Using MetaboAnalyst 6.0 for exposomics data analysis – from LC-MS/MS spectra processing to dose response modeling and causal estimate

Zhiqiang Pang[1,2], Yao Lu[1,2], Guangyan Zhou[1,2], Huiting Ou[3,4], Charles Viau[1,2], Fumihiko Matsuda[4], Niladri Basu[2], and Jianguo Xia[1,2,3*]

[1] Department of Microbiology and Immunology, Faculty of Medicine and Health Sciences, McGill University, Montreal, Québec, Canada

[2] Faculty of Agricultural and Environmental Sciences, McGill University, Ste-Anne-de-Bellevue, Québec, Canada

[3] Department of Human Genetics, McGill University, Montreal, Québec, Canada

[4] Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan

*Correspondence: J. Xia

Email: jeff.xia@mcgill.ca (J.X.)

Abstract

Exposomics is an emerging field of research that aims to comprehensively investigate individuals' environmental exposures and how these exposures relate to health outcomes. Liquid chromatography-tandem mass spectrometry (LC-MS/MS) is widely used in exposomics studies. MetaboAnalyst (https://www.metaboanalyst.ca/) is a widely used platform for statistical and functional analysis of metabolomics data. The recent MetaboAnalyst 6.0 release incorporates updates designed to address the unique analytical requirements arising from exposomics studies focusing on identification and characterization of exposures and their biological impacts. This protocol extends our 2022 Nature Protocol by providing step-by-step instructions on how to use MetaboAnalyst 6.0 for exposomics data analysis, including: LC-MS/MS spectra processing and compound identification (Stage 1); exposomics data processing and exploratory analysis (Stage 2); dose-response modeling to study metabolic responses to exposure levels (Stage 3); and leveraging known genetic associations for causal inference (Stage 4). We demonstrate Stages 1-3 using data from a recent blood exposomics study concerning electronic waste exposure. Stage 4 is illustrated through an investigation of the potential causal link between L-Isoleucine and type 2 diabetes. Stage 1 may take approximately two hours to complete depending on server load, and the remaining stages may be executed in a total of approximately 90 minutes.

Key points

The protocol provides instructions for four key tasks in LC-MS/MS based exposomics studies

1. Raw Spectra Processing – Enables sensitive peak detection and compound annotation
2. Exploratory Data Analysis – Provides quality control, data cleaning, statistical analysis, and visualization
3. Dose Response Analysis – Quantifies the relationship between exposure levels and biological or phenotypic effects
4. Causal Inference – Estimates causal relationships between exposures and health outcomes using genetic variants as instrumental variables

Key references

- Pang, Z. et al. *Nucleic Acids Res*. 52, W398–W406 (2024): https://doi.org/10.1093/nar/gkae253
- Pang, Z. et al. *Nat Commun.* 15, 3675 (2024): https://doi.org/10.1038/s41467-024-48009-6
- Chang, L. et al. *Exposome*. 4, osae005 (2024): https://doi.org/10.1093/exposome/osae005
- Pang, Z. et al. *Metabolites.* 14, 671 (2024): https://doi.org/10.3390/metabo14120671

Introduction

Deep sequencing and genome-wide association studies (GWAS) have identified numerous associative relationships between genetic variants and various health outcomes [1]. However, not all individuals who have inherited a particular genotype develop its associated phenotype. A particular phenotype is the product of complex interactions between genetics and environmental factors.

Exposomics is an emerging field of research that aims to comprehensively investigate the environmental factors that are associated with health outcomes [2,3]. The exposome is defined as the entire set of environmental exposures throughout the life course [2]. In practice, however, most current exposomics studies employ cross-sectional or short-term designs focusing on phenotypic differences associated with specific exposures. Exposomics studies employ various technologies such as mass spectrometry (MS), next-generation sequencing, remote sensors, and wearables to measure exposures at both the individual level (e.g., dietary intake, microbiome composition, lifestyle variables) and the population level (e.g., air quality, water contamination, soil composition). Modern high-resolution MS instruments can simultaneously detect hundreds to thousands of chemical compounds in biological samples, providing comprehensive insights into both environmental exposures and their associated metabolic responses [4-6].

The field of exposomics draws researchers from three distinct fields - biomedical science, epidemiology and environmental science. After data collection, different schools of researchers often take different paths in their data analysis. Biomedical researchers are generally interested in understanding molecular responses to determine the pathways by which exposures lead to health outcomes. Epidemiologists perform exposome-wide association studies to identify potential risk factors that are significantly associated with phenotypes of interest. Environmental scientists focus on identifying exposures that are associated with adverse outcomes and further characterize their relationships for quantitative risk assessment. There is an urgent demand for bioinformatics tools that support approaches developed from different fields to enable comprehensive analysis and interpretation of data from exposomics studies.

Addressing unique needs for exposomics data analysis

Most current methods developed for exposomics data analysis are implemented as R packages[7-11]. MetaboAnalyst (https://www.metaboanalyst.ca) is a well-established, web-based platform for comprehensive metabolomics data analysis. While exposomics and metabolomics share many important procedures in data acquisitions and analysis, exposomics presents distinct analytical requirements. To address these needs, MetaboAnalyst 6.0 introduces several new modules alongside substantial enhancements to existing modules.

Sensitive LC-MS/MS spectra processing and compound identification.

Raw spectral processing is often the first and most critical step in exposomics data analysis. Key tasks include feature detection and peak alignment, spectra deconvolution and compound annotation. Compared to metabolomics, exposomics places greater emphasis on compound identification and detection of low-abundance exposures. MetaboAnalyst 6.0 integrates Asari [12] into its LC-MS spectra processing pipeline for feature detection and annotation. The novel algorithm represents a shift from traditional sample-wise peak picking to a track-centric approach. It first identifies consistent mass features across the full experimental cohort, significantly improving the consistency of elution peak detection and annotation [13]. For MS/MS spectra processing, MetaboAnalyst 6.0 provides efficient peak deconvolution for both DDA and SWATH-DIA data [14], with comprehensive built-in databases for compound annotation and exposome classification (see **Box 1**).

Enhanced workflows for quality checking, data processing and exploration

Large-scale exposomics studies often suffer from retention-time and intensity drift that obscure biologically important features. To address this issue, quality control (QC) and blank samples are critical components during data acquisition. QC samples, prepared by pooling aliquots from all study samples, are injected repeatedly as technical replicates to assess analytical system stability and reproducibility. Blank samples, prepared by substituting biological material with solvent, facilitate identification and removal of background contaminants from sample preparation and spectra collection. During spectral processing, these samples are processed together with other biological samples. Downstream data analysis must be able to use these samples to improve data quality. Additionally, large cohort studies are usually associated with complex metadata requiring flexible data analysis framework. MetaboAnalyst 6.0 addresses these requirements through enhanced metadata validation, quality assessment, feature filtering, and missing value imputation capabilities, coupled with advanced statistical and visualization tools to enable flexible, comprehensive data exploration (see **Box 2**).

Dose response analysis for modelling response profiles

Identifying exposures that differ significantly between phenotypic groups is a critical first step in exposomics research. However, simply noting a compound's presence or significant change is insufficient for health risk assessment, as biological effects are fundamentally dependent on exposure levels. Dose response analysis, developed in environmental science, provides a quantitative framework for establishing the relationship between exposure levels and biological responses. This approach is widely used in traditional toxicology studies across different dose levels and time points [15], and has been increasingly applied to high-throughput omics data [16-18]. MetaboAnalyst 6.0 supports dose response analysis for data from both traditional repeated-dosing

experiments and exposomics studies with continuous exposures, accommodating categorical and continuous outcome variables (see **Box 3**).

Mendelian randomization for causal analysis

Epidemiological studies have identified myriads of associations between various exposures and outcomes. A critical challenge is distinguishing which modifiable exposures are causally, rather than merely associatively, linked to health outcomes. However, randomized controlled trials, the gold standard for establishing causality, are usually unfeasible, unethical, or prohibitively expensive in human exposomics research. When genetic variants that affect exposure levels are available, Mendelian randomization (MR) offers a powerful alternative. MR uses these genetic variants as instrumental variables (IVs) to estimate causal relationships between exposures and health outcomes[19]. MetaboAnalyst 6.0 offers two-sample MR (2SMR) utilizing summary statistics curated from public epidemiological studies (see **Box 4**).

Together, these modules form a comprehensive workflow spanning raw spectra processing, data processing, statistical modeling, and interpretation for LC-MS/MS-based exposomics. The underlying R package, MetaboAnalystR 4.0 is also released to ensure transparency and reproducibility by enabling users to execute all analyses locally [14]. This R package mirrors the web server's functionality while providing greater flexibility and extensibility for advanced users. **Table 1** summarizes the key updates from the previous 2022 Nature Protocol based on MetaboAnalyst 5.0.

Limitations of this protocol and software

Raw spectra processing is computationally intensive. To ensure fair access and platform stability, MetaboAnalyst currently limits uploads to a maximum of 200 LC-MS1 files and 50 MS2 files per session. Researchers with larger data can submit their data in multiple batches and are encouraged to install MetaboAnalystR for local processing [14]. MetaboAnalyst currently does not support processing spectra from GC-MS, which is increasingly used in exposomics for better detection of volatile and semi-volatile molecules [20]. Several powerful tools such as MZmine, MS-DIAL or MS-Finder can be used for this task [21-23]. In addition, environmental chemical annotations are limited, though users can access such information through resources like the US EPA's CompTox Dashboard or ECHA's Chemicals Database [24,25]. The MR implementation is currently restricted to 2SMR with a single exposure, lacking support for multivariable MR [26]. We intend to address these limitations in the next release. Finally, MetaboAnalyst is under continuous development based on user feedback and community requests. We encourage users to post their questions on the user forum (https://omicsforum.ca/) to get timely support, and to visit the MetaboAnalyst website for the latest application notes, tutorials, and other learning resources. Users are also welcome to contact the corresponding authors directly with specific queries.

Overview of the Procedure

Figure 1 depicts the overall workflow for LC-MS/MS-based exposomics data analysis in MetaboAnalyst 6.0. There are four stages:

1) LC-MS/MS spectra processing and compound annotation (Steps 1 - 32);

2) data processing and exploratory data analysis (Steps 33 - 90);

3) dose-response modeling (Steps 91 - 102); and

4) causal analysis and interpretation (Steps 103 - 113).

Each stage corresponds to an independent module in MetaboAnalyst. This design allows researchers to upload data generated from other tools rather than relying entirely on MetaboAnalyst. For instance, many researchers perform raw spectra processing locally and upload the data tables to MetaboAnalyst for comprehensive statistical and functional analysis. In this protocol, Stage 1 is used to process raw spectra data, and the result table generated from Stage 1 can be used as input for Stage 2 and Stage 3. Stage 4 is demonstrated based on a different example.

Applications

MetaboAnalyst 6.0 can be readily applied to exposomics and metabolomics data analysis, supporting the full analytical workflow from raw spectra processing to statistical exploration and biological interpretation. The workflow was recently applied to an electronic waste exposomics study[13]. It successfully detected more than 23,000 MS features, including many trace-level signals, and identified over 600 compounds from the blood samples of e-waste workers. These compounds were further classified into distinct exposome categories to infer their potential sources. Dose response analysis identified many metabolic features with distinct response profiles to different metal exposures, indicating their potential mechanisms and toxicological relevance. MetaboAnalyst has been used for other exposomics studies involving household dust [27] and aquatic environments [28]. As one of the most comprehensive platforms for targeted and untargeted metabolomics, MetaboAnalyst is a core component of our omics tool suite to enable systematic integration of multiple omics layers for exposomics research[29-32].

Experimental Design

Electronic waste (e-waste) generation has increased rapidly worldwide, raising serious environmental and public health concerns due to the release of toxic substances during informal recycling processes. Agbogbloshie in Ghana is one of the largest and most intensively studied e-waste recycling sites globally, where dismantling and open burning of electronic materials are common [33]. Numerous toxic elements, particularly heavy metals, have been detected in the local soil, air, and water, posing significant exposure risks to nearby populations. Our previous exposomics study comprehensively characterized the influence of heavy metal exposure on human health and metabolic profiles[13]. To illustrate the workflow implemented in that study, we included one of the e-waste LC–MS/MS datasets as an example. The dataset was acquired in C18 negative ion mode and contains metadata describing 18 metal concentrations and 15 additional physiological or environmental descriptors of the participants. This example dataset is used to demonstrate Stages 1–3 of the analytical workflow.

MR requires exposures associated with genetic determinants and is therefore unsuitable for exposures primarily driven by socioeconomic factors, such as occupational heavy metal exposure among e-waste workers. To demonstrate the causal analysis procedures, we selected L-isoleucine as the exposure and type 2 diabetes (T2D) as the outcome. L-isoleucine is a branched-chain amino acid (BCAA) crucial for energy metabolism and protein synthesis, with its primary blood source being dietary [34,35]. Moreover, the human gut microbiome is known to influence host plasma concentrations of L-isoleucine and the risk of T2D [36,37]. In this protocol, we use this example to demonstrate how to perform 2SMR to estimate the potential causal relationship between L-isoleucine, a modifiable exposure, and T2D.

MATERIALS

Computer requirements

- Hardware requirements: >4 GB of RAM and a screen resolution of at least 1,080 × 960 is preferred. At least 8 GB available hard drive space is needed to store the raw spectra files
- Browser requirements: MetaboAnalyst 6.0 runs on all modern web browsers. For the best results, we recommend Google Chrome 100+, Firefox 92+, Safari 12+ and Microsoft Edge v93+. JavaScript must be enabled in your browser.
- Internet connection requirements: a fast connection is highly recommended. At least 1 MB per second is required for uploading raw spectra.

Data files

Stages 1 - 3 utilize a subset of an exposomics dataset derived from electronic waste (e-waste) workers [13,33]. LC-MS1 and MS2 spectra were acquired using an ultra-high-performance liquid chromatography (UHPLC) Q-Exactive Orbitrap system. The metadata table contains blood concentration levels of various heavy metals, including cadmium (Cd), lanthanum (La), and lead (Pb), along with age, body mass index (BMI), smoking status, and alcohol consumption. No input file is required for Stage 4.

- Stage 1: Exposomics raw spectra processing and compound annotation:

https://www.xialab.ca/api/download/metaboanalyst/NP_exposome_raw.zip

https://www.xialab.ca/api/download/metaboanalyst/NP_exposome_raw_complete.zip (Optional)

This demo dataset comprises a subset of LC-MS spectra of whole blood samples from 10 e-waste workers and 6 individuals who served as controls. Users can optionally download the complete

LC-MS spectra from 100 e-waste workers and 6 QC samples. Three DDA-MS2 files, acquired from the pooled QC samples, were included for compound identification.

- Stage 2: Data processing and exploratory data analysis

https://www.xialab.ca/api/download/metaboanalyst/ewaste_data_QC.csv

https://www.xialab.ca/api/download/metaboanalyst/ewaste_metadata.csv

This is a subset of LC-MS peak intensity data from the whole blood samples of individuals working in e-waste recycling sites. The "ewaste_data_QC.csv" contains 100 samples including 6 QCs. Its associated metadata is "ewaste_metadata.csv". The primary exposure of interest is "BCd_group" describing categorical blood concentration level of cadmium (low/medium/high).

- Stage 3: Dose response analysis:

https://www.xialab.ca/api/download/metaboanalyst/ewaste_data.csv

https://www.xialab.ca/api/download/metaboanalyst/ewaste_metadata_dose.csv

The "ewaste_data_dose.csv" has gone through data filtering and missing values imputation already, with QCs samples removed. Its associated metadata is "ewaste_metadata_dose.csv". The primary exposure of interest is "BCd" describing the blood cadmium concentrations.

PROCEDURE

Stage 1: LC-MS/MS Spectra Processing and Compound Annotation (Timing 1.5 – 2.5 h)

<CRITICAL> The general workflow for this stage consists of four sequential tasks: 1) spectra preparation and uploading, 2) LC-MS feature detection and alignment, 3) MS2 spectra deconvolution and compound identification, and 4) results exploration.

1. *Starting up.* Visit the MetaboAnalyst home page (https://www.metaboanalyst.ca/) and click 'Click here to start' button to go to the "Module Overview" page. Click the top button "Spectra Processing [LC-MS w/wo MS2]" to enter the module.

Data uploading.

<CRITICAL> Alternatively, you can load the example data directly by selecting the "E-Waste dataset" from the example datasets at the bottom of the page. If you do this, proceed directly to Step 7.

<CRITICAL> LC-MS1 and MS2 spectra files should be centroided first and uploaded separately in open format (see **Box 1**). If MS1 and MS2 spectra were acquired concurrently, please use ProteoWizard [38] to split them into different files based on their MS levels prior to uploading. MS2 spectra files must be labeled as group "MS2" in the metadata file ("sample.txt" in this example) to be recognized for MS2 spectra processing.

