Using MetaboAnalyst 5.0 Part I: Optimizing parameters for LC-HRMS spectra processing

Zhiqiang Pang
McGill University
Guangyan Zhou
McGill University
Jessica Ewald
McGill University https://orcid.org/0000-0002-7555-736X
Le Chang
McGill University
Orcun Hacariz
McGill University
Niladri Basu
McGill University
Jianguo Xia (jeff.xia@mail.mcgill.ca)
McGill University

Method Article

Keywords: Global metabolomics, exposomics, HRMS, spectra processing, parameter optimization, meta-analysis, multi-omics, metadata, covariate adjustment

Posted Date: June 17th, 2022

DOI: https://doi.org/10.21203/rs.3.pex-1917/v1

Abstract

Liquid-chromatography coupled with high-resolution mass spectrometry (LC-HRMS) has become a workhorse in global metabolomics studies with growing applications across biomedical and environmental sciences. However, outstanding bioinformatics challenges in terms of data processing, statistical analysis and functional interpretation remain critical barriers to the wider adoption of this technology. To help the user community overcome these barriers, we have made major updates to the well-established MetaboAnalyst platform (www.metaboanalyst.ca). This protocol is part one of four and provides stepwise instructions on how to use MetaboAnalyst to optimize parameters for LC-HRMS spectra processing. It may take ~2 hours to complete depending on the server load.

Introduction

Introduction

The goal of metabolomics is to comprehensively study all metabolites in biological samples. For research concerning pre-defined lists of compounds (targeted metabolomics), various protocols have been established and an increasing number of commercial kits are becoming available. However, unbiased comprehensive metabolome profiling (global metabolomics) remains a critical bottleneck due to several complex analytical and bioinformatics challenges [1-3]. Developing high-throughput global metabolomics technologies has become a high priority task in metabolomics[4] as well as the burgeoning fields of exposomics[5] and precision medicine[6]. Among different technologies available, high-resolution mass spectrometry (HRMS) has shown great promise[3,7]. HRMS instruments such as Orbitrap or time-of-flight (TOF) systems coupled with gas or liquid chromatography (GC/LC) can simultaneously measure a wide range of endogenous and exogenous compounds to characterize an individual's metabolic phenotype, environmental exposures, and associated biological responses. LC-HRMS is arguably the most widely used platform as it can measure a broad range of compounds with minimal sample preparation. However, the associated data processing and analysis steps remain particularly challenging to most researchers.

The computational workflow for LC-HRMS-based global metabolomics can be summarized in three general steps: 1) *spectra processing* to convert raw spectra into a peak intensity table; 2) *peak annotation* to characterize peak relationships and to assign putative compound identities (ID); 3) *data analysis* to identify important features, patterns of variation, and their functional interpretation. Although these steps are conceptually similar to that of other omics data analysis such as gene expression or microbiome data analysis[8,9], the inherent characteristics of the data generated in LC-HRMS-based global metabolomics present unique challenges especially for those without statistical programming skills and/or a deep understanding of the analytical instrumentation. To address this gap, we have recently made significant updates to MetaboAnalyst to support LC-HRMS-based global metabolomics data analysis and interpretation[10]. This protocol provides an overview of these new features followed by step-wise instructions through several example datasets using MetaboAnalyst 5.0.

MetaboAnalyst and other web-based tools

MetaboAnalyst was launched in 2009, and since then there have been major update releases every three years [10-14]. MetaboAnalyst versions 1.0 - 3.0 were mainly designed for general statistical and functional analysis of targeted metabolomics data. Starting with v4.0, the development of MetaboAnalyst has gradually shifted towards addressing more complex bioinformatics and statistical needs arising from global metabolomics data, including raw spectra processing, functional analysis, and integration with other omics data. According to Google Analytics, the public platform (www.metaboanalyst.ca) is currently being accessed by ~2,000 users daily worldwide.

