OXFORD

Briefings in Bioinformatics, 2022, **23(6)**, 1–16

https://doi.org/10.1093/bib/bbac455

Advance access publication date 21 October 2022

Problem Solving Protocol

LargeMetabo: an out-of-the-box tool for processing and analyzing large-scale metabolomic data

Qingxia Yang, Bo Li, Panpan Wang, Jicheng Xie, Yuhao Feng, Ziqiang Liu and Feng Zhu

Corresponding authors: Qingxia Yang, School of Geographic and Biologic Information, Nanjing University of Posts and Telecommunications, Nanjing, 210023, China. E-mail: yangqx@njupt.edu.cn; Feng Zhu, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, 310058, China. E-mail: zhufeng@zju.edu.cn

Abstract

Large-scale metabolomics is a powerful technique that has attracted widespread attention in biomedical studies focused on identifying biomarkers and interpreting the mechanisms of complex diseases. Despite a rapid increase in the number of large-scale metabolomic studies, the analysis of metabolomic data remains a key challenge. Specifically, diverse unwanted variations and batch effects in processing many samples have a substantial impact on identifying true biological markers, and it is a daunting challenge to annotate a plethora of peaks as metabolites in untargeted mass spectrometry-based metabolomics. Therefore, the development of an out-of-the-box tool is urgently needed to realize data integration and to accurately annotate metabolites with enhanced functions. In this study, the LargeMetabo package based on R code was developed for processing and analyzing large-scale metabolomic data. This package is unique because it is capable of (1) integrating multiple analytical experiments to effectively boost the power of statistical analysis; (2) selecting the appropriate biomarker identification method by intelligent assessment for large-scale metabolic data and (3) providing metabolite annotation and enrichment analysis based on an enhanced metabolite database. The LargeMetabo package can facilitate flexibility and reproducibility in large-scale metabolomics. The package is freely available from https://github.com/LargeMetabo/LargeMetabo.

Keywords: large-scale metabolomics, data processing, data integration, marker identification, metabolite annotation

Introduction

Metabolomic profiling of complete sets of metabolites in biological samples has been widely applied for identifying biomarkers in the diagnosis and prediction of disease [1]. Numerous samples are essential in metabolomics to increase statistical power and to address crucial issues related to heterogeneity in disease biology [2, 3]. Because it is such a powerful technique, large-scale metabolomics has attracted widespread attention and is applied to measure hundreds of compounds simultaneously in large-scale studies, involving thousands of samples or more [4]. Large-scale metabolomics can capture the 'fingerprints' of specific biological processes critical to precise medical applications, such as studying complicated disease mechanisms and discovering new therapeutic targets [5–7]. Several large-scale cohort studies have been conducted using metabolomic analyses. For example, the Consortium of Metabolomics Studies comprised 47 cohorts and >136 000 blood samples collected from 1985 to 2017 for tracking and analyzing metabolomic data [8].

Despite a rapid increase in the number of large-scale metabolomic studies, the analysis of metabolomic data is increasingly vital to achieving consistent results [9, 10]. Because of the need to analyze massive numbers of samples, all samples are divided into multiple batches, spanning months to years of sample collection [11]. However, in sequential studies, the accurate measurement of biochemical signals drifts over extended sampling periods, and intra-batch or inter-batch variations inevitably occur. Thus, such unwanted variations in large-scale metabolomics have a substantial impact on downstream analytics in attempting to identify true biological markers [12]. Therefore, large-scale metabolomic studies are seriously hampered by the inefficiency of integrating data from separate batches of samples or analytical experiments [13]. Moreover, it is fundamentally necessary for mass spectrometry (MS) peaks of interest to be recognized and linked to biological functions. However, it has been reported that <2% of the peaks detected in untargeted MS-based metabolomics can be annotated accurately [14]. Therefore, the foremost challenge for meaningful metabolite annotation requires the ability to decipher raw peak features into patterns of specific metabolites with biologically interpretable functions in large-scale metabolomics [15, 16].

Several tools and applications have been designed and developed for analyzing metabolomic data. As shown in Supplementary Table S1, available online at http://bib.oxfordjournals.org/,

Qingxia Yang is an associate professor of the School of Geographic and Biologic Information in Nanjing University of Posts and Telecommunications, China. Her current focus includes bioinformatics, computational biology and omics data analysis.

Bo Li is an associate professor worked in the College of Life Sciences, Chongqing Normal University, China. His current focus includes bioinformatics and computational biology.

Panpan Wang is a lecturer of the College of Chemistry and Pharmaceutical Engineering in Huanghuai University, China. Her current focus includes bioinformatics, drug discovery and medicinal chemistry.

Jicheng Xie, Yuhao Feng and Ziqiang Liu are undergraduate students of the School of Geographic and Biologic Information in Nanjing University of Posts and Telecommunications, China. They are interested in bioinformatics.

Feng Zhu is a tenured professor of the College of Pharmaceutical Sciences in Zhejiang University, China. His research laboratory (https://idrblab.org/) has been working in the fields of bioinformatics, OMIC-based drug discovery, system biology and medicinal chemistry.

Received: June 4, 2022. **Revised:** September 6, 2022. **Accepted:** September 24, 2022

© The Author(s) 2022. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bib/article/23/6/bbac455/6768054 by Zhejiang University user on 23 November 2022

hRUV [5], metaX [17], MMEASE [18], MetaboAnalyst [19], MRMkit [9], Norm ISWSVR [20], Workflow4Metabolomics [21] and XCMS online [22, 23] can realize data integration based on the embedded sample replicates, quality control samples, pathway-level integration, peak pooling by combining complementary measurements and shared metabolic features. The result is that using these tools can involve limitations, such as the loss of specific metabolites, and data integration must rely on the embedded quality control samples, internal standards or exact feature names. In addition, the use of online servers comes with inherent limitations. For example, it is difficult to reproduce the same results when re-analyzing the dataset using an online server. The upload speed and calculation speed are severely limited by the online server when handling large-scale or long-term datasets. To address the problems of reproducibility and flexibility, developing a stand-alone tool is highly necessary for large-scale metabolomics [24–26]. Moreover, when peaks are annotated as specific metabolites, it is essential to identify the precise functions in the complex matrix of untargeted metabolomics [27]. Several tools, including the metaP-server [28], metaX, MMEASE, MetaboAnalyst, Workflow4Metabolomics and XCMS online, provide annotation functions for metabolites, but there is little functional annotation information in these tools. Along with the rapid accumulation of metabolites, it is feasible to significantly enhance the metabolite annotation and enrichment. Therefore, for large-scale metabolomics to succeed, there is an urgent need to develop a new and effective tool that is able to realize data integration from multiple analytical experiments with significantly enhanced functions in accurately annotating metabolites.