2. Unzip the dataset #1 (NP_exposome_raw.zip) into individual files.
3. Click "Select" button to open a File Chooser dialog. Locate and select all 19 files (*.zip) together with the metadata file (sample.txt). Confirm the selection and close the dialog. All files will be displayed in the file upload panel, and a "Upload" button will appear beside the "Select" button.
4. Click "Upload" to start file uploading.
5. When all files are uploaded, make sure the "Select Mode" option is set to "MS1+DDA", and click "Proceed" to continue.

<CRITICAL> This example dataset includes 16 LC-MS spectra files and 3 DDA MS2 files. Choose "MS1+SWATH-DIA" if MS2 spectra were acquired in SWATH-DIA mode.

6. (*Optional*) If you are using the complete dataset, which contains 109 samples, perform similar operations as above to upload all files. Please note that the data processing time is proportional to the size of the dataset; large datasets may have very long processing times.

Troubleshooting

Data integrity check.

7. The page automatically performs data integrity check and displays all uploaded spectra in a table, where each row represents a spectral file. The first five columns (Spectra, Centroid, Size (MB), MS Level, Group) summarize their basic information. If appropriate, include/exclude some of the spectra by using the checkbox in the "Include" column. Click the "Next" button.
    - The "MS Level" column will only appear when MS2 spectra are uploaded.
    - All centroid spectra are labelled as "True" in the "Centroid" column. Any spectra flagged as "False" will be automatically excluded from the processing list. Users can click the "Convert" button to see if it can be fixed using our built-in method [39]. The conversion is slower than ProteoWizard and should only be used as a backup.

Troubleshooting

Parameter customization.

8. Navigate to the "LC-MS/MS Spectra Processing" page to customize parameters.
9. Specify the platform to "UPLC-Q/E" to load the default parameters. Several parameters need to be updated for LC-MS/MS-based exposomics data analysis. The complete raw spectral processing pipeline is described in **Extended Figure 1**.
10. By default, peak detection is based on "centWave-auto" [40] which performs automatic parameter optimization for the centWave algorithm [41]. The algorithm detects peaks based

on the chromatographic profile of individual MS features and performs well for relatively abundant endogenous metabolites. Change the option to "Asari" which identifies peaks based on composite mass tracks across samples[12], with higher sensitivity to trace-level MS features compared to centWave.

11. (optional) Where applicable to your specific dataset, manually adjust the peak alignment and annotation parameters in accordance with the previously established protocol [39]. To view parameter explanations, hover over their associated question marks. For the e-waste dataset, keep these parameters as default.
12. Customize parameters for MS2 spectra processing and compound annotation. A total of nine parameters can be customized. For this exposomics data, set the "Target Peaks" to "All Features", "MS2 Database" to "All Database", and "Target Omics" to "Exposomics". Keep the rest as default.

<CRITICAL> Using the "All Database" option will search the complete reference spectra database for compound identification. When the "Target Omics" is set to "Exposomics", an extra step will be performed for exposome annotation and classification as discussed in Step 29.

Job submission.

13. Click the "Submit Job" button at the bottom. On the confirmation dialog, click "Confirm" to submit your raw spectra job to the processing queue.
    - Each job is allocated 48GB of RAM and 4 CPU cores, with resource orchestration managed by the SLURM (Simple Linux Utility for Resource Management) workload manager.

Job status monitoring.

14. In the “Job Status View” page, you can see that a Job ID has been assigned to this job. Click the “Create Job URL” link to generate a unique URL link.
15. On the pop-up dialog, click the “Copy” button and save the URL to a file. You can now close the browser and come back later using the URL. The job status will be displayed as the progress bar with all the progress details listed in the output box as plain text.

Troubleshooting

16. *Job completion*. When the spectra processing job is complete (i.e., the progress bar reaches 100%), a status message “Everything of this LC-MS/MS dataset has been completed successfully!” will appear at the last line of the text output. Click the “Proceed” button.
17. *(optional) Job cancellation*. To modify job parameters after job submission, first cancel the current job by clicking the “Cancel Job” button at the bottom of the page. You can now modify parameters and submit a new job from the previous page.

LC-MS result exploration.

18. Inspect the results. The result page is divided into two main sections with graphical summaries displayed on the top section and tabular outputs at the bottom section.
19. Interactive 3D PCA visualization. This graphic offers an intuitive summary of all spectra samples and their underlying MS features (**Figure 2A**).
    - By default, the scores plot will be displayed in the main view while the loadings plot as the inset view. Switch between these two views using the double arrow icon on the vertical toolbar at the top left.
    - Rotate the view by clicking and dragging with the mouse or use the scroll wheel to zoom in and out.

- Customize the visualization by adjusting the background, shapes, and other visual elements using the dedicated tool buttons in the top menu bar.

20. Explore other graphical outputs including intensity plot, TIC (total ion chromatogram) plot and BPI (base peak ion) plot. The intensity plot provides a summary of intensities of all detected MS features of individual samples. The TIC and BPI plots show the overall profile of mass spectral signals across all scans.
21. Explore the summary pie charts for metabolome and exposome (**Figure 2B**).
    - The metabolome pie chart summarizes chemical taxonomies of all identified compounds based on the HMDB classification [42].
    - The exposome pie chart summarizes all annotated exposome-related compounds based on the categories from NORMAN Suspect List Exchange database [43]. More details will be discussed on Step 29.
22. Examine the sample or spectra table. This table contains seven columns. The column "Spectra" and "Group" summarize the information of processed spectral files. "Peak No." refers to the total number of peaks detected in the corresponding spectra. "Missing (%)" refers to the percentage of missing features in the acquisition data file. "RT range" and "mz range" refers to the range of retention time and m/z of detected features. The last "View" column allows users to view the TIC of individual spectral file.
23. Explore the feature or peak table. The table contains all detected MS1 features including m/z, retention time, average intensity, coefficient of variation (CV), p-values, false discovery rate (FDR), adduct and isotopes annotations, as well as the putative chemical identity based on MS1 level data. The last "View" column allows users to examine extracted ion chromatogram (EIC) of the corresponding feature.

- The p-values are based on t-tests or ANOVA to offer basic guidance during feature exploration. For comprehensive statistical analysis, please follow the data processing, normalization and linear modeling as described in **Stage 2.**
- If a MS1 feature is associated with MS2 spectra, the "View" dialog will contain two tabs. The first tab shows the EIC and box plot, while the second tab includes the detailed compound identification results from the corresponding MS2 spectra.

24. Check the MS2 results table. This table contains complete MS2-based compound annotation results. More details will be discussed on Step 28.

EIC creation and comparison.

25. Make sure the 3D PCA loading plot is in the main view and double click a data point of interest. A dialog will pop up showing a box plot summary of the feature intensities across different groups. Alternatively, users can locate the corresponding feature from the peak table and click the "View" button.
26. Click any data points on the box plot to view the corresponding EIC on the right side. Users can overlay and visually assess the peak quality based on EICs.
27. On the PCA loading plot, select the "MS2 only" checkbox to highlight all MS1 features associated with MS2 spectral metadata (**Figure 2A**). This streamlines biomarker discovery by integrating comparative abundance visualization with putative chemical identification (**Figures 2C–F**).

Compound annotation.

28. MS2-based compound annotation is summarized in the MS/MS result table tab. The identified MS features are displayed in expandable rows. Click the triangle to expand a row for details. There are seven columns in this table. Compound, Formula, InChIKey and

Database indicate the corresponding chemical information of the annotated chemical candidates of the feature. Matching score indicates the similarity of the experimental MS2 spectra to the reference spectra, with 0 means no matching at all, while 100 indicates a perfect match.

- The MS2 spectra matching patterns can be viewed as mirror plots in the "View" column. A mirror plot is an interactive graph consisting of top and bottom sections. The top part shows experimental spectra in blue, while the bottom part shows the reference spectra in red. All matched fragments are labelled with red diamond. Hover mouse over a fragment of interest to view its details including m/z, relative intensity and potential formula of the fragment. The matching results, together with the mirror plot can be downloaded by using the icons in the mirror plot dialog (**Figure 2F**).
- When the "Target Omics" option has been set to "Exposomics", an extra column, "Classification" will be displayed to show detailed exposome classifications as described in the next step.

Troubleshooting

29. *Exploration of the annotation and classification results.* To better characterize the chemical exposures captured within biological samples, exposome-specific annotation and classification have been integrated into the MS/MS results table. For compounds successfully matched to known exposome databases, a "Details" link is provided in the final column. This feature enables users to investigate the compound's classification hierarchy, potential environmental sources, and biological relevance.

- Exposome classification is based on a curated ontology that groups chemicals into 21 distinct categories reflecting their primary sources or exposure routes (**Table 2)**. These categories encompass a broad spectrum of environmental and anthropogenic sources, including environmental chemicals, personal care products, pesticides, foods, air pollutants, and drugs, among others.

Troubleshooting

30. *Result download.* Click the "Download" link on the left navigation tree to view all figures and tables generated during the session. There are five important tables:
    - *metaboanalyst_input.csv*. This table contains complete MS features with m/z and retention time (rt) labelled as feature name.
    - *metaboanalyst_input_compounds.csv (optional)*. This table contains MS features that have been identified as compounds based on MS2 spectra. Feature names are derived from the compounds with the highest matching scores. The minimum acceptable top score threshold is 50 out of 100. If multiple peaks are assigned to the same compound, only the one with the highest average abundance is retained.
    - *metaboanalyst_input_clean.csv.* This table contains parent ions of MS features, with all related ions such as adducts, isotopes, and isomers removed. Features identified based on MS2 data are labeled with their corresponding compound names, as provided in the *metaboanalyst_input_compound.csv* file. Features with annotated empirical formulas are labeled with their respective formula, m/z, and rt. The remaining parent ions are labeled based on their m/z and rt values. This table essentially combines information from the previous two tables while removing all redundant MS features.

- *Compound_msn_results*.csv This file provides a summary of all identified compounds, along with the corresponding MS1 feature information (including m/z and rt ranges). For each feature, the top five scoring compound matches are included. The table contains detailed chemical identification information, including compound names, InChIKeys, database sources, molecular formulas, and similarity scores. The scores range from 0 to 100, where 0 indicates no match and 100 represents a perfect match.
- *metaboanalyst_feature_reference.csv.* This is a comprehensive reference table that includes complete information such as MS1 features, adducts, isotopes, MS2-based compound identification, metabolome and exposome classifications. It serves as a master reference table and is intended for reference purposes only. It cannot be re-uploaded into MetaboAnalyst.

31. *(optional) MS2 annotation for a single MS2 spectrum.* Users can perform compound annotation for a single MS2 spectrum. To do this:
    - Start from MetaboAnalyst homepage and click "Peak Annotation [MS2-DDA/DIA]" to access the module.
    - Copy and paste the spectrum into the input box,
    - Enter the precursor ion mass and adjust the required search parameters, then click "Submit" to query the database for potential candidates
32. *(optional) MS2 annotation for multiple MS2 spectra.* Users can annotate multiple MS2 spectra using the same module. On the data upload page, select the second tab, "Multiple Tandem Spectra", to upload a .mgf or .msp file. After setting up all parameters, click

"Submit". Following the data integrity check, a job will be submitted for database searching. All potential chemical candidates will be displayed once the search is complete.

Troubleshooting

Stage 2: Data Processing and Exploratory Data Analysis (Timing 30 min ~ 1h)

<CRITICAL>This stage consists of four sequential tasks: 1) data upload and integrity check, 2) data processing consisting of data filtering, missing value imputation, and normalization, 3) data overview and exploratory data analysis, and 4) advanced statistical, functional, and predictive analysis. These tasks are divided into two sections: the first section focuses on quality checking and data processing, while the second section demonstrates how to identify main patterns, significant features and functions. This stage provides detailed instructions for processing the result table generated from Stage 1.

Data processing and quality check

33. *Starting up.* Go to the MetaboAnalyst homepage (https://www.metaboanalyst.ca/) and select the "Click here to start" button to enter the "Modules Overview" page. Click on "Statistical Analysis [metadata table]" to enter this module.
34. *Data upload.* Set data type to "Peak intensities", keep the default options for study design ("Multiple factors / covariates") and data format ("Samples in columns"). Use the corresponding File Choosers to select data file ("ewaste_data_QC.csv") and metadata file ("ewaste_metadata.csv"). Click "Submit".
    - This data file contains a total of 100 samples from e-waste workers including 6 QC samples and a total of 30,395 peaks. It was generated from **Stage** 1 ("metaboanalyst_input.csv") using the complete raw spectral files.
    - The primary metadata of interest should be in the first column immediately after the sample names in the metadata file. Due to historical reasons, MetaboAnalyst requires the primary factor to be categorical. If your primary metadata is a

continuous variable, add an extra column by converting it into discrete categories. For instance, the "BCd_group" was created by converting the main exposure of interest ("BCd") into three equal sized bins (low, medium and high). Please note that the main purpose here is to meet the input requirement to pass data integrity check, users should select the "BCd" for analysis whenever continuous responses are supported by the method such as linear modelling, random forests, etc.

<CRITICAL> In MetaboAnalyst, QC and blank samples must begin with the prefixes "QC_" and "BLANK_" respectively in the metabolomics/exposomics data file. The metadata table should not include rows for QC or blank samples since they lack other metadata information.

Troubleshooting

35. *Data integrity check.* This page displays detailed text summary of the uploaded data. Spend some time carefully reading the text summary to confirm that the correct parameters were specified in the previous step. Pay attention to any highlighted texts which indicate potential issues detected in the data. Click the "Proceed" button.
    - MetaboAnalyst automatically performs a series of checks and displays a summary of the data. Particular attention should be paid to sample naming conventions, class labels, and the data format. When missing values are present, users can further check whether the missing percentages differ significantly among groups based on Kruskal-Wallis tests. This example dataset contains a high proportion (54.3%) of missing values with a p-value of 0.0467.
    - Since our dataset contains QC replicates, the workflow automatically computes relative standard deviation (RSD) to help assess technical precision. In this case, it

reports a median RSD of 15% and around 74.7% features have less than 30% RSD, indicating a good instrument stability.

36. *Metadata integrity check.* MetaboAnalyst automatically infers whether each column in the metadata table contains categorical or continuous values. They may not always be accurate. The "Metadata Check" page allows users to review all metadata to make sure that they are accurate before further analysis. Carefully inspect the assigned data type for each metadata factor and make corrections when necessary. Click the "Proceed" button.
    - Unlike the data table, no missing values are allowed in the metadata table. You can exclude the metadata column or manually enter the missing information by clicking the corresponding "Edit" link.

Troubleshooting

Data filtering.

<CRITICAL> This step aims to exclude features that are unlikely to contribute to downstream analysis while improving statistical power and computing efficiencies. MetaboAnalyst currently offers four types of filters: low-quality filter, low-repeatability filter, low-variance filter and low-abundance filter (**Box 2**). Since the blank subtraction was performed during raw spectra processing, no blank samples are included in this data.

<CRITICAL> Based on the recommended practices in omics data processing [44] and analytical characteristics in LC-MS based metabolomics [45,46], MetaboAnalyst permits a maximum of 10,000 features to pass the data filtering step for downstream analysis. The number will be enforced using the low-variance filter based on interquartile range (IQR). Users are strongly recommended to apply proper data filter(s) to control the total feature number.

37. Select “Enable missing-value exclusion” under “low-quality filter” and leave the threshold value at 50%. When missing value distributions are significantly different between groups, users may consider selecting the “Group-wise threshold” option, which will retain features whenever it meets the threshold in at least one group. This is useful when the groups reflect true experimental groups. We use the default global threshold since the “BCd_group” was created artificially.
38. Apply “low-repeatability filter” and keep QC RSD cutoff as default (20%). The low-repeatability filter removes features based on their variance in QC replicates. You can examine the overall feature-level RSD distribution by clicking on “View” on the right of the row. The violin plot confirms that most of the features in QC samples have RSD percent less than 25%, indicating good analytical precision.
39. Turn off the low-variance filter and low-abundance filter by adjusting their respective sliders to 0%.
40. Click “Submit”, a message will show up stating 14,863 features were excluded based on low-quality filter, and 3,488 features were excluded based on low-repeatability filter. A total of 7,251 features remain.
41. Click the “Proceed” button.

Missing value imputation.

42. The “Missing value imputation” page allows users to select a method for estimating missing values based on their diagnostic plots. Due to the data filtering performed in the previous step, the percentage of missing values has now dropped to 11.6%, a substantial reduction from 54.3% in the original data. The missing value distribution plot shows the distribution of missing values as well as the averages of non-missing values across primary

metadata groups. The missing value heatmap shows missing features for each sample. Missing values in LC-MS data are usually left-censored due to low abundance.

43. Select the "quantile regression (QRILC)" option under "Left-censored data estimation".

<CRITICAL> An alternative option is to substitute each missing value with limit of detection (LOD), but this could artificially narrow the feature's distribution. The Quantile Regression Imputation of Left-Censored data (QRILC) algorithm introduces stochastic variability around low-intensity values to better preserve their original variance structure [47].

44. Click "Submit". After the process is finished, click "Proceed" button.
45. Review the results. Pronounced differences in missing-value distribution or average signal intensity across sample groups could suggest batch-related effects. It is recommended to apply sample-level normalization method to correct these systematic differences.

