Journal of Cheminformatics

SOFTWARE

Open Access

TraceMetrix: a traceable metabolomics interactive analysis platform

Wei Chen[1], Yanpeng An[2], Ziru Chen[3], Ruijin Luo[4], Qinwei Lu[2], Cong Li[2], Chenhan Zhang[2], Qingxia Huang[2], Qinsheng Chen[2], Lianglong Zhang[2], Xiaoxuan Yi[2], Yixue Li[1,3*], Huiru Tang[2*] and Guoqing Zhang[1*]

Abstract

Metabolomics data analysis is a multifaceted process often constrained by limited data sharing and a lack of transparency, which hinders reproducibility of results. While existing bioinformatics tools address some of these challenges, achieving greater simplicity and operational clarity remains essential for fully leveraging the potential of metabolomics. Here, we introduce TraceMetrix, a web-based platform designed for interactive traceability in metabolomics data analysis. TraceMetrix provides a flexible management system for both raw and derived data, enabling comprehensive tracking of file origins and destinations throughout the whole analysis pipeline. The platform documents the software and parameters used across four key modules, from raw data preprocessing, data cleaning, statistical analysis to functional analysis, enabling users to easily track critical factors influencing result accuracy. By mapping upstream and downstream relationships for nearly 19 analytical functions, TraceMetrix ensures end-to-end traceability, viewable interactively online or exportable as detailed reports. To address the limitations of single-machine environments in processing large-scale datasets, TraceMetrix is deployed on a high-performance computing cluster for efficient batch processing. Using a non-targeted metabolomics dataset, we demonstrated its traceability function to optimize parameter selection, successfully reproducing the analysis process and validating the original study's findings. TraceMetrix integrates traceability across data, software, and processes, significantly enhancing reproducibility in metabolomics research. The platform supports diverse applications and is freely available at https://www.biosino.org/tracemetrix.

Scientific contribution

TraceMetrix introduces a novel web-based platform for metabolomics data analysis, offering interactive traceability that ensures comprehensive tracking of the entire analysis process. Unlike existing tools, TraceMetrix enables traceability of data files, processes (analysis methods), and parameters through efficient data management, significantly enhancing transparency and reproducibility. Additionally, by deploying on high-performance computing clusters, it addresses the challenges of large-scale metabolomics data analysis.

Keywords Metabolomics, Interactive analysis platform, Traceable, Reproducibility

*Correspondence:
Yixue Li
li_yixue@gzlab.ac.cn
Huiru Tang
huiru_tang@fudan.edu.cn
Guoqing Zhang
gqzhang@sinh.ac.cn
Full list of author information is available at the end of the article

Introduction

Metabolomics, the systematic study of small-molecule metabolites within biological systems, has become an essential approach in modern life sciences, offering valuable insights into cellular metabolism and physiological states. Its applications span diverse fields, ranging from unraveling disease mechanisms and advancing clinical diagnostics [1–3] to revolutionizing plant science and agricultural research [4] and food science investigations [5]. However, as metabolomics technologies advance, ensuring the reproducibility of results has emerged as a critical challenge. This issue is particularly evident in mass spectrometry-based techniques, such as liquid chromatography-mass spectrometry (LC–MS), which has gained widespread use due to its high sensitivity, broad coverage of metabolites, and relatively straightforward operational procedures. Despite its advantages, the increasing scale and complexity of LC–MS data create significant barriers to ensuring reproducibility, as minor variations in analysis parameters at each step can significantly impact the results.

Over the years, the metabolomics community has developed a variety of computational tools to address these challenges. These tools can be broadly categorized into functional groups such as: (1) feature extraction tools, including XCMS [6], MZmine [7], MS-DIAL [8], and MAVEN [9]; (2) feature annotation tools such as MS-FINDER [10], MetDNA [11] and SIRIUS [12]; (3) data cleaning tools represented by MetFlow [13] and MetaboDiff [14]; (4) statistical analysis tools including R packages muma [15] and ropls [16], (5) functional interpretation tools like FILLA [17] and MetPA [18]. Additionally, integrated platforms such as MetaboAnalyst [19], and XCMS Online [20] aim to streamline the entire metabolomics data analysis workflow. While these tools offer significant functionality, limited interoperability between them and the challenges surrounding data standardization remain substantial hurdles. These issues can hinder data integration, reduce workflow efficiency, and ultimately compromise the reproducibility of results.

Reproducibility, a cornerstone of scientific research, is particularly difficult in metabolomics [21]. Nuclear magnetic resonance (NMR) is highly regarded for its excellent intrinsic reproducibility [22], while LC–MS techniques are comparatively limited in this regard. These limitations primarily stem from the poor intrinsic reproducibility of chromatography and platform-dependent inter-batch variations, making cross-laboratory data comparisons extremely challenging. Additionally, metabolomics datasets often involve large cohorts, generating complex spectra from thousands of metabolites. This massive scale makes single-machine processing unfeasible, and further complicating reproducibility efforts [23]. Some LC–MS data analysis tools,such as TidyMass [24] and asari [25], have been developed to improve traceability by documenting processing steps and allowing for easier verification of results. However, these solutions typically require programming expertise, limiting their accessibility to many researchers. Moreover, the lack of standardized documentation and poor record-keeping practices have contributed to what has been termed a "reproducibility crisis" in the field.

To address these issues, we present TraceMetrix, a web-based platform designed to enhance traceability and reproducibility in LC–MS based untargeted metabolomics data analysis. TraceMetrix provides an intuitive, flexible management system for organizing raw and processed data, analysis workflows, and the parameters used across four key modules, which span 19 distinct analysis functions. The platform incorporates a novel three-pronged approach to traceability—file traceability, software traceability, and process traceability—which allows real-time, interactive tracking of all analysis steps. This comprehensive traceability ensures that every decision made during the analysis is well-documented, allowing for full reproducibility of results. Additionally, TraceMetrix is deployed on a high-performance computing infrastructure, enabling efficient processing of large datasets through its one-click batch analysis capabilities. By integrating these features into a unified platform, TraceMetrix aims to provide an accessible, user-friendly, traceable, reproducible, and efficient solution to support the growing needs of metabolomics research and accelerate discovery in the field.