Many software tools have been developed in the past decade for processing and analyzing metabolomics data [15,16]. Most of these tools need to be locally installed by the users. For web-based platforms, there are several popular options including MetaboAnalyst, XCMS online[17], W4M[18] and NOREVA[19]. These web-based tools have been compared with MetaboAnalyst 5.0 in our recent publication[10]. In summary, MetaboAnalyst offers the most features for statistical and functional analysis in comparison to XCMS online and W4M (providing more comprehensive support for raw data processing), and NOREVA (dedicated for metabolomics data normalization and quality assessment). A hallmark of MetaboAnalyst is its ease of use which enables researchers from diverse backgrounds to perform various complex tasks of data analysis. For instance, MetaboAnalyst 5.0 offers an efficient parameter optimization pipeline to allow automated high-quality raw LC-HRMS spectra processing, as compared to many other tools which generally requires users to manually adjust multiple parameters to achieve satisfactory performance. In addition, the results of spectra processing can be directly transferred to other compatible modules within MetaboAnalyst for streamlined statistical and functional analysis. The overall workflow of MetaboAnalyst is depicted in **Figure 1**.

Driven by user feedback, we have published several comprehensive tutorials and protocols for each major release of MetaboAnalyst [28-33]. These tutorials provide detailed instructions for tasks related to data processing, filtering, normalization, statistical analysis of datasets with single experimental factors, as well as functional analysis for targeted metabolomics. Our 2019 protocol[32] is still up to date on these topics. Accordingly, this current protocol details four major updates that have been made to MetaboAnalyst 5.0 including: Part 1) Optimized processing of LC-HRMS spectra (Steps 1-18); Part 2) Deriving functional insights from LC-HRMS peaks (Steps 19 - 31); Part 3) Meta-analysis and integration with transcriptomics data (Steps 32 - 45); and Part 4) Statistical analysis and exploration with complex metadata (Steps 46 - 65). More background information for Part 1 is given below:

Spectra processing

This is the first task in global metabolomics. The continuum/profile data of a single raw spectrum obtained from a LC-HRMS instrument (such as Q-Exactive) is typically 1-2 gigabytes (GB) in mzXML/mzML format. A common practice is to first perform peak centroiding to condense the Gaussian shaped mass peak to a single mass centroid. This step can significantly reduce the file size to ~100 MB. The next step is to detect peaks (also known as peak picking) from the centroid data. Multiple algorithms have been developed to identify peaks in different dimensions such as retention time[26,34,35]. Please see **Figure 4** for more details. Among them, the *centWave* algorithm[34] implemented in XCMS has been shown to perform well in processing LC-HRMS spectra. However, a practical difficulty associated with using XCMS is to decide appropriate values for several key parameters, which requires a relatively deep understanding of both MS instrumentation and the peak picking algorithm. Automating or standardizing parameter specification has become critical to help facilitate broad adoption by the diverse user communities[36,37]. After peak detection, peak alignment is performed to address retention time variations across spectra. These aligned peaks form a peak intensity table with varying proportions of missing values. These missing values indicate that either peak detection failed, or that the corresponding feature is absent from the respective sample. Therefore, the final step is “gap filling” by re-performing direct peak extraction on corresponding regions in the raw spectra.

Peak annotation

A typical LC-HRMS spectrum of common biofluids (such as blood or urine samples) can often produce over 10,000 peaks. However, a critical point is that this number is not equivalent to the number of compounds detected. The correspondence between these peaks and the actual metabolites remains elusive. Many peaks are derived from the same compounds in the forms of adducts, isotopes, fragmentations, *etc.* (real or biological peaks), and other peaks may reflect background noise (artifacts or noise peaks)[38]. Therefore, the first step in peak annotation aims to identify real peaks, and to clarify the relationships among them. Many empirical and statistical rules have been developed to address this problem, including CAMERA and CliqueMS[39,40] which are two popular R packages. The next step is to assign putative compound IDs to those peaks. This is a challenging task as even with high resolution instruments, a single peak can potentially match multiple compounds[41]. It is important to keep in mind that, unlike targeted metabolomics, the primary goal of global metabolomics is to understand the overall patterns and to identify promising features to inform the design of follow-up studies for more targeted analysis. As discussed in the sections below, accurate peak annotation, although important, is not an absolute prerequisite for functional interpretation of global metabolomics data.

Reagents

Equipment

Computer requirements

· Browser requirements: MetaboAnalyst 5.0 runs on all modern web browsers. For the best results, we recommend Google Chrome 92+, Firefox 92+, Safari 12+, and Microsoft Edge v93+. JavaScript must be enabled in your browser.