Data normalization.

46. Data normalization aims to reduce heterogeneities due to variations in sample preparation, instrument response, or biological variability. Start by selecting "Normalization by median" under "Sample Normalization".

<CRITICAL> "Normalization by median" is applied because we have observed differences in abundance distribution across groups in the previous step. Probabilistic quotient normalization (PQR) normalization is another method commonly used in metabolomics to address systematic differences during sample preparations.

47. Select the "Variance stabilizing normalization" (VSN) option under "Data Transformation" and click the "Normalize" button.

<CRITICAL> Log transformation is not recommended for this example due to the presence of a large number of left-censored data values, and applying log will amplify their influence. VSN is

better suited here because it models the mean-variance relationship across the entire dynamic range and applies a transformation that flattens the heteroscedasticity - variance that depends on signal intensity [48] (see **Box 2**).

48. *Assessing normalization effect.* Click the "View Result" button. The dialog compares the overall data distribution before and after the normalization in the forms of density plots and boxplots at the feature and sample levels, respectively. Close the dialog and click the "Proceed" button.
    - It's important to keep in mind that the goal of normalization is to address systematic difference to make samples/features more comparable. Achieving a multivariate normal distribution is neither required in downstream analysis nor practically feasible for high-dimensional omics data. QC clustering, biologically meaningful separations, or classification performance are better indicator of normalization effects.
49. *Quality check with PCA.* On the method selection page, click on "Interactive PCA Visualization". A pairwise PCA scores plot is shown on the new page. QC samples (light blue) cluster tightly in the top components. No clear separation is seen based on "BCd_group" - high (red), medium (blue), and low (green).
    - PCA visualization can provide important insights into data. Users can visually assess the technical variability based on QC samples. If QC replicates appear scattered within the main sample cluster, it may indicate issues with the LC-MS instrument during spectral data collection. PCA can also help evaluate effectiveness of normalization procedures (i.e. improved group separation) as proper normalization often improves biological signals while reducing random variance.

- Exposures such as blood cadmium (BCd) typically vary across individuals, and their metabolic effects are often subtle compared with the total inter-individual variance driven by age, BMI, etc. Because PCA is an unsupervised method that maximizes global variance rather than exposure-related variance, these weaker BCd-associated signals can be masked by larger, unrelated sources of variation.

Excluding QC samples.

<CRITICAL> After evaluating QC clustering using PCA, it is important to remove the QC samples before proceeding to downstream analyses. This step is necessary as removing samples alters the global intensity distribution and variance structure, which can influence the outcome of normalization

50. Click on "Data Editor" in the left navigation tree.
51. In the Data Editor page, type "QC" in the search bar on top of the left panel.
52. Select QC samples 1 to 6, then click the right-pointing arrow to move them to the exclusion list.
53. Once selected, click "Submit" to finalize the removal. You will be redirected to the Normalization page, select "Normalization by median" followed by "Variance stabilizing normalization" and click on "Normalize".

Identification of main patterns, significant features and functions

54. *Exploratory data analysis with PCA.* Click on "iPCA" on the navigation tree. Without the influence of QC samples, the PCA plot more accurately reflects biological variations. Although the sample separation does not seem to be evident upon visual inspection,

PERMANOVA results indicate marginal improvement in p-values between top components (**Extended Figure 2**).

- The p-values displayed in the upper-right half of the panels are based on PERMANOVA tests on the sample scores for each PC pair. For discrete metadata, it assesses group centroid separation. For continuous metadata, it evaluates whether the data exhibits a gradient aligned with the metadata. The p-values can vary slightly between runs due to random permutation.

<CRITICAL> PERMANOVA requires sufficient sample size for reliable permutation testing. At least 15-20 samples per group is recommended based on our empirical experience.

Interactive visualization.

55. Click "Synchronized PCA 3D" tab. The default PCA view shows synchronized scores plot (main view) and loadings plot (inset view).
56. Click the ellipse icon (the third one from the top of the vertical tool bar) to apply confidence ellipsoids around different groups.
57. Click the double-arrow icon at the top of the tool bar to put loadings as the main view.
58. Double-click any node with a high loading value (e.g. those at the peripheral) to view the distribution of its intensity profile.

Troubleshooting

Metadata visualization.

59. Click the "Metadata Visualization" link to compute relationships between metadata variables and visually explore them in an interactive hierarchical clustering heatmap.
60. To view the pairwise correlations between the metadata factors, click on the "Correlation Heatmap" tab. Each cell shows the correlation coefficient between that pair of metadata.

61. Examine the data. For the worked example, the linear correlation between "BCd" (continuous) and "BCd_group" (discrete) is not very high. In this case, this is because the binning process simplifies the continuous values into three discrete levels, resulting in information loss. The negative correlation between them is the result from how the "BCd_group" categories were numerically encoded based on their alphabetic order (e.g., high 0, low 1, and medium 2). If this happens, manually set the order using the "Metadata Editor" [39].
62. By default, the correlation coefficients are based on Pearson correlation, which captures both direct and indirect relationships between variables. The recently added partial correlation can help uncover direct relationships by adjusting for the influence of all other variables in the dataset. Switch to "Pearson (partial)" and click on "Update"**.** As shown in **Figure 3**, the weak positive correlation between BCd and BMI (r = 0.055) has become negative (r = –0.157) based on partial correlation.

<CRITICAL> Many environmental exposures and clinical outcomes are related to each other. Identifying such dependencies helps make informed decisions about variable selection and covariate adjustment in the next steps. Notice that "BCd" and "Smoking" demonstrate a notable positive correlation. This correlation is likely due to the accumulation of Cadmium from smoking [49]. Knowing this, it would be important to adjust for "Smoking" as covariate when performing linear modeling on BCd. Similarly, using both "BCd" and "Smoking" as predictors may lead to multicollinearity, potentially inflating variance estimates and reducing the stability of the model.

Linear modeling with covariate adjustment.

63. Click the “Linear Model” node in the left navigation tree. The approach supports both discrete and continuous response variables. We aim to identify metabolic features associated with blood cadmium (BCd) exposure.
64. Select “BCd” instead of its categorical version “BCd_group” as the primary metadata.
65. Include “Age”, “BMI”, “Smoking”, and “Alcohol” as covariates.
66. Select “Raw” option next to “P-value cutoff” and click “Submit”.
67. Examine the result (**Figure 4**). The scatter plot illustrates the relationship between $-\log_{10}$(p-values) calculated using unadjusted models (x-axis) and those adjusted for covariates (y-axis). Each point represents a feature, allowing users to assess how adjustment for confounders (e.g., age, BMI, smoking, alcohol) affects its association with the BCd levels. Features that remain significant after adjustment, become significant only after adjustment, or lose significance when adjusted are distinguished in different colors. A total of 453 features passed the significance threshold.
68. Double-click any features on the top-right corner. It will open a dialog displaying violin plots (for categorical metadata) or regression plots (for continuous metadata) as shown on the right side of **Figure 4**. Switching the primary metadata to “BCd” to examine how the feature intensity varies across exposure levels.

(Optional) Functional interpretation of LC-MS peaks.

<CRITICAL> This Steps 69-76 is based on a different module - do this step only after you have finished the current stage.

<CRITICAL> Although the “Functional Analysis” module accepts a LC-MS peak intensity table as input, its built-in statistical analysis does not support covariate adjustment. The

following steps show how to prepare a ranked peak list obtained from other methods for more accurate functional analysis.

69. The results from covariate adjusted linear modeling can be used to perform functional analysis. Click "Download" to download the result table ("covariate_result.csv").
70. Edit the file to keep only the feature names, raw p-values, and t-scores.
71. Rename these three columns as "m.z", "p.value", "t.score" respectively.
72. You are now ready to upload this dataset to "Functional Analysis [LC-MS]" module. Make sure the first tab "A peak list profile" is selected. Update "Retention time" to "Yes - Seconds" and keep the rest of parameters as default.
73. Perform the data processing and method selection steps using the default parameters.
74. Once you reach the "Mummichog Pathway Activity Profile" page, select "Enrichment Network". The enrichment network reveals significant metabolic pathways and their underlying compounds.
75. In enrichment network, two pathways are connected if the overlaps between their members are above certain threshold. Adjust the threshold by using the "Overlap threshold" dialog in the "Edge" drop-down menu on the top menu bar.
76. Expand the pathway nodes using the "Bipartite View" to show their underlying metabolites (**Figure 5**). The most significant pathway associated with exposure to cadmium is "Tyrosine Metabolism", which is consistent with literature [50]

Interactive heatmap visualization

77. Click the 'Heatmap2' node in the left navigation tree. Interactive heatmap is a powerful approach for visualizing high-dimensional omics data. It displays the relative abundance

of features (rows) across individual samples (columns), with color gradients indicating intensity levels. This allows for quick identification of feature clusters or sample clusters regarding different exposures. The default heatmaps show the first four columns of metadata.

78. *Heatmap customization.* Users can further customize the heatmap to observe exposure-related trends or subgroup differences. Various options are available for data scaling, distance metrics, and clustering algorithms. Update the "Metadata in annotation" option to include "BCd", "Age", "BMI", "Smoking", and "Alcohol". Set "Sample arrangement" to "BCd" so that samples are grouped by their BCd exposure levels.
79. Click "Submit". We don't see any clusters with clear monotonic responses to BCd levels. Note there is a cluster of features with similar intensity patterns at the bottom of the heatmap.

(Optional) Functional analysis LC-MS peak clusters.

<CRITICAL> This Steps 80 - 84 are based on a different module - do this step only after you have finished the current stage.

80. We can use the "Functional Analysis [LC-MS]" module to explore the functional annotation of the feature clusters. Click on "Download" located at the bottom of navigation tree.
81. Download the file "data_processed_input.csv", which includes an additional metadata row indicating the class assignment for each sample ("BCd_group" in this case).

82. On the data upload page, click the second tab “A peak intensity table”. For the parameters, set “Retention Time” to seconds and “Samples in columns” under “Data Format”. Keep the rest of the parameters and upload the “data_processed_input.csv” file.
83. On the “Parameter Setting” page, select “Heatmaps” option under “Visual Analytics” option.
84. The new page will show an interactive heatmap that supports functional enrichment analysis. Drag-and-select regions of interest on the heatmap at the left-side panel (“Overview”) and perform functional enrichment analysis on the selected region/cluster (“Focus View”). Please refer to our previous protocol for more details.[39]

Classification analysis using Random Forest.

85. Click the “Random Forest” node in the left navigation tree. “BCd_group” should be already set as the outcome variable of interest. Users may optionally include other metadata variables as predictors. In this case, we can add “Age” as an additional predictor to examine whether this helps improve cadmium exposure classification.
86. Leave the remaining parameters as default and click “Update”. The confusion matrix shows that the overall OOB error is approximately 0.5, with the “low” group exhibiting the lowest misclassification rate at around 0.35. This poor performance indicates that the algorithm failed to learn effective patterns for distinguishing between the three groups.
    - Keep in mind that the three groups were created through equal binning, a process that inherently causes a loss of information and discriminative power.
87. (optional) The results may be slightly different due to the randomness of the algorithm, especially when the sample size is small. Choose the option “Use a constant (123456)” under “Randomness” to turn off this feature.

Regression analysis using Random Forest.

88. Random Forest can also perform regression analysis when the response variable is continuous. Set “Primary metadata” to “BCd” and click “Update”. The resulting plots show the OOB Mean Squared Error (MSE) stabilizes after 200 trees. The scatter plot (bottom left), which compares predicted versus observed values, demonstrates that the predicted values are largely confined towards the center of the distribution. This is a known characteristic of Random Forests regression, resulting from the averaging of predictions from many individual decision trees. In contrast, the scatter plot (bottom right), illustrating the correlation between observed and predicted ranks, shows a better relationship with a Spearman correlation coefficient (~ 0.46), indicating moderate predictive power.
89. Explore the top ranked features in the dot plot under the “Var. Importance” tab and access the full ranked table by clicking the table icon. Feature importance ranking is based on the percent increase in MSE, where features that cause a larger increase in error when permuted are deemed more influential.
90. *Result download*. Click the “Download” node from the navigation tree on the left panel. On the download page, you can download the result files including tables and graphical outputs. Click “Exit” to finalize the analysis.

Stage 3: Dose Response Analysis (Timing ~ 30 min)

<CRITICAL> After data integrity check and processing, the workflow for dose response analysis consists of three sequential tasks: 1) identification of omics features with potential dose-dependent responses; 2) curve fitting using a set of linear and nonlinear models; and 3) result exploration. These steps are presented in detail below.

91. *Starting up*. Go to the MetaboAnalyst homepage (https://www.metaboanalyst.ca/) and select the "Click here to start" button to enter the modules overview page. Click the "Dose Response Analysis" button to enter this module.
92. *Data uploading.* Upload the feature abundance table (ewaste_data.csv) and a metadata table (ewaste_metadata_dose.csv). Make sure the data type is set to "Peak intensities", and dose type to "Continuous Dosing". This dataset is also available under the "Try our example data" section. Select the "E-waste" example and click "Submit".
    - The "Dose Type" setting should match your study design: choose "Repeated Dosing" for controlled experiments with discrete dosage groups or time points, or "Continuous Exposure" for observational data with varying concentration measurements. The first column in the metadata table is treated as the primary exposure variable by default.
93. *Data processing.* These data were generated from Stage 2 after data filtering and missing value imputation. Do not perform any data filtering. Normalize using "Normalization by median" and "Variance stabilizing normalization" as described in Steps 46-47.
94. *Feature selection.* A key step in dose response analysis is to identify features with potential dose-dependent behavior. MetaboAnalyst uses linear regression with covariate adjustment to evaluate the association between omics features and the exposure variable [51]. The

analysis will output p-values along with regression coefficients which indicate the strength and direction of the association. For exposomics data, it is important to adjust for covariates to help better estimate the exposure effects of interest. Add “Age”, “BMI”, “Alcohol” and “Smoking” under “Covariates (control for)” option, uncheck “Adjusted p-value (FDR)” check box, and turn off regression coefficient to 0.0. Click “Submit”.

- This step mainly serves as a gentle data pre-filter step to select the features that are likely to exhibit dose-dependent behavior before proceeding to the computationally intensive curve fitting step. Using the default thresholds will lead to very few features for downstream exploratory analysis in this case.

Troubleshooting

95. After the calculations are complete, a scatter diagram will appear displaying the features with their corresponding regression coefficient (x-axis) along with their log (p-values) in y-axis. A total of 453 features has passed the specified thresholds. Click the table icon located on the top right corner to view details. When you have completed exploration, click the “Proceed” button.
96. *Curve fitting.* This step currently offers 17 statistical models for continuous exposure (**Extended Figure 3A**) including variants of logistic (L), log-logistic (LL), Weibull (W), Michaelis-Menten (MM), Brain-Cousens (BC) and Aranda-Ordaz (AR) models[52]. Each feature is independently fit to the selected set of candidate models. A goodness-of-fit filter based on Neill’s lack-of-fit test is applied to remove models with poor fitting. Among the remaining models, the best one is selected using the Akaike Information Criterion (AIC). The selected model is then used to compute the BMD for that feature, along with lower (BMDl) and upper (BMDu) confidence bounds. Leave the parameter as default and click

"Submit". The process may take a while to complete depending on the number of models and features.

- Note that Neill's lack-of-fit test operates differently than traditional hypothesis tests: here, a high p-value signifies a good model fit, whereas a low p-value suggests the model does not adequately represent the data. AIC balances model complexity and goodness-of-fit. Lower AIC values indicate better model fit.

<CRITICAL> Avoid selecting more than 5 models, which could dramatically increase computational time without guaranteed improvements in model quality. After performing curve fitting using the default recommended models, consider refining your selection based on model performance summaries or prior toxicological knowledge.

Troubleshooting

97. *Review model fitting results*. After the computing is finished, you should see a total of 183 features with converged BMD. A bar plot is displayed summarizing the number of times each statistical model was selected as the best-fitting model across all features. This visualization provides a quick overview of which models performed best overall and how often they yielded valid BMDs. Each bar is divided into two segments, with blue for the number of features with BMD calculated, and gray for those a valid BMD could not be derived. Click the "Proceed" button.
    - To ensure proper model fitting, the exposure variable must cover a biologically relevant and sufficiently broad range to induce detectable metabolic responses.
98. *Examine the overall distribution and characteristics of BMD values* (**Figure 6A**). The vertical lines indicate estimated thresholds of system-level metabolic shift based on three

commonly used criteria: the 20th ranked feature (feat.20), the 10th percentile of all BMDs (feat.10th), and the mode of the distribution (mode).

- This approach is adapted from the concept of transcriptomic point-of-departure (tPOD), an approach used in toxicogenomics field to define the dose at which concerted molecular changes occur. While tPODs are typically derived from controlled animal studies, these adapted estimates offer a practical means of approximating system-level sensitivity using real-world exposomics data.