Implementation

Data analysis pipeline

The workflow of TraceMetrix (Fig. 1) begins with raw data preprocessing (Fig. 1A) where mass spectrometry (MS) raw data from various vendors are automatically converted into standardized formats utilizing ProteoWizard's MSConvert [26] and ThermoRawFileParser [27]. Currently, we support four vendor formats (.d,.raw,.wiff, and.wiff2) for conversion to open formats (.mzML and.mzXML).The standardized data then undergoes peak detection, peak alignment, and peak grouping via XCMS, enabling the identification of key features. Comprehensive metabolite annotation follows, using multiple metabolite databases, including those providing only accurate mass information (e.g., Blood Exposome [28], KEGG [29], T3DB [30], DrugBank [31]) as well as those providing both accurate mass and MS/MS spectral information (HMDB [32], MassBank [33], and MoNA [33]). TraceMetrix adheres to the Metabolomics Standards Initiative (MSI) guidelines [34], implementing both adduct annotation and database searches through accurate mass

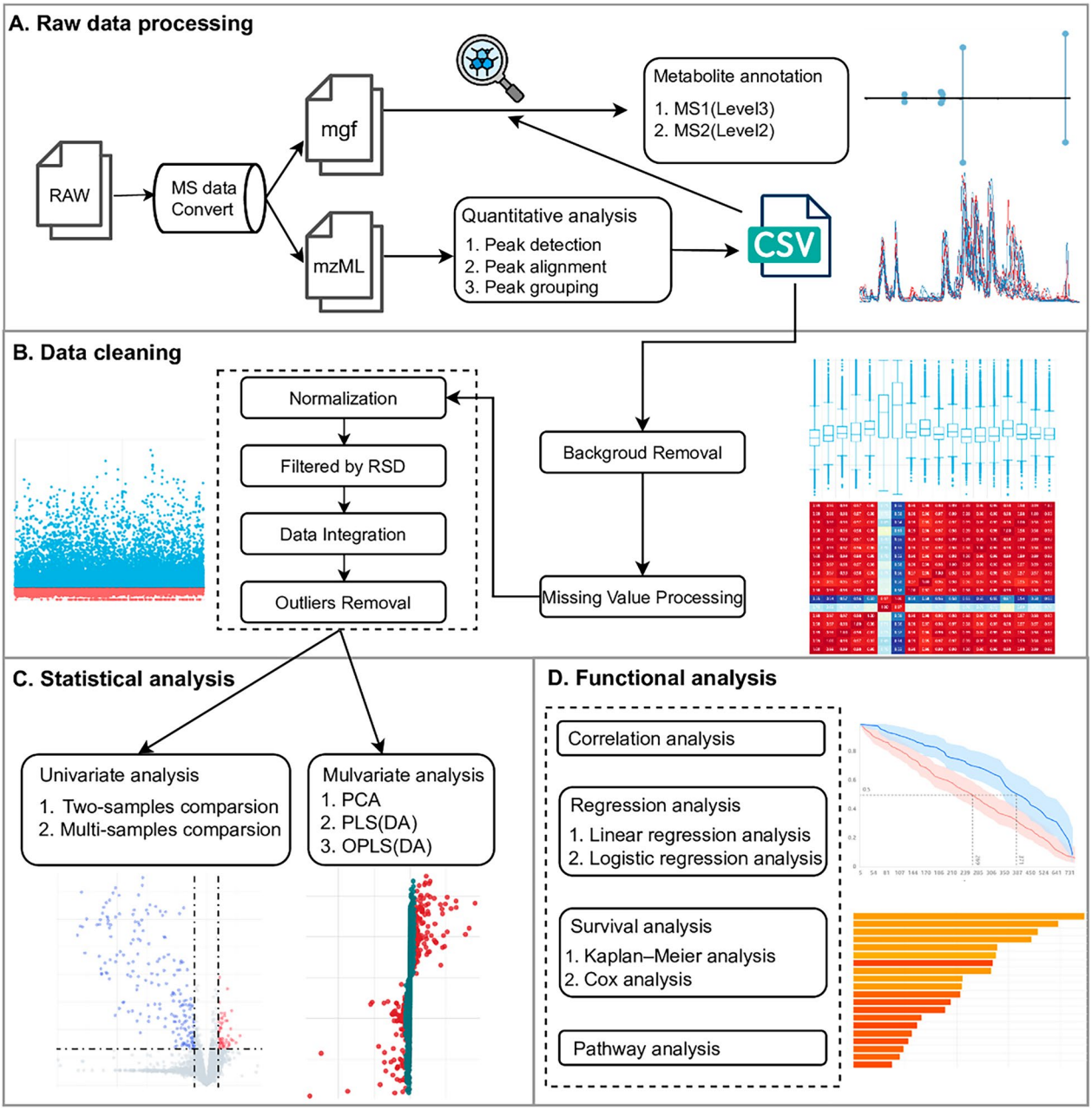

Fig. 1 Analysis workflow of TraceMetrix. **A** Raw data processing is used for peak extraction and metabolite annotation from LC–MS raw data. When converting raw data to mzML format, centroiding is performed by default. MS1(level 3) refers to metabolite identification using only the feature's accurate mass matched with public databases, while MS2 (level 2) involves combining accurate mass with MS/MS spectral information for identification. **B** Different data cleaning methods are employed for quality control of metabolomics data. **C** Statistical analyses are used to screen for differential metabolites. **D** Functional analysis is used for the integrated analysis of metabolomics data with other data, such as clinical information, to explore biological significance

measurement and tandem mass spectrometry spectral matching.

Following preprocessing, data cleaning (Fig. 1B) is performed to address variability and enhance the reliability of subsequent analyses. This step includes background removal, missing value processing, data normalization, filtered by the relative standard deviation (RSD), data integration and outlier removal. Notably, each processing step provides extensive visualizations to display data results, enabling users to interactively assess data quality

through interactive figures. During "Missing Value Processing", users can first set various parameters within the same results interface. They can then compare how different imputation methods impact subsequent statistical analyses, with the effects immediately shown through updated visualizations. Additionally, in the "Outlier Removal" step, users can identify anomalous points through interactive figures. By referring to the actual context of the samples, they can assess whether these are true outliers and make real-time decisions to either exclude or retain the corresponding data points.

The statistical analysis module (Fig. 1C) aims to identify potentially differential metabolites through both univariate analysis (including two-sample and multi-sample comparisons) and multivariate analysis (including PCA, PLS-DA, and OPLS-DA). For univariate analysis, p-values are further adjusted using both False Discovery Rate (FDR) and Bonferroni correction methods. These analyses provide insights into the variability and significance of metabolic differences across conditions.

Finally, functional analysis (Fig. 1D) supports biological interpretation by integrating metabolomics data with other data, such as clinical or phenotypic information. This module includes correlation analysis, regression analysis, survival analysis, and pathway enrichment analysis, enabling researchers to uncover the biological significance of their findings and connect metabolite changes to broader biological processes.

Data management

To support the iterative and data-intensive nature of metabolomics research, TraceMetrix offers a comprehensive data management system that allows researchers to efficiently organize, trace, and document both experimental and derived data (Fig. 2). The system is structured hierarchically into four levels: project, experiment, analysis module, and task. A single project can contain multiple experimental datasets, each comprising a full suite of analysis modules. Within each module, users can define multiple tasks to support flexible workflows, such as

