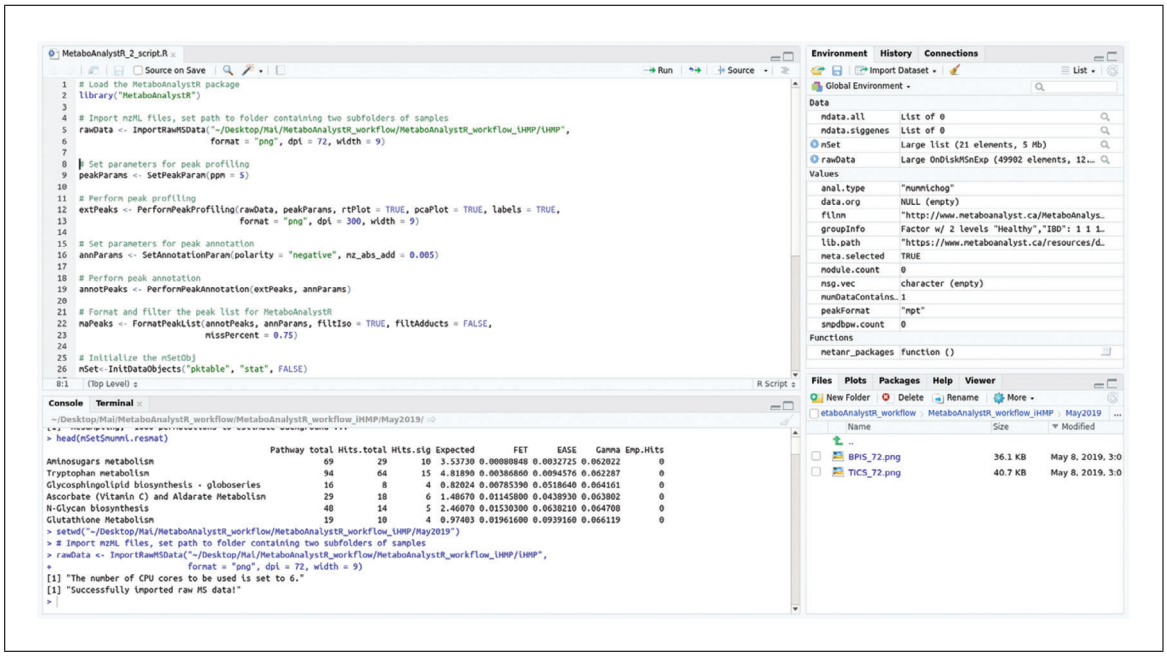


Figure 133 A screenshot of the import of raw MS spectra. All resulting plots are output to the current working directory. By default, the number of CPU cores is set to 6 (half of all available cores).

the `PerformPeakProfiling` function for peak picking and alignment, and (iii) the `PerformPeakAnnotation` function for peak annotation. Next, the resulting peak list is formatted to the correct structure for MetaboAnalystR and filtered based upon user's specifications using the `FormatPeakList` function. This function permits the filtering of adducts (i.e., removal of all adducts except for [M+H]+/[M-H]−) and filtering of isotopes (i.e., removal of all isotopes except for monoisotopic peaks). The goal of filtering peaks is to remove degenerative signals and reduce the file size.

18. Next, enter the command at the end of this step to read in the raw MS data. Set the path to the unzipped folder (`iHMP`), which contains two subfolders labelled `Healthy` and `IBD`. Note that the names of the subfolders will be used as the group names for the samples. The `ImportRawMSData` function reads in raw MS data files and saves it as an `OnDiskMSnExp` object. The function outputs two plots—the Total Ion Chromatogram (TIC), which provides an overview of all spectra, and the Base Peak Chromatogram (BPC) which is a cleaner profile of the spectra based on the most abundant signals. As preprocessing raw MS spectra can be memory intensive, the `ImportRawMSData` function will limit the number of cores used. By default, it will determine the number of cores on a user's computer and set the number of cores used to half that number (Fig. 133).

```
rawData <- ImportRawMSData("~/Desktop/iHMP", format =
"png", dpi = 72, width = 9)
```

19. After importing the raw spectra, set the parameters for peak picking using the `SetPeakParam` function. This function sets all the parameters used for downstream pre-processing of user's raw MS data such as the mass error (ppm), signal-to-noise threshold, and bandwidth. In this case, set the mass error to 5. Following the setting of parameters, use the `PerformPeakProfiling` function to perform peak

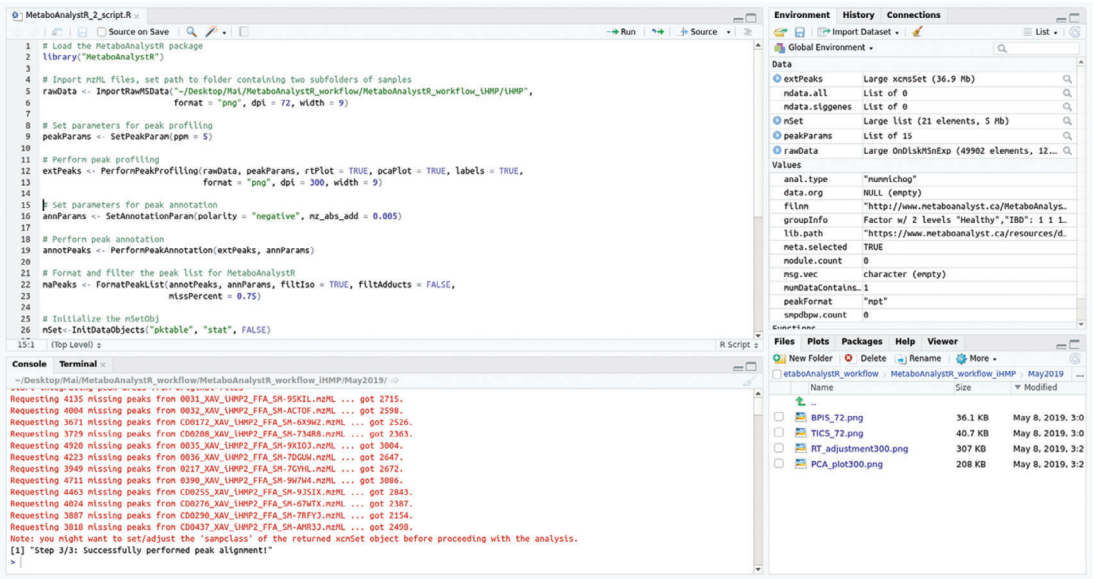


Figure 134 A screenshot of the peak picking, grouping, retention time correction and alignment of the example MS data.