99. *Exploring the detailed result table*. Use the "View" button in each row to visually inspect the fitted curve for individual features (**Extended Figure 3B**). **Figure 6B** shows an example curve fitting model of the feature "253.0182__210.05". Note this fitted curve only captures the major trend of the response around the common exposure levels, illustrating the critical limitations in exposomics research as discussed in **Box 3**.

100. Cross reference this data with the *compound_msn_results.csv* table generated from **Stage 1**. In this example feature "253.0182__210.05" has been matched with *menadione sodium bisulfite*, a synthetic precursor of vitamin K. Although there is no direct link with cadmium in human studies, several studies show menadione sodium bisulfite can mitigate Cd toxicity in plant [53,54].

101. (optional) *Exploring the effects of Age.* Age is an important continuous variable that can influence metabolic profiles. Use "Age" as the primary exposure variable for dose-response modeling. A total of 850 features passed the feature selection step (raw p-value threshold 0.05 while adjusting for BMI, Smoking and Alcohol) and 462 were successfully fitted with dose-response models. The high convergence rate highlights that many

metabolites undergo age-associated changes. For example, glutathione and pyruvic acid displayed interesting trends, consistent with literature [55,56].

102. Download the complete results table ("curvefit_detailed_table.csv") by clicking on "Download" button on the top. Click the "Download" node to visit the Download page and download all text and graphics generated during the analysis session. Click "Exit" to finalize the analysis.

Stage 4: Causal analysis and interpretation (Timing 20 ~ 30 min)

<CRITICAL> Two-sample MR (2SMR) is conducted using summary-level genetic association data from independent cohorts. The general workflow for conducting 2SMR analysis in MetaboAnalyst consists of three tasks: 1) identify candidate SNPs; 2) perform SNP filtering and harmonization; 3) conduct MR tests and interpret results.

103. *Starting up*. Go to the MetaboAnalyst homepage (https://www.metaboanalyst.ca/) and click the "Click here to start" button to enter the "Modules Overview" page. Click the "Causal Analysis [Mendelian randomization]" button to enter this module.

104. *Selecting an exposure.* Type "L-isoleucine" in the top search box to filter the list of available metabolites. Click the item to select.

- SNP-exposure association summary statistics (beta coefficients, standard errors, p-values) were obtained from our internal database, curated from 65 published metabolomics genome-wide association studies (mGWAS) [57].

Troubleshooting

105. *Selecting an outcome.* Type "type 2 diabetes" into the search box at the bottom to view the available outcome datasets. In this example, we choose "finn-b-E4_DM2". Click "Proceed" button.

- The SNP-outcome association summary statistics were sourced from the IEU OpenGWAS database [58]. The "finn-b-E4_DM2" data comes from a T2D GWAS conducted on a large, homogeneous population of European ancestry with comprehensive registry data.

106. *Examining candidate SNPs*. The SNPs extracted from the two databases are displayed in a table along with their nearest genes, p-values associated with exposure or outcome, population, and biofluid information. Spend some time examining the main sample metadata such as studies and populations

- Independence of samples: The two samples (e.g. exposure study and outcome study) should be entirely independent (non-overlapping) to avoid bias towards the confounded observational estimate.
- Population similarity: The two samples should come from the same underlying population (e.g., similar ancestry, age distribution, sex ratio) to ensure that the genetic effects on the exposure are generalizable and consistent across both datasets. Significant population differences can introduce bias.
- The "Advanced Filter" above the table allows users to further include or exclude SNPs based on specific p-values or other criteria.

107. *SNP filtering and harmonization.* MetaboAnalyst provides comprehensive options to help ensure that the selected SNPs are valid IVs. These options are briefly described below.

- *Linkage Disequilibrium (LD) clumping* is a process used to select independent genetic variants (i.e. SNPs) for MR. By removing variants in high LD (highly correlated), this process prevents the inclusion of redundant SNPs that tag the same genetic signal, which would otherwise lead to biased or inflated causal estimates.
- *Using proxy SNP*. When a SNP from the exposure dataset is not found in the outcome GWAS (e.g., due to different genotyping arrays), a proxy SNP that is in high LD with the missing SNP can be used instead. This helps ensure continuity of analysis even

when exact SNP matches are not available across datasets. However, using proxies may increase the uncertainty of estimates compared to using the original SNPs which need to be applied with caution.

- *Allele harmonization* ensures that the genetic effect estimates for the exposure and outcome are aligned to the same reference allele. This step is crucial for 2SMR because mismatched allele coding can lead to incorrect interpretation of the direction of effect. Harmonization corrects such discrepancies by flipping effect alleles, when necessary.
- *Remove pleiotropic SNPs.* This step excludes SNPs associated with multiple exposures to mitigate horizontal pleiotropy and confounding, thereby enhancing the specificity and validity of the causal inference.
- *Steiger filtering* assesses the direction of causality by comparing how strongly genetic variants explain the exposure versus the outcome [59]. It removes SNPs that are more strongly associated with the outcome than the exposure, which would suggest reverse causality.

In this case, we apply the following settings: perform LD clumping to prune SNPs, do not use proxy variants, harmonize alleles based on allele frequency information, and only include SNPs that pass the Steiger filtering. Click "Submit".

Troubleshooting

108. The result table shows that 4 SNPs were retained for MR analysis. Click "Proceed".

- Note that all four SNPs show significantly stronger associations with the exposure than the outcome (**Extended Figure 4**), suggesting that the genetic instrument is unlikely to directly influence the outcome through pleiotropic pathways independent of the exposure (L-Isoleucine).

109. *MR method selections.* Users can select from multiple MR methods to estimate causal relationships. Because each estimator relies on specific assumptions, the choice of test should be informed by its robustness against pleiotropy, outlier heterogeneity, or instrumental strength. These methods are briefly described below.

- *Wald ratio* is the simplest approach, suitable when only one strong genetic variant is available. It cannot account for pleiotropy or heterogeneity.
- *Inverse variance weighted (IVW) methods*, including IVW-MRE (multiplicative random effects), IVW-FE (fixed effects), and IVW radial, are commonly used when multiple SNPs are available. IVW-FE assumes all SNPs are valid instruments, while IVW-MRE accounts for heterogeneity across SNPs, and IVW-radial allows outlier detection and visual diagnostics.
- *MR-Egger* extends IVW by estimating an intercept that captures unbalanced pleiotropy, useful for identifying directional pleiotropy but sensitive to outliers [60].
- *Median-based estimators*, such as Simple median and Weighted median, are robust to invalid instruments. They provide valid estimates even when up to 50% of the SNPs are invalid.
- *Mode-based estimators* including Simple mode, Weighted mode, and their NOME (No Measurement Error) variants assume that the largest cluster of instruments with similar effects are valid and are helpful when pleiotropy is widespread but consistent across a subset of instruments.
- *Maximum likelihood estimator* assumes a joint normal distribution for SNP effects and can improve efficiency under ideal conditions but is sensitive to model misspecification.

- *Unweighted regression* and *Sign concordance test* are more exploratory tools: the former treats all SNPs equally without weighting by variance, and the latter tests whether the direction of SNP-exposure and SNP-outcome effects align more than expected by chance.

In this example, we selected six complementary methods including two IVW-based approaches (fixed effects and multiplicative random effects) for high efficiency under no pleiotropy, one robust estimator (weighted median) that tolerates invalid instruments, two mode-based estimators (simple and weighted mode) for identifying consistent causal effects, and MR-Egger, which accounts for directional pleiotropy. Click "Proceed".

110. *Exploring MR result table.* The result of causal effect estimates across the selected methods are summarized in a table including the number of SNPs used, effect sizes, standard errors, p-values, and diagnostic statistics for heterogeneity and horizontal pleiotropy. The analysis demonstrated a consistent positive causal effect of L-isoleucine on T2D risk across multiple methods including IVW, MR-Egger, and weighted median. Overall, this MR results support the hypothesis that elevated L-isoleucine level contributes to the pathogenesis of T2D.

111. *Exploring MR graphical outputs.* MetaboAnalyst provides multiple plots to help researchers assess the reliability of the MR findings and to interpret the results:

- *Forest plot* summarizes the causal effect estimates for both individual SNPs and pooled MR methods. This visualization facilitates a direct comparison of effect sizes and their corresponding confidence intervals across different analytical approaches, helping to identify potential outliers or inconsistencies (**Figure 7A**). Specifically, our results indicate that SNPs like rs2941456 (SLC16A11) and rs1420601 (near BCAT1), which

are functionally linked to BCAA metabolism[47], are associated with an increased risk of T2D. These findings reinforce the argument that L-isoleucine levels may causally contribute to heightened T2D risk, potentially via dietary or gut microbiome interactions.

- *Scatter plot* illustrates the association between SNP-specific effects on the exposure (L-isoleucine; x-axis) and the outcome (T2D; y-axis). Each point represents a genetic variant, with the horizontal and vertical error bars denoting the standard errors of the respective effect estimates (**Figure 7B**). Overlaid on this plot are regression lines, each representing the overall causal effect estimated by a specific MR method, with its slope indicating the causal association. The observation that these SNPs exhibit small but consistently directional effects, reinforces the conclusion that elevated L-isoleucine levels causally increase T2D risk.
- *Funnel plot* serves to assess potential directional pleiotropy or bias. A symmetrical distribution of SNPs around the causal estimate suggests low pleiotropic distortion. The data points are approximately symmetric in our plot, indicating that pleiotropy is likely minimal. This is consistent with our early-stage SNP harmonization, which included Steiger filtering to ensure all instruments primarily influenced the exposure rather than the outcome
- *Sensitivity analysis:* All individual estimates are closely aligned with the overall estimate (red line), and their confidence intervals largely overlap, indicating that the association between L-isoleucine and T2D is not unduly driven by any single SNP.

112. *Exploring potential mechanistic links:* The "Literature Evidence" tab offers potential mechanistic connections linking L-isoleucine to the development of the T2D

(**Extended Figure 5A**). The curated pathways from published studies reveal that L-isoleucine is mechanistically linked to T2D through multiple intermediate molecules and processes. For example, L-isoleucine may influence serine levels, which in turn modulates *IRS1* (insulin receptor substrate 1), a key regulator of insulin signaling implicated in T2D. Other pathways suggest potential immune-metabolic roles via defensins and infection susceptibility [61-63]. The "Variant Effect (AlphaGenome)" tab shows the predicted functional consequences of each instrumental SNPs based on AlphaGenome[64]. In the current example, all four SNPs are predicted to alter the expression or regulation of multiple genes (**Extended Figure 5B**).

- Literature mining is based on triangulation approach cross-referencing findings from multiple independent sources to enhance the validity and robustness of the conclusions [65,66]. Functional predictions from AlphaGenome serve as an additional independent line of evidence. Finally, the "Evidence Comparison" tab integrates MR statistics, literature mining, and variant functional predictions, providing a comprehensive assessment of evidence strength and mechanistic insights for each SNP instrument.

Troubleshooting

113. *Result downloading.* Click the "Download" node on the navigation tree or at the bottom of the page to enter the Download page. All processed results are shown in a table on the "Results Download" section. Click the "Download.zip" to download all files generated during the analysis.

Table 1. Comparison of MetaboAnalyst 5.0 and 6.0

	MetaboAnalyst 5.0	**MetaboAnalyst 6.0**
Raw spectra scopes	LC-MS	LC-MS and MS/MS (DDA and SWATH-DIA)
Spectra processing algorithms	centWave, MatchedFilter, Massifquant	centWave, MatchedFilter, Massifquant, Asari; spectral deconvolution for both DDA and SWATH-DIA
Peak annotation & compound identification	CAMERA	CAMERA, Khipu and MS2 database matching (dot-product or spectral entropy similarity)
Spectra processing results	Basic visualization and summary	Interactive exploration of MS2 results, integrative visualization of MS1 and MS2 results, and summary of metabolome/ exposome composition
Data processing	Missing value imputation, data filtering, and normalization	Enhanced workflow with improved support for feature filtering based on QC and BLANK samples, enhanced missing value imputation and normalization support
Dose response analysis	N/A	A new module accommodating both repeated and continuous exposure variables using multiple curve-fitting models.
Causal analysis & interpretation	N/A	A module providing end-to-end support for IV selection, harmonization, MR testing, and result interpretation

Table 2. Exposome Classifications

	Classes	Descriptions
1	Biocides	Biocides compounds, biocidal active substances.
2	Drugs	Pharmaceuticals and containing illicit drugs.
3	Environment Contaminants	Contaminations found from the environment, including air, water, soil, etc.
4	Foods	Compounds from food, food contact chemicals and packings.
5	Indoor environment	Compound related to indoor dust and environment
6	Industrial toxins	Industrial chemicals and toxins.
7	Microbes	Microbial metabolites, mycotoxins and antibiotics
8	Natural toxins	Natural toxins, including naturally occurring insecticides
9	Neuro toxins	Chemicals are associated with neurotoxicity.
10	Personal care products	Chemicals related to personal care, including cosmetic products, ingredients for tattoo ink and permanent make up.
11	PFAS	Per- and polyfluoroalkyl substances
12	Phenols	Phenol substances, including bisphenols and tert-butyl phenols.
13	Plants	Plants related chemicals, including plant protection products and toxic plant-phytotoxin.
14	Plastics	Plastic chemicals, including plastic additives.
15	PMTs	A series of persistent, mobile, and toxic substances
16	Smokes	Tobacco and Cannabis-related chemicals.
17	Surfactants	Surfactant-related chemicals
18	Water contaminants	Potential contaminants, chemicals from water
19	Other health-related	Other health or disease-related chemicals
20	Other hazards	Other hazardous substances
21	Others	All other exposome-related compounds, which have not been included above

TROUBLESHOOTING

Troubleshooting advice can be found in Table 3.

Table 3. Troubleshooting guide

Step	Problems	Possible reasons	Possible solutions
6	Data upload failed	The files are not zipped, or the file names do not match the metadata	Make sure every file is zipped individually. The file names must strictly match those in the metadata.
7	MS level information not displayed	Data mode selection is incorrect	Go back to the data upload page, and select "MS1+DDA" or "MS1+SWATH-DIA" based on the MS2 acquisition method
15	Spectra processing failed	The algorithm could not find any peaks using the specified parameters, or the algorithm is requiring more RAM than allocated	Use a different algorithm, like Asari, or *centWave-auto* to avoid optimizing parameters manually
28	MS2 fragment unknown	This fragment has not been annotated from the reference database	This fragment is unknown. Try to search for other databases or use "All Database"
29	Exposome classification figure not generated	The omics type was incorrect in the parameter setting page	Choose "Exposomics" as the omics type from the parameter setting page
32	MGF or MSP file size limitation error	The .msp or .mgf file is too big	Split the spectra into multiple files and perform search in multiple batches
34	Upload or integrity check failed	Incorrect "Data format" option selected; Incorrect data format; Sample names in data file and metadata file do not match.	Double check data format option; Make sure sample names are identical in both files; Remove special characters; Do not include QC or BLANK samples in metadata file
36	Not all metadata are "OK"	Missing values, not enough replicates per group, or continuous metadata values are not numeric	Manually correct the values using the "Edit" column on the metadata table; Remove that metadata.
58	Legend too large	A continuous metadata has been assigned as categorical.	Go to "Metadata Check" page and update it to "Continuous"
94	No significant features identified	The threshold may be overly restrictive, or the biological effect could be subtle and masked by data noise or batch effects	Relax threshold, try different covariate adjustment or try different filtering, missing value imputation and normalization options.

96	Function unresponsive	The algorithm could not finish because of lacking computing resources	Set a more stringent threshold in the feature filtering step; Lower the number of models selected
104	No metabolites found in the autocomplete list	The metabolites have not been found to be significantly associated with any SNPs; the metabolites may be referred to by different names.	We will expand the database when new studies are published; In the second case, use resources like HMDB, KEGG, or PubChem to search for alternative or standardized names.
107	All the SNPs are excluded	All identified SNPs that are associated with multiple metabolites are excluded	Check if the associated metabolites are functionally related, such as being part of the same pathway; manually include SNPs based on context and study objectives.
112	The number of SNPs displayed in the forest plot is fewer than the number initially input.	Some SNPs appear multiple times due to differing statistics from different studies. SNPs that yield extremely high standard errors or invalid estimates may be excluded from the plot to improve clarity.	Double check the input list on the filtering and harmonization page and include only the SNPs of interest.

TIMING

Steps 1 - 32, LC-MS/MS Spectra Processing and Compound Annotation: 1.5 – 2.5 h

Steps 33 - 90, Data Processing and Exploratory Data Analysis: 30 min ~ 1h

Steps 91 - 102, Dose Response Analysis: ~ 30 min

Steps 103 - 113, Causal analysis and interpretation: 20 ~ 30 min

ANTICIPATED RESULTS

Utilizing e-waste and T2D as illustrative case studies, this protocol has detailed the comprehensive workflow for LC-MS/MS raw spectra processing, exploratory statistical analysis, dose-response modeling, and causal inference. The following sections summarize the results obtained from these stages.