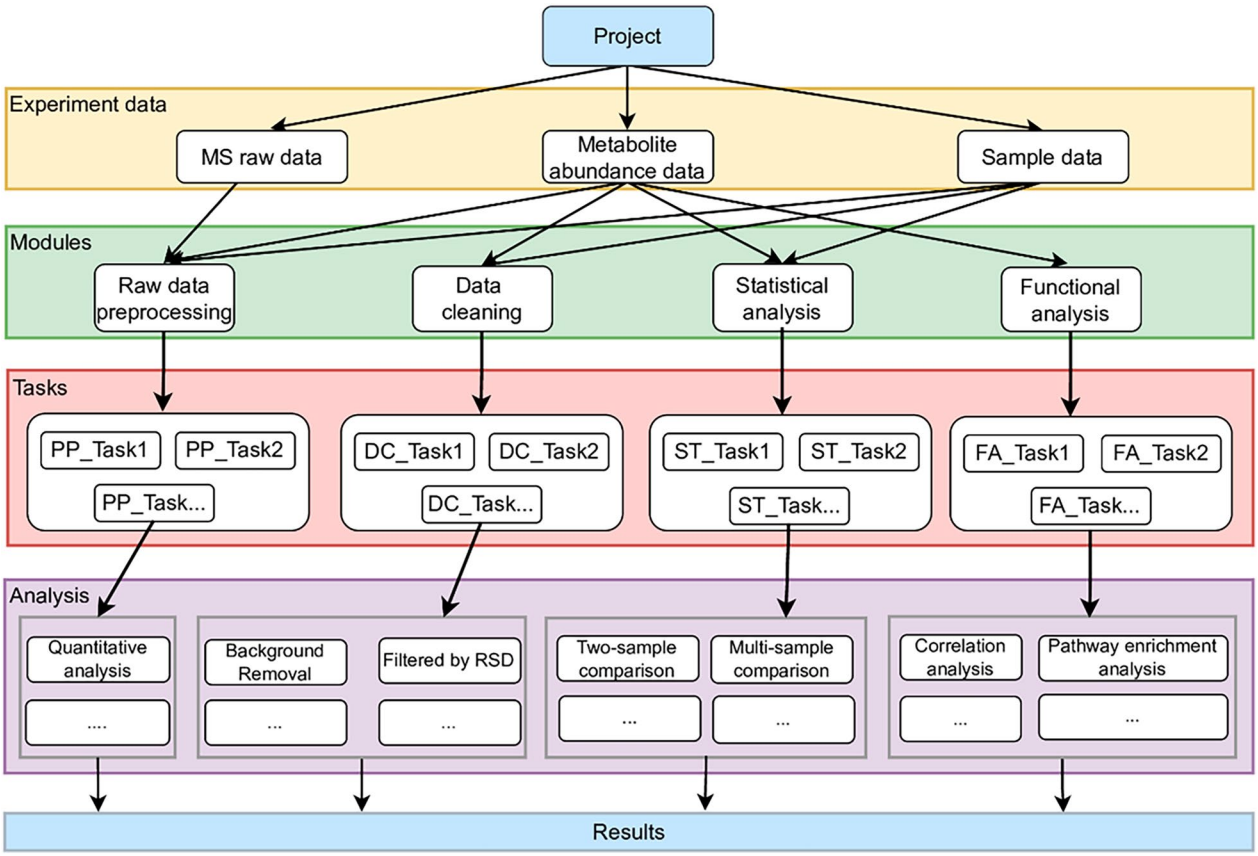

Fig. 2 Data management in TraceMetrix**.** Data management architecture of TraceMetrix. A project contains different types of metabolomics data. These data types feed into four analytical modules (raw data preprocessing, data cleaning, statistical analysis, and functional analysis) based on their specific applications. Different results are generated under various tasks using different parameters. The bottom layer details the specific analyses available within each module. All analysis results are systematically tracked and stored, ensuring complete analytical reproducibility

varying parameter settings or exploring alternative analysis strategies.

Upon project creation, users can upload diverse types of data, such as raw mass spectrometry (MS) data, metabolite abundance matrices, and sample metadata. Each data type supports multiple uploads. For instance, raw MS data acquired under different ionization modes can be uploaded separately within the same project and processed independently. Each experimental dataset includes the full suite of TraceMetrix analysis modules, which are interconnected to support end-to-end metabolomics workflows—from raw data preprocessing and peak extraction to downstream statistical analysis and data mining.

To support flexible and reproducible analysis, we introduce the concept of tasks within each analysis module. A task represents a specific execution instance of an analysis step, allowing users to test different parameter settings or apply alternative analytical strategies within the same module. All tasks are integrated and visualized on the module result page corresponding to the specific experimental dataset, enabling users to directly compare the outcomes of different tasks. For each task, users can access detailed information including input data, selected parameters, and output results.

To enhance the traceability and centralized management of both the analytical process and results, all uploaded data, analysis steps, and outputs associated with an experimental dataset are organized within a unified interface. Users can easily navigate between different analysis modules, inspect task-specific results, and trace the full analytical provenance—including inputs, parameters, and outputs—of any given task.

Furthermore, TraceMetrix supports the management of various types of experimental datasets. Raw MS data are compatible with a wide range of formats, including vendor-specific formats (e.g., Thermo, Agilent, Waters) and open standard formats, ensuring broad compatibility with mainstream mass spectrometry platforms while supporting efficient storage of large-scale datasets. Metabolite abundance matrices are uploaded in standardized CSV format, containing both chemical annotations and quantitative values. Sample metadata files include essential sample identifiers, experimental design attributes (e.g., group, sample type), clinical information (e.g., age, sex, disease status), and technical parameters (e.g., acquisition batch, internal standards, injection order). These data types are strategically utilized across different analytical modules: raw MS data and sample metadata serve as inputs for upstream preprocessing, while metabolite abundance matrices are applied in downstream steps such as data cleaning, statistical analysis, and functional interpretation.

Data traceability

Traceability framework

TraceMetrix implements a comprehensive traceability framework based on three interrelated dimensions: task dependency, parameter configuration, and data lineage. The task dependency dimension captures the logical relationships among analytical steps; the parameter configuration dimension records all user-defined or default parameters; and the data lineage dimension preserves the provenance and transformation paths of all input and output data. Together, these dimensions enable TraceMetrix to maintain a coherent, modular, and fully traceable analytical ecosystem. By interconnecting different modules through these unified dimensions, the system ensures that any step in the analysis can be transparently linked to its origins and downstream consequences, thereby supporting both retrospective auditing and forward traceability.

To demonstrate how this framework operates in practice, we use the data cleaning module as a representative example (Fig. 3A). First, after the raw data is analyzed in the Raw data preprocessing module, a series of results are generated, among which data.csv and sample_info.csv are used as input data of the data cleaning module and enter the data cleaning module. The data cleaning module contains multiple analysis methods. The user first uses background removal and then performs missing value processing. At this time, missing value processing and background removal establish a dependency relationship. Each processing step is associated with a specific parameter configuration, which is automatically captured and stored. The system not only tracks the execution path of the analysis method (background removal to missing value processing), but also maintains a complete record of the input and output file relationship and parameter specifications of each step. It should be emphasized that the framework has good flexibility. Users can selectively perform data cleaning steps according to specific analysis needs without strictly following the preset complete process. The system can still accurately maintain the corresponding dependency and traceability records.