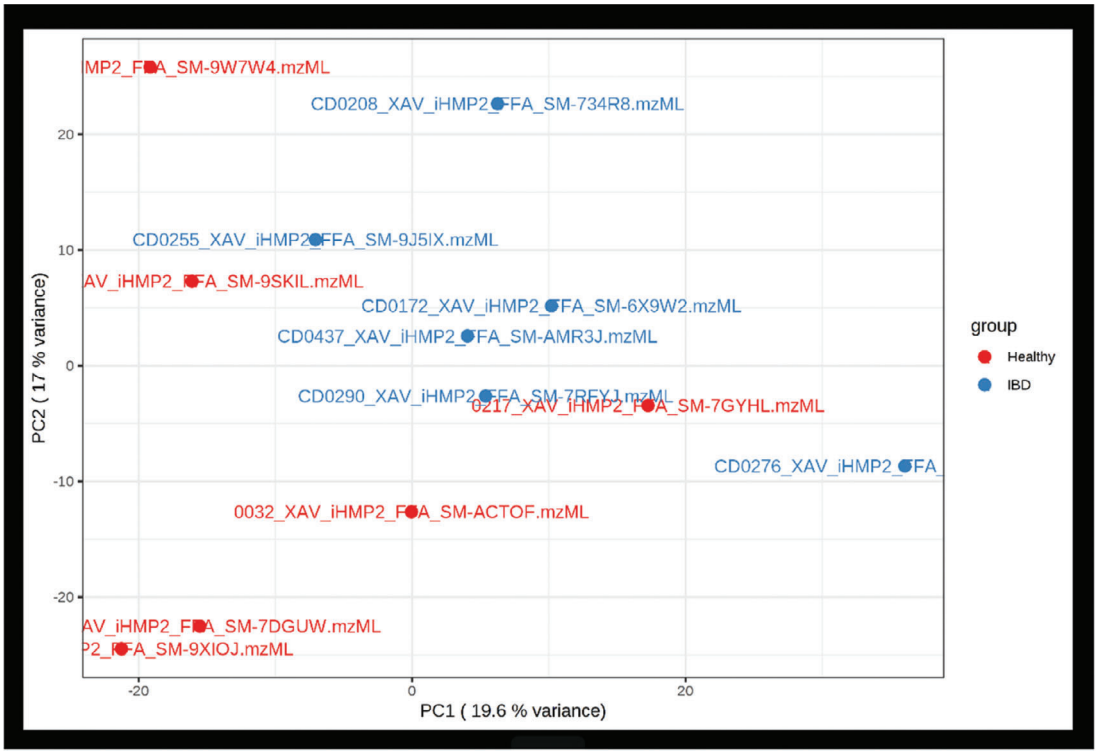


Figure 135 A screenshot of the PCA plot of all samples from the example MS data.

picking, grouping, retention time correction, and alignment (Fig. 134). Of all the preprocessing steps, this takes the most time. The resulting peaks are returned as an `XCMSnExp` object. The function also generates two diagnostic plots, including a

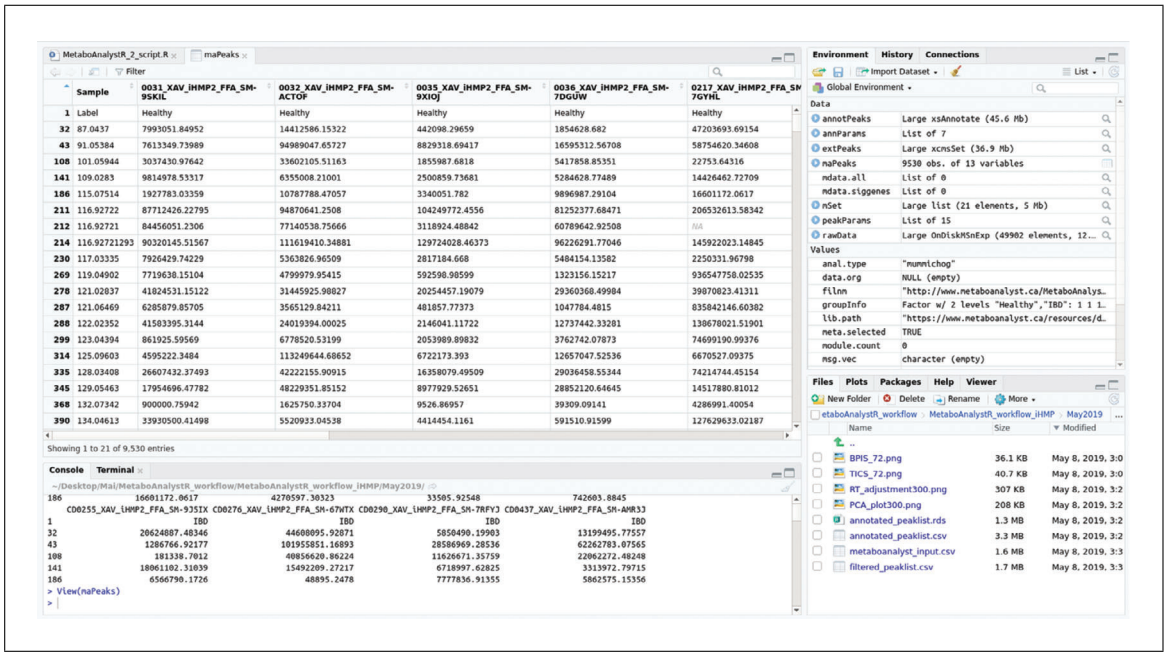


Figure 136 A screenshot of the peak annotation step, showing the resulting peak table.

retention time adjustment map and a PCA plot showing the overall sample clustering prior to data cleaning and statistical analysis (Fig. 135).

```
peakParams <- SetPeakParam(ppm = 5)
extPeaks <- PerformPeakProfiling(rawData, peakParams,
rtPlot = TRUE, pcaPlot = TRUE, labels = TRUE, format =
"png", dpi = 300, width = 9)
```