Stage 1: LC-MS/MS Spectra Processing and Compound Annotation

LC-MS/MS spectra processing generates multiple data tables, including an MS1 feature table (metaboanalyst_input.csv), which contains 16,021 MS1 features, and MS2 compound identification results table (compound_msn_results.csv), which contains 728 identified MS features and their corresponding chemical candidates. Most identified compounds (>98%) can be matched to at least one exposome classification (Step 29). Other tables are also available from the Result download page (Step 30). Based on the classification, e-waste workers (labelled as "Worker") contain various exposome compounds across almost all exposome classes, especially environmental contamination related compounds, toxins and other hazardous chemicals. This is expected based on the study context of the dataset.

Stage 2: Data processing and exploratory data analysis

The protocol identifies 453 metabolic features significantly associated with blood cadmium levels after adjusting for age, BMI, smoking, and alcohol intake (Step 67), with detailed result table downloaded (covariate_result.csv). PCA scores plot demonstrated tight clustering of QC samples and a subtle separation based on "BCd_group" (Step 49). Heatmap analysis revealed an interesting cluster of abundance patterns (Step 79). Finally, a Random Forest model was used to classify cadmium exposure groups, with the associated error rates and a list of high-importance features provided (randomforests_sigfeatures.csv) (Steps 85 and 86).

Stage 3: Dose response analysis

In the feature selection for BCd exposure (Steps 94 and 95), 453 features met the specified significance thresholds. Subsequent dose-response modeling (Steps 96–97) resulted in 183 successfully converged models. The output includes a BMD distribution plot providing system-level thresholds (Step 98), alongside a comprehensive summary table (Step 99) detailing best-fit models, BMD values, and confidence intervals for each feature. Dose response analysis using Age as continuous variable yielded 850 features passing through the initial feature selection and 462 fitted models (Step 101).

Stage 4: Causal Estimate based on Mendelian Randomization

The MR analysis demonstrated an overall consistent positive causal effect of the input metabolite (L-Isoleucine) on the selected outcome (type 2 diabetes) across the selected methods (MR Egger, Inverse variance weighted, Inverse variance weighted, Weighted median, Simple mode and Weighted mode) (Step 110). No strong evidence of heterogeneity and directional pleiotropy was observed. Two SNPs (rs58101275 and rs1420601) were identified as positively associated with the outcome, suggesting their potential causal involvement in the observed effect (Step 111).

Literature evidence based on triangulation analysis suggested potential pathways linking L-isoleucine to T2D, involving its effect on serine levels and subsequent modulation of IRS1 (Step 112).

Data Availability Statement

All data supporting this protocol has been included in the material, data files section of this protocol. The original dataset and metadata table used as the e-waste example in this protocol is available from our previously published paper[13].

Code Availability Statement

MetaboAnalyst 6.0 is freely available at (https://metaboanalyst.ca/). The underlying R package is available at GitHub (https://github.com/xia-lab/MetaboAnalystR). The current protocol is based on MetaboAnalystR v4.2.0 (https://github.com/xia-lab/MetaboAnalystR/releases/tag/v4.2.0). It is available from the Zenodo repository at: https://doi.org/10.5281/zenodo.17393838. MetaboAnalyst-Pro Enterprise Solution is available for local installation (https://www.xialab.ca/pro/protools.xhtml)

Acknowledgements

This work is supported by the Canadian Foundation for Innovation (CFI), Genome Canada, the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canadian Institutes of Health Research (CIHR), the Canada Research Chairs (CRC) program, and the Kyoto-McGill International Collaborative Program

Author contributions

Z.P., G.Z., Y.L. and J. X. developed and updated MetaboAnalyst and MetaboAnalystR tools. Z.P., G.Z., Y.L. and J. X. designed the protocols and performed the data analysis Z.P., G.Z., Y.L., H.O., C.V. and J. X. tested the entire workflow. Z.P., G.Z., Y.L. and J. X. wrote the manuscript. H.O., C.V. and F.M. provided critical comments for the manuscript. N.B. helped with the preparation of example data. J.X. supervised the study. All authors read and approved the final manuscript.

Competing interests

J.X. is the founder of XiaLab Analytics, a startup created to support the long-term maintenance and sustainability of MetaboAnalyst and related omics tools. The remaining authors declare no competing interests.

Figures

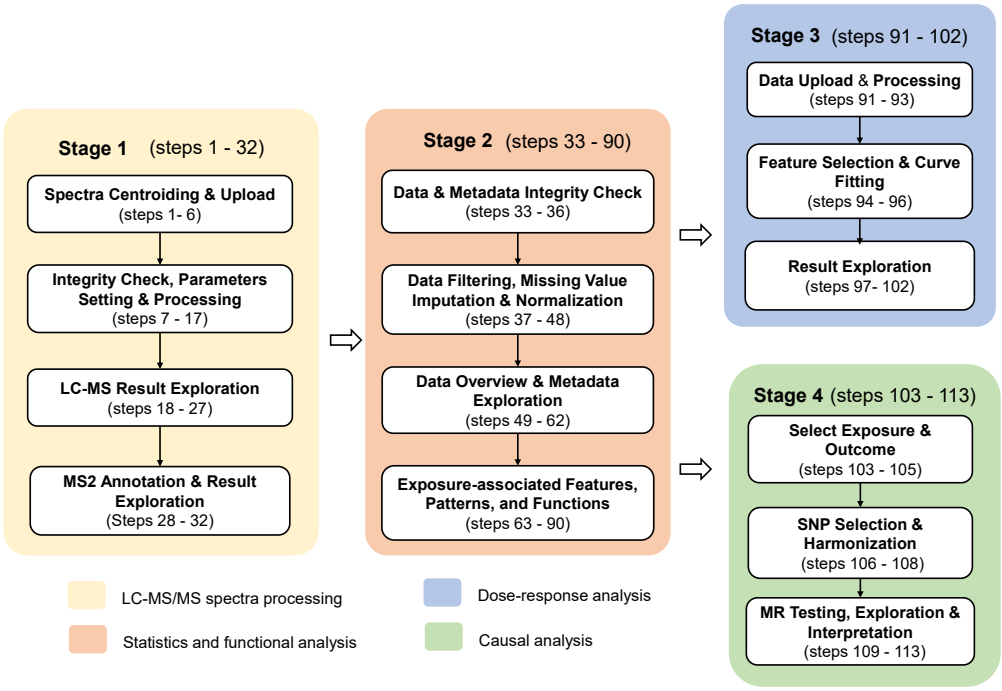


Figure 1 | Overall workflow of exposomics data analysis in MetaboAnalyst 6.0.

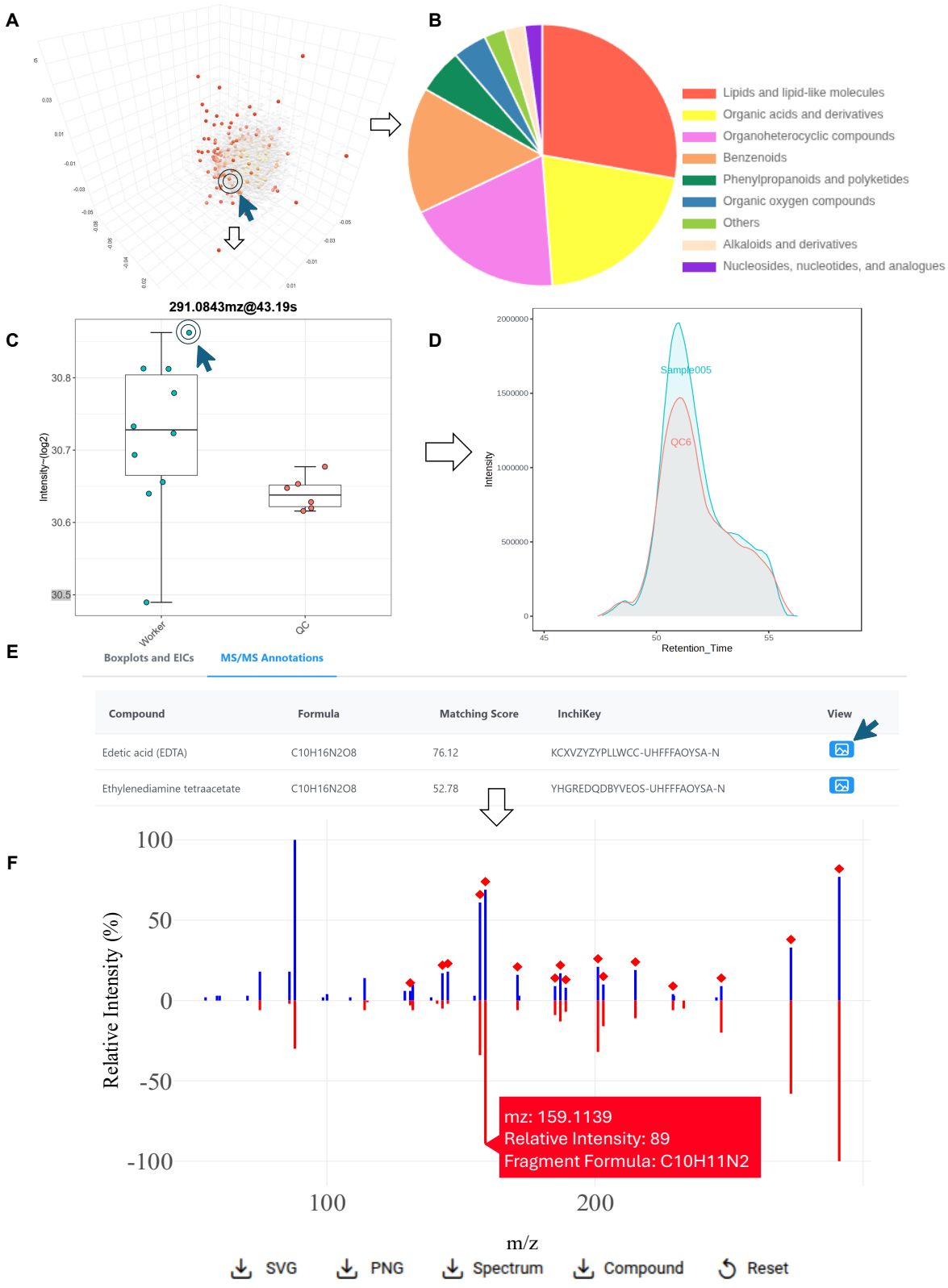

A
B
Lipids and lipid-like molecules
Organic acids and derivatives
Organoheterocyclic compounds
Benzenoids
Phenylpropanoids and polyketides
Organic oxygen compounds
Others
Alkaloids and derivatives
Nucleosides, nucleotides, and analogues
291.0843mz@43.19s
C
Intensity~(log2)
30.8
30.7
30.6
30.5
Worker
QC
D
Intensity
2000000
1500000
1000000
500000
0
Sample005
QC6
45
50
55
Retention_Time
E
Boxplots and EICs
MS/MS Annotations
Compound
Formula
Matching Score
InchiKey
View
Edetic acid (EDTA)
C10H16N2O8
76.12
KCXVZYZYPLLWCC-UHFFFAOYSA-N
Ethylenediamine tetraacetate
C10H16N2O8
52.78
YHGREDQDBYVEOS-UHFFFAOYSA-N
F
Relative Intensity (%)
100
50
0
-50
-100
100
200
m/z
mz: 159.1139
Relative Intensity: 89
Fragment Formula: C10H11N2
SVG
PNG
Spectrum
Compound
Reset

Figure 2 | Interactive exploration of LC-MS/MS spectral processing results. (A) An interactive 3D PCA loading plot. The highlighted features contain MS2-based compound annotations. (B) Interactive pie charts provide a high-level overview of the metabolome or exposome, organized by chemical taxonomy and exposome classifications, respectively. Double-clicking a feature in the PCA loading plot automatically opens a summary box plot displaying the intensity distribution across samples and groups (C), as well as the corresponding extracted ion chromatogram (EIC) (D). When available, potential chemical candidates (E) and MS2 mirror plots (F) can also be interactively explored, offering deeper insight into compound identity and spectral quality.

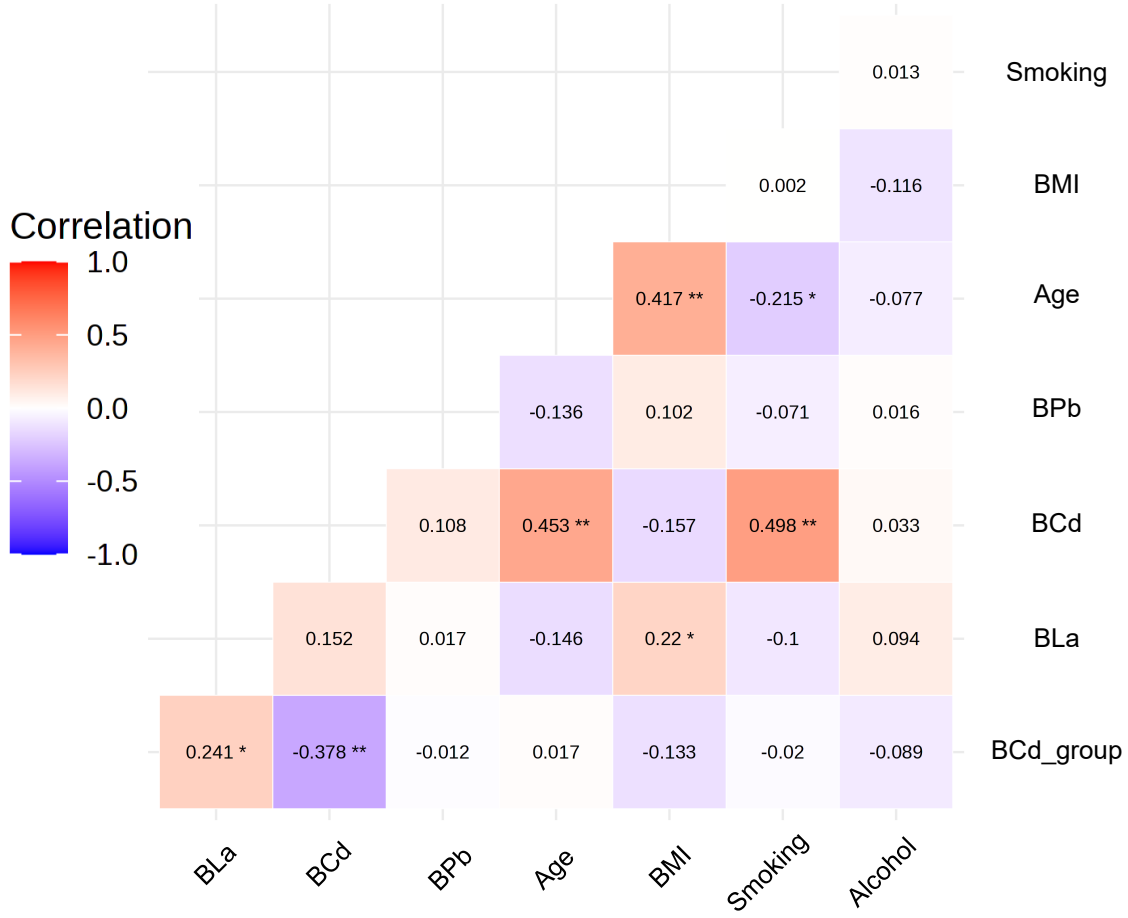


Figure 3 | Metadata Correlation Heatmap. The heatmap displays partial Pearson correlation coefficients between all pairs of metadata variables. Statistical significance is denoted by asterisks (*: $p < 0.05$; **: $p < 0.01$).

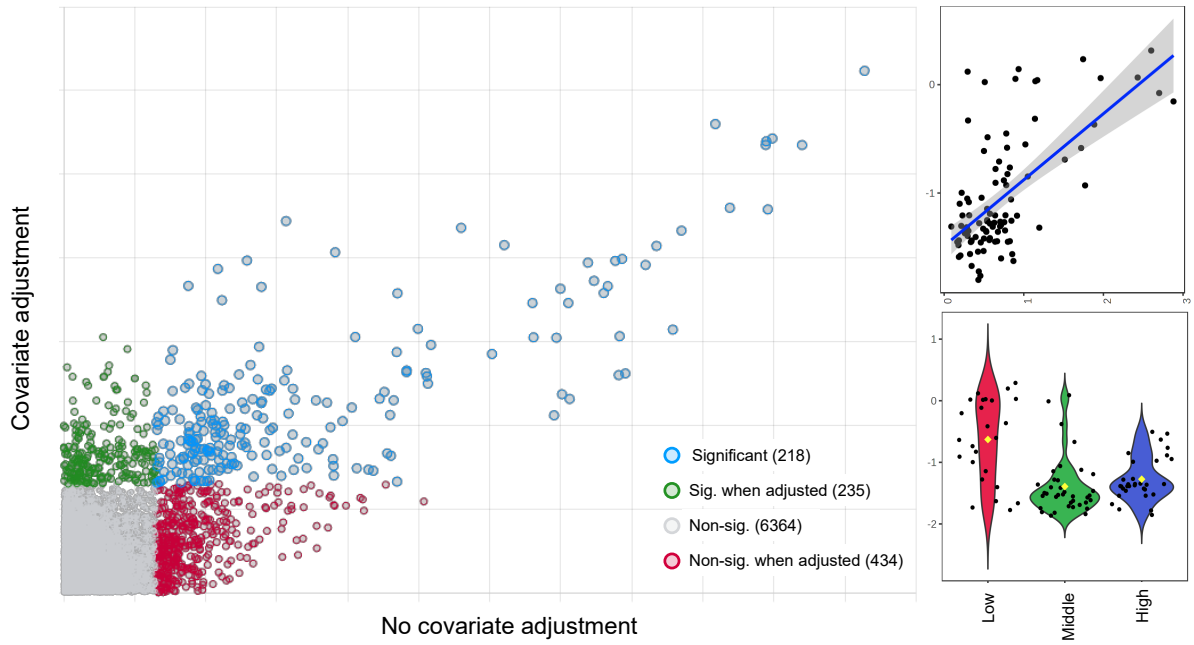


Figure 4 | Significant features with covariate adjustment. This scatter plot compares the $-\log_{10}$ p-values derived from linear models before (unadjusted; x-axis) and after (adjusted; y-axis) covariate correction. Each point represents a molecular feature, color-coded based on its significance status following adjustment. Selecting a feature launches a detailed visualization of its abundance profile relative to the chosen metadata: a violin plot is displayed for discrete variables, while a scatter plot is used for continuous variables.