Based on the records of the above task links and execution information, TraceMetrix can achieve powerful traceability. The system uses the dependencies between tasks to build a bidirectional data lineage track based on Experiment Data across analysis modules, supporting tracing back to the source of the data from any node, or tracing down to all derived results. As shown in Fig. 3A, the data flow progresses from raw preprocessing through data cleaning to downstream statistical and functional analysis, and each conversion point maintains complete source information. Users can view the detailed information of each task in the traceability system, including

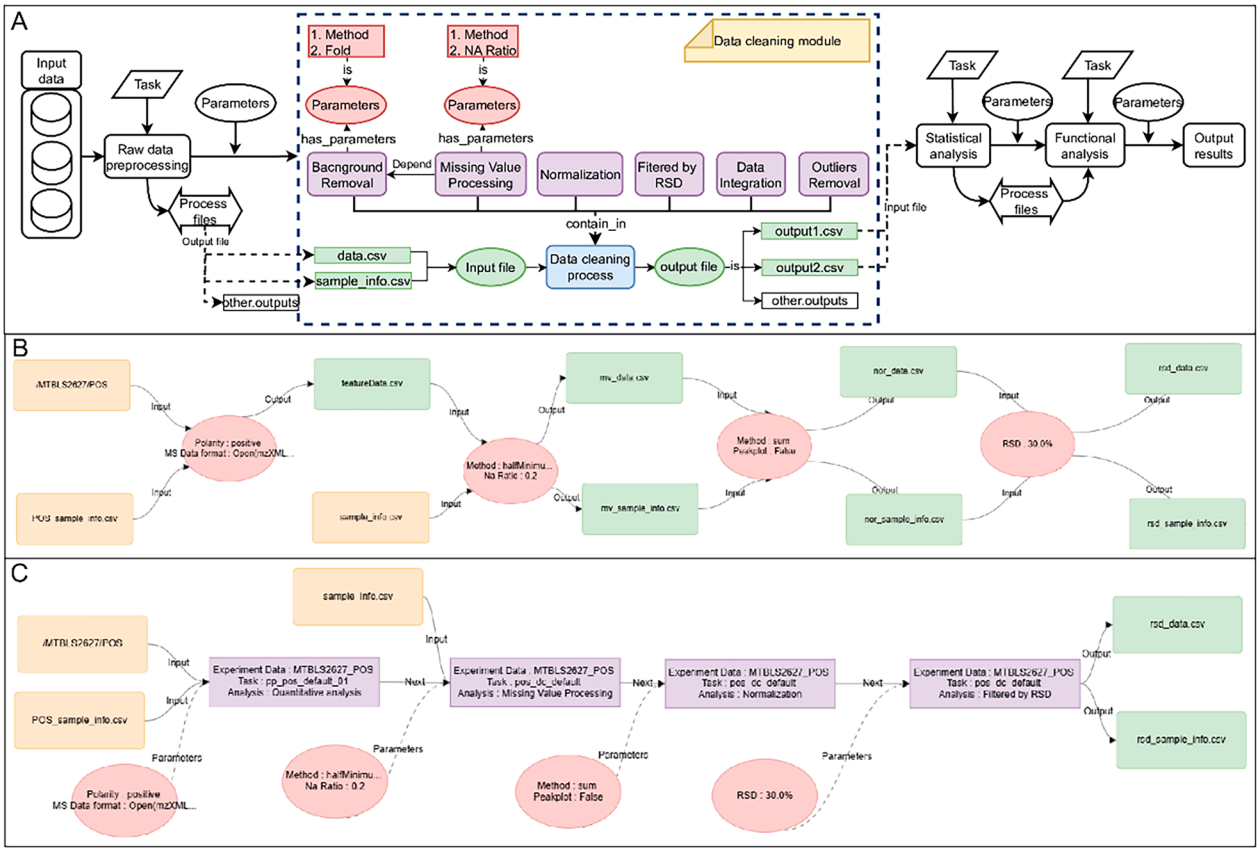

Fig.3 Three traceability in TraceMetrix**. A** The data traceability is categorized into three types: file traceability, process traceability, and software traceability, based on analysis data, analysis parameters, process files, and the analysis process. **B** A web-based example of file traceability. **C** A web-based example of process traceability

input and output files, algorithms used, parameter configurations, etc., to provide complete transparency for the analysis process.

Through this complete traceability system, TraceMetrix provides traceability support for metabolomics research in three key dimensions. In terms of file traceability, the system fully records the generation and conversion process of data files; in terms of process traceability, the system tracks every step of the analysis process in detail; in terms of software traceability, the system records the algorithms, tools and parameter configurations used. The system ensures complete transparency of the analysis process and reproducibility of the results, providing important technical guarantees for the reliability of scientific research.

File traceability

In metabolomics research, raw data passes through several analysis steps before reaching the final result, with various files generated to support each stage of the analysis. For example, raw mass spectrometry data files are used in data preprocessing, two-dimensional matrix data files are generated in quantitative analysis, and subsequent statistical analyses produce result files containing key statistical information and visualizations. TraceMetrix's file traceability stores and records data files from each analysis step, while using links like "input" and "output" to create a continuous file chain between different analysis steps. This chain clearly shows how the result files were transformed step by step from raw data, including the intermediate file processing (Fig. 3B). Furthermore, the web-based interactive file traceability feature allows users to click on any data file involved in the traceability process, enabling quick navigation to the analysis step where the file is located.

Software traceability

The diversity of methods in metabolomics analysis introduces a large number of parameters, which can significantly influence the outcomes. For example, in raw data preprocessing, the signal intensity threshold in peak detection directly affects the identification of valid peaks. Similarly, data cleaning methods, such as missing value imputation, can impact subsequent statistical analyses.

Tools like XCMS-based quantitative analysis require a multitude of parameters for mass spectrometry peak identification and alignment, increasing both complexity and the risk of reducing reproducibility.

To address these challenges, TraceMetrix ensures comprehensive software traceability. It visually displays analysis methods and their parameters at each step, allowing users to track the settings used. Additionally, the 'report' function enables a one-click export of all analysis parameters and steps into a detailed document (Supplementary Fig. 1). This not only simplifies parameter management but also facilitates the reproducibility and transparency of the analysis process.

Process traceability

TraceMetrix's process traceability records the entire analysis workflow, from raw data preprocessing to biological interpretation (Fig. 3C). It systematically documents the sequence of functions and their upstream or downstream relationships. For instance, it shows how data cleaning methods influence differential metabolite screening and how metabolite annotation results contribute to pathway enrichment analysis. The interactive flow diagram provided by TraceMetrix allows researchers to visualize the analysis process, helping them understand key steps and identify optimization opportunities. This structured record enhances the transparency and reproducibility of the research, facilitating validation and data sharing.

Other features

Resuming analysis

Metabolomics data analysis is a multi-step, interdependent process, where each stage builds on the previous one. However, disruptions or errors typically require restarting the entire workflow. TraceMetrix overcomes this challenge through a robust data management system and modular design, which logs and preserves every operation. The platform's "Create next step" feature (Supplementary Fig. 3) ensures seamless progression between stages. Additionally, users can select and build upon previous results (Supplementary Fig. 3), maintaining continuity across modules. This approach improves the resilience and efficiency of the analysis pipeline, minimizing the impact of interruptions and streamlining the workflow.

Historical analysis

To enhance reproducibility, TraceMetrix includes a "Historical Analysis" feature that records and saves results for each parameter adjustment. Users can easily review and compare outcomes under different settings, facilitating more flexible and efficient analysis (Supplementary Fig. 4). This feature ensures users can accurately trace and compare results across various conditions, aiding in deeper insights and better decision-making.

One-click analysis

TraceMetrix offers a one-click analysis feature that automates the data cleaning and statistical analysis workflows. Users can perform metabolomics data analysis with a one-click function using ether the platform's recommended parameters or parameters they have created. The intuitive interface enables users to process and analyze data without requiring advanced technical knowledge. By streamlining data upload and parameter selection, this functionality accelerates analysis, enabling researchers to handle larger datasets more efficiently (Supplementary Fig. 5). The integration of automated workflows with a user-friendly interface optimizes resource allocation, allowing researchers to focus more on interpretation and hypothesis generation rather than the intricacies of data processing.

