**2. Analysis Utilities**

MetaboAnalystR has been designed to synchronize with MetaboAnalyst website for comprehensive metabolomics analysis. LC-MS/MS Raw Spectral Analysis, Functional Analysis of GLBiomarker Analysis, Enrichment Analysis, Meta-Analysis, Pathway Analysis, Integrated Pathway Analysis, Power Analysis Module, Time Series or Two Factor Design, Network Explorer Module, MS Peaks to Pathways, Batch effect correction etc. can be easily achieved in website.  
More importantly, all kinds of analysis results including figures generation can be repeated with MetaboAnalystR for further advanced editing. MetaboAnalyst website also provides a real-time R codes for users to finish this process. Therefore, it is highly recommended to use website for further processing after the raw data processing. In order to provide a easy way for advanced user to perform the analysis locally.  
We summarized all vignettes of different modules as below as the guidance for in-depth data mining at R end.

**2.1 LC-MS/MS Raw Spectra Processing**

Liquid chromatography coupled to high-resolution mass spectrometry platforms are increasingly employed to comprehensively measure metabolome changes in systems biology and complex diseases. Over the past decade, several powerful computational pipelines have been developed for spectral processing, annotation, and analysis. However*, significant obstacles remain with regard to parameter settings, computational efficiencies, spectral deconvolution and compound identification*.  
  
In previous version, MetaboAnalystR adopts an optimization strategy based on *regions of interest (ROI)* to avoid the time-consuming step of recursive peak detection using complete spectra. Briefly, the algorithm first scans the whole spectra across m/z and retention time dimensions to select several ROIs that are enriched for real peaks. Second, these ROIs are then extracted as new synthetic spectra. Finally, a DoE model is used to optimize peak picking parameters based on the synthetic spectra.  
  
MetaboAnalystR **4.0 mainly focus on the raw LC-MS/MS data processing**. As the most widely used technique, LC-MS/MS is typically performed with MS1 full scan coupled with MS/MS to achieve high-throughput quantification and compound annotation. MS/MS spectra can be generated using data-dependent acquisition (DDA) or data-independent acquisition (DIA) methods.  
1). DDA acquires MS/MS spectra by fragmentation of precursor ions selected using a relatively narrow MS/MS isolation window (e.g., 1 m/z). Although DDA spectra are directly linked to precursors, recent [studies](https://www.nature.com/articles/s41592-021-01195-3) show that > 50% of them are ‘chimeric’ and need to be deconvolved before searching any reference database.  
2). DIA usually fragments all ions in a wider m/z range (e.g., >15 m/z) with multiple cycles to improve the coverage on the metabolome. SWATH-MS (sequential window acquisition of all theoretical fragment ion spectra mass spectrometry) is a common DIA approach for both metabolomics and proteomics. Spectral deconvolution is essential for DIA to relink precursors with fragment ions.  
  
In brief, LC-MS/MS data processing workflow in MetaboAnalystR involves several steps, including raw spectral data import, MS data processing (auto-optimized peak picking, alignment, gap filling and annotation), DDA/SWATH-DIA data deconvolution, spectrum consensus from replicates, MS/MS reference library searching, results export, and integration into functional prediction.

**A detailed vignette has been prepared**[**here**](https://www.metaboanalyst.ca/resources/vignettes/LCMSMS_Raw_Spectral_Processing.html)**to showcase how to perform LC-MS/MS Raw Spectra Processing with MetaboAnalystR 4.0.**

[Back to Content](https://www.metaboanalyst.ca/docs/RTutorial.xhtml#top)

**2.2 Functional Analysis of Global Metabolomics**

Tools for functional interpretation of global metabolomics data is in general lacking or poorly addressed. A prerequisite for metabolomics data interpretation is metabolite identification, thereby permitting the contextualization of annotated peaks in metabolic pathways and their integration with other omics data.  
  
However, even with high mass accuracy afforded by the current high-resolution MS platforms, it is often impossible to uniquely identify a given peak based on its mass alone. Researchers usually need to manually search compound databases and then perform further experimental validations such as tandem MS. Novel bioinformatics tools are urgently needed to enable researchers to gain biological insights with a minimum amount of manual efforts. To get around this bottleneck, a key concept is to shift the unit of analysis from individual compounds to individual pathways or a group of functionally related compounds (i.e., metabolite sets). The general assumption is that the collective behavior of a group is more robust against a certain degree of random errors of individuals.  
  