20. Following peak profiling, the next step is to perform peak annotation. First, set the peak annotation parameters using the `SetAnnotationParam` function. Here, specify that the samples were collected in negative ion mode—"negative," and the allowed tolerance of *m/z* for the search (for adduct annotation) to 0.005. Next, use the `PerformPeakAnnotation` function to annotate the detected peaks to potential isotopes and adducts. This will output a `.csv` file named `annotated_peaklist.csv`, which contains the entire list of all identified *m/z* features, their retention time, and their annotations. Finally, use the `FormatPeakList` function to format this annotated peak list into the correct format to be used by MetaboAnalystR (and the web version). In this function, users can choose to filter the peaks, i.e., remove features that are missing in more than 75% of samples per group and filter out all isotopes. The function will produce a file named `metaboanalyst_input.csv`, which will be used as input for further data analysis functions downstream (Fig. 136). It also outputs the `filtered_peaklist.csv`, which contains a column named `Pklist_inx`, which provides users with the row number of where to find the *m/z* feature in the `annotated_peaklist.csv`.

```
annParams <- SetAnnotationParam(polarity = "negative",
mz_abs_add = 0.005)
annotPeaks <- PerformPeakAnnotation(extPeaks, annParams)
maPeaks <- FormatPeakList(annotPeaks, annParams, filtIso =
TRUE, filtAdducts = FALSE, missPercent = 0.75)
```

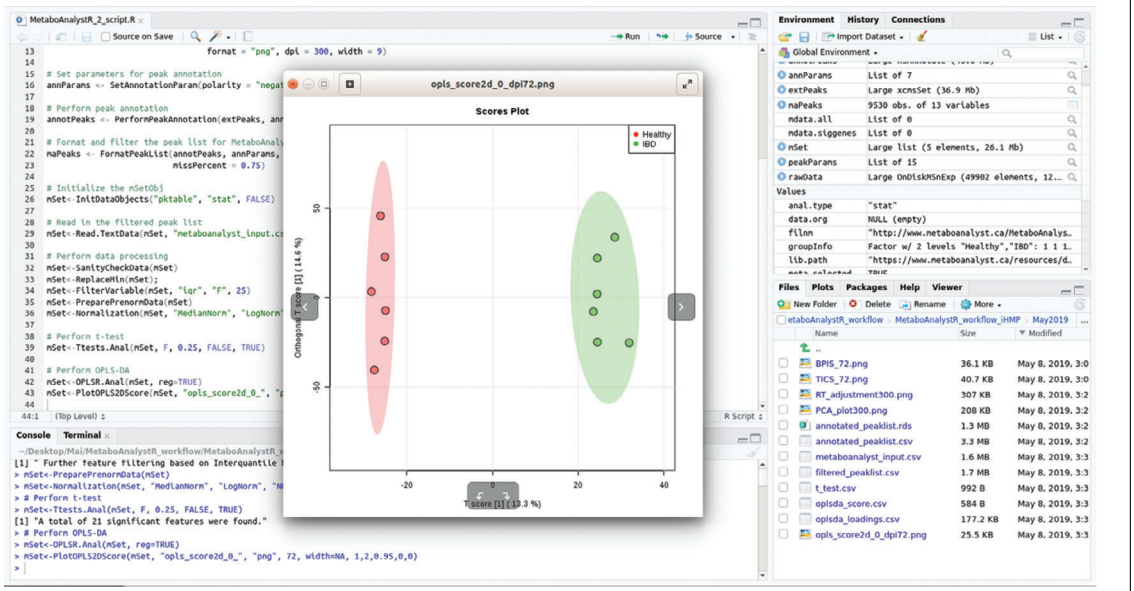


Figure 137 A screenshot of the OPLS-DA plot showing good separation between the healthy controls and IBD patients.

21. After peak annotation, read in the filtered peak table using the `Read.TextData` function. As always, first initiate an `mSetObj` with `InitDataObjects`. Then, paste the following R commands from the vignette to ultimately perform an orthogonal partial least-squares discriminate analysis (OPLS-DA; Fig. 137):

GUIDELINES FOR UNDERSTANDING RESULTS

Basic Protocol 1

This protocol is designed to show users how to prepare multiple peak list files for uploading to MetaboAnalyst, how to use various data processing options to "clean up" their data, how to normalize their data, and how to perform a standard data quality check. The central goal of data processing in MetaboAnalyst is to convert various data inputs into a suitable data matrix compatible with most statistical and machine learning tools. Raw NMR or LC-MS or GC-MS spectra should be first processed using locally installed software to produce multiple peak list files (or a single peak intensity table), which are of much smaller size and can be easily uploaded (via the internet) to MetaboAnalyst. After converting these files to their proper format, users need to organize these files into different folders and create a zip file (step 2), then upload them to MetaboAnalyst using the appropriate options (steps 3 to 4). MetaboAnalyst will perform a peak list alignment to convert the multiple files into a single data table (step 5) and then conduct a basic data integrity check (step 6). Untargeted metabolomic data (from LC-MS runs) usually contains a significant quantity of missing values, false peaks, or baseline "noise." MetaboAnalyst provides various methods for missing value imputation (step 7) as well as for low quality data filtering (step 8). These steps can usually improve data quality and reduce the incidence of false discoveries in the downstream statistical analysis or data-interpretation stages. The data are then subject to various data-normalization procedures to adjust systemic bias, as well as to make the data set more compatible (i.e., normally distributed) for proper handling by many common statistical procedures (steps 9 to 10). Finally, the protocol uses an artificial data set to illustrate how to perform a simple data quality check and how to deal with outliers using PCA, a heatmap, and MetaboAnalyst's Data Editor (steps 11-15).

Basic Protocol 2

The selection or identification of important variables or informative metabolites is probably the most common task in metabolomic data analysis. MetaboAnalyst provides a number of methods to support this kind of task. This particular protocol outlines the procedures for variable selection using three common scenarios. In the first case, the protocol outlines how one should upload and process CSV data (steps 1 to 4). For the selection of variables that are significantly different in different groups, this protocol demonstrates how to use both ANOVA and post-hoc tests (steps 5 to 6). It further shows how to set the Delta value to control the FDR and discusses the selection of important features using SAM (step 7). For the identification of variables showing particular patterns of change, the protocol gives several examples to illustrate how to use different options and how to fine-tune the parameters used in the PatternHunter method to obtain the desired results (steps 8 to 11). Finally, it demonstrates how to identify variables that are significantly associated with a known biomarker using correlation analysis (steps 12 to 13). The protocol concludes with a description on various files that are available for download (step 14).