System architecture and deployment

TraceMetrix employs a browser/server architecture as a web-based analytical tool that supports access through different browsers, including Chrome and Safari. The backend is built with Spring Boot, while the frontend utilizes Vue and ElementUI frameworks. The underlying analytical workflows are implemented in R (3.6.1) and encapsulated within Singularity containers to ensure version stability. Data storage is managed through MongoDB, with graphical data visualization powered by ECharts. The system can be deployed on high-performance computing clusters with job scheduling managed by Slurm and runs on CentOS (7.9.2009). For large-scale cohort data, FTP protocols (such as XFTP) ensure rapid and stable data upload to user accounts.

Results

Interactive visualization

In TraceMetrix, interactive visualization is a key feature throughout the analysis process, enabling users to explore and modify parameters for reanalysis. Hover functionality provides detailed information when users position the cursor over data elements (Fig. 4).

In metabolite annotation, TraceMetrix enhances efficiency by allowing users to select metabolites from the annotation table and instantly view their MS/MS spectral matches from integrated databases (Supplementary Fig. 2). This eliminates manual searching, reduces human error, and accelerates workflow. For data cleaning, TraceMetrix provides comprehensive quality assessment tools with interactive visualizations. The scatter plot for missing value assessment (Fig. 4A) dynamically shows each sample's missing value

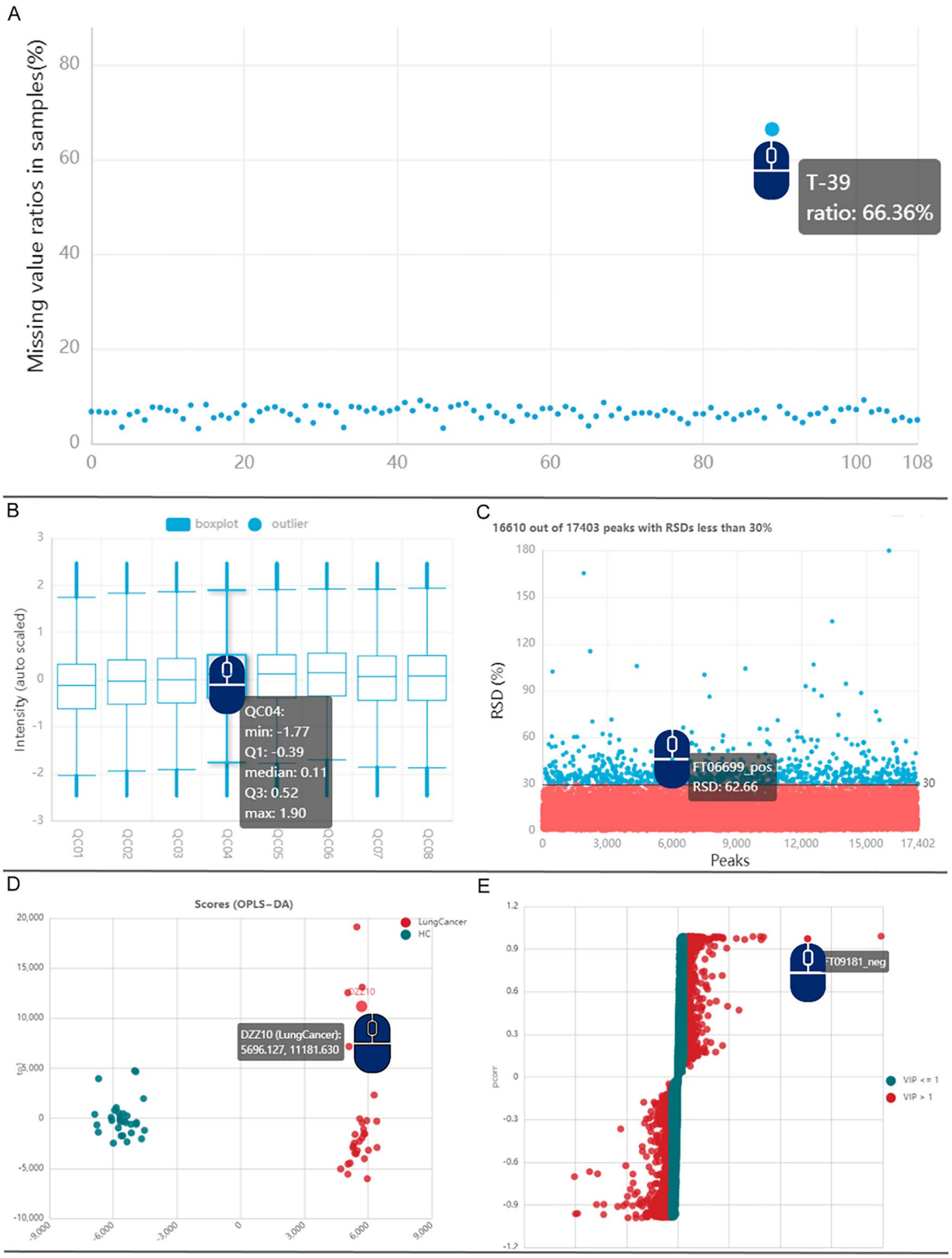

Fig. 4 Examples of interactive operation in TraceMetrix. **A** Interactive scatter plot of missing values summary of all samples. **B** Interactive boxplot of feature accumulation profile of quality control (QC) samples. **C** Interactive scatter plot of RSD values summary of all features. Blue dots represent 30% or more, while red dots represent less than 30%. **D** Interactive plot of OPLS-DA score distribution, showing the sample distribution for different groups. Hovering over the points displays the specific sample names and scores. **E** Interactive plot of OPLS-DA 's VIP values

percentage, allowing quick identification of problematic samples. The interactive box plot for QC samples (Fig. 4B) displays metabolite distributions, enabling real-time exploration of sample differences and batch effects. The RSD analysis (Fig. 4C) allows users to dynamically explore metabolites exceeding RSD thresholds, facilitating more efficient screening. In univariate analysis, the interactive box plot (Supplementary Fig. 1) visualizes data distribution, enabling real-time exploration of statistical significance. Both the OPLS-DA score plot (Fig. 4D) and VIP (Variable Importance in Projection) scatter plot (Fig. 4E) allow users to explore sample classification and identify key metabolites, simplifying multivariate analysis interpretation.

Case study

To illustrate the reproducibility and traceability capabilities of TraceMetrix, we present two case studies based on publicly available untargeted LC–MS metabolomics datasets from published research. In both cases, we successfully reconstructed the complete analysis workflows and reproduced the key findings reported in the original publications.

Case study 1

We re-analyzed publicly available non-targeted LC–MS metabolomics data from lung cancer patients, including CTC-positive lung cancer patients and 30 healthy controls (including 8 QC samples), acquired using a UPLC-Triple-TOF–MS-based platform (AB SCIEX, USA) under both positive and negative ionization modes [35]. Through the traceability of TraceMetrix, we successfully reproduced the results reported in the literature (Fig. 5). After uploading the mzXML data files to TraceMetrix, we set the initial parameters for raw data quantification and metabolite identification based on the literature descriptions (Supplementary Table 1). After running the analysis workflow, we found that some of the metabolite identification results did not fully match those in the literature. Through iterative adjustments and re-runs of the analysis workflow, we optimized the parameters and successfully reproduced the main findings.