In this study, the LargeMetabo package based on R code was developed for processing and analyzing large-scale metabolomic data. Compared to the existing tools and packages, there are three factors of functional improvement to tackle the challenges of large-scale metabolomics in the LargeMetabo package. First, one of the greatest challenges is how to effectively integrate the metabolites in hundreds/thousands of samples when there are no embedded quality control samples and internal standards. The LargeMetabo package can integrate multiple datasets from different experiments based only on aligning the metabolites using mass and retention time (RT). The detailed validation for data integration using the metabolic data is shown in case study 1. Second, one of the greatest challenges is that it is difficult to select an appropriate one from various methods of marker identification in large-scale metabolomics. In the LargeMetabo package, the most appropriate method can be discovered based on the performance assessment of the classification model. Based on the metabolic data, the detailed application for selecting marker identification method is shown in case study 2. Third, one of the greatest challenges is that there are few functional annotations in the existing tools for metabolomics. The LargeMetabo package can significantly enhance metabolite annotation and enrichment using a new metabolite database. The detailed application for functional annotations of metabolites is shown in case study 3.

The LargeMetabo package is unique in processing and analyzing large-scale metabolomic data because it is capable of (1) integrating multiple datasets from different analytical experiments based on mass and RT to effectively boost the power of statistical analysis if there are no quality control samples, internal standards and exact feature names; (2) discovering the most appropriate method based on the performance assessment of classification among various biomarker identification methods and (3) providing enhanced metabolite annotation and enrichment analysis based on a new metabolite database. The LargeMetabo package can facilitate flexible and reproducible analysis in large-scale metabolomic studies. The LargeMetabo package is freely available from https://github.com/LargeMetabo/LargeMetabo.

Materials and methods

The development version of the LargeMetabo package is hosted on GitHub, and the stable release will soon be available as an R package on CRAN. It builds upon the R code with extensive modifications to ensure the accurate identification of functions. Several R packages are utilized in the background processes, including AUC, CluMSID, corrplot, d3heatmap, dplyr, e1071, factoextra, FSelector, genefilter, ggfortify, ggplot2, igraph, magrittr, MASS, mixOmics, readr, ropls, sampling, siggenes, SOMbrero and varSelRF. The analytical workflow in this study, including five operational steps, is shown in Figure 1.

Methods for data integration and batch effect removal

Large-scale metabolomics can effectively boost the power of statistical analysis and can accurately identify biological markers. In large-scale metabolomic studies, hundreds/thousands of samples are used and frequently separated into multiple experiments [11]. In step 1 of the LargeMetabo package, multiple datasets from different analytical experiments can be integrated into a combined dataset, and three methods are provided to remove unwanted variations and batch effects.

To integrate multiple datasets, a useful strategy for data integration was applied in this study by permitting tolerance of RT and mass-to-charge ratio (m/z) for specific metabolite peaks [29]. The detailed description and sketch map of this strategy for data integration are shown in Figure 2. Using this strategy, the peaks in all samples of multiple datasets can be aligned as a metabolite if both the differences in their RT values are small enough and the differences in their m/z values are also small enough. There were three steps in this strategy. First, a peak with the strongest intensity was set as peak reference 1, and other peaks in all samples were placed in an aligned group if the differences in values (both RT and m/z) between these peaks and peak reference 1 were in the first tolerance setting. Second, the peak with the median RT value in the aligned group was then set as peak reference 2. The peaks were selected in the second group if the differences of values (both RT and m/z) between these peaks and the peak reference 2 were in the second tolerance setting. These selected peaks in the second group were aligned as the same metabolite. Third, other unaligned peaks for the metabolites can be aligned by repeating the above stepwise process similarly.

In the process of data integration, various unwanted variations and batch effects resulted from different experimental conditions in multiple datasets. It was essential to remove these unwanted variations for the integrated data from different analytical experiments [30, 31]. After data integration, the methods of batch effect removal can be applied to the integrated data directly rather than to each dataset independently. This application can ensure the removal of the batch effects among different datasets. There were three methods used for removing unwanted variations and batch effects among different analytical experiments in the LargeMetabo package. These methods used for removing batch effects included batch mean-centering (BMC/PAMR) [32], the empirical Bayes method (ComBat/EB) [33] and global normalization (GlobalNorm) [34].

Downloaded from https://academic.oup.com/bib/article/23/6/bbac455/6768054 by Zhejiang University user on 23 November 2022