· Internet connection requirements: A fast connection is highly recommended.

· Hardware requirements: >4 GB of RAM and a screen resolution of at least 1200 x 800 is preferred. At least 8 GB hard drive space is needed to store the raw spectra files.

Data files

· *Input Files*. MetaboAnalyst accepts multiple data formats for different modules. For raw spectral processing, the supported data formats include mzML, mzXML, mzData and NetCDF, while other modules usually require comma-separated values (.csv) or tab delimited (.txt) file format. More details of these data formats are provided in **Box 1**.

· *Example Datasets*. Multiple built-in example datasets are offered in each MetaboAnalyst module. Users can simply choose an example dataset and click the 'Submit' button to start exploring the tool. Six datasets are used in this protocol:

1) A raw spectra data (*malaria_raw.zip*) consisting of 12 plasma samples from healthy and malaria semi-immune subjects along with 3 Quality Controls (QC) [53].

2) A feature table (*malaria_feature_table.csv*) generated from processing the above raw spectra for functional analysis.

3) A gene list (*integ_genes_1.txt*) and a compound list (*integ_cmpds.txt*) to demonstrate the integration of targeted metabolomics data with transcriptomics data.

4) A gene list (*integ_genes_2.txt*) and a peak list (*integ_peaks.txt*) to demonstrate the integration of global metabolomics data with transcriptomics data.

5) Three peak intensity tables (*A1_pos.csv*, *B1_pos.csv*, *C1_pos.csv*) from global metabolomics studies of serum samples from COVID-19 patients to demonstrate the integration of multiple global metabolomics datasets[54].

6) A feature table (*TCE_feature_table.csv*) and its associated metadata table (*TCE_metadata.csv*) from an exposomics study on metabolic changes associated with occupational exposure to trichloroethylene (TCE) [55].

Equipment Setup

(Optional) Download the example datasets. Go to the MetaboAnalyst home page and click 'Data Formats' link from the left panel. Users can download all the example datasets used in the protocol. These datasets are also available as built-in examples in each corresponding module. Users can directly select those datasets and follow the protocols without downloading and uploading them.

Procedure

Part I: LC-MS Raw Spectra Processing (Timing 1.5 ~ 2 hours)

1. *Starting up*. Go to the MetaboAnalyst home page and click the 'Click here to start' button in the middle of the page. The module page displays all modules as a pyramid organized into four tiers. The top tier contains one button 'LC-MS Spectra Processing'. Click the button to enter the raw spectra processing module.

Critical Step: *Almost all modern browsers support multiple tabs. Please keep MetaboAnalyst open in one tab only. Accessing MetaboAnalyst from multiple tabs could interfere with each other leading to unpredictable results.*

2. *Data uploading*. Unzip the dataset #1 (*malaria_raw.zip*) into individual files. Click the 'Select' button to open a File Chooser dialog. Locate the and select all spectra files (.zip) and the metadata file (.txt). Click the 'Upload' button to start uploading. At any time, users can cancel the upload by clicking the 'Reset' button below the uploading progress bars.

Critical Step: *All raw spectra files need to be compressed to *.zip files individually and uploaded together with a metadata file (.txt). QC files should start with "QC_". The formats of raw spectra files and metadata list are introduced in* ***Figure 7****. If your spectra files are not centroided, please do it locally with ProteoWizard*[20]*. This will significantly reduce the file size and increase upload speed.*

Troubleshooting (see Figure 5)

3. Once the upload has finished, click the 'Proceed' button. Alternatively, users can use the built-in example dataset without uploading spectra. To do this, locate the table under "Try our example data", select the 2nd option and click "Submit".

4. *Data integrity check*. Data integrity check results are shown on this page. The spectra file names, centroid status, sizes, and group information are displayed in the first four columns. Only centroid data are supported for further processing. Users can simply click the wrench icon to centroid their data on the fly. By default, all samples will be processed. Users can choose to exclude certain samples by deselecting the corresponding checkboxes in the 'Include' column. Click the "Next" button.