We compared the classification of differential metabolites in HMDB and performed pathway enrichment analysis in KEGG. The differential metabolites were predominantly classified into lipids, lipid-like molecules, and organic acids (Fig. 5A, Supplementary Table 2). In KEGG pathways, significant enrichment was observed in lipid

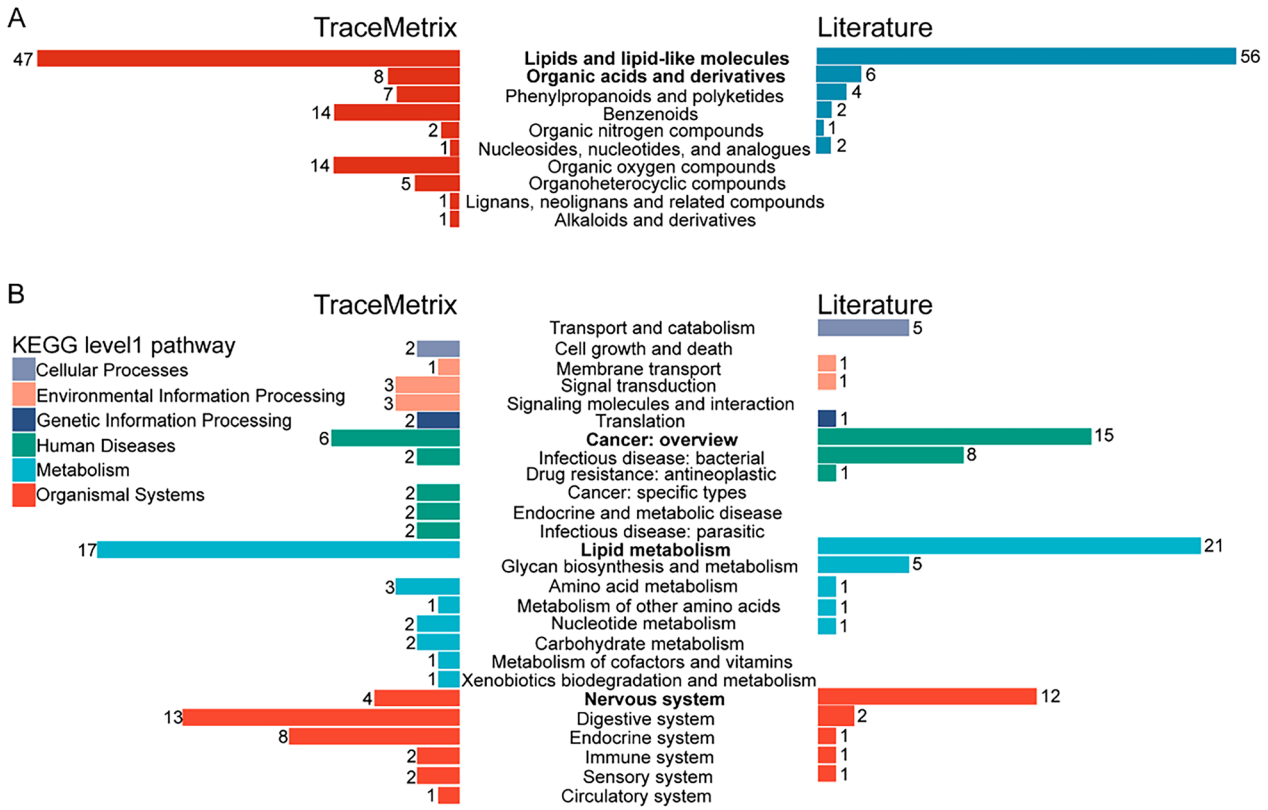

Fig. 5 Comparison of case study results between TraceMetrix and literature. **A** Comparison results of metabolite superclass in HMDB database. **B** Comparison results of KEGG Level 2 Pathway

metabolism and cancer-related pathways (Fig. 5B). Additionally, PCA plots (Supplementary Fig. 7), OPLS-DA score plots, and permutation plots (Supplementary Fig. 6) were generated and compared with the corresponding figures in the literature, showing good consistency.

By utilizing TraceMetrix's traceability feature, we documented the entire analysis process (Supplementary Fig. 8) and exported the complete workflow along with the involved parameters (Supplementary Fig. 9). This ensured the reproducibility and transparency of the analysis, highlighting TraceMetrix's capability in supporting robust and reproducible metabolomics data analysis for broader applications.

Case study 2

We also re-analyzed publicly available untargeted LC–MS metabolomics data from intrahepatic cholestasis of pregnancy (ICP) studies, including serum samples from 57 healthy pregnant women and 52 ICP patients (with 18 QC samples), acquired on AB Sciex TripleTOF 6600 [1] and successfully reproduced the findings reported in the literature using TraceMetrix, establishing a publicly accessible and transparent online traceable workflow. Following data download from MetaboLights (MTBLS2627), we converted the raw data to mzML format using ProteoWizard and uploaded it to TraceMetrix via XFTP. Subsequently, we employed the preprocessing module to extract metabolic features and perform metabolite annotation. Data filtering was conducted using the data cleaning module, including missing value imputation, normalization, and RSD filtering. We then utilized the two-sample comparison function within the statistical analysis module to identify differentially expressed metabolites between the ICP group and normal controls, followed by KEGG pathway analysis in the functional analysis section.

Through ESI + and ESI- modes, we identified 116 differentially expressed metabolites (Supplementary Table 3), including bile acids such as taurocholic acid, glycocholic acid, allocholic acid, and 6-ethylchenodeoxycholic acid; lipid metabolites including sphingosine, sphinganine, and glycerol 3-phosphate; as well as amino acids and their derivatives, carbohydrates, organic acids, and steroid hormones. The KEGG pathway analysis revealed consistent findings with the original publication, including bile secretion, tryptophan metabolism, steroid hormone biosynthesis, and biosynthesis of unsaturated fatty acids (Supplementary Table 4). Additionally, we identified several pathways such as ABC transporters, which are associated with bile acid transport dysfunction [2], and Cushing syndrome, which may be related to hormonal changes during pregnancy [3] (Supplementary Table 4). Furthermore, we generated principal component analysis (PCA) plots, orthogonal partial least squares discriminant analysis (OPLS-DA) score plots, and permutation plots for both ESI + and ESI- modes, which demonstrated good consistency when compared with corresponding figures in the literature (Supplementary Fig. 10).

We leveraged TraceMetrix's traceability functionality to document the entire analytical process (Project Name: MTBLS2627) and exported the complete workflow with associated parameters (Supplementary Fig. 11). This ensures reproducibility and transparency of the analysis, highlighting TraceMetrix's capability in supporting robust and reproducible metabolomics data analysis for broader applications.