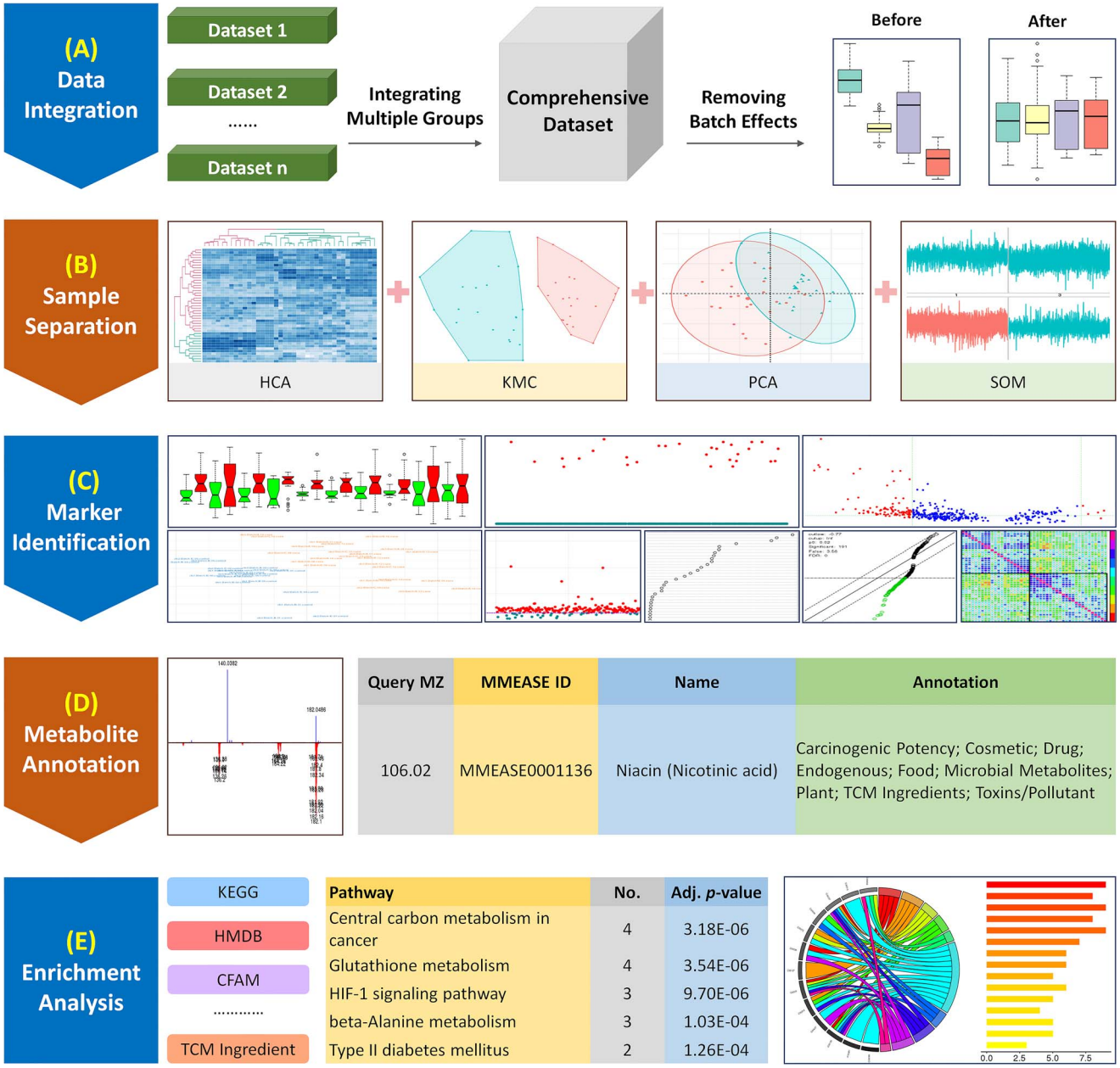

Figure 1. The analytical workflow of the LargeMetabo package. There are five major functions: (**A**) data integration, (**B**) sample separation, (**C**) biomarker identification, (**D**) metabolite annotation and (**E**) enrichment analysis.

Methods for sample separation and biomarker identification

Four sample separation methods were applied for visualizing the clustering and separation of different samples in step 2. After data integration and batch effect removal, sample separation methods were used for the visualization of different samples in large-scale metabolomics [35]. In the LargeMetabo package, four methods were provided for sample separation, including hierarchical clustering analysis (HCA) [36], *k*-means clustering (KMC) [37, 38], self-organizing map (SOM) [39] and principal component analysis (PCA) [40, 41].

Subsequently, 13 methods of biomarker identification were used to discover metabolic markers for the given dataset in step 3. In the LargeMetabo package, the popular methods for marker identification included fold change (FC) [42], partial least squares discrimination analysis (PLS-DA) [43], orthogonal PLS-DA (OPLS-DA) [44], Student's *t*-test [45], Chi-squared test [46], correlation-based feature selection (CFS) [47], entropy-based filter method [47], linear models and empirical Bayes method [48], Relief [49], random forest-recursive feature elimination (RF-RFE) [50], significance analysis for microarrays (SAM) [51], support vector machine-recursive feature elimination (SVM-RFE) [52] and Wilcoxon rank sum test (WRST) [53]. Methods of sample separation and biomarker identification are described in detail in the Supplementary Methods.

Because there were great variations in the statistical theories and model assumptions for biomarker identification methods, different methods could lead to contradictory results even when using the same dataset [54–56]. The appropriate application of a biomarker identification method is heavily dependent on the

Downloaded from https://academic.oup.com/bib/article/23/6/bbac455/6768054 by Zhejiang University user on 23 November 2022