Basic Protocol 3

This protocol provides detailed instructions on how to use several well-established multivariate methods implemented in MetaboAnalyst, including PCA, heatmap, PLS-DA, and random forest, for exploratory data analysis. These methods offer excellent support for data visualization, pattern discovery, feature selection, and classification. For PCA analysis, this protocol shows how to detect patterns of separation from a PCA score plot and how to identify significant features from the corresponding loading plot (steps 2 to 5). The protocol then describes data analysis and visualization using heatmaps to facilitate data summarization and pattern discovery (steps 6 to 7). For PLS-DA, this protocol shows how to use the basic scores and loading plots to search for patterns and important variables (steps 8 to 9). It then provides detailed instructions on how to use cross validation for model selection (step 10), how to use permutation tests for model validation (step 11), and how to select important variables using VIP scores or regression coefficients (step 12). Finally, this data analysis protocol shows how to perform classification and feature selection using the random forest algorithm and how to interpret the results (steps 12 to 14). The protocol concludes with some tips on how to further investigate the patterns of change for certain selected features (step 15).

Basic Protocol 4

This protocol outlines various procedures for high-level functional interpretation using MSEA and MetPA. Given different data types, MESA can perform over-representation analysis (ORA), SSP, and QEA. For ORA, this protocol demonstrates how users can upload a list of significant compound names and perform name standardization (steps 1 to 3). It then gives detailed instructions on setting parameters for ORA and understanding the results (steps 4 to 6). The protocol further explains the procedures for conducting SSP, with particular focus on how to perform metabolite concentration comparisons (steps 7 to 9). For QEA using a concentration table, the data-processing steps are identical to those for processing CSV data for other statistical analyses, with the extra step of compound name standardization. The protocol focuses on how to set parameters and how to interpret the results (i.e., a metabolite set plot; steps 10 to 11). As noted in this protocol, pathway analysis is an extension of the MSEA, with extra procedures to calculate the impact on the pathways based on network topology analysis. The protocol gives detailed instructions on how to set parameters for pathway analysis and how to explore the results interactively to obtain a better understanding of the pathways involved (steps 12 to 15).

Basic Protocol 5

This protocol is designed to show users how to perform a range of different biomarker analyses and how to use ROC curves to evaluate their performance. Compared to exploratory statistical analysis, biomarker analysis requires larger sample sizes and is only applicable for two-group comparisons. It is recommended that users have a balanced design with at least 12 samples per group. Users also need to be aware of two different biomarker analysis objectives—biomarker discovery and biomarker performance evaluation. The former aims to maximize the chance to identify the most promising biomarkers, while the latter tries to obtain the most objective measure of the performance of the biomarker(s). Optimizing both aims simultaneously requires much larger sample sizes and the implementation of double or triple cross-validation procedures. This advanced feature is not currently supported in MetaboAnalyst, as we have found that the majority of sample sizes are too small to benefit from such an analysis. As a result, users need to decide on their primary objective (discovery or validation/evaluation) prior to using the module. Using compound ratios at the normalization stage (steps 3 and 9) and manually selecting features to build biomarker models (step 16) tends to increase the chance of over-fitting (i.e., the actual performance in later validation studies will not be as good as initially achieved). In addition, users need to be aware that the metabolites contained in multivariate biomarker models are not fixed (step 13). Therefore, the performance measure is only an average of multiple biomarker models created using different subsets of data. Users need to employ an independent (or validation) data set to obtain a truly objective measure of performance. Finally, users may notice that, frequently, the results from conventional univariate ROC analysis are better than those obtained from multivariate ROC analysis—this is because the latter is based on cross validation (a more objective evaluation), while the former is computed on the same data (a less objective evaluation).

Basic Protocol 6

This protocol provides detailed instructions on how to use several well-established methods for visualizing and analyzing time-series and two-factor metabolomics data sets. Note that time-series data are a special case of two-factor design. MetaboAnalyst expects that the overall design will involve examining the primary experimental factor (i.e., wild type versus mutant) across different time points (the secondary factor). It also supports data analysis with only the time factor. We recommend that users integrate high-level PCA (step 4) with low-level heatmap (steps 5 to 6) visualization to identify interesting patterns, corresponding features, and critical factors, followed by univariate (two-way ANOVA; steps 7 to 8) and multivariate (ASCA, step 9; and MEBA, step 10) approaches for further statistical validation. These methods work particularly well together to reveal significant patterns or interactions hidden in these kinds of complex data sets.

Basic Protocol 7

This protocol outlines the basic steps for power analysis and sample size estimation for metabolomics data. From a user's point of view, the key step is to obtain pilot data either from a preliminary study or from a public data set generated from a study very similar to the study to be carried out. During this sort of analysis, users need to make sure the diagnostic plots (step 3) show the expected distributions so that the predicted statistical power estimates will be more accurate. Users are encouraged to try different data-processing procedures to see if the results can be improved. However, it is important to keep in mind that the same data-processing procedures must be applied when analyzing any future data to minimize any potential systemic bias.

Basic Protocol 8

This is a relatively simple data analysis module for joint pathway analysis. Users need only supply two lists, one for genes and one for metabolites. One potential issue with this module is the compound name mapping step, in which some compounds may not be recognized correctly. In this case, we suggest that users directly supply KEGG compound IDs instead of common names or HMDB IDs, as the underlying pathways (for now) are based on the KEGG database. A shift to include the SMPDB database will be occurring over the coming months.

Basic Protocol 9

This protocol provides detailed instructions on how to predict pathway activity from an MS peak list generated from untargeted metabolomics. The module is relatively straightforward, requiring users to upload a pre-ranked MS peak list, to specify the mass accuracy, to indicate the analytical mode of their MS instrument, and to select the format of their peak list and algorithm(s) to perform. If the *mummichog* algorithm is to be performed, users must select a *p*-value cutoff. Using the *mummichog* and GSEA algorithms individually and combined is recommended to fully explore the MS peak list data. Further, there can be several options for selecting a pathway library for a specific organism. It is therefore suggested that users explore all options to gain a comprehensive insight into their data. The outputs of the "MS Peak to Pathways" module include two tables, the first consisting of the top-ranked enriched pathways, and the second containing the compound-matching information for all user-uploaded *m/z* features. While compound identification was generally de-emphasized in the original *mummichog* implementation, post hoc analysis of the matched compounds is critical for downstream validation and interpretation. Therefore, it is highly suggested that users use the KEGG global metabolic network to visualize the overall peak matching patterns of their data. This will also aid in selecting candidate compounds by examining all matched isotopic and adduct forms.

Basic Protocol 10

This protocol outlines the basic steps for performing Biomarker Meta-Analysis. Again, this is a straightforward module, where users must upload their datasets ($\geq$3 datasets) to perform meta-analysis. Users need to make sure that the datasets were collected under similar or comparable experimental conditions. Otherwise, the inherent bias and heterogeneity of the datasets from different sources could mask the true biological effects. The data should be analyzed on the same scale or range (either keep the datasets in their raw form or normalize them in the same way), and it is suggested that metabolomic data be re-scaled with a logarithmic transformation. The performance of the selected method for meta-analysis will depend on the quality and consistency of the data. Users are therefore encouraged to explore different parameters and methods to obtain further insights into their data.