Function comparison with other tools

TraceMetrix excels in metabolomics data processing, statistical analysis, functional interpretation, and result visualization, providing a comprehensive and integrated solution for the entire data analysis workflow (Table 1). Unlike other platforms, TraceMetrix supports large-scale LC–MS data analysis without file size limitations, thanks to its external upload tools such as XFTP or FileZilla. This feature makes it well-suited for large datasets, which is a significant advantage over platforms that have file size constraints.

The platform also offers comprehensive quality control features, allowing users to assess data quality at each step. Through its interactive and visual approach, TraceMetrix allows users to dynamically explore data distributions and identify potential issues with missing values, sample variance, and batch effects. This flexibility enhances the overall efficiency and accuracy of data preprocessing.

TraceMetrix's integration of multiple analysis methods, such as univariate and multivariate statistical techniques, allows for seamless exploration of metabolomics data. Users can perform *t*-tests, ANOVA, PCA, PLS-DA, and OPLS-DA analysis directly within the platform, enabling them to derive meaningful insights quickly. Additionally, TraceMetrix supports advanced functional analysis, including pathway enrichment and joint analysis with other data types like clinical and omics data. This multidimensional analysis capability provides deeper insights into the biological context of the data.

Conclusions

TraceMetrix introduces a comprehensive traceability framework that addresses critical challenges in research reproducibility. The platform's innovative three-dimensional traceability system—encompassing file, software, and process tracking—provides unprecedented transparency in metabolomics analytical workflows. This integrated approach not only ensures methodological rigor but also facilitates the validation

Table 1 Comparison of TraceMetrix with other metabolomics tools

Tool name	TraceMetrix	MetaboAnalyst	XCMS Online	W4M	TidyMass
Platform	Web	Web, R	Web	Web	R
Raw data preprocess					
Peak detection	√	√	√	×	√
MS1 feature annotation	√	√	√	√	√
MS2 spectra annotation	√	√	√	×	√
Data cleaning					
Data quality check	√	×	√	×	√
Missing value processing	√	√	√	×	√
Normalization	√	√	√	√	√
Outliers check	√	√	×	×	√
Statistical analysis					
Univariate analysis	√	√	√	√	√
Multivariate analysis	√	√	√	√	√
Functional analysis					
Enrichment analysis	√	√	×	×	√
Correlation analysis	√	√	×	×	√
Regression analysis	√	×	×	×	×
Survival analysis	√	×	×	×	×
High-performance cluster	√	×	√	×	/
Data reproducibility					
Data management	√	×	√	×	/
Data traceability	√	×	×	×	√
Data transparency	√	√	√	√	√
Analysis reports	√	√	√	√	√

Symbols used for feature evaluations with '√' for present, '×' for absent, and '/' for unsuitable.W4M: Workflow4Metabolomics

and reproduction of metabolomics findings, addressing a fundamental need in the field.

A distinguishing feature of TraceMetrix is its robust interactive visualization capabilities, which transform complex metabolomics data into intuitive, explorable representations. This significantly reduces the technical barriers traditionally associated with metabolomics data analysis, enabling researchers to focus primarily on biological interpretation. The integration of real-time parameter adjustment with comprehensive traceability documentation ensures that users can thoroughly explore different methodological parameters without compromising reproducibility. The successful replication of published non-targeted metabolomics results in our case study validates the platform's effectiveness in supporting reproducible research while maintaining analytical rigor. For users requiring high-throughput analysis capabilities, TraceMetrix provides automated metabolomics data analysis workflows. The platform's deployment on a high-performance computing cluster significantly enhances its capability to process large-scale metabolomics datasets efficiently.

While TraceMetrix offers substantial advantages, we acknowledge certain limitations, particularly in the scope of analytical method integration. Future developments will focus on implementing advanced machine learning algorithms for automated parameter optimization, expanding the range of supported analytical techniques, and developing more sophisticated data integration capabilities for multi-omics analysis. We also plan to incorporate public metabolite databases such as GNPS, and commercial libraries like NIST, subject to licensing policies.

The impact of TraceMetrix extends beyond its immediate analytical capabilities. By establishing a framework for transparent and reproducible metabolomics analysis, the platform contributes to the broader goal of standardizing metabolomics research practices. This standardization is crucial for advancing the field and ensuring the reliability of metabolomics findings across various research contexts, from basic science to clinical applications.

Availability of data and materials

The Trace Metrix source code and case study 1 datasets supporting the findings of this study are publicly available in the GitHub repository (https://github.com/cherwell118/tracemetrix). Public metabolomics data for case study 2 were obtained from the MetaboLights repository under accession number MTBLS2627 (https://www.ebi.ac.uk/metabolights/MTBLS2627). The corresponding web-based platform can be accessed at https://www.biosino.org/tracemetrix , where all resources and documentation are provided.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-025-01095-0.

Additional file 1 (DOCX 1568 KB)

Acknowledgements
We would like to acknowledge Dr. Qingwei Xu from the Ezhou Industrial Technology Research Institute at Huazhong University of Science and Technology for supporting the website development.

Author contributions
G.Q.Z., H.R.T., Y.X.L., conceived and designed the study. W.C., Z.R.C., provide the analysis code. Y.P.A., Q.W.L., C.L., C.H.Z., Q.X.H., Q.S.C., L L.Z., and X.X.Y. provided advice on the analysis process. W.C., Z.R.C., and Y.P.A., designed the TraceMetrix. R.J.L., developed the website. W.C., Z.R.C., wrote the manuscript with revising feedbacks from G.Q.Z., H.R.T., Y.X.L.

Funding
We acknowledge financial supports from the National Key R&D Program of China (2023YFA0915501), Shanghai Municipal Science and Technology Major Project (2017SHZDZX01, 2023SHZDZX02, GTP), Self-supporting Program of Guangzhou Laboratory (SRPG22-007), Major Project of Guangzhou National Laboratory (GZNL2024A01002), National Key R&D Program of China (2022YFC3400700, 2022YFA0806400), STI2030-Major Projects (2022ZD0211600) and the National Natural Science Foundation of China (31821002).

Data availability
Project name: tracemetrix. Project home page: https://github.com/cherwell118/tracemetrix. Operating system(s): Platform independent (accessible via modern web browsers). Programming language: R (version 3.6.1). License: MIT License.

Declarations

Competing interests
The authors declare no competing interests.

Author details
[1]Bio-Med Big Data Center, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China. [2]State Key Laboratory of Genetics and Development of Complex Phenotypes, School of Life Sciences, Human Phenome Institute, Zhangjiang Fudan International Innovation Center, Metabonomics and Systems Biology Laboratory at Shanghai International Centre for Molecular Phenomics, Zhongshan Hospital, Fudan University, Shanghai 200438, China. [3]Guangzhou National Laboratory, No. 9 XingDaoHuanBei Road, Guangzhou International Bio Island, Guangzhou 510005, China. [4]Shanghai Southgene Technology Co., Ltd., Shanghai 201203, China.