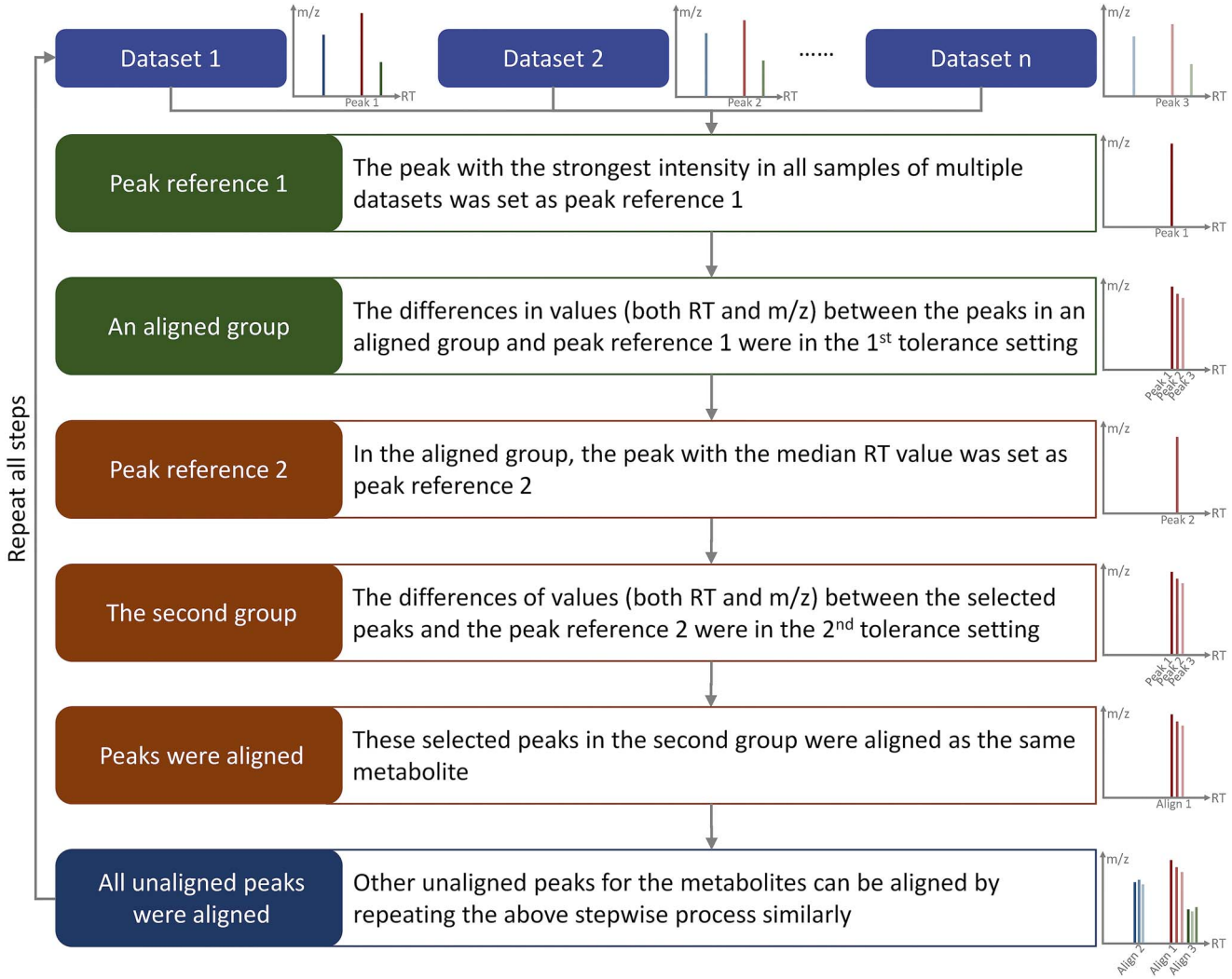

Figure 2. The detailed description and sketch map of the data integration strategy for multiple datasets from different analytical experiments based on m/z and RT.

natural properties of certain studies [57–59]. To achieve systematic selection, the Marker_Assess function was provided to assess the performance of the biomarker identification methods in the LargeMetabo package. The assessment for these methods was performed based on the receiver operator characteristic (ROC) curve and area under curve (AUC) values in the SVM classification model using metabolic markers [60–62]. In the model, 2-fold cross-validation was used due to the limited number of samples. The classification model was constructed with the default parameters by applying the SVM function of the e1071 package, and the AUC value was calculated by applying the roc function of the AUC package.

Methods for metabolite annotation and enrichment analysis

The MS peaks of interest need to be annotated based on their mass values, but <2% of the detected peaks can be annotated in untargeted MS-based metabolomics [14]. Moreover, the conversion from a raw peak into a specific metabolite with biological interpretation remains a major challenge [15]. Herein, the metabolites detected by MS (MS1) and tandem MS (MS/MS) are annotated in step 4. The metabolite annotation was based on a new metabolite database, which was constructed by systematic literature reviews and searching from the metabolite databases, including HMDB [63], MMCD [64], METLIN [65], LMSD [66] and MoNA [67]. The presumptive metabolites could be annotated using MS1 spectra with mass information, and the metabolites could also be annotated with the MS/MS information using the CluMSID package. In this metabolite database, information on the name, mass and adduct list was provided for peak annotation. Moreover, the detailed biological functions were added for these metabolites, including the endogenous and exogenous factors [68–71]. These factors were obtained by performing literature reviews and searching different databases, including CFAM [72], Drugbank [73], ECMDB [74], FooDB (www.foodb.ca), HMDB, Kyoto Encyclopedia of Genes and Genomes (KEGG) [75], PMDB [76], T3DB [77], TCMID [78] and YMDB [79]. The exogenous factors referred to the source of metabolites consisting of agricultural chemicals, cosmetics, drugs, drug metabolites, foods, microbes, plants, TCM (traditional Chinese medicine) ingredients and environmental toxins/pollutants. For the LargeMetabo package, one of the advantages was that it provided metabolite annotation with significantly enhanced functions.

Downloaded from https://academic.oup.com/bib/article/23/6/bbac455/6768054 by Zhejiang University user on 23 November 2022

When performing metabolite annotation using MS1 spectra, the presumptive metabolites can be annotated using mass spectra based on the adducts and multiply charged ions as the popular method. The detected peaks will be compared with the predicted adduct ions. There were various types of adduct ions, including $[M+H]^+$, $[M+Na]^+$, etc. in positive mode and $[M-H]^-$, $[M+Cl]^-$, etc. in negative mode. First, the theoretical m/z values were calculated based on the differences between the molecular weight of the detected MS peaks and each predicted adduct ion. Second, these theoretical m/z values were compared with the molecular weight of metabolites in the metabolite database, and a tolerance of 0.05 Da was set as the default. For metabolites meeting the tolerance, the information of the input m/z, LargeMetabo ID, mass, name and detailed biological functions was provided in the form of a table. The results of annotation using MS1 spectra were presumptive and were only the reference for the enquired peaks because of a lack of the experimental validation.

Moreover, to perform metabolite annotation of MS/MS in the LargeMetabo package, the m/z value of the parent ion and the matrix consisting of m/z values and intensities of MS/MS were needed. The annotation was performed to identify metabolites by MS/MS spectral similarity and unsupervised learning methods using the CluMSID package. The m/z value of the parent ion was compared with the parent ions of the metabolite database using a tolerance of 0.1 Da as the default. The m/z values of MS/MS peaks were compared with those in the metabolite database using a tolerance of 0.5 Da as the default. The fit value was applied to calculate the spectral similarity for the matched metabolite, and the largest value (fit = 1) indicated the most appropriate match for the specific metabolite. For metabolite annotation of MS/MS, the information of the ID, name, mass, detailed biological functions and fit value was provided in the form of a table. The mirror plot of the annotation result between the input spectrum and the specific metabolite spectrum of the metabolite database is shown in the LargeMetabo package.

In step 5 of the LargeMetabo package, there were eight categories of functional metabolites used for enrichment analysis based on the metabolite database. Because metabolomics has been demonstrated to play important roles in complex diseases by altering endogenous and exogenous sources, enrichment analysis could reveal the functional roles and exogenous sources of metabolites [80–82]. The enrichment analysis in the LargeMetabo package involved three parts: (1) biological pathways (including KEGG pathways and metabolic and disease pathways); (2) biological functions and structures (including biological function classes and structural categories) and (3) exogenous sources (including food components and food additives, therapeutic classes of traditional medicine, species taxonomy from traditional medicine and toxins and environmental pollutants). Specifically, these eight categories contained the information of a large collection of metabolites, including (1) KEGG pathways for metabolites, (2) human metabolic and disease pathways from the SMPDB database, (3) classes of biological function to reflect the biological roles of metabolites, (4) categories of chemical structure in chemical families for metabolites, (5) food sources consisting of food components and food additives, (6) therapeutic classes of secondary metabolites from traditional medicine, (7) species taxonomy from traditional medicine for metabolites and (8) exogenous sources from toxins and environmental pollutants for metabolites. All metabolites with the information of these eight categories are open source in the LargeMetabo package.