The [*mummichog*](https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003123) algorithm is the first implementation of this concept to infer pathway activities from a ranked MS peaks. The original algorithm implements an over-representation analysis (ORA) method to evaluate pathway-level enrichment based on significant peaks. An alternative approach is the [Gene Set Enrichment Analysis (GSEA)](https://www.pnas.org/doi/abs/10.1073/pnas.0506580102) method, which is widely used to test enriched functions from ranked gene lists. Unlike ORA, GSEA considers the overall ranks of features without using a significance cutoff. It can detect subtle and consistent changes which could be missed from using ORA methods. Despite its widespread applications in gene expression profiling, it has not yet been applied to global metabolomics.  
  
There two options of *mummichog* provided in MetaboANalystR 4.0. Mummichog version 2 has incorporated retention time in grouping ions and introduced the concept of empirical compounds (ECs). ECs are putative metabolites as measured by LC-HRMS, possibly containing a mixture of enantiomers, stereoisomers, and positional isomers that are not resolved by the instruments. Thus, ECs are similar to the "feature groups". More details have been described in our previous [publication](https://www.mdpi.com/2218-1989/10/5/186).

**A step by step vignette has been prepared**[**here**](https://www.metaboanalyst.ca/resources/vignettes/Functional_Analysis_global_metabolomics.html)**on how to perform functional analysis for global metabolomics data with MetaboAnalystR 4.0.**

[Back to Content](https://www.metaboanalyst.ca/docs/RTutorial.xhtml#top)

**2.3 Statistical Analysis (one-factor)**

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 and MetaboAnalystR 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.  
  
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.  
  
Here, we are providing a detailed vignette on the common metablomics statistical analysis. This tutorial is mainly for one-factor design of metabolomics dataset.

**Read**[**here**](https://www.metaboanalyst.ca/resources/vignettes/Statistical_Analysis_Module.html)**for more details on a step-by-step statistical analysis with MetaboAnalystR 4.0.**

[Back to Content](https://www.metaboanalyst.ca/docs/RTutorial.xhtml#top)

**2.4 Enrichment Analysis of Targeted Metabolomics**

The enrichment analysis module performs metabolite set enrichment analysis (MSEA) for human and mammalian species based on several libraries containing ~6300 groups of metabolite sets. Users can upload either  
1) a list of compounds,  
2) a list of compounds with concentrations, or  
3) a concentration table.

**Read**[**here**](https://www.metaboanalyst.ca/resources/vignettes/Enrichment_Analysis.html)**for more details on how to perform Enrichment Analysis with MetaboAnalystR 4.0.**

[Back to Content](https://www.metaboanalyst.ca/docs/RTutorial.xhtml#top)

**2.5 Pathway Analysis of Targeted Metabolomics**

The pathway analysis module supports pathway analysis (integrating enrichment analysis and pathway topology analysis) and visualization for 21 model organisms, including Human, Mouse, Rat, Cow, Chicken, Zebrafish, Arabidopsis thaliana, Rice, Drosophila, Malaria, S. cerevisae, E.coli, and others, with a total of ~1600 metabolic pathways.

**Read**[**here**](https://www.metaboanalyst.ca/resources/vignettes/Pathway_Analysis.html)**for more details on the basic design and rules of MetaboAnalystR 4.0.**

[Back to Content](https://www.metaboanalyst.ca/docs/RTutorial.xhtml#top)

**2.6 Biomarker Analysis**

The metabolome is well-known to be a sensitive measure of health and disease, reflecting alterations to the genome, proteome, and transcriptome, as well as changes in life-style and environment. As such, one common goal of metabolomic studies is biomarker discovery, aiming to identify a metabolite or a set of metabolites capable of classifying conditions or disease, with high sensitivity (true-positive rate) and specificity (true negative rate). This is achieved through building predictive models of one or multiple metabolites and evaluating the performance, or robustness of the model, to classify new patients into diseased or healthy categories. The main steps for biomarker analysis are as follows:

 1) Biomarker selection

 2) Performance evaluation

 3) Model creationMetaboAnalystR provides users with several options for single (classical) or multiple biomarker analysis, as well as for predictive biomarker model creation and evaluation. For a comprehensive introductory tutorial and further details concerning biomarker analysis, please refer to Xia et al. 2013 (PMID: 23543913).