Troubleshooting (see Figure 5)

5. *Customize profiling parameters*. The LC-MS Spectra Processing page displays all important parameters to process raw spectra. Several platform-specific parameters are provided by default. Three algorithms are available for peaks picking and two algorithms for peak alignment. Users need to first select a LC-MS platform. For this dataset, select 'UPLC-Q/E'. There are two options for parameter setting - *Default/Manual* and *Auto-optimized*. Here we select 'Auto-optimized' option for spectra processing. Please refer to **Figure 8** for more details on parameter optimization in MetaboAnalyst. For advanced users, they could manually configure the parameters by using the first option.

Critical Step: *The Auto-optimized procedure has been developed for XCMS centWave algorithm which performs generally well for HRMS spectra. For low resolution MS spectra, users may want to use MatchedFilter. For users interested in detecting low-intensity peaks, Massifquant algorithm could perform better*[56]*.* Implementation of different algorithms for peak picking and alignment are summarized in **Table 1**.

6. *Customize annotation parameters*. The peak annotation parameters (including ion mode, potential adducts *etc.*) must be specified manually based on the experimental conditions. Make sure the "Polarity" is set to "positive" and keep other options as default.

7. (Optional) *Spectral inspection for potential contaminants*. If 'Auto-optimized' option is selected for peak profiling, the '*Contaminants*' option will be enabled. This option allows users to exclude potential instrumental noise or contaminants present during the chromatographic run. This option is selected by default. Users can visually inspect the spectral data by clicking the '*View*' link to bring up the 'Spectral Inspection' dialog (**Figure S1. A**). By default, a random QC sample (if provided) is displayed in 3D style. Users can switch to other spectra files or different spectral regions based on m/z and RT ranges (**Figure S1. B**). Alternatively, users can use a 2D heatmaps to view the same information (**Figure S1. C**). Here, we could see multiple peaks persist over half of the chromatogram, therefore we keep the '*Remove*' checked.

Critical Step: *In addition to removal of contaminant peaks, spectra files with general signal intensity below 5% of the average of all spectra files will be excluded as these outliers will cause error during peak alignment.*

8. *Job submission*. Once the parameter configuration is done, click '*Submit Job*' and '*Confirm*' to proceed. Once the job is submitted, users cannot come back to this page and modify parameters until the job is complete or cancelled.

9. *Create a job URL*. Raw spectra processing could take hours to finish depending on the server load. After the job is submitted, users will be directed to the '*Job Status View*' page. A '*Job ID*' will be assigned to this job automatically. Users can view the running status of the submitted job from the 'Current Status'. The status will show '*Pending*' if the job is in queue. After the job starts running, the actual processing takes ~75 min for this data using the 'Auto-optimized' mode, or ~10 min using Default/manual mode. We suggest that users create a URL bookmark and save the link in a text file. Then click '*Exit*' from the navigation tree of the left panel. Users can return via their bookmark URLs at any time to check the status of their job.

10. *Job progress and status*. The job execution process is displayed as a progress bar in the middle of the '*Job Status*' box. The data processing details are displayed in the 'Text Output' box. The job status is refreshed every 3 seconds. Click 'Cancel Job' to cancel the current job at any time. Once the job is complete, click 'Proceed' to view the results.

Tip: the optimized parameters can be downloaded in a text file as an experimental record or for further re-use. Users can manually enter these parameters using the Manual mode next time when they process the same or similar data to achieve faster analysis.

Troubleshooting (see Figure 5)

11. *Result visualization*. The processed results are summarized using several graphics (PCA, TIC, BPI, *etc.*) in separate tabs (**Figure 3A**). User can rotate the PCA 3D plot to view data from different angles or zoom in and out with a mouse scroll. By default, all features are shown in the loading plot. Drag the '*PCA Loadings*' slider to keep the top features based on their mean intensities. The PCA score plot of samples can be displayed based on all peaks or current peaks in the loading plot.

Tip: *PCA summarizes the main patterns of variation of the data. The top 25~50% most abundant features can usually reveal the same general pattern without overcrowding the loading plot*.