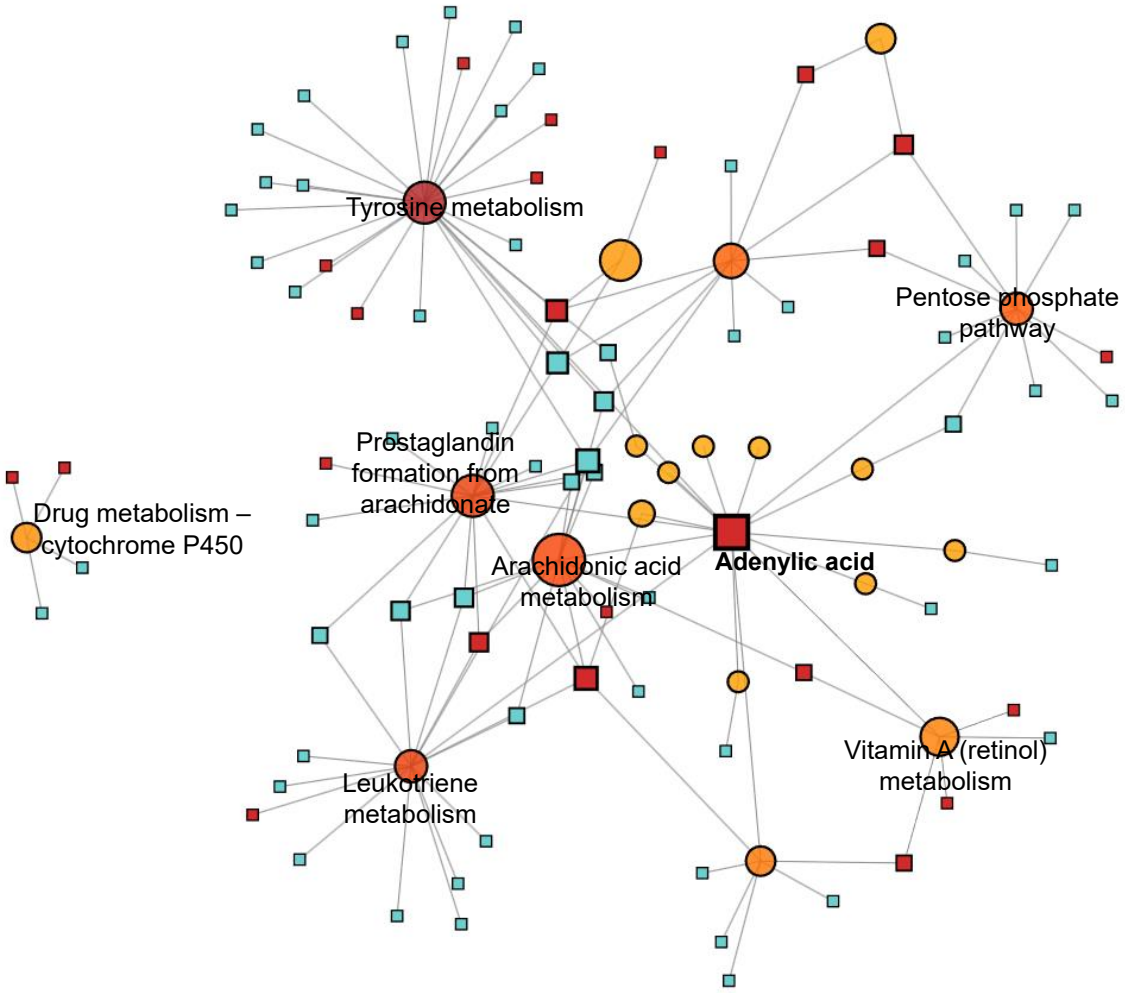


Figure 5 | Mummichog Enrichment Network. Pathways are depicted as circles; larger circles indicate a higher count of matched peaks, while darker red shades represent greater statistical significance. Square nodes denote the specific compounds matched within those pathways. If a compound is statistically significant, its square node is highlighted in red.

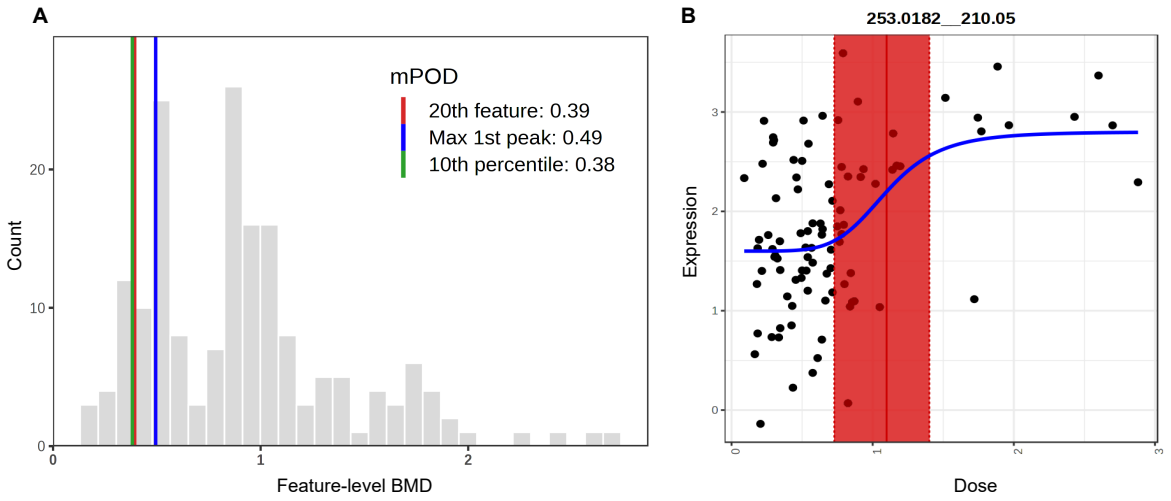


Figure 6 | Dose response analysis. A) Histogram showing the overall distribution of feature-level benchmark doses that were computed and passed the statistical flags for a given exposure during the curve fitting step; B) An example curve fitting plot showing the fitted feature intensity curve across exposure levels. The vertical red line indicates the estimated BMD, with the shaded red region representing its 95% confidence interval.

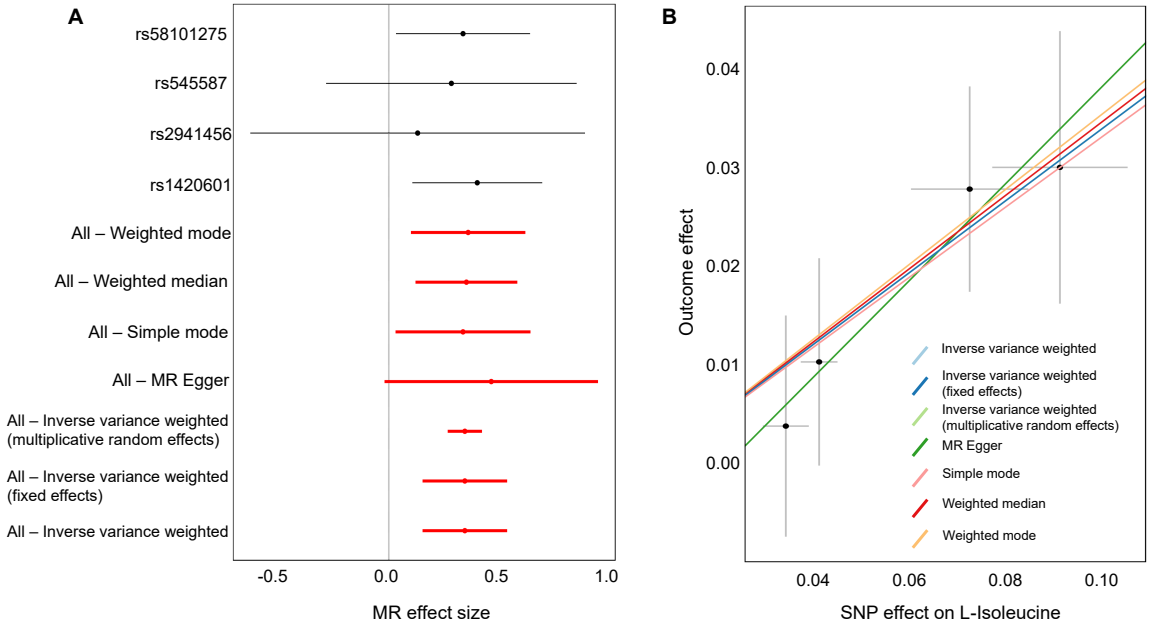
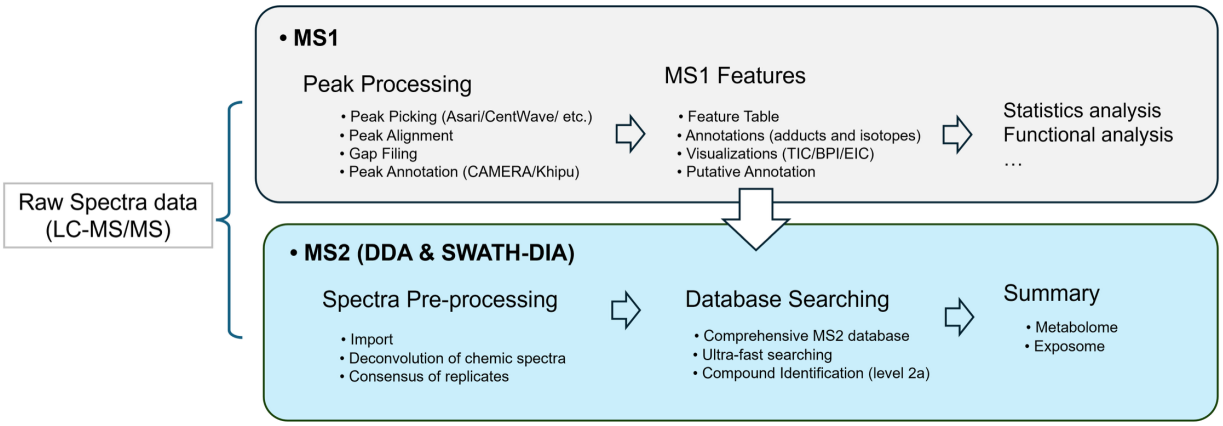
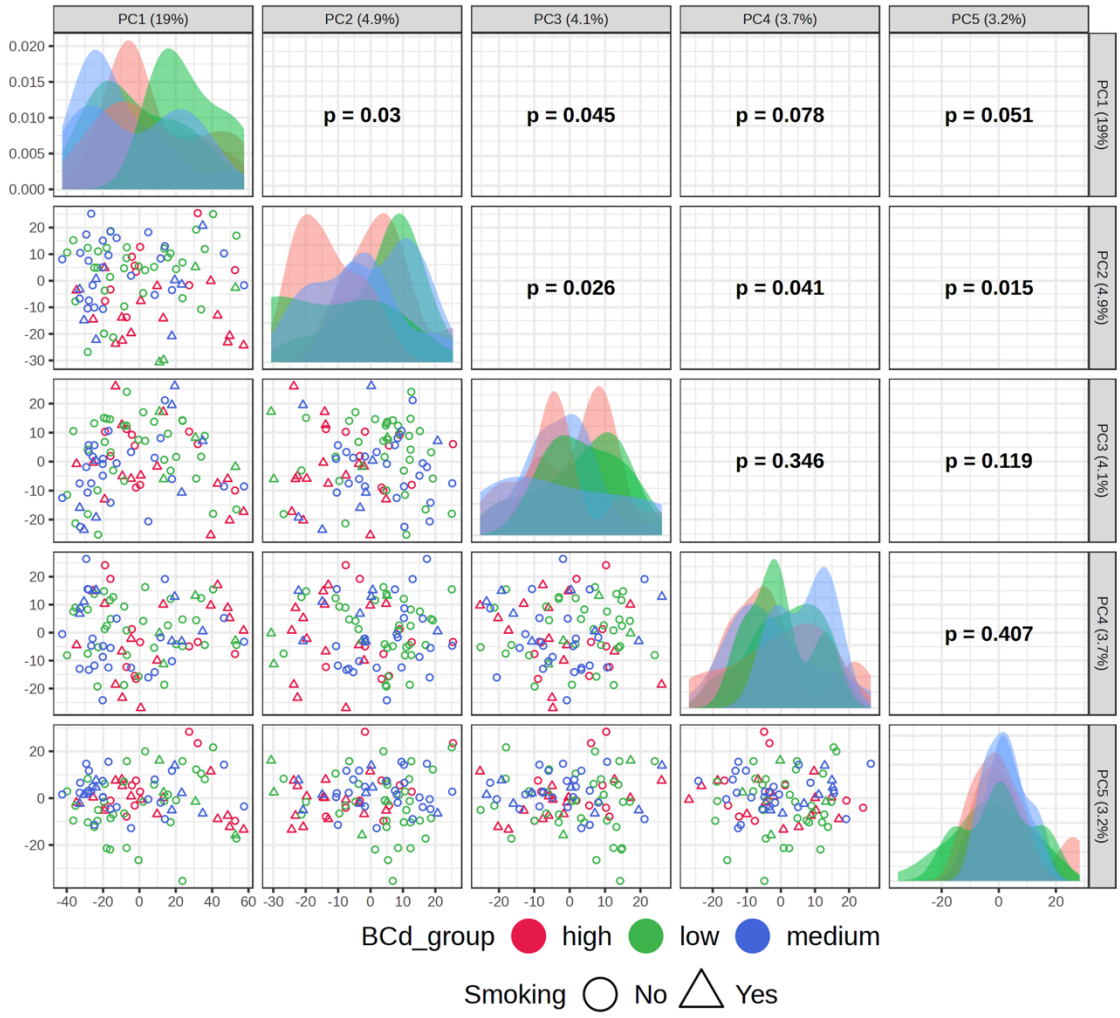


Figure 7 | Causal analysis based on two-sample Mendelian randomization. A) Forest plot illustrating individual SNP-specific causal estimates and the pooled estimates from each MR method. B) Scatter plot showing SNP effects on exposure (x-axis) versus outcome (y-axis), overlaid with regression lines representing causal estimates from each MR method.

Extended Data Figs

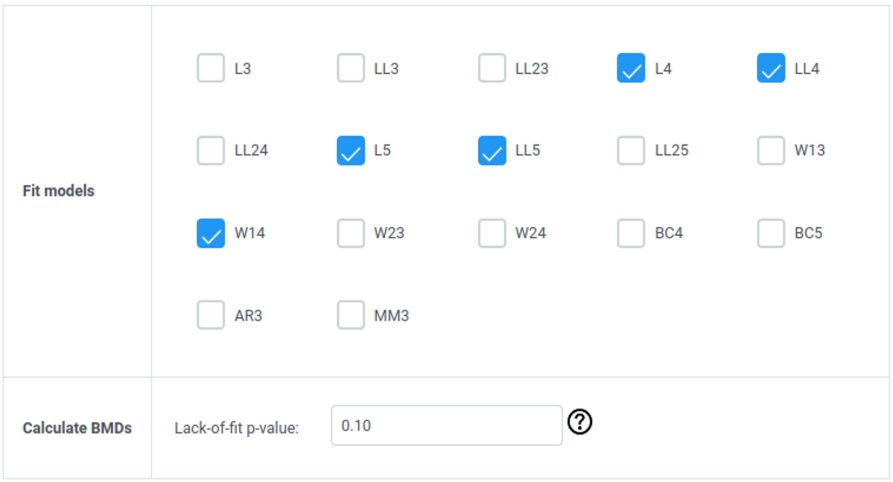


Extended Data Fig. 1 | Workflow of raw spectral processing. It includes MS1 peak profiling and MS2 compound annotation.



Extended Data Fig. 2 | PCA overview coupled with PERMANOVA results. Numbers in upper right panels are PERMANOVA p-values for group separation along each component pair.