Basic Protocol 11

This protocol is designed to give a comprehensive overview on how to use the Network Explorer module. Users need to enter a list of genes or metabolites (or both), with optional fold-change values. One potential issue is with the compound and gene name mapping to the internal MetaboAnalyst database, where the user's data may have no matches. Given the comprehensiveness of the MetaboAnalyst knowledgebase (derived from HMDB, STRING, and KEGG), this may well indicate that there is a general lack of information regarding those compounds and genes. Further, the interaction networks were created primarily from human-only data. Users with data from other organisms may use these networks to explore their data; however, any insights should be treated with caution due to inherent differences in their biology (i.e., insect versus human).

Basic Protocol 12

This protocol is designed to provide users with a brief introduction to MetaboAnalystR, the companion R package to the MetaboAnalyst web server. The protocol first details the steps to install the R package, from downloading package dependencies to installation of the R package itself, as well as tips for troubleshooting installation and general use. If installation was unsuccessful using the method discussed in the protocol, the MetaboAnalystR GitHub has detailed step-by-step instructions for alternative methods. Next, the protocol demonstrates how MetaboAnalystR promotes reproducible research by replicating the web-server results locally on a user's computer (steps 5 to 11). The flexibility of the package is also shown by customizing a user's analysis in a way that is not achievable using the web server (step 12). The protocol then provides an example of using MetaboAnalystR to perform batch-processing of big metabolomics data, which is not feasible using the web server due to data upload size restrictions (steps 13 to 15). Finally, the protocol shows how to pre-process raw LC/MS data into the correct tabular format for MetaboAnalystR and the web server, from which various statistical analyses can be performed. As mentioned above, the aim of this protocol is to introduce the R package. For users who wish to know more details on how to use MetaboAnalystR, the package contains several comprehensive vignettes (one per module) that go over a typical analytical workflow with example data. Further, users who run into any problems are encouraged to open an issue on the MetaboAnalystR GitHub.

COMMENTARY

Metabolomics is commonly considered as "the new kid on the 'omics' block," having really only emerged in the last 18 years (Fiehn, 2002). Genomics, transcriptomics, and proteomics are generally much more mature, having had more time to develop the necessary hardware and software infrastructure to enable rapid data collection, robust data analysis, and facile data interpretation. However, there are certain advantages to being the last to arrive on the scene. For instance, many of the statistical and computational methods needed to process and interpret high-dimensional data were developed, tested, and refined long before metabolomics really hit the mainstream of "omics" research. In particular, techniques such as SAM (Tusher et al., 2001), EBAM (Efron et al., 2001), heatmaps with hierarchical clustering (Eisen, Spellman, Brown, & Botstein, 1998), gene set enrichment analysis (Subramanian et al., 2005), pathway analysis (Goffard & Weiller, 2007), and others were initially developed to serve the microarray community. However, the utility and applicability of these methods proved to be so broad that other "omics" fields (i.e., proteomics and then metabolomics) quickly adopted them. This simple approach of borrowing well-tested ideas from other "omics" platforms and then adopting them to the specific needs of metabolomic data has allowed metabolomics to quickly catch up to its more mature cousins.

In developing MetaboAnalyst, many of the ideas and data processing strategies originally developed for microarray analysis were liberally borrowed and adopted to fit the specific needs or requirements of metabolomic data (Xia et al., 2009). Indeed, key concepts such as interactive web-based data processing, data pipelining, name normalization, data normalization, SAM, volcano plots, heatmaps with hierarchical clustering, pathway topology analysis, and gene-set enrichment analysis technology are all examples of ideas or tools developed originally for microarray analysis that are now embedded in MetaboAnalyst. However, to adopt these tools for metabolomic data analysis, it was necessary to assemble large and well-annotated databases of metabolites, metabolite synonyms, metabolite concentrations, metabolic pathways, and metabolite sets. This aspect was both time-consuming and difficult. While MetaboAnalyst's methodologies may not be completely unique or original, its knowledgebases certainly are.

Metabolomic data processing also draws on certain ideas or concepts that appear to be unique to this field. In particular, spectral binning and peak alignment are examples of data processing concepts that grew from the field of chemometrics (Deming, 1986) and are

not commonly used or seen in other "omics" fields. Likewise, PCA, partial least square discriminant analysis (PLS-DA), sparse partial least square discriminant analysis (sPLS-DA), and orthogonal partial least square discriminant analysis (orthoPLS-DA) also have their origins in chemometric data processing. Interestingly, some of these ideas that originated in metabolomics or chemometrics are now finding some traction in analyzing transcriptomic and proteomic data. In other words, metabolomics is starting to pay back some of the intellectual debt it owes to transcriptomics and proteomics.

The 12 basic protocols presented here were designed to take readers through some of the most typical or common scenarios in metabolomic data processing and data interpretation. Obviously, it is impossible to describe every conceivable scenario that can be handled by MetaboAnalyst. Likewise, it is almost impossible to devise a scenario where every function, routine, or algorithm within MetaboAnalyst can be called or demonstrated. Indeed, we would estimate that less than half of all MetaboAnalyst's available functionalities have been demonstrated with these twelve protocols. Of those functions or routines not seen or described in this chapter, many are discussed in MetaboAnalyst's online tutorials, help pages, description hyperlinks, or FAQ pages. In other cases, readers can see how MetaboAnalyst was used to analyze specific kinds of metabolomic data or specific kinds of data processing scenarios through the many publications that have cited MetaboAnalyst. If an answer cannot be found or a methodology seems too confusing, we would encourage readers to contact MetaboAnalyst's support team (*jasmine.chong@mail.mcgill.ca* or *jeff.xia@mcgill.ca*).