Example data for the input files of the LargeMetabo package

In the LargeMetabo package, the example data of the input file are provided in the format of RData and csv files in the Github repository. For data integration, the MutileGroup object, including three datasets, are the input data for the Integrate_Data function, and these three datasets (batch_data_1.csv, batch_data_2.csv and batch_data_3.csv) and a data list (MutileGroup.RData), including these three datasets, can be downloaded as the input files for users. In the step of batch effect removal for multiple datasets, the MutileAlign object is the integrated dataset and it can be as the input data for Removal_Batch function, and the files (MutileAlign.csv and MutileAlign.RData) can be downloaded for users. In the step of sample separation and biomarker identification for large-scale metabolomic data, the MarkerData object is a data list consisting of the metabolite intensities and sample labels, which can be as the input data for the Sample_Separation, Marker_Identify and Marker_Assess functions. The files (MarkerData.RData and MarkerData.csv), including the metabolite intensities and sample labels, can be downloaded for users correspondingly. For metabolite annotation, from the AnnotaData list (AnnotaData.RData), the AnnotaMS object (including the peak list of m/z values) is provided for MS1, and the ParentMass object (including the m/z value of parent ion) and TandomData object (including the matrix of m/z values and intensities of MS/MS) are provided for MS/MS. For enrichment analysis, from the EnrichData list (EnrichData.RData), the sampleDatakegg object is the input file of enrichment of KEGG pathways, and sampleDatacas and enrichDB object are provided for other categories of enrichment analysis.

Results and discussion

To test the usability and flexibility of the LargeMetabo package, three case studies were performed using the example data in this study. These case studies included (1) case study 1: data integration from three analytical experiments; (2) case study 2: sample separation and biomarker identification for metabolomic data and (3) case study 3: metabolite annotation and enrichment analysis for the specific metabolite.

Case study 1: data integration from three analytical experiments

In the LargeMetabo package, multiple datasets from different analytical experiments can be integrated into a comprehensive dataset. As the input, the csv files of the feature-by-sample matrix were needed for data integration. There were four essential parts, including the metabolite names, peak intensities, mass and RT in the columns. The metabolite names must be kept in the first column and the sample names must be kept in the first row. The sample labels indicating different sample groups (case versus control) were placed in the second row. Using the example data provided in the LargeMetabo package, the result after data integration is shown in Supplementary Table S2 available online at http://bib.oxfordjournals.org/. After data integration, three methods were provided to remove batch effects among different analytical experiments, including BMC/PAMR, ComBat/EB and GlobalNorm. Using the example data provided in the LargeMetabo package, the result after data integration and batch effect removal is shown in Supplementary Table S3 available online at http://bib.oxfordjournals.org/.

Downloaded from https://academic.oup.com/bib/article/23/6/bbac455/6768054 by Zhejiang University user on 23 November 2022

Table 1. The number (No.) of metabolites and true markers in the benchmark dataset MTBLS59; in this dataset, each spike-in compound generated one or more features defined as the intensity of the MS-signals (peaks) identified by RT and m/z values; the markers were identified using the PLS-DA and Student's *t*-test methods with the cut-off of VIP >1 and adjusted *P*-value <0.05; the true markers referred to the spike-in compounds

	The No. of peaks of all metabolites	The No. of peaks of markers	The No. of peaks of true markers	The No. of spike-in compounds
Dataset 1	649	13	6	2
Dataset 2	641	2	2	2
Dataset 3	648	14	2	2
Integrated dataset	612	46	12	6

The function of data integration was tested by the benchmark dataset (MTBLS17) [83] from MetaboLights [84]. In this benchmark dataset, 78 hepatocellular carcinoma (HCC) patients and 184 cirrhotic (CIR) controls were detected by metabolomic profiling. The changes in the metabolite levels of HCC and CIR samples were compared using ultra-performance liquid chromatography-MS (UPLC–MS). All samples were divided into three analytical experiments. There were 129 CIR controls and 60 HCC patients in the first experiment, 50 CIR controls and 13 HCC patients in the second experiment and 5 CIR controls and 5 HCC patients in the third experiment. The datasets from three analytical experiments were integrated using the data integration function. After data integration, the BMC/PAMR method was applied for batch effect removal. After integrating three datasets and removing batch effects, these comprehensive data were used for the downstream analysis.

As shown in Figure 3**A** and **B**, boxplots were applied to visualize the m/z and RT values of the raw data from different analytical experiments. As shown, there were specific variations in the m/z and RT values among the separate datasets, especially for the third dataset. After data integration, the boxplots of the intensities in each sample before and after batch effect removal are shown in Figure 3**C** and **D**, respectively. The intensities of the samples ranged from 4 to 8 before removing batch effects, while the range distribution was from −4 to 0 after removing batch effects. From these two boxplots, the range distribution could be significantly reduced by batch effect removal. The PCA plots of samples before and after batch effect removal are shown in Figure 3**E** and **F**, respectively. PCA plots were applied to visualize the sample separation from different datasets before and after batch effect removal. From the PCA plots, the samples in different datasets were separated from each other before batch effect removal, while the samples in different datasets could be clustered together after removing batch effects. The dendrograms of the samples in three different datasets before and after batch effect removal are shown in Figure 4**A** and **B**, respectively. The samples in different datasets are colored red, green and blue. The samples of the integrated dataset colored in the same color before batch effect removal were clustered together (Figure 4**A**). After batch effect removal using the BMC/PAMR method, the samples in the same color were dispersed in the whole dendrogram (Figure 4**B**). Therefore, most of the batch effects were removed from the integrated data from the PCA plots and dendrograms.

To validate the effectiveness of the data integration method, a benchmark spike-in dataset (MTBLS59) [85] was applied in this study. Herein, a real metabolomic dataset was measured by spiking 10 out of 40 apple samples with a mixture of nine known compounds using UPLC–MS. This dataset consisted of 10 control samples and three spiked sets of the same size, where naturally occurring compounds were added at different concentrations. In this dataset, each spike-in compound generated one or more features defined as the intensity of the MS-signals (peaks) identified by RT and m/z value. This benchmark dataset can serve as a test bed to assess the performance of the new algorithm. The peak table of raw data (CDF file) was obtained using the xcms package with the default parameters. An integrated dataset was combined from three datasets using the Integrate_Data function in the LargeMetabo package. As shown in Table 1, the number of peaks of all metabolites was 649, 641, 648 and 612 for dataset 1, dataset 2, dataset 3 and integrated dataset, respectively. The metabolic markers were identified using both PLS-DA and Student's *t*-test methods with the cut-off of variable importance in the projection (VIP) >1 and adjusted *P*-value < 0.05. As a result, there were 13, 2, 14 and 46 peaks of markers for dataset 1, dataset 2, dataset 3 and the integrated dataset, respectively. Among these peaks of markers, the number of peaks of true markers was 6, 2, 2 and 12 for these four datasets, respectively. In these peaks of true markers, there were 2, 2, 2 and 6 spike-in compounds for these four datasets, respectively. The number of peaks of true markers and spike-in compounds was significantly increased based on the integrated dataset. Therefore, the effectiveness of the data integration strategy was validated based on the benchmark dataset.