12. *Exploring samples and feature using PCA.* Double click a node in the PCA score plot to generate the Total Ion Chromatogram (TIC) plot of the corresponding sample (**Figure 3B**). Double-clicking a node in the PCA loading plot will display a dialog box with two panels (**Figure 3C**). The left panel shows box plots summarizing the intensity distributions of the selected feature across groups. Double-clicking a data point on the box plot will generate its corresponding EIC plot. Clicking more data points will overlay their EIC plots. The function allows users to create and visually compare typical EICs in different groups. Click the reset icon on the right-top corner to re-initiate the process.

13. *Result tables*. The second half of the results consist of two tables. 'Spectra / Sample Table' summarizes peak detection results from different samples, while the 'Feature/Peak Table' displays all peaks (m/z, retention time) and their annotations (isotopes, adducts, formula, potential compounds, *etc.*). Users can search and view specific samples or features for more detailed inspection.

Tip: *The character 'M' represents the precursor ion. All blanks in the isotopic and adducts' annotation columns are the either precursors ions or ones without annotations.*

14. *Analysis report generation and results downloading*. Click the 'Download Page' button to enter the corresponding page. All processed results are shown in a table on the 'Results Download' tab. Click the "Generate Analysis Report" to create a PDF report of the current analysis.

15. (Optional) User can compare their results with those obtained using the default parameter settings. Click the 'Spectra Processing' node from the navigation tree to return to the Parameter Setting page. Select 'UPLC-Q/E' platform, make sure the "Default/manual" mode is selected, set the "Polarity" as 'positive' and submit the job. **Figure 6** shows the results based on the 'Optimized' and the 'Default' options, in which optimized parameters produced greater number of peaks and higher proportions of peaks with putative annotations. In addition, the top two PCs capture more variations of the data. These results clearly validate the effectiveness of the optimization strategies described in **Figure 8**.

16. After finishing the download, click 'Logout' to end the analysis. Users can also select the "Start New Journey" tab to explore results in other compatible modules.

Troubleshooting (see Figure 5)

17. (Optional) *Create a user account*. Users can set up an account to better manage their jobs. Click '*Log in*' button from the left panel of the data uploading page (**Step 2**). If users do not have an account, they can register for one by clicking 'Create account' to create a new account or clicking 'Forgot password?' to set a new password.

18. (Optional) *Manage projects and account.* After logging into their account, all projects are displayed on the Project page. A maximum of 10 projects are allowed per user. Each project will be saved for at most 180 days. At any time, users can click the 'Delete Account' button from the projects management page to completely remove their accounts from MetaboAnalyst.

Troubleshooting

Time Taken

Anticipated Results

LC-MS raw spectra processing

Raw spectra processing produces a peak intensity table (*metaboanalyst_input.csv*). The peaks annotation result is saved in the *annotated_peaklist.csv* file. A total of 3,125 peaks have been found in the dataset #1 with 'Auto-optimized' pipeline, and of these 1,442 peaks (46.1% of the total) were annotated as isotopes or adducts. Base Peak Ion Chromatogram (BPI), Total Ion Chromatogram (TIC) and specific Extracted Ion Chromatogram (EIC) are all generated and downloadable to show the chromatographical results.

References

1 Alseekh, S. *et al.* Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices. *Nat Methods* **18**, 747-756, doi:10.1038/s41592-021-01197-1 (2021).

2 Alseekh, S. & Fernie, A. R. Metabolomics 20 years on: what have we learned and what hurdles remain? *Plant J* **94**, 933-942, doi:10.1111/tpj.13950 (2018).

3 Doerr, A. Global metabolomics. *Nature Methods* **14**, 32-32, doi:10.1038/nmeth.4112 (2017).

4 Cajka, T. & Fiehn, O. Toward Merging Untargeted and Targeted Methods in Mass Spectrometry-Based Metabolomics and Lipidomics. *Anal Chem* **88**, 524-545, doi:10.1021/acs.analchem.5b04491 (2016).

5 Vermeulen, R., Schymanski, E. L., Barabasi, A. L. & Miller, G. W. The exposome and health: Where chemistry meets biology. *Science* **367**, 392-396, doi:10.1126/science.aay3164 (2020).

6 Beger, R. D. *et al.* Metabolomics enables precision medicine: "A White Paper, Community Perspective". *Metabolomics* **12**, 149, doi:10.1007/s11306-016-1094-6 (2016).