Detailed tutorial of Pathway Analysis can be downloaded [**here**](https://drive.google.com/file/d/1DjLDE9IGvU_rjfdIDdlUtS7FGahfcsov/view?usp=sharing).

[Back to Content](https://www.metaboanalyst.ca/docs/RTutorial.xhtml#top)

**2.7 Statistical Analysis (Metadata table)**

Metadata describes the data, and contains details on the experimental conditions, sample sources (i.e., species, tissue), sample collection (i.e., location, time) and other factors. Such metadata are critical for data interpretation, because they allow researchers to account for the biological and environmental context when they analyze the data, and facilitate data reuse by allowing other researchers to search for, and meaningfully compare and potentially integrate, results from across diverse studies. Details on the context and sample source are becoming increasingly important as observational studies that collect omics data from human populations or animals outside laboratory settings are becoming more common.  
  
Epidemiologic studies in biomedical or environmental sciences generally involve a primary variable of interest, such as presence/absence of a certain disease or exposure to a specific chemical, as well as variables such as age, sex or other potential factors that covary with the primary metadata. Statistical analyses that take these covariates into account can lead to substantial increases in power and draw more robust conclusions about the relationships between the primary variable and the omics data.

Detailed tutorial of Statistical Analysis (Metadata table) will be available soon

[Back to Content](https://www.metaboanalyst.ca/docs/RTutorial.xhtml#top)

**2.8 Joint-Pathway Analysis**

This module performs integrated metabolic pathway analysis on results obtained from combined metabolomics and gene expression studies conducted under the same experimental conditions. This approach exploits KEGG metabolic pathway models to complete the analysis. The underlying assumption behind this module is that by combining evidence from both changes in gene expression and metabolite concentrations, one is more likely to pinpoint the pathways involved in the underlying biological processes. To this end, users need to supply a list of genes and metabolites of interest that have been identified from the same samples or obtained under similar conditions. The metabolite list can be selected from the results of a previous analysis downloaded from MetaboAnalyst. Similarly, the gene list can be easily obtained using many excellent web-based tools such as GEPAS or INVEX. After users have uploaded their data, the genes and metabolites are then mapped to KEGG metabolic pathways for over-representation analysis and pathway topology analysis. Topology analysis uses the structure of a given pathway to evaluate the relative importance of the genes/compounds based on their relative location. Clicking on the name of a specific pathway will generate a graphical representation of that pathway highlighted with the matched genes/metabolites. Users must keep in mind that unlike transcriptomics, where the entire transcriptome is routinely mapped, current metabolomic technologies only capture a small portion of the metabolome. This difference can lead to potentially biased results. To address this issue, the current implementation of this omic integration module allows users to explore the enriched pathways based either on joint evidence or on the evidence obtained from one particular omic platform for comparison.

Detailed tutorial of Integrated Pathway Analysis can be downloaded [**here**](https://drive.google.com/file/d/1HVoXNX98CZLcpr7MkpVZP5DueXcnZeIj/view?usp=sharing).

[Back to Content](https://www.metaboanalyst.ca/docs/RTutorial.xhtml#top)

**2.9 Functional Meta-Analysis**

It is notoriously challenging to integrate untargeted metabolomics data across different studies, because different extraction methods, chromatographic conditions and mass spectrometry platforms all lead to heterogeneity of HRMS data. This issue has precluded the use of untargeted metabolomics datasets for large-scale meta-analysis using conventional statistical methods. To address this gap, we have developed a new module to enable researchers to perform functional meta-analysis of global metabolomics datasets.