Received: 11 February 2025 Accepted: 8 September 2025
Published online: 29 September 2025

References

1. Liu W, Wang Q, Chang J et al (2022) Circulatory metabolomics reveals the association of the metabolites with clinical features in the patients with intrahepatic cholestasis of pregnancy. Front Physiol 13:848508. https://doi.org/10.3389/fphys.2022.848508
2. Wasmuth HE, Glantz A, Keppeler H et al (2007) Intrahepatic cholestasis of pregnancy: the severe form is associated with common variants of the hepatobiliary phospholipid transporter ABCB4 gene. Gut 56:265–270. https://doi.org/10.1136/gut.2006.092742
3. Petrescu AD, Kain J, Liere V et al (2018) Hypothalamus-pituitary-adrenal dysfunction in cholestatic liver disease. Front Endocrinol 9. https://doi.org/10.3389/fendo.2018.00660
4. Razzaq A, Sadia B, Raza A et al (2019) Metabolomics: a way forward for crop improvement. Metabolites 9:303. https://doi.org/10.3390/metabo9120303
5. Zhang J, Sun M, Elmaidomy AH et al (2023) Emerging trends and applications of metabolomics in food science and nutrition. Food Funct 14:9050–9082. https://doi.org/10.1039/D3FO01770B
6. Smith CA, Want EJ, O'Maille G et al (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal Chem 78:779–787. https://doi.org/10.1021/ac051437y
7. Schmid R, Heuckeroth S, Korf A et al (2023) Integrative analysis of multimodal mass spectrometry data in MZmine 3. Nat Biotechnol 41:447–449. https://doi.org/10.1038/s41587-023-01690-2
8. Tsugawa H, Cajka T, Kind T et al (2015) MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. Nat Methods 12:523–526. https://doi.org/10.1038/nmeth.3393
9. Clasquin MF, Melamud E, Rabinowitz JD (2012) Lc-ms data processing with MAVEN: a metabolomic analysis and visualization engine. Curr Protoc Bioinformatics 37:14.11.1-14.11.23. https://doi.org/10.1002/0471250953.bi1411s37
10. Mallmann LP, O Rios A, Rodrigues E (2023) MS-FINDER and SIRIUS for phenolic compound identification from high-resolution mass spectrometry data. Food Res Int Ott Ont 163:112315. https://doi.org/10.1016/j.foodres.2022.112315
11. Shen X, Wang R, Xiong X et al (2019) Metabolic reaction network-based recursive metabolite annotation for untargeted metabolomics. Nat Commun 10:1516. https://doi.org/10.1038/s41467-019-09550-x
12. Dührkop K, Fleischauer M, Ludwig M et al (2019) Sirius 4: a rapid tool for turning tandem mass spectra into metabolite structure information. Nat Methods 16:299–302. https://doi.org/10.1038/s41592-019-0344-8
13. Shen X, Zhu Z-J (2019) Metflow: an interactive and integrated workflow for metabolomics data cleaning and differential metabolite discovery. Bioinforma Oxf Engl 35:2870–2872. https://doi.org/10.1093/bioinformatics/bty1066
14. Mock A, Warta R, Dettling S et al (2018) Metabodiff: an R package for differential metabolomic analysis. Bioinforma Oxf Engl 34:3417–3418. https://doi.org/10.1093/bioinformatics/bty344
15. Gaude E, Chignola F, Spiliotopoulos D, et al. muma, An R package for metabolomics univariate and multivariate statistical analysis. http://www.eurekaselect.com
16. Thévenot EA, Roux A, Xu Y et al (2015) Analysis of the human adult urinary metabolome variations with age, body mass index, and gender by implementing a comprehensive workflow for univariate and OPLS statistical analyses. J Proteome Res 14:3322–3335. https://doi.org/10.1021/acs.jproteome.5b00354
17. Picart-Armada S, Fernández-Albert F, Vinaixa M et al (2017) Null diffusion-based enrichment for metabolomics data. PLoS ONE 12:e0189012. https://doi.org/10.1371/journal.pone.0189012
18. Xia J, Wishart DS (2010) MetPA: a web-based metabolomics tool for pathway analysis and visualization. Bioinforma Oxf Engl 26:2342–2344. https://doi.org/10.1093/bioinformatics/btq418

19. Pang Z, Lu Y, Zhou G et al (2024) Metaboanalyst 6.0: towards a unified platform for metabolomics data processing, analysis and interpretation. Nucleic Acids Res 52:W398–W406. https://doi.org/10.1093/nar/gkae253
20. Tautenhahn R, Patti GJ, Rinehart D, Siuzdak G (2012) XCMS online: a web-based platform to process untargeted metabolomic data. Anal Chem 84:5035–5039. https://doi.org/10.1021/ac300698c
21. Kapoor S (Eds) (2000) FST TCS 2000: Foundations of Software Technology and Theoretical Computer Science [electronic resource]: 20th Conference New Delhi, India, December 13–15, 2000 Proceedings
22. Huang Q, Chen Q, Yi X et al (2024) Reproducibility of NMR-based quantitative metabolomics and HBV-caused changes in human serum lipoprotein subclasses and small metabolites. J Pharm Anal. https://doi.org/10.1016/j.jpha.2024.101180
23. Ganna A, Salihovic S, Sundström J et al (2014) Large-scale metabolomic profiling identifies novel biomarkers for incident coronary heart disease. PLoS Genet 10:e1004801. https://doi.org/10.1371/journal.pgen.1004801
24. Shen X, Yan H, Wang C et al (2022) Tidymass an object-oriented reproducible analysis framework for LC–MS data. Nat Commun 13:4365. https://doi.org/10.1038/s41467-022-32155-w
25. Li S, Siddiqa A, Thapa M et al (2023) Trackable and scalable LC-MS metabolomics data processing using asari. Nat Commun 14:4113. https://doi.org/10.1038/s41467-023-39889-1
26. Adusumilli R, Mallick P (2017) Data conversion with ProteoWizard msConvert. Methods Mol Biol Clifton NJ 1550:339–368. https://doi.org/10.1007/978-1-4939-6747-6_23
27. Hulstaert N, Shofstahl J, Sachsenberg T et al (2020) ThermoRawFileParser: modular, scalable, and cross-platform RAW file conversion. J Proteome Res 19:537–542. https://doi.org/10.1021/acs.jproteome.9b00328
28. Rappaport SM, Barupal DK, Wishart D et al (2014) The blood exposome and its role in discovering causes of disease. Environ Health Perspect 122:769–774. https://doi.org/10.1289/ehp.1308015
29. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30. https://doi.org/10.1093/nar/28.1.27
30. Wishart D, Arndt D, Pon A et al (2015) T3DB: the toxic exposome database. Nucleic Acids Res 43:D928-934. https://doi.org/10.1093/nar/gku1004
31. Wishart DS, Knox C, Guo AC et al (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res 36:D901-906. https://doi.org/10.1093/nar/gkm958
32. Wishart DS, Guo A, Oler E et al (2022) HMDB 5.0: the human metabolome database for 2022. Nucleic Acids Res 50:D622–D631. https://doi.org/10.1093/nar/gkab1062
33. Horai H, Arita M, Kanaya S et al (2010) MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom 45:703–714. https://doi.org/10.1002/jms.1777
34. Fiehn O, Robertson D, Griffin J et al (2007) The metabolomics standards initiative (MSI). Metabolomics 3:175–178. https://doi.org/10.1007/s11306-007-0070-6
35. Wan L, Liu Q, Liang D et al (2021) Circulating tumor cell and metabolites as novel biomarkers for early-stage lung cancer diagnosis. Front Oncol. https://doi.org/10.3389/fonc.2021.630672

Publisher's Note