Case study 2: sample separation and biomarker identification for metabolomic data

After integrating three analytical experiments and removing batch effects for the benchmark dataset MTBLS17, a combined dataset consisting of 78 HCC patients and 184 CIR controls was obtained. The separation of all samples for this combined dataset could be visualized using four methods. When applying all metabolites in this dataset, the samples might be gathered and not be separated distinctly. Therefore, the markers were identified with a cut-off of VIP >1 for the PLS-DA method. Fifty-nine markers were applied for the separation of all samples for sample separation. The dendrograms of the samples in different groups (case versus control) before and after batch effect removal are shown in Figure 4**C** and **D**, respectively. Before batch effect removal, the samples of different groups colored in the same color were dispersed in the whole dendrogram (Figure 4**C**). After batch effect removal using the BMC/PAMR method, a majority of samples from the same group using the markers were clustered together (Figure 4**D**). Therefore, the differences in samples from different groups can be captured after removing batch effects.

For sample separation, fifty-nine markers were applied using different methods. As shown in Figure 4**E**, hierarchical clustering for the samples and metabolites was performed using the

Downloaded from https://academic.oup.com/bib/article/23/6/bbac455/6768054 by Zhejiang University user on 23 November 2022

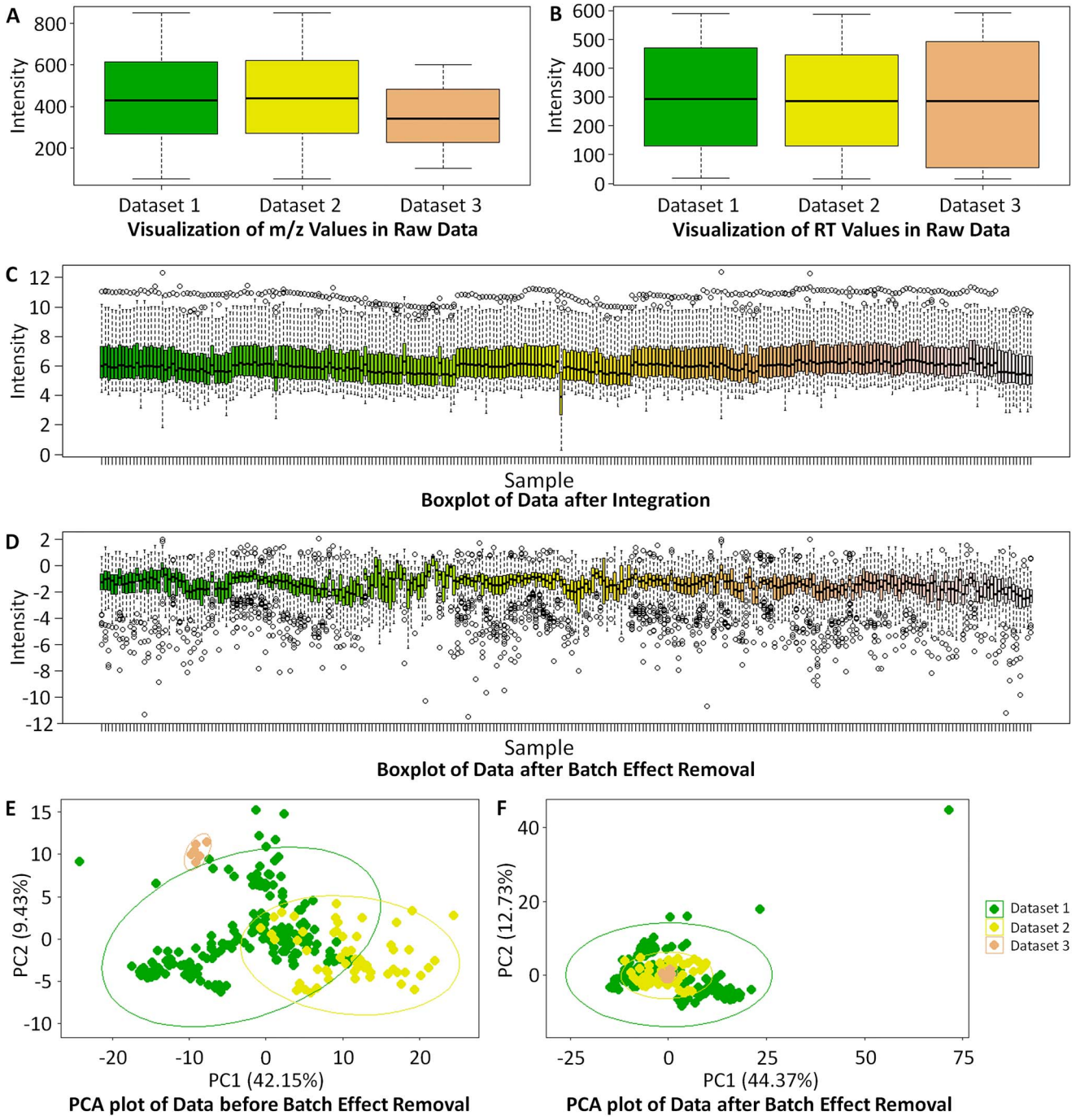

Figure 3. The data integration for three analytical experiments in the benchmark dataset MTBLS17. The boxplots were applied to visualize the raw data for (**A**) m/z and (**B**) RT values in three analytical datasets. Boxplots of intensities in each sample were used for integrated datasets (**C**) before and (**D**) after batch effect removal. PCA plots were applied to visualize the distribution of samples using all metabolites in different analytical datasets (**E**) before and (**F**) after batch effect removal.

parameter (clusters = 3) of HCA for separating the samples of the case group (HCC patients) and the control group (CIR controls). The lowest and highest values are shown in orange and blue, respectively, and the others are shown in gradient color. There were three clusters in rows for all samples; the samples in the green and blue clusters were mainly HCC patients, and the samples in the red cluster were mainly CIR controls. The sample clusters for KMC and SOM in different groups (case versus control) using the parameter (clusters = 2) are shown in Figure 4**F** and **G**, respectively. As shown in Figure 4**H**, the samples were separated using a PCA plot, and samples in two groups (case versus control) were divided into two clusters. Moreover, the results of the example data embedded in the LargeMetabo package are shown in Supplementary Figure S1, available online at http://bib.oxfordjournals.org/, using four sample separation methods.