Detailed tutorial of Functional Meta-Analysis will be available soon

[Back to Content](https://www.metaboanalyst.ca/docs/RTutorial.xhtml#top)

**2.10 Network Analysis**

Biological processes are driven by a complex web of interactions amongst numerous molecular entities of a biological system. The classical method of pathway analysis is unable to identify important associations or interactions between molecules belonging to different pathways. Network analysis is therefore commonly used to address this limitation. Here, the aim of the Network Explorer module is to provide an easy-to-use tool to that allows users to map their metabolites and/or genes onto different networks for novel insights or development of new hypotheses. Mapping of both metabolites and genes are supported in this module (including KOs), whereby either entity can be projected onto five existing biological networks including the KEGG global metabolic network, the gene-metabolite interaction network, the metabolite-disease interaction network, the metabolite-metabolite interaction network, and the metabolite-gene-disease interaction network. The last four networks are created based on information gathered from HMDB and STITCH and are applicable to human studies only. Users can upload either a list of metabolites, a list of genes, or a list of both metabolites and genes. MetaboAnalystR currently accepts compound names, HMDB IDs, KEGG compound IDs as metabolite identifiers. As well, we only accept Entrez IDs, ENSEMBL IDs, Official Gene Symbols, or KEGG Orthologs (KO) as gene identifiers. The uploaded list of metabolites and/or genes is then mapped using our internal databases of metabolites and gene annotations. Following this step, users can select which of the five networks to begin visually exploring their data.

Detailed tutorial of Network Explorer Module can be downloaded [**here**](https://drive.google.com/file/d/1fZ364APP8pqemFg0oD0kIhAzH9Zm9p-0/view?usp=sharing).

[Back to Content](https://www.metaboanalyst.ca/docs/RTutorial.xhtml#top)

**2.11 Power Analysis**

The Power analysis module supports sample size estimation and power analysis for designing population-based or clinical metabolomic studies. As metabolomics is becoming a more accessible and widely used tool, methods to ensure proper experimental design are crucial to allow for accurate and robust identification of metabolites linked to disease, drugs, environmental or genetic differences. Traditional power analysis methods are unsuitable for metabolomics data as the high-throughput nature of this data means that it is highly dimensional and often correlated. Further, the number of metabolites identified greatly outnumbers the sample size. Thus, modified methods of power analysis are needed to address such concerns.  
  
One solution is to use the average power of all metabolites, and to correct for multiple testing using methods such as the false discovery rate (FDR) instead of raw p-values. MetaboAnalystR uses the SSPA R package to perform power analysis (van Iterson et al. 2009, PMID: 19758461).

Detailed tutorial of Power Analysis Module can be downloaded [**here**](https://drive.google.com/file/d/1UE7592V1vvpaeeyCfVPD_692Coxww238/view?usp=sharing).

[Back to Content](https://www.metaboanalyst.ca/docs/RTutorial.xhtml#top)

**2.12 Meta-Analysis**

A major challenge in biomarker discovery for disease detection, classification, and monitoring is the validation of potential metabolic markers. Questions have been raised about biomarker consistency and robustness across individual metabolomic studies of the same disease, and the importance of external validation to improve statistical power to validate biomarkers has been recently reviewed. Therefore to address the lack of user-friendly tools for the horizontal integration of metabolomics data, we present a second new module called “Meta-Analysis”. The primary goal of the Meta-Analysis module is to provide a user-friendly and comprehensive tool for the integration of individual metabolomic studies to identify biomarkers of disease. The steps for Meta-Analysis occur as follows:

 1) Users must upload individual data, which must be in tabular form. Prior to uploading the data, the user must clean the datasets to ensure consistency amongst named metabolites, spectral bins, or peaks, as well as consistency in the included metadata across included studies.

 2) The module performs differential enrichment analysis for each individual study to compute summary level-statistics for each metabolite feature (e.g. p-value). The summary level-statistical results from all studies are combined, and meta-analysis is performed using one of several statistical options: combining of p-values, vote counting, or direct merging of data into a mega-dataset.

 3) The results can be visualized as a Upset diagram to view all possible combinations of shared features between the datasets.

Detailed tutorial of Meta Analysis can be downloaded [**here**](https://drive.google.com/file/d/1nCH980icadGcQJwtQ8yHDKcuM32V7BLG/view?usp=sharing).

[Back to Content](https://www.metaboanalyst.ca/docs/RTutorial.xhtml#top)