It is important to note that MetaboAnalyst is not the only metabolomic data analysis platform available to today's metabolomics researchers. One popular stand-alone commercial tool is SIMCA-P+ (Umetrics, Sweden). SIMCA-P+ is a desktop application with a very nicely designed graphical user interface that supports a wide variety of data transformations and multivariate statistical analyses, including PCA, PLS-DA, and OPLS. However, SIMCA-P+ does not support metabolomic-specific data processing (for NMR and/or MS data), nor does it offer high-level functional interpretation through automated metabolite annotation, biomarker analysis, power analysis, joint pathway analysis, MSEA, or MetPA. Outside of MetaboAnalyst, there are several other freely available web-based metabolomic data processing tools—MeltDB (Kessler et al., 2013), MetabolomeExpress (Carroll, Badger, & Harvey Millar, 2010), and XCMS Online (Huan et al., 2017; Tautenhahn et al., 2012). MeltDB and MetabolomeExpress have been primarily developed for MS-based metabolomics data storage, administration, analysis, and annotation; XCMS Online is a cloud-based informatics platform designed to process and visualize MS-based, untargeted metabolomic data for dependent (paired), two-group comparisons, meta-analysis, and multi-group comparisons, with various statistical outputs, interactive PCA plots, and pathway analysis. For untargeted MS-based metabolomics data, we recommend using XCMS Online (Huan et al., 2017) to process the raw MS spectral data and then uploading the resulting peak intensity tables for statistical analysis and functional interpretation using MetaboAnalyst. For users who are familiar with R, we have discussed in Basic Protocol 12 how to use the MetaboAnalyst R package to perform raw spectral preprocessing.

Critical Concepts and Troubleshooting

MetaboAnalyst 4.0 was implemented using the latest PrimeFaces library (*http://www.primefaces.org/*) based on the JavaServer Faces (JSF) Technology. Most of the backend computations and visualizations are carried out by >800 functions written in R (v3.5.1). MetaboAnalyst has been tested on all major web browsers with JavaScript enabled. Users should enable JavaScript or upgrade their browser if the MetaboAnalyst home page is not displayed properly. One important concept to remember is that MetaboAnalyst is not "bookmarkable"—users should always first visit the home page and then go to the modules and follow the steps listed in our protocols. This ensures that all the resources will be properly loaded based on the user's data types and parameter settings.

For first-time users, the most common problem is proper data preparation. This is often manifested in the form of difficulties passing the MetaboAnalyst's data integrity check. To address this problem, a comprehensive resource is provided in the "Data Formats" and "FAQs" pages. Several functions have also been implemented to automatically correct minor issues such as double quotes and empty lines, as well as to

generate more informative error messages. Most importantly, example datasets for all the modules are provided to allow users to directly download and open these data in a text editor or spreadsheet program to view their detailed structures.

MetaboAnalyst has seen a continuous increase in user traffic over the past decade. The public server now routinely processes metabolomics data submitted from ~1200 users daily, with over >2.5 million data analysis jobs completed during the last 12 months. To address the increasing computational demand, version 4.0 is now hosted on a high-performance Google Compute Engine with 52GB RAM and eight virtual CPUs with 2.6 GHz each. Users with large datasets (i.e., from high-resolution untargeted metabolomics) or metabolomic centers are encouraged to install a local copy of MetaboAnalyst, using the provided WAR file or Docker image, or use the MetaboAnalystR package. Detailed instructions on how to install MetaboAnalyst on a local computer running Linux or Mac operating systems with Java and R installed are provided on the web server. Users are required to have some basic bioinformatics and general computer knowledge to perform the installation. Furthermore, advanced users of MetaboAnalyst may have felt constrained by the analysis boundaries placed by the web interface. Therefore, the recently released MetaboAnalystR 2.0 package now permits a complete and flexible metabolomics data workflow as well as support for local batch-processing capabilities. Users will be able to create a workflow (R script) using the web application, customize the workflow to their data, and then execute the customized workflow in batch mode using the MetaboAnalystR package. More advanced users will be able to directly modify the underlying R code to suit their needs, or use it in conjunction with other R packages. The MetaboAnalystR package is available from the GitHub (*https://github.com/xia-lab/MetaboAnalystR*). All instructions can be found on the "Resources" page.

As with any software tool, MetaboAnalyst becomes easier to use as one gains more experience working with it. However, if a user believes that there is something wrong with the program or the server, they are encouraged to contact the primary MetaboAnalyst authors (*jasmine.chong@mail.mcgill.ca* or *jeff.xia@mcgill.ca*). Due to the large number of users and the limited resources, users are required to follow all steps listed under the "Troubleshooting/Contact" page regarding how to report such issues.

Acknowledgments

The authors wish to thank McGill University, the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canadian Institutes for Health Research (CIHR), Genome Canada (GC), and the Canada Research Chairs (CRC) Program for financial support.

Literature Cited

Ametaj, B. N., Zebeli, Q., Saleem, F., Psychogios, N., Lewis, J. L., Dunn, S. M., ... Wishart, D. M. (2010). Metabolomics reveals unhealthy alterations in rumen metabolism with increased proportion of cereal grain in the diet of dairy cows. *Metabolomics*, *5*(4), 375–386.

Bijlsma, S., Bobeldijk, I., Verheij, E. R., Ramaker, R., Kochhar, S., Macdonald, I. A., ... Smilde, A. K. (2006). Large-scale human metabolomics studies: A strategy for data (pre-) processing and validation. *Analytical Chemistry*, *78*(2), 567–574. doi: 10.1021/ac051495j.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. doi: 10.1023/A:1010933404324.

Carroll, A. J., Badger, M. R., & Harvey Millar, A. (2010). The MetabolomeExpress Project: Enabling web-based processing, analysis and transparent dissemination of GC/MS metabolomics datasets. *BMC Bioinformatics*, *11*, 376. doi: 10.1186/1471-2105-11-376.

Chong, J., Soufan, O., Li, C., Caraus, I., Li, S., Bourque, G., ... Xia, J. (2018). MetaboAnalyst 4.0: Towards more transparent and integrative metabolomics analysis. *Nucleic Acids Research*, *46*(W1), W486–W494. doi: 10.1093/nar/gky310.

Chong, J., & Xia, J. (2018). MetaboAnalystR: An R package for flexible and reproducible analysis of metabolomics data. *Bioinformatics*, *34*(24), 4313–4314. doi: 10.1093/bioinformatics/bty528.

Chong, J., Yamamoto, M., & Xia, J. (2019). MetaboAnalystR 2.0: From raw spectra to biological insights. *Metabolites*, *9*(3), 57. doi: 10.3390/metabo9030057.

Craig, A., Cloarec, O., Holmes, E., Nicholson, J. K., & Lindon, J. C. (2006). Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Analytical Chemistry*, *78*(7), 2262–2267. doi: 10.1021/ac0519312.

Deming, S. N. (1986). Chemometrics: An overview. *Clinical Chemistry*, *32*(9), 1702–1706.

Dieterle, F., Ross, A., Schlotterbeck, G., & Senn, H. (2006). Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Analytical Chemistry*, *78*(13), 4281–4290. doi: 10.1021/ac051632c.