Moreover, 13 biomarker identification methods were provided to discover metabolic markers for the integrated dataset in the LargeMetabo package. These methods have been widely used in identifying markers for metabolomic data [86–88]. As shown in Table 2, the top 20 metabolites with the highest VIP values

Downloaded from https://academic.oup.com/bib/article/23/6/bbac455/6768054 by Zhejiang University user on 23 November 2022

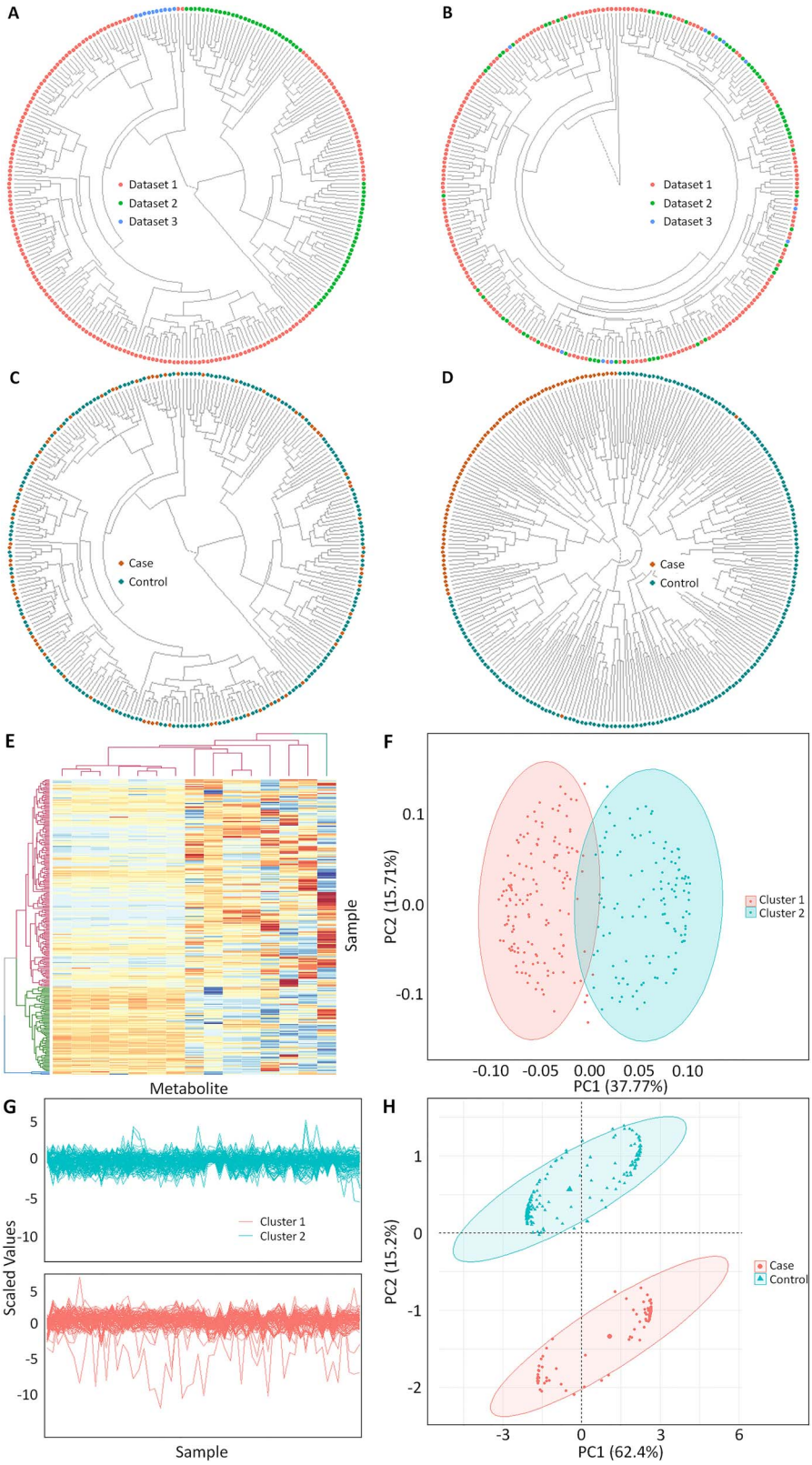

Figure 4. The sample separation of all samples in different datasets or different groups (case versus control) from the benchmark dataset MTBLS17. The dendrograms were applied for different datasets including dataset 1, dataset 2 and dataset 3 (A) before and (B) after batch effect removal. Using metabolic metabolites, the dendrograms were applied for case group and control group (C) before and (D) after batch effect removal. (E) Based on metabolic markers, sample separation in different groups was performed using HCA with parameters (clusters = 3 for samples and clusters = 2 for metabolites). Sample clustering in different groups was performed by (F) KMC and (G) SOM using parameter (clusters = 2). (H) Sample separation for the case group and control group was performed using PCA.

Downloaded from https://academic.oup.com/bib/article/23/6/bbac455/6768054 by Zhejiang University user on 23 November 2022

Table 2. Thirteen methods of biomarker identification were applied to discover metabolic markers; the top 20 metabolites with the highest VIP values using the PLS-DA method were included; herein, the results of other biomarker identification methods are also provided; PLS-DA, FC, OPLS-DA, CFS, correlation coefficient (Cor), Relief, RF-RFE, mean decrease accuracy (ACC), WRST, SAM, linear models and empirical Bayes (LMEB), Chi-squared test (CHIS), importance (Import), entropy-based filter method (ENTROPY), InfGain Order (InfGain), SVM-RFE