7 Want, E. J. *et al.* Global metabolic profiling of animal and human tissues via UPLC-MS. *Nat Protoc* **8**, 17-32, doi:10.1038/nprot.2012.135 (2013).

8 Zhou, G. *et al.* NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. *Nucleic Acids Res* **47**, W234-W241, doi:10.1093/nar/gkz240 (2019).

9 Chong, J., Liu, P., Zhou, G. & Xia, J. Using MicrobiomeAnalyst for comprehensive statistical, functional, and meta-analysis of microbiome data. *Nat Protoc* **15**, 799-821, doi:10.1038/s41596-019-0264-1 (2020).

10 Pang, Z. *et al.* MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights. *Nucleic Acids Research* **49**, W388-W396, doi:10.1093/nar/gkab382 (2021).

11 Xia, J., Psychogios, N., Young, N. & Wishart, D. S. MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Res* **37**, W652-660, doi:10.1093/nar/gkp356 (2009).

12 Xia, J., Mandal, R., Sinelnikov, I. V., Broadhurst, D. & Wishart, D. S. MetaboAnalyst 2.0–a comprehensive server for metabolomic data analysis. *Nucleic Acids Res* **40**, W127-133, doi:10.1093/nar/gks374 (2012).

13 Xia, J., Sinelnikov, I. V., Han, B. & Wishart, D. S. MetaboAnalyst 3.0–making metabolomics more meaningful. *Nucleic Acids Res* **43**, W251-257, doi:10.1093/nar/gkv380 (2015).

14 Chong, J. *et al.* MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. *Nucleic Acids Res* **46**, W486-W494, doi:10.1093/nar/gky310 (2018).

15 Stanstrup, J. *et al.* The metaRbolomics Toolbox in Bioconductor and beyond. *Metabolites* **9**, doi:10.3390/metabo9100200 (2019).

16 Gardinassi, L. G., Xia, J., Safo, S. E. & Li, S. Bioinformatics Tools for the Interpretation of Metabolomics Data. *Current Pharmacology Reports* **3**, 374-383, doi:10.1007/s40495-017-0107-0 (2017).

17 Tautenhahn, R., Patti, G. J., Rinehart, D. & Siuzdak, G. XCMS Online: a web-based platform to process untargeted metabolomic data. *Anal Chem* **84**, 5035-5039, doi:10.1021/ac300698c (2012).

18 Giacomoni, F. *et al.* Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics. *Bioinformatics* **31**, 1493-1495, doi:10.1093/bioinformatics/btu813 (2015).

19 Yang, Q. *et al.* NOREVA: enhanced normalization and evaluation of time-course and multi-class metabolomic data. *Nucleic Acids Res* **48**, W436-W448, doi:10.1093/nar/gkaa258 (2020).

20 Chambers, M. C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol* **30**, 918-920, doi:10.1038/nbt.2377 (2012).

21 Pluskal, T., Castillo, S., Villar-Briones, A. & Oresic, M. 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 (2010).

22 Du, X. X., Smirnov, A., Pluskal, T., Jia, W. & Sumner, S. Metabolomics Data Preprocessing Using ADAP and MZmine 2. *Computational Methods and Data Analysis for Metabolomics* **2104**, 25-48, doi:Book_Doi 10.1007/978-1-0716-0239-3 (2020).

23 Tsugawa, H. *et al.* A lipidome atlas in MS-DIAL 4. *Nat Biotechnol* **38**, 1159-1163, doi:10.1038/s41587-020-0531-2 (2020).

24 Tsugawa, H. *et al.* Hydrogen Rearrangement Rules: Computational MS/MS Fragmentation and Structure Elucidation Using MS-FINDER Software. *Anal Chem* **88**, 7946-7958, doi:10.1021/acs.analchem.6b00770 (2016).

25 Rost, H. L. *et al.* OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat Methods* **13**, 741-748, doi:10.1038/nmeth.3959 (2016).

26 Kenar, E. *et al.* Automated label-free quantification of metabolites from liquid chromatography-mass spectrometry data. *Mol Cell Proteomics* **13**, 348-359, doi:10.1074/mcp.M113.031278 (2014).

27 Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47, doi:10.1093/nar/gkv007 (2015).