Draghici, S., Khatri, P., Tarca, A. L., Amin, K., Done, A., Voichita, C., ... Romero, R. (2007). A systems biology approach for pathway level analysis. *Genome Research*, *17*(10), 1537–1545. doi: 10.1101/gr.6202607.

Efron, B., Tibshirani, R., Storey, J. D., & Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, *96*(456), 1151–1160. doi: 10.1198/016214501753382129.

Efron, B., & Tibshirani, R. (2007). On testing the significance of sets of genes. *The Annals of Applied Statistics*, *1*(1), 107–129. doi: 10.1214/07-AOAS101.

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(25), 14863–14868. doi: 10.1073/pnas.95.25.14863.

Eisner, R., Stretch, C., Eastman, T., Xia, J., Hau, D., Damaraju, S., ... Vickie, B. (2010). Learning to predict cancer-associated skeletal muscle wasting from 1H-NMR profiles of urinary metabolites. *Metabolomics*, *3*(2), 207–214.

Eisner, R., Stretch, C., Eastman, T., Xia, J. G., Hau, D., Damaraju, S., ... Baracos, V. E. (2011). Learning to predict cancer-associated skeletal muscle wasting from H-1-NMR profiles of urinary metabolites. *Metabolomics*, *7*(1), 25–34. doi: 10.1007/s11306-010-0232-9.

Fahrmann, J. F., Kim, K., DeFelice, B. C., Taylor, S. L., Gandara, D. R., Yoneda, K. Y., ... Miyamoto, S. (2015). Investigation of metabolomic blood biomarkers for detection of adenocarcinoma lung cancer. *Cancer Epidemiology and Prevention Biomarkers*, *24*(11), 1716–1723. doi: 10.1158/1055-9965.EPI-15-0427.

Fiehn, O. (2002). Metabolomics—The link between genotypes and phenotypes. *Plant Molecular Biology*, *48*(1–2), 155–171. doi: 10.1023/A:1013713905833.

Frolkis, A., Knox, C., Lim, E., Jewison, T., Law, V., Hau, D. D., ... Wishart, D. S. (2010). SMPDB: The small molecule pathway database. *Nucleic Acids Research*, *38*(Database issue), D480–D487. doi: 10.1093/nar/gkp1002.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., ... Zhang, J. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, *5*(10), R80. doi: 10.1186/gb-2004-5-10-r80.

Goeman, J. J., van de Geer, S. A., de Kort, F., & van Houwelingen, H. C. (2004). A global test for groups of genes: Testing association with a clinical outcome. *Bioinformatics*, *20*(1), 93–99. doi: 10.1093/bioinformatics/btg382.

Goffard, N., & Weiller, G. (2007). PathExpress: A web-based tool to identify relevant pathways in gene expression data. *Nucleic Acids Research*, *35*(Web Server issue), W176–W181. doi: 10.1093/nar/gkm261.

Hackstadt, A. J., & Hess, A. M. (2009). Filtering for increased power for microarray data analysis. *BMC Bioinformatics*, *10*, 11. doi: 10.1186/1471-2105-10-11.

Hendriks, M. M., Smit, S., Akkermans, W. L., Reijmers, T. H., Eilers, P. H., Hoefsloot, H. C., ... Smilde, A. K. (2007). How to distinguish healthy from diseased? Classification strategy for mass spectrometry-based clinical proteomics. *Proteomics*, *7*(20), 3672–3680. doi: 10.1002/pmic.200700046.

Huan, T., Forsberg, E. M., Rinehart, D., Johnson, C. H., Ivanisevic, J., Benton, H. P., ... Siuzdak, G. (2017). Systems biology guided by XCMS Online metabolomics. *Nature Methods*, *14*(5), 461. doi: 10.1038/nmeth.4260.

Hummel, M., Meister, R., & Mansmann, U. (2008). GlobalANCOVA: Exploration and assessment of gene group effects. *Bioinformatics*, *24*(1), 78–85. doi: 10.1093/bioinformatics/btm531.

Integrative HMP (iHMP) Research Network Consortium. (2014). The Integrative Human Microbiome Project: Dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell Host & Microbe*, *16*(3), 276–289.

Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, *8*(1), 118–127. doi: 10.1093/biostatistics/kxj037.

Kessler, N., Neuweger, H., Bonte, A., Langenkamper, G., Niehaus, K., Nattkemper, T. W., & Goesmann, A. (2013). MeltDB 2.0-advances of the metabolomics software system. *Bioinformatics*, *29*(19), 2452–2459. doi: 10.1093/bioinformatics/btt414.

Kuhl, C., Tautenhahn, R., Bottcher, C., Larson, T. R., & Neumann, S. (2012). CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Analytical Chemistry*, *84*(1), 283–289. doi: 10.1021/ac202450g.

Kuhn, M., von Mering, C., Campillos, M., Jensen, L. J., & Bork, P. (2007). STITCH: Interaction networks of chemicals and proteins. *Nucleic Acids Research*, *36*(suppl_1), D684–D688. doi: 10.1093/nar/gkm795.

Li, S., Park, Y., Duraisingham, S., Strobel, F. H., Khan, N., Soltow, Q. A., ... Pulendran, B. (2013). Predicting network activity from high throughput metabolomics. *PLoS Computational Biology*, *9*(7), e1003123. doi: 10.1371/journal.pcbi.1003123.

Lommen, A., & Kools, H. J. (2012). MetAlign 3.0: Performance enhancement by efficient use of advances in computer hardware. *Metabolomics*, *8*(4), 719–726. doi: 10.1007/s11306-011-0369-1.

Meinicke, P., Lingner, T., Kaever, A., Feussner, K., Goebel, C., Feussner, I., ... Morgenstern, B. (2008). Metabolite-based clustering and visualization of mass spectrometry data using one-dimensional self-organizing maps.

Algorithms for Molecular Biology, *3*, 9. doi: 10.1186/1748-7188-3-9.

Murakami, Y., Kubo, S., Tamori, A., Itami, S., Kawamura, E., Iwaisako, K., ... Taguchi, Y. H. (2015). Comprehensive analysis of transcriptome and metabolome analysis in intrahepatic cholangiocarcinoma and hepatocellular carcinoma. *Scientific Reports*, *5*, 16294. doi: 10.1038/srep16294.

Pavlidis, P., & Noble, W. S. (2001). Analysis of strain and regional variation in gene expression in mouse brain. *Genome Biology*, *2*(10). RESEARCH0042. doi: 10.1186/gb-2001-2-10-research0042.