A



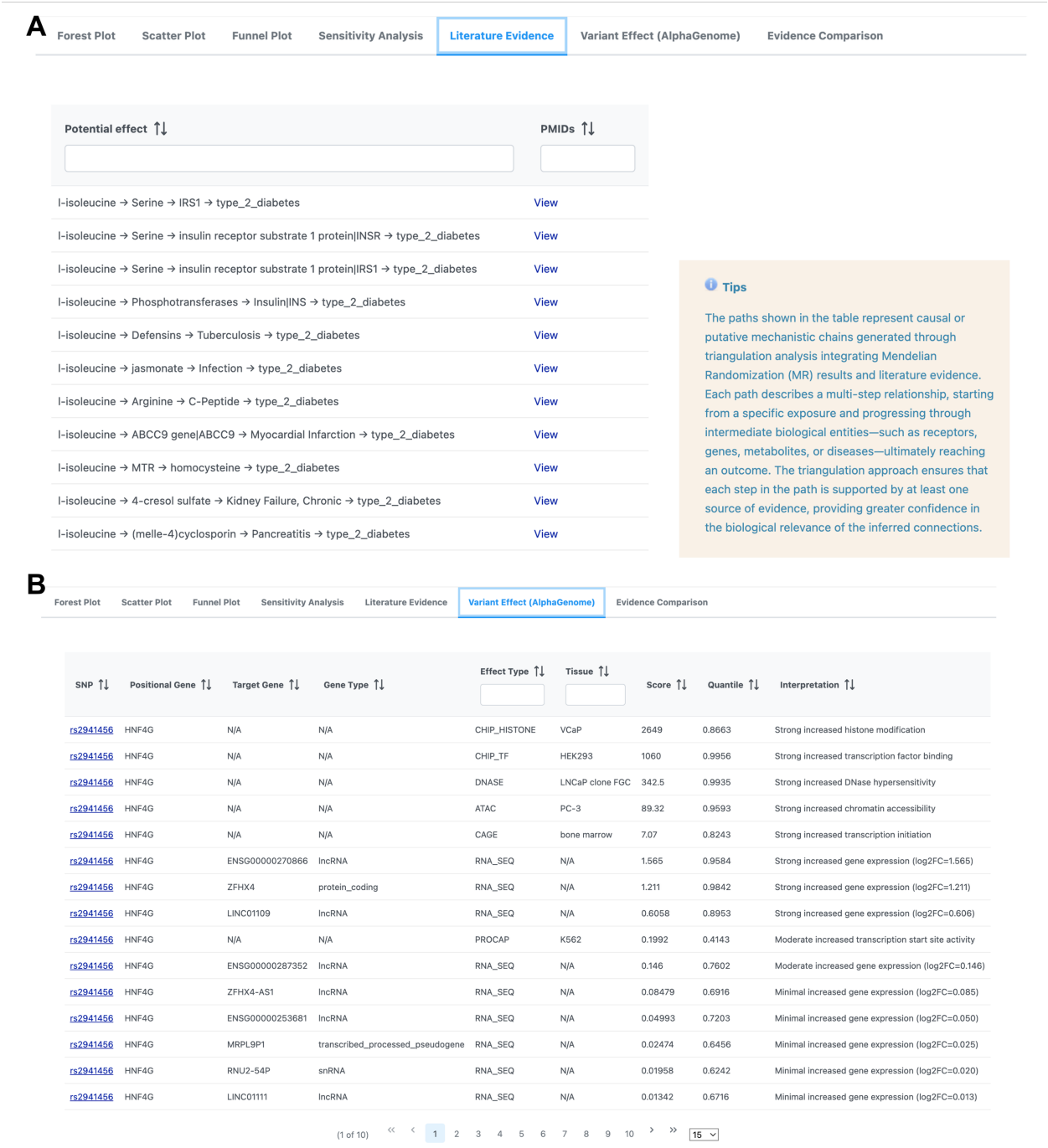
B

ID	P-val	BMDl	BMD	BMDu	AIC	Model name
317.0673__141.42	1.0	0.12	0.18	0.23	132.47	ll4
898.5249__336.66	1.0	0.073	0.18	0.28	-19.5	w14
634.3023__216	1.0	0.016	0.21	0.4	298.76	w14
428.9777__41.5	1.0	0.12	0.25	0.39	265.53	ll5
346.056__42.57	1.0	0.097	0.26	0.42	171.22	ll4
260.0497__189.62	1.0	0.21	0.29	0.37	237.81	ll4
634.3023__209.78	1.0	0.16	0.3	0.43	328.84	ll5
149.0599__189.62	1.0	0.26	0.32	0.37	-11.66	ll4
331.0611__143.83	1.0	0.3	0.33	0.37	146.38	ll4

Extended Data Fig. 3 | Dose response analysis. A screenshot showing the available statistical models for curve fitting.

Include	SNP ID	Nearest Gene	Pval (outcome)	Exposure	Pval (exposure)	Biofluid	Population	Study	Pval (steiger)
☑	rs2941456	HNF4G	0.7371	L-Isoleucine	9.10e-09	Blood	European	Borges_UKBB_2020	0.009525
☑	rs1420601	CBLN1	0.007379	L-Isoleucine	3.70e-08	Blood	European	27898682	1.45e-05
☑	rs58101275	TRMT61A	0.02959	L-Isoleucine	2.78e-08	Blood	European	27898682	3.933e-07
☑	rs545587	HSD17B14	0.3295	L-Isoleucine	2.30e-19	Blood	European	Borges_UKBB_2020	0.000241
☐	rs7302925	SPRYD4	0.8385	Ala/Gln, D-Alanine,	3.60e-16	Blood	European	Borges_UKBB_2020	8.988e-05
☐	rs1260326	GCKR	8.442e-07	Concentration of m	2.93e-14	Blood	European	22286219	1.718e-06
☐	rs1260326	GCKR	8.442e-07	Concentration of m	2.93e-14	Blood	European	22156771	1.991e-10
☐	rs117643180	SLC2A4	0.1253	D(-)-beta-hydroxy	2.90e-15	Blood	European	Borges_UKBB_2020	0.0083
☐	rs1260326	GCKR	8.442e-07	Concentration of m	2.72e-17	Blood	European	27005778	7.229e-10
☐	rs1260326	GCKR	8.442e-07	Concentration of m	2.30e-29	Blood	European	Borges_UKBB_2020	0.1642

Extended Data Fig. 4 | SNP selection and harmonization for Mendelian randomization. The table displays candidate genetic instruments for L-isoleucine, including SNP IDs, nearest genes, and p-values for both exposure and outcome. Metadata regarding biofluid, population, and source study are provided to ensure data provenance. The "Include" checkboxes indicate the final subset of SNPs selected after harmonization and quality control filtering



Extended Data Fig. 5 | Supporting evidence for causal inference. A) Literature-based triangulation from the MelodiPresto knowledge base identifies curated mechanistic pathways

linking L-isoleucine to type 2 diabetes through intermediate biological entities such as serine, IRS1, and defensins. Each path is supported by published studies with links to PubMed identifiers for manual verification. B) Variant functional evidence from querying AlphaGenome reveals the predicted regulatory consequences of each instrumental SNP at single-base resolution. A larger quantile score indicates stronger evidence for the functional impact of the SNP.

Box 1 | LC-MS/MS Spectra Processing and Compound Annotation

LC-MS/MS is widely used in metabolomics and exposomics studies. For spectra collection, a common strategy is performing LC-MS for each sample, and MS/MS (MS2) on pooled quality check (QC) samples [67]. Both data-dependent acquisition (DDA) and data-independent acquisition (DIA) are often employed for MS2 spectra acquisition. The main tasks in this stage include LC-MS feature detection, relative quantification and MS2-based compound annotation.

Spectra Conversion and Centroiding

Raw mass spectra are initially acquired in proprietary vendor formats as profile mode data. Converting these files to standardized open formats such as mzML, mzXML, or CDF is essential for cross-platform compatibility and high-throughput processing. This conversion incorporates a centroiding algorithm that compresses the data by eliminating redundant spectral information while preserving peak characteristics. ProteoWizard is the standard tool for format conversion and centroiding operations [38,68].

LC-MS Feature Detection and Peak Annotation

Feature detection involves extracting meaningful chromatographic elution profiles from consecutive scans in raw LC-MS data. Among available algorithms, centWave is the most widely used approach. It functions by modeling and extracting Gaussian-shaped features from individual samples before the subsequent step of peak alignment across the entire dataset [41]. MetaboAnalystR 3.0 and its subsequent versions enhance centWave's performance by incorporating a parameter optimization step to streamline the whole process [40]. A recent innovation, Asari, takes a fundamentally different approach. It first aggregates mass features from all samples into composite mass tracks which effectively stabilizes the mass signal across the full dataset and bypasses the variability of single-sample peak detection [12]. This global signal aggregation boosts the sensitivity

for detecting trace-level features [13], making Asari particularly valuable for exposomics applications. LC-MS features often contain significant redundancy due to the presence of multiple adducts, isotopologues, and isomers [69,70]. MetaboAnalyst integrates CAMERA[69] to annotate MS features detected by centWave, MatchedFilter, and Massifquant. [69,70]. Khipu is used to annotate features detected by Asari [12,71].

MS2 Spectra Deconvolution and Compound Identification

Environmental contaminants often occur at trace concentrations, resulting in spectra that are frequently confounded by signals from abundant, co-eluting isobaric interferents. Spectral deconvolution addresses this challenge by resolving overlapping peaks and extracting pure ion chromatograms, thereby enhancing the reliability of downstream compound annotation. MetaboAnalyst supports MS2 spectral deconvolution for both DDA and Sequential Window Acquisition of all Theoretical fragment (SWATH)-DIA datasets [14]. Compound identification based on MS2 spectra is based on the similarity between query spectra with reference spectra. MS2 spectra similarity can be evaluated with dot-product or spectra entropy [72,73]. Dot-product similarity evaluates MS2 spectra based on the weighted cosine similarity of peak intensities, favoring spectra with shared high-intensity peaks. While spectral entropy considers the overall distribution of fragment intensities, providing a more holistic measure of spectral similarity. A higher matching score usually means more confidence in the compound identification. MetaboAnalyst provides access to ten public MS2 spectra databases, complemented by a specialized exposome database containing approximately 100,000 compounds curated from multiple exposome databases. This is considered level 2 identification based on public MS2 reference spectra. Further validation via in-house reference standards is required to achieve level 1 annotation [74]. Finally, to provide high-level overview and functional categories of the annotated compounds, MetaboAnalyst offers

metabolome annotation based on HMDB annotation, and exposome classification as defined by the NORMAN Suspect List Exchange[43].

End of Box 1

Box 2 | Data Processing Strategies for High-dimensional Omics Data

LC-MS/MS exposomics datasets are inherently high-dimensional and present multiple analytical challenges, including substantial missing values and noises with signal intensities spanning several orders of magnitude. Data processing systematically addresses these issues through three sequential steps – data filtering, missing value imputation and data normalization.

Data Filtering

Data filtering is crucial for improving data quality and statistical power by removing features that are unlikely to contribute to downstream analysis. MetaboAnalyst provides four complementary filters: low-quality filter, low-repeatability filter, low-variance filter and low-abundance filter. The low-quality filter removes features identified as background or contaminants (when blank samples are provided), or containing high proportions of missing values; low-repeatability filter removes features exhibiting high relative standard deviation (RSD) among QC replicates; low-variance filter discards near-constant features; and low-abundance filter excludes features with baseline-level intensities.

Missing Value Imputation

Exposomics datasets typically contain high proportions of missing values. While initial data filtering reduces this issue, the remaining missing values must be addressed before proceeding with statistical analysis. MetaboAnalyst provides three imputation strategies – left-censored data estimation, univariate methods, and multivariate methods. The default approach assumes missing

values are due to values falling below the detection limit (left-censored data). Users can replace missing values with their estimated detection limits (1/5 of the minimum positive value observed for each individual feature). A drawback of this method is the introduction of many identical, small constant values. The quantile regression imputation of left-censored data (QRILC) method models the low tail of each feature as log-normal and samples replacements, thereby preserving inherent variance without introducing downward bias [47]. Users can also explore other univariate (min, mean, median) or multivariate (KNN, PCA, Random Forest) methods which leverage correlations between features or samples to estimate the missing entries.

Data Normalization

MetaboAnalyst provides three normalization categories: sample normalization, data transformation, and data scaling. Sample normalization corrects systematic technical variations between samples, such as differences in loading volume or weight. These batch effects can also be detected using PCA and through examining distributions of missing values and sample intensities. Log transformation, while a common and effective data transformation for targeted metabolomics, can disproportionately amplify subtle variations at the lower end of the intensity spectrum. This is problematic, especially in untargeted metabolomics/exposomics where small, noisy measurements are prevalent. For these datasets, more robust alternatives like variance stabilizing normalization (VSN)[48] or Pareto scaling are often preferred. VSN aims to stabilize the variance across all intensity levels, while Pareto scaling reduces the dominance of large variables, thus preventing noisy low-abundance variables from unduly influencing data analysis [75].

End of Box 2

Box 3 | Dose Response Analysis and Benchmark Dose

The central toxicological maxim, "the dose makes the poison," underscores that the toxicity of a chemical depends on its concentration in a biological system. Dose response analysis aims to quantitatively relate exposure (dose) to a biological effect (response). In the absence of specific biological priors, this analysis employs parametric regression models to evaluate whether a statistically significant relationship exists between the paired data points [52]. The process will also identify best fit models to describe the relationship.

Experimental Studies

Conventional dose-response studies are performed in animal models (and increasingly, in *in vitro* systems) where a graded range of defined chemical concentrations are tested. Once the data are fitted to a curve fitting model, researchers derive the benchmark dose (BMD) which is defined as the minimum concentration of a substance that produces a measurable, low-level response [76]. The lower confidence limit of the BMD (BMDl) is often used as the point of departure (POD) for risk assessment. In high-throughput omics studies, BMDs are calculated for individual features (e.g., genes or metabolites) and subsequently aggregated to estimate the dose triggering a systems-level response. These aggregate values are termed the transcriptomic POD (tPOD) or metabolic POD (mPOD), depending on the data type.

Observational Studies

Dose response analysis is also applicable in exposomics studies, where individuals naturally experience varying levels of exposures and associated responses. Response variables may include omics data (usually continuous) or phenotypic traits (continuous or categorical). To minimize potential confounding in such observational studies, researchers should adjust for covariates when selecting features for analysis. The resulting dose response (or, more appropriately termed

exposure response) curve provides valuable insights into the relationships between exposures and biological responses, under real-world scenarios. Similar to experimental studies, the BMDs are estimated from the fitted dose-response curves using mathematical formula. However, due to significant inter-individual variations in exposomics studies, dose-response curves typically reflect only the broad population trend, rather than individual nuances. The scarcity of high-level exposure data further compounds this challenge. The interpretation of BMDs from such exposomics studies requires careful attention (e.g., identification of biases including selection, measurement, and recall) and is context dependent (especially the definition of the baseline or control group). These findings should be viewed as hypothesis-generating to help inform the design of targeted experimental studies or more sophisticated observational ones.

End of Box 3

Box 4 | Causal Estimate and Mendelian Randomization

MR leverages the fact that individuals naturally inherit different "doses" of specific genetic variants (e.g., 0, 1, or 2 copies of a particular allele). These "doses" of genetic variants often lead to different average levels of exposure within the population. Because these genetic variants are randomly allocated at conception, their "dose" is largely independent of environmental and lifestyle factors that could otherwise confound observational studies. MR then assesses whether these genetically determined differences in exposure levels (the "treatment" effect induced by the genetic "dose") are associated with differences in the outcome. If they are, and the core IV assumptions hold (see below), it provides strong evidence for a causal link.

Instrumental Variables (IV) Selection

IV selection is key to the success of MR analysis. The genetic variants must fulfill three assumptions:

i) strongly associated with the exposure of interest (relevance);

ii) independent of any confounders that affect both exposure and outcome (independence); and

iii) influence the outcome only through the exposure of interest (exclusion).

After identification of SNP candidates associated with exposure and outcome (relevance), users should perform SNP filtering and harmonization to ensure that effect alleles are consistent across different datasets (independence). The exclusion assumption is the most challenging to satisfy, primarily due to reverse causality and horizontal pleiotropy (SNP affecting outcome independently of exposure). Steiger filtering mitigates reverse causality by assessing whether the SNP's explained variance is significantly greater for the exposure than the outcome [59].

MR Tests and Interpretation

The simplest causal effect estimate for a single genetic variant is the Wald ratio, calculated as the ratio of its effect on the outcome to its effect on the exposure, intuitively interpreted as the outcome change per unit exposure change. When using multiple SNPs, individual Wald ratios are combined via simple median or the more robust inverse-variance weighted (IVW) method to address MR assumption deviations. MR analysis results are also summarized in various graphics (e.g., forest plot, scatter plot or funnel plot) to help assess IV heterogeneity and potential assumption violations. Finally, leave-one-out sensitivity analysis is typically employed to determine if the overall causal effect is disproportionately driven by any single variant.