28 Xia, J. & Wishart, D. S. Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst. *Nat Protoc* **6**, 743-760, doi:10.1038/nprot.2011.319 (2011).

29 Xia, J. & Wishart, D. S. Metabolomic data processing, analysis, and interpretation using MetaboAnalyst. *Curr Protoc Bioinformatics* **Chapter 14**, Unit 14 10, doi:10.1002/0471250953.bi1410s34 (2011).

30 Xia, J. & Wishart, D. S. Using MetaboAnalyst 3.0 for Comprehensive Metabolomics Data Analysis. *Curr Protoc Bioinformatics* **55**, 14 10 11-14 10 91, doi:10.1002/cpbi.11 (2016).

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

32 Chong, J., Wishart, D. S. & Xia, J. Using MetaboAnalyst 4.0 for Comprehensive and Integrative Metabolomics Data Analysis. *Curr Protoc Bioinformatics* **68**, e86, doi:10.1002/cpbi.86 (2019).

33 Chong, J. & Xia, J. Using MetaboAnalyst 4.0 for Metabolomics Data Analysis, Interpretation, and Integration with Other Omics Data. *Methods Mol Biol* **2104**, 337-360, doi:10.1007/978-1-0716-0239-3_17 (2020).

34 Tautenhahn, R., Bottcher, C. & Neumann, S. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics* **9**, 504, doi:10.1186/1471-2105-9-504 (2008).

35 Yu, T., Park, Y., Johnson, J. M. & Jones, D. P. apLCMS–adaptive processing of high-resolution LC/MS data. *Bioinformatics* **25**, 1930-1936, doi:10.1093/bioinformatics/btp291 (2009).

36 Alboniga, O. E., Gonzalez, O., Alonso, R. M., Xu, Y. & Goodacre, R. Optimization of XCMS parameters for LC-MS metabolomics: an assessment of automated versus manual tuning and its effect on the final results. *Metabolomics* **16**, 14, doi:10.1007/s11306-020-1636-9 (2020).

37 Pang, Z., Chong, J., Li, S. & Xia, J. MetaboAnalystR 3.0: Toward an Optimized Workflow for Global Metabolomics. *Metabolites* **10**, doi:10.3390/metabo10050186 (2020).

38 Sindelar, M. & Patti, G. J. Chemical Discovery in the Era of Metabolomics. *J Am Chem Soc* **142**, 9097-9105, doi:10.1021/jacs.9b13198 (2020).

39 Senan, O. *et al.* CliqueMS: a computational tool for annotating in-source metabolite ions from LC-MS untargeted metabolomics data based on a coelution similarity network. *Bioinformatics* **35**, 4089-4097, doi:10.1093/bioinformatics/btz207 (2019).

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

41 Kind, T. & Fiehn, O. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *Bmc Bioinformatics* **8**, doi:10.1186/1471-2105-8-105 (2007).

42 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550, doi:10.1073/pnas.0506580102 (2005).

43 Xia, J. & Wishart, D. S. MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res* **38**, W71-77, doi:10.1093/nar/gkq329 (2010).

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

45 Li, S. *et al.* Predicting network activity from high throughput metabolomics. *PLoS Comput Biol* **9**, e1003123, doi:10.1371/journal.pcbi.1003123 (2013).

46 Xia, J. *et al.* INMEX--a web-based tool for integrative meta-analysis of expression data. *Nucleic Acids Res* **41**, W63-70, doi:10.1093/nar/gkt338 (2013).

47 Zhou, G. & Xia, J. OmicsNet: a web-based tool for creation and visual analysis of biological networks in 3D space. *Nucleic Acids Res* **46**, W514-W522, doi:10.1093/nar/gky510 (2018).

48 Schriml, L. M. *et al.* COVID-19 pandemic reveals the peril of ignoring metadata standards. *Sci Data* **7**, 188, doi:10.1038/s41597-020-0524-5 (2020).

49 Nygaard, V., Rodland, E. A. & Hovig, E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics* **17**, 29-39, doi:10.1093/biostatistics/kxv027 (2016).

50 Kahan, B. C., Jairath, V., Dore, C. J. & Morris, T. P. The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. *Trials* **15**, 139, doi:10.1186/1745-6215-15-139 (2014).