Pluskal, T., Castillo, S., Villar-Briones, A., & Oresic, M. (2010). MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics*, *11*, 395. doi: 10.1186/1471-2105-11-395.

Psihogios, N. G., Kalaitzidis, R. G., Dimou, S., Seferiadis, K. I., Siamopoulos, K. C., & Bairaktari, E. T. (2007). Evaluation of tubulointerstitial lesions' severity in patients with glomerulonephritides: An NMR-based metabonomic study. *Journal of Proteome Research*, *6*(9), 3760–3770. doi: 10.1021/pr070172w.

Ravanbakhsh, S., Liu, P., Bjordahl, T. C., Mandal, R., Grant, J. R., Wilson, M., ... Wishart, D. S. (2015). Accurate, fully-automated NMR spectral profiling for metabolomics. *PloS One*, *10*(5), e0124219. doi: 10.1371/journal.pone.0124219.

Röst, H. L., Sachsenberg, T., Aiche, S., Bielow, C., Weisser, H., Aicheler, F., ... Kohlbacher, O. (2016). OpenMS: A flexible open-source software platform for mass spectrometry data analysis. *Nature Methods*, *13*(9), 741. doi: 10.1038/nmeth.3959.

Saghatelian, A., Trauger, S. A., Want, E. J., Hawkins, E. G., Siuzdak, G., & Cravatt, B. F. (2004). Assignment of endogenous substrates to enzymes by global metabolite profiling. *Biochemistry*, *43*(45), 14332–14339. doi: 10.1021/bi0480335.

Scheltema, R. A., Jankevics, A., Jansen, R. C., Swertz, M. A., & Breitling, R. (2011). PeakML/mzMatch: A file format, Java library, R library, and tool-chain for mass spectrometry data analysis. *Analytical Chemistry*, *83*(7), 2786–2793. doi: 10.1021/ac2000994.

Smilde, A. K., Jansen, J. J., Hoefsloot, H. C., Lamers, R. J., van der Greef, J., & Timmerman, M. E. (2005). ANOVA-simultaneous component analysis (ASCA): A new tool for analyzing designed metabolomics data. *Bioinformatics*, *21*(13), 3043–3048. doi: 10.1093/bioinformatics/bti476.

Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R., & Siuzdak, G. (2006). XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry*, *78*(3), 779–787. doi: 10.1021/ac051437y.

Stacklies, W., Redestig, H., Scholz, M., Walther, D., & Selbig, J. (2007). pcaMethods—A bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, *23*(9), 1164–1167. doi: 10.1093/bioinformatics/btm069.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., ... Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545–15550. doi: 10.1073/pnas.0506580102.

Tai, Y. C., & Speed, T. P. (2006). A multivariate empirical Bayes statistic for replicated microarray time course data. *Annals of Statistics*, *34*(5), 2387–2412. doi: 10.1214/009053606000000759.

Tautenhahn, R., Patti, G. J., Rinehart, D., & Siuzdak, G. (2012). XCMS Online: A web-based platform to process untargeted metabolomic data. *Analytical Chemistry*, *84*(11), 5035–5039. doi: 10.1021/ac300698c.

Tsugawa, H., Cajka, T., Kind, T., Ma, Y., Higgins, B., Ikeda, K., ... Arita, M. (2015). MS-DIAL: Data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nature Methods*, *12*(6), 523–526. doi: 10.1038/nmeth.3393.

Tusher, V. G., Tibshirani, R., & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(9), 5116–5121. doi: 10.1073/pnas.091062498.

van den Berg, R. A., Hoefsloot, H. C., Westerhuis, J. A., Smilde, A. K., & van der Werf, M. J. (2006). Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genomics*, *7*, 142. doi: 10.1186/1471-2164-7-142.

van Iterson, M., 't Hoen, P. A., Pedotti, P., Hooiveld, G. J., den Dunnen, J. T., van Ommen, G. J., ... Menezes, R. X. (2009). Relative power and sample size analysis on gene expression profiling data. *BMC Genomics*, *10*, 439. doi: 10.1186/1471-2164-10-439.

Westerhuis, C. A., Hoefsloot, C. J. H., Smit, S., Vis, J. D., Smilde, A. K., van Velzen, E. J. J., ... van Dorsten, F. A. (2007). Assessment of PLSDA cross validation. *Metabolomics*, *4*(1), 81–89. doi: 10.1007/s11306-007-0099-6.

Wickham, H., & Chang, W. (2016). Devtools: Tools to make developing r packages easier. R package version. 1(0). Available at https://cran.r-project.org/web/packages/devtools/index.html.

Xia, J., Broadhurst, D. I., Wilson, M., & Wishart, D. S. (2013). Translational biomarker discovery in clinical metabolomics: An introductory tutorial. *Metabolomics*, *9*(2), 280–299. doi: 10.1007/s11306-012-0482-9.

Xia, J., Mandal, R., Sinelnikov, I. V., Broadhurst, D., & Wishart, D. S. (2012). MetaboAnalyst 2.0—A comprehensive server for

metabolomic data analysis. *Nucleic Acids Research*, *40*(Web Server issue), W127–W133. doi: 10.1093/nar/gks374.

Xia, J., Psychogios, N., Young, N., & Wishart, D. S. (2009). MetaboAnalyst: A web server for metabolomic data analysis and interpretation. *Nucleic Acids Research*, *37*(Web Server issue), W652–W660. doi: 10.1093/nar/gkp356.

Xia, J., Sinelnikov, I. V., Han, B., & Wishart, D. S. (2015). MetaboAnalyst 3.0—Making metabolomics more meaningful. *Nucleic Acids Research*, *43*(W1), W251–W257. doi: 10.1093/nar/gkv380.

Xia, J., Sinelnikov, I. V., & Wishart, D. S. (2011). MetATT: A web-based metabolomics tool for analyzing time-series and two-factor datasets. *Bioinformatics*, *27*(17), 2455–2456. doi: 10.1093/bioinformatics/btr392.

Xia, J., & Wishart, D. S. (2010). MSEA: A web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Research*, *38*(Web Server issue), W71–W77. doi: 10.1093/nar/gkq329.

Xia, J., & Wishart, D. S. (2010). MetPA: A web-based metabolomics tool for pathway analysis and visualization. *Bioinformatics*, *26*(18), 2342–2344. doi: 10.1093/bioinformatics/btq418.

Yao, Q., Xu, Y., Yang, H., Shang, D., Zhang, C., Zhang, Y., ... Li, X. (2015). Global prioritization of disease candidate metabolites based on a multi-omics composite network. *Scientific Reports*, *5*, 17201. doi: 10.1038/srep17201.