MR Implementation in MetaboAnalyst

MR can be performed in a single study where the exposure, outcome, and IVs are all measured in the same individuals, or in two separate studies (2SMR) for exposure and outcome assuming they are from the same underlying population or are sufficiently similar [57,58,77]. The MR analysis in MetaboAnalyst is based on 2SMR. The SNP - metabolite summary statistics were curated from 65 mGWAS [78]. The SNP - outcome summary statistics are based on OpenGWAS [58]. MetaboAnalyst offers 14 MR tests, including MR-Egger, Weighted Median, and Mode-based methods. Finally, MetaboAnalyst offers semantic triples (subject–predicate–object) from literature mining to help interpret potential mechanisms linking genetic variants, exposures, and phenotypes [66].

End of Box 4

REFERENCES

1 Abdellaoui, A., Yengo, L., Verweij, K. J. H. & Visscher, P. M. 15 years of GWAS discovery: Realizing the promise. *Am J Hum Genet* **110**, 179–194 (2023). https://doi.org/10.1016/j.ajhg.2022.12.011

2 Wild, C. P. Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol Biomarkers Prev* **14**, 1847–1850 (2005). https://doi.org/10.1158/1055-9965.Epi-05-0456

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

4 Escher, B. I. *et al.* From the exposome to mechanistic understanding of chemical-induced adverse effects. *Environ Int* **99**, 97–106 (2017). https://doi.org/10.1016/j.envint.2016.11.029

5 Hu, X. *et al.* A scalable workflow to characterize the human exposome. *Nat Commun* **12**, 5575 (2021). https://doi.org/10.1038/s41467-021-25840-9

6 Maitre, L. *et al.* Multi-omics signatures of the human early life exposome. *Nat Commun* **13**, 7024 (2022). https://doi.org/10.1038/s41467-022-34422-2

7 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 (2012). https://doi.org/10.1093/nar/gks374

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

9 Chong, J. *et al.* MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. *Nucleic Acids Res* **46**, W486–w494 (2018). https://doi.org/10.1093/nar/gky310

10 Pang, Z. *et al.* MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights. *Nucleic Acids Res* **49**, W388–w396 (2021). https://doi.org/10.1093/nar/gkab382

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 (2009). https://doi.org/10.1093/nar/gkp356

12 Li, S., Siddiqa, A., Thapa, M., Chi, Y. & Zheng, S. Trackable and scalable LC-MS metabolomics data processing using asari. *Nat Commun* **14**, 4113 (2023). https://doi.org/10.1038/s41467-023-39889-1

13 Pang, Z., Viau, C., Fobil, J. N., Basu, N. & Xia, J. Comprehensive Blood Metabolome and Exposome Analysis, Annotation, and Interpretation in E-Waste Workers. *Metabolites* **14** (2024). https://doi.org/10.3390/metabo14120671

14 Pang, Z. *et al.* MetaboAnalystR 4.0: a unified LC-MS workflow for global metabolomics. *Nat Commun* **15**, 3675 (2024). https://doi.org/10.1038/s41467-024-48009-6

15 Holland-Letz, T. & Kopp-Schneider, A. Optimal experimental designs for dose-response studies with continuous endpoints. *Arch Toxicol* **89**, 2059–2068 (2015). https://doi.org/10.1007/s00204-014-1335-2

16 Ewald, J., Soufan, O., Xia, J. & Basu, N. FastBMD: an online tool for rapid benchmark dose-response analysis of transcriptomics data. *Bioinformatics* **37**, 1035–1036 (2021). https://doi.org/10.1093/bioinformatics/btaa700

17 Phillips, J. R. *et al.* BMDExpress 2: enhanced transcriptomic dose-response analysis workflow. *Bioinformatics* **35**, 1780–1782 (2019). https://doi.org/10.1093/bioinformatics/bty878

18 Yao, C. H. *et al.* Dose-Response Metabolomics To Understand Biochemical Mechanisms and Off-Target Drug Effects with the TOXcms Software. *Anal Chem* **92**, 1856–1864 (2020). https://doi.org/10.1021/acs.analchem.9b03811

19 Sanderson, E. *et al.* Mendelian randomization. *Nat Rev Methods Primers* **2** (2022). https://doi.org/10.1038/s43586-021-00092-5

20 Misra, B. B. Advances in high resolution GC-MS technology: a focus on the application of GC-Orbitrap-MS in metabolomics and exposomics for FAIR practices. *Anal Methods* **13**, 2265–2282 (2021). https://doi.org/10.1039/d1ay00173f
21 Tsugawa, H. *et al.* A lipidome atlas in MS-DIAL 4. *Nat Biotechnol* **38**, 1159–1163 (2020). https://doi.org/10.1038/s41587-020-0531-2
22 Tsugawa, H. *et al.* Hydrogen Rearrangement Rules: Computational MS/MS Fragmentation and Structure Elucidation Using MS-FINDER Software. *Anal Chem* **88**, 7946–7958 (2016). https://doi.org/10.1021/acs.analchem.6b00770
23 Schmid, R. *et al.* Integrative analysis of multimodal mass spectrometry data in MZmine 3. *Nat Biotechnol* **41**, 447–449 (2023). https://doi.org/10.1038/s41587-023-01690-2
24 Williams, A. J. *et al.* The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. *J Cheminform* **9**, 61 (2017). https://doi.org/10.1186/s13321-017-0247-6
25 Glüge, J., McNeill, K. & Scheringer, M. Getting the SMILES right: identifying inconsistent chemical identities in the ECHA database, PubChem and the CompTox Chemicals Dashboard. *Environmental Science: Advances* **2**, 612–621 (2023). https://doi.org/10.1039/D2VA00225F
26 Sanderson, E. Multivariable Mendelian Randomization and Mediation. *Cold Spring Harb Perspect Med* **11** (2021). https://doi.org/10.1101/cshperspect.a038984
27 Talavera Andújar, B. *et al.* Exploring environmental modifiers of LRRK2-associated Parkinson's disease penetrance: An exposomics and metagenomics pilot study on household dust. *Environ Int* **194**, 109151 (2024). https://doi.org/10.1016/j.envint.2024.109151
28 Huang, Z. *et al.* Longitudinal Mapping of Personal Biotic and Abiotic Exposomes and Transcriptome in Underwater Confined Space Using Wearable Passive Samplers. *Environ Sci Technol* **58**, 5229–5243 (2024). https://doi.org/10.1021/acs.est.3c09379
29 Ewald, J. D. *et al.* Web-based multi-omics integration using the Analyst software suite. *Nat Protoc* **19**, 1467–1497 (2024). https://doi.org/10.1038/s41596-023-00950-4
30 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 (2021). https://doi.org/10.1093/nar/gkab394
31 Lu, Y. *et al.* MicrobiomeAnalyst 2.0: comprehensive statistical, functional and integrative analysis of microbiome data. *Nucleic Acids Res* **51**, W310–W318 (2023). https://doi.org/10.1093/nar/gkad407
32 Liu, P. *et al.* ExpressAnalyst: A unified platform for RNA-sequencing analysis in non-model species. *Nat Commun* **14**, 2995 (2023). https://doi.org/10.1038/s41467-023-38785-y
33 Srigboh, R. K. *et al.* Multiple elemental exposures amongst workers at the Agbogbloshie electronic waste (e-waste) site in Ghana. *Chemosphere* **164**, 68–74 (2016). https://doi.org/10.1016/j.chemosphere.2016.08.089
34 Vanweert, F., Schrauwen, P. & Phielix, E. Role of branched-chain amino acid metabolism in the pathogenesis of obesity and type 2 diabetes-related metabolic disturbances BCAA metabolism in type 2 diabetes. *Nutr Diabetes* **12**, 35 (2022). https://doi.org/10.1038/s41387-022-00213-3
35 Yu, D. *et al.* The adverse metabolic effects of branched-chain amino acids are mediated by isoleucine and valine. *Cell Metab* **33**, 905–922.e906 (2021). https://doi.org/10.1016/j.cmet.2021.03.025
36 Liu, R. *et al.* Gut microbiome and serum metabolome alterations in obesity and after weight-loss intervention. *Nat Med* **23**, 859–868 (2017). https://doi.org/10.1038/nm.4358
37 Pedersen, H. K. *et al.* Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature* **535**, 376–381 (2016). https://doi.org/10.1038/nature18646
38 Kessner, D., Chambers, M., Burke, R., Agus, D. & Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **24**, 2534–2536 (2008). https://doi.org/10.1093/bioinformatics/btn323

39 Pang, Z. *et al.* Using MetaboAnalyst 5.0 for LC-HRMS spectra processing, multi-omics integration and covariate adjustment of global metabolomics data. *Nat Protoc* **17**, 1735–1761 (2022). https://doi.org/10.1038/s41596-022-00710-w

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

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

42 Wishart, D. S. *et al.* HMDB 5.0: the Human Metabolome Database for 2022. *Nucleic Acids Research* **50**, D622–D631 (2021). https://doi.org/10.1093/nar/gkab1062

43 Mohammed Taha, H. *et al.* The NORMAN Suspect List Exchange (NORMAN-SLE): facilitating European and worldwide collaboration on suspect screening in high resolution mass spectrometry. *Environ Sci Eur* **34**, 104 (2022). https://doi.org/10.1186/s12302-022-00680-6

44 Bourgon, R., Gentleman, R. & Huber, W. Independent filtering increases detection power for high-throughput experiments. *Proc Natl Acad Sci U S A* **107**, 9546–9551 (2010). https://doi.org/10.1073/pnas.0914005107

45 Chi, Y. *et al.* Constructing a consensus serum metabolome. *bioRxiv*, 2025.2005.2007.652782 (2025). https://doi.org/10.1101/2025.05.07.652782

46 Mahieu, N. G. & Patti, G. J. Systems-Level Annotation of a Metabolomics Data Set Reduces 25 000 Features to Fewer than 1000 Unique Metabolites. *Anal Chem* **89**, 10397–10406 (2017). https://doi.org/10.1021/acs.analchem.7b02380

47 Wei, R. *et al.* Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. *Sci Rep* **8**, 663 (2018). https://doi.org/10.1038/s41598-017-19120-0

48 Li, B. *et al.* Performance Evaluation and Online Realization of Data-driven Normalization Methods Used in LC/MS based Untargeted Metabolomics Analysis. *Sci Rep* **6**, 38881 (2016). https://doi.org/10.1038/srep38881

49 Kim, J. *et al.* Smoking and passive smoking increases mortality through mediation effect of cadmium exposure in the United States. *Sci Rep* **13**, 3878 (2023). https://doi.org/10.1038/s41598-023-30988-z

50 Xu, M. Y. *et al.* Metabolomics analysis and biomarker identification for brains of rats exposed subchronically to the mixtures of low-dose cadmium and chlorpyrifos. *Chem Res Toxicol* **28**, 1216–1223 (2015). https://doi.org/10.1021/acs.chemrestox.5b00054

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

52 Ritz, C., Baty, F., Streibig, J. C. & Gerhard, D. Dose-Response Analysis Using R. *PLoS One* **10**, e0146021 (2015). https://doi.org/10.1371/journal.pone.0146021

53 Akbar, A. *et al.* Exogenous menadione sodium bisulphite alleviates detrimental effects of alkaline stress on wheat (Triticum aestivum L.). *Physiol Mol Biol Plants* **28**, 1889–1903 (2022). https://doi.org/10.1007/s12298-022-01250-z

54 Rasheed, R., Arslan Ashraf, M., Kamran, S., Iqbal, M. & Hussain, I. Menadione sodium bisulphite mediated growth, secondary metabolism, nutrient uptake and oxidative defense in okra (Abelmoschus esculentus Moench) under cadmium stress. *J Hazard Mater* **360**, 604–614 (2018). https://doi.org/10.1016/j.jhazmat.2018.08.043

55 Rebrin, I. & Sohal, R. S. Pro-oxidant shift in glutathione redox state during aging. *Adv Drug Deliv Rev* **60**, 1545–1552 (2008). https://doi.org/10.1016/j.addr.2008.06.001

56 Uthayakumar, B. *et al.* Age-associated change in pyruvate metabolism investigated with hyperpolarized (13) C-MRI of the human brain. *Hum Brain Mapp* **44**, 4052–4063 (2023). https://doi.org/10.1002/hbm.26329

57 Chang, L., Zhou, G. & Xia, J. mGWAS-Explorer 2.0: Causal Analysis and Interpretation of Metabolite-Phenotype Associations. *Metabolites* **13** (2023). https://doi.org/10.3390/metabo13070826

58 Elsworth, B. *et al.* The MRC IEU OpenGWAS data infrastructure. *bioRxiv*, 2020.2008.2010.244293 (2020). https://doi.org/10.1101/2020.08.10.244293
59 Hemani, G., Tilling, K. & Davey Smith, G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS Genet* **13**, e1007081 (2017). https://doi.org/10.1371/journal.pgen.1007081
60 Burgess, S. & Thompson, S. G. Interpreting findings from Mendelian randomization using the MR-Egger method. *Eur J Epidemiol* **32**, 377–389 (2017). https://doi.org/10.1007/s10654-017-0255-x
61 Ito, K., Bahry, M. A., Hui, Y., Furuse, M. & Chowdhury, V. S. Acute heat stress up-regulates neuropeptide Y precursor mRNA expression and alters brain and plasma concentrations of free amino acids in chicks. *Comp Biochem Physiol A Mol Integr Physiol* **187**, 13–19 (2015). https://doi.org/10.1016/j.cbpa.2015.04.010
62 Shibata, T., Takaguri, A., Ichihara, K. & Satoh, K. Inhibition of the TNF-α-induced serine phosphorylation of IRS-1 at 636/639 by AICAR. *J Pharmacol Sci* **122**, 93–102 (2013). https://doi.org/10.1254/jphs.12270fp
63 Awazawa, M. *et al.* A microRNA screen reveals that elevated hepatic ectodysplasin A expression contributes to obesity-induced insulin resistance in skeletal muscle. *Nat Med* **23**, 1466–1473 (2017). https://doi.org/10.1038/nm.4420
64 Avsec, Z. *et al.* Advancing regulatory variant effect prediction with AlphaGenome. *Nature* **649**, 1206–1218 (2026). https://doi.org/10.1038/s41586-025-10014-0
65 Lawlor, D. A., Tilling, K. & Davey Smith, G. Triangulation in aetiological epidemiology. *Int J Epidemiol* **45**, 1866–1886 (2016). https://doi.org/10.1093/ije/dyw314
66 Elsworth, B. & Gaunt, T. R. MELODI Presto: a fast and agile tool to explore semantic triples derived from biomedical literature. *Bioinformatics* **37**, 583–585 (2020). https://doi.org/10.1093/bioinformatics/btaa726
67 Broeckling, C. D. *et al.* Current Practices in LC-MS Untargeted Metabolomics: A Scoping Review on the Use of Pooled Quality Control Samples. *Anal Chem* **95**, 18645–18654 (2023). https://doi.org/10.1021/acs.analchem.3c02924
68 Adusumilli, R. & Mallick, P. Data Conversion with ProteoWizard msConvert. *Methods Mol Biol* **1550**, 339–368 (2017). https://doi.org/10.1007/978-1-4939-6747-6_23
69 Kuhl, C., Tautenhahn, R., Böttcher, 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 (2012). https://doi.org/10.1021/ac202450g
70 Giné, R. *et al.* HERMES: a molecular-formula-oriented method to target the metabolome. *Nat Methods* **18**, 1370–1376 (2021). https://doi.org/10.1038/s41592-021-01307-z
71 Li, S. & Zheng, S. Generalized Tree Structure to Annotate Untargeted Metabolomics and Stable Isotope Tracing Data. *Anal Chem* **95**, 6212–6217 (2023). https://doi.org/10.1021/acs.analchem.2c05810
72 Li, Y. *et al.* Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification. *Nat Methods* **18**, 1524–1531 (2021). https://doi.org/10.1038/s41592-021-01331-z
73 Tsugawa, H. *et al.* MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat Methods* **12**, 523–526 (2015). https://doi.org/10.1038/nmeth.3393
74 Sumner, L. W. *et al.* Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* **3**, 211–221 (2007). https://doi.org/10.1007/s11306-007-0082-2
75 van den Berg, R. A., Hoefsloot, H. C., Westerhuis, J. A., Smilde, A. K. & van der Werf, M. J. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics* **7**, 142 (2006). https://doi.org/10.1186/1471-2164-7-142

76 Farmahin, R. *et al.* Recommended approaches in the application of toxicogenomics to derive points of departure for chemical risk assessment. *Arch Toxicol* **91**, 2045–2065 (2017). https://doi.org/10.1007/s00204-016-1886-5
77 Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *Elife* **7** (2018). https://doi.org/10.7554/eLife.34408
78 Chang, L., Zhou, G., Ou, H. & Xia, J. mGWAS-Explorer: Linking SNPs, Genes, Metabolites, and Diseases for Functional Insights. *Metabolites* **12** (2022). https://doi.org/10.3390/metabo12060526