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

52 Chong, J., Yamamoto, M. & Xia, J. MetaboAnalystR 2.0: From Raw Spectra to Biological Insights. *Metabolites* **9**, doi:10.3390/metabo9030057 (2019).

53 Gardinassi, L. G. *et al.* Integrative metabolomics and transcriptomics signatures of clinical tolerance to Plasmodium vivax reveal activation of innate cell immunity and T cell signaling. *Redox Biol* **17**, 158-170, doi:10.1016/j.redox.2018.04.011 (2018).

54 Pang, Z., Zhou, G., Chong, J. & Xia, J. Comprehensive Meta-Analysis of COVID-19 Global Metabolomics Datasets. *Metabolites* **11**, doi:10.3390/metabo11010044 (2021).

55 Walker, D. I. *et al.* High-resolution metabolomics of occupational exposure to trichloroethylene. *Int J Epidemiol* **45**, 1517-1527, doi:10.1093/ije/dyw218 (2016).

56 Conley, C. J. *et al.* Massifquant: open-source Kalman filter-based XC-MS isotope trace feature detection. *Bioinformatics* **30**, 2636-2643, doi:10.1093/bioinformatics/btu359 (2014).

57 Vaughan, A. M. *et al.* Type II fatty acid synthesis is essential only for malaria parasite late liver stage development. *Cell Microbiol* **11**, 506-520, doi:10.1111/j.1462-5822.2008.01270.x (2009).

58 Shears, M. J., Botte, C. Y. & McFadden, G. I. Fatty acid metabolism in the Plasmodium apicoplast: Drugs, doubts and knockouts. *Mol Biochem Parasitol* **199**, 34-50, doi:10.1016/j.molbiopara.2015.03.004

(2015).

59 Cumnock, K. *et al.* Host Energy Source Is Important for Disease Tolerance to Malaria. *Curr Biol* **28**, 1635-1642 e1633, doi:10.1016/j.cub.2018.04.009 (2018).

60 Libiseller, G. *et al.* IPO: a tool for automated optimization of XCMS parameters. *BMC Bioinformatics* **16**, 118, doi:10.1186/s12859-015-0562-8 (2015).

61 McLean, C. & Kujawinski, E. B. AutoTuner: High Fidelity and Robust Parameter Selection for Metabolomics Data Processing. *Anal Chem* **92**, 5724-5732, doi:10.1021/acs.analchem.9b04804 (2020).

62 Zhou, G., Ewald, J. & Xia, J. OmicsAnalyst: a comprehensive web-based platform for visual analytics of multi-omics data. *Nucleic Acids Res* **49**, W476-W482, doi:10.1093/nar/gkab394 (2021).

63 Rappoport, N. & Shamir, R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Res* **47**, 1044, doi:10.1093/nar/gky1226 (2019).

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

65 Prince, J. T. & Marcotte, E. M. Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping. *Anal Chem* **78**, 6140-6152, doi:10.1021/ac0605344 (2006).

Acknowledgements

Acknowledgements

The authors would like to thank Genome Canada, Génome Québec, Natural Sciences and Engineering Research Council of Canada (NSERC), and Canada Research Chairs (CRC) Program for funding support.

Author Contributions

Z.P., J.E., N.B. and J.X. prepared the manuscript. Z.P., G.Z., J.E., L.C., O.H. and J.X. contributed to the development and testing of the MetaboAnalyst. All authors read and approved the final manuscript.

Competing Financial Interests

The authors declare no competing financial interests.

Data Availability Statement

All example datasets used in the protocol are integrated as example datasets in their respective modules and are also available for download from the “Format” page of MetaboAnalyst (https://www.metaboanalyst.ca/MetaboAnalyst/docs/Format.xhtml). There are no restrictions on their use.

Code Availability Statement

MetaboAnalyst is freely accessible as a web-based application. The underlying R code is freely available at GitHub as the MetaboAnalystR (https://github.com/xia-lab/MetaboAnalystR) and OptiLCMS (https://github.com/xia-lab/OptiLCMS) packages under the GNU General Public License version 2 or later.

Figures

Figure 1

Figure 2

Figure 3

Figure 4

Figure 5

Figure 6

Figure 7

Figure 8