MZ/RT	PLS-DA VIP	FC FC	OPLS-DA VIP	*t*-test *P*-value	CFS Cor	Relief VIP	RF-RFE ACC	WRST *P*-value	SAM *P*-value	LMEB *P*-value	CHIS Import	ENTROPY InfGain	SVM-RFE Order
448.31/202.65	2.21	−0.76	3.29	0.02	−0.25	2.21	0.00	0.00	0.00	0.00	0.28	0.00	85
467.31/203.39	2.20	−0.71	3.24	0.02	−0.26	2.20	0.00	0.00	0.00	0.00	0.30	0.00	41
134.01/21.24	2.20	0.13	1.77	0.36	0.15	2.20	0.00	0.02	0.04	0.37	0.00	0.00	6
327.05/19.13	2.08	0.14	1.69	0.67	0.19	2.08	0.00	0.00	0.05	0.42	0.00	0.00	187
449.31/202.74	2.07	−0.62	3.07	0.02	−0.24	2.07	0.00	0.00	0.00	0.01	0.32	0.00	166
271.97/23.07	1.98	0.21	2.01	0.22	0.18	1.98	0.00	0.00	0.01	0.19	0.00	0.00	88
466.31/204.05	1.93	−0.72	2.90	0.03	−0.24	1.93	0.00	0.00	0.00	0.01	0.27	0.00	43
431.30/202.71	1.91	−0.67	2.89	0.04	−0.22	1.91	0.00	0.00	0.00	0.01	0.28	0.00	127
430.30/202.73	1.81	−0.68	2.78	0.06	−0.22	1.81	0.00	0.00	0.00	0.02	0.29	0.00	78
144.10/24.25	1.71	−0.70	2.40	0.08	−0.18	1.71	0.00	0.00	0.00	0.07	0.00	0.00	163
412.28/202.74	1.66	−0.65	2.52	0.10	−0.19	1.66	0.00	0.00	0.00	0.05	0.00	0.00	143
132.10/34.04	1.59	0.15	1.00	0.92	0.13	1.59	0.00	0.04	0.16	0.81	0.00	0.00	102
246.16/102.03	1.59	0.19	1.26	0.90	0.15	1.59	0.00	0.02	0.07	0.80	0.00	0.00	156
531.00/20.06	1.58	0.08	0.96	0.95	0.19	1.58	0.00	0.00	0.25	0.81	0.00	0.00	7
243.62/202.37	1.49	−0.43	2.38	0.09	−0.18	1.49	0.00	0.00	0.00	0.07	0.00	0.00	144
464.20/163.09	1.49	−0.54	2.20	0.13	−0.16	1.49	0.00	0.01	0.00	0.11	0.00	0.00	121
218.97/21.73	1.43	0.02	0.40	0.97	0.07	1.43	0.00	0.24	0.68	0.90	0.00	0.00	177
227.96/19.97	1.43	0.02	0.42	0.97	0.07	1.43	0.00	0.26	0.70	0.90	0.00	0.00	45
130.97/581.84	1.38	0.02	0.33	0.97	0.04	1.38	0.00	0.48	0.74	0.94	0.00	0.24	64
218.14/36.87	1.37	0.12	0.85	0.92	0.08	1.37	0.00	0.17	0.24	0.84	0.00	0.00	71

of the PLS-DA method were discovered. Herein, other methods were also applied for biomarker identification. There were significantly different results when using different biomarker identification methods even for the same dataset. Therefore, it was very important to select the most appropriate method for a specific dataset [89]. Moreover, the plots of example data embedded in the LargeMetabo package by applying the biomarker identification methods are shown in Supplementary Figure S2 available online at http://bib.oxfordjournals.org/.

The appropriate application of these biomarker identification methods was heavily dependent on the performance assessment [90]. The performances of these methods were assessed using the Marker_Assess function in the LargeMetabo package. The internal data (MarkerData object) were used for the performance assessment, which included 398 metabolites in samples of 21 controls and 24 cases. Based on the discovered markers, the ROC curve and AUC value were applied in the SVM classification model. Considering the limited number of samples, the classification model with 2-fold cross-validation was constructed using the SVM function of the e1071 package and the AUC value was calculated using the ROC function of the AUC package. As shown in Figure 5, the AUC values were evidently different when using these 13 biomarker identification methods. In particular, the AUC value of the classification model equaled 1 when using the markers of PLS-DA, Chi-squared test, CFS, entropy-based filter, Relief, RF-RFE and SVM-RFE. The AUC value was close to 1 when using the markers of OPLS-DA, Student's *t*-test, LMEB and WRST. The AUC values were low when using the markers of FC and SAM. The appropriate methods, including PLS-DA, Chi-squared test, CFS, entropy-based filter, Relief, RF-RFE and SVM-RFE, were suggested to be applied for this dataset by performance assessment.

Moreover, a benchmark dataset (the study of Xiao *et al.*) [91] was applied to identify the metabolic markers of triple-negative breast cancer. This dataset included 330 samples with triple-negative breast cancer and 149 paired normal breast tissues. In this study, both FC and Student's *t*-test methods were used to discover biomarkers of the raw data. The cut-off (logFC > 1 and *P*-value < 0.05) was set for FC and Student's *t*-test method, respectively. As a result, there were 302 markers at the intersection of these two methods, comprising 98 upregulated and 204 downregulated metabolites between triple-negative breast cancer and paired normal breast tissues. The boxplots of the top 10 most upregulated and downregulated markers are shown in Figure 6**A** and **B**, respectively. Taking phenylalanyl-threonine as an example, the value in samples with triple-negative breast cancer was five times that in normal breast tissues. The value of uridine 5′-diphosphoglucuronic acid (UDP-D-glucuronate) in samples of normal breast tissues was five times that of samples with triple-negative breast cancer. The scatter diagram of negative logarithmic transformation of the adjusted *P*-values for each metabolite using Student's *t*-test is shown in Figure 6**C**. The points in red indicate the metabolites with adjusted *P*-values < 0.05 between the two sample groups (triple-negative breast cancer versus paired normal breast tissues). These metabolic markers can be used for the downstream analysis in case study 3.

Case study 3: metabolite annotation and enrichment analysis for specific metabolite

In the LargeMetabo package, the function of metabolite annotation was provided based on an enhanced metabolite database. The metabolite database was constructed by systematic literature reviews and searching various databases. Information on the name, mass, adduct list and biological functions, including the endogenous and exogenous factors for each metabolite, was provided. This metabolite database was applied for enhanced annotation of metabolites, and the metabolite data can be downloaded from the LargeMetabo package. For metabolite annotation of MS1, the input data were the peak list of m/z values. For metabolite annotation of MS/MS, the input data consisted of the

Downloaded from https://academic.oup.com/bib/article/23/6/bbac455/6768054 by Zhejiang University user on 23 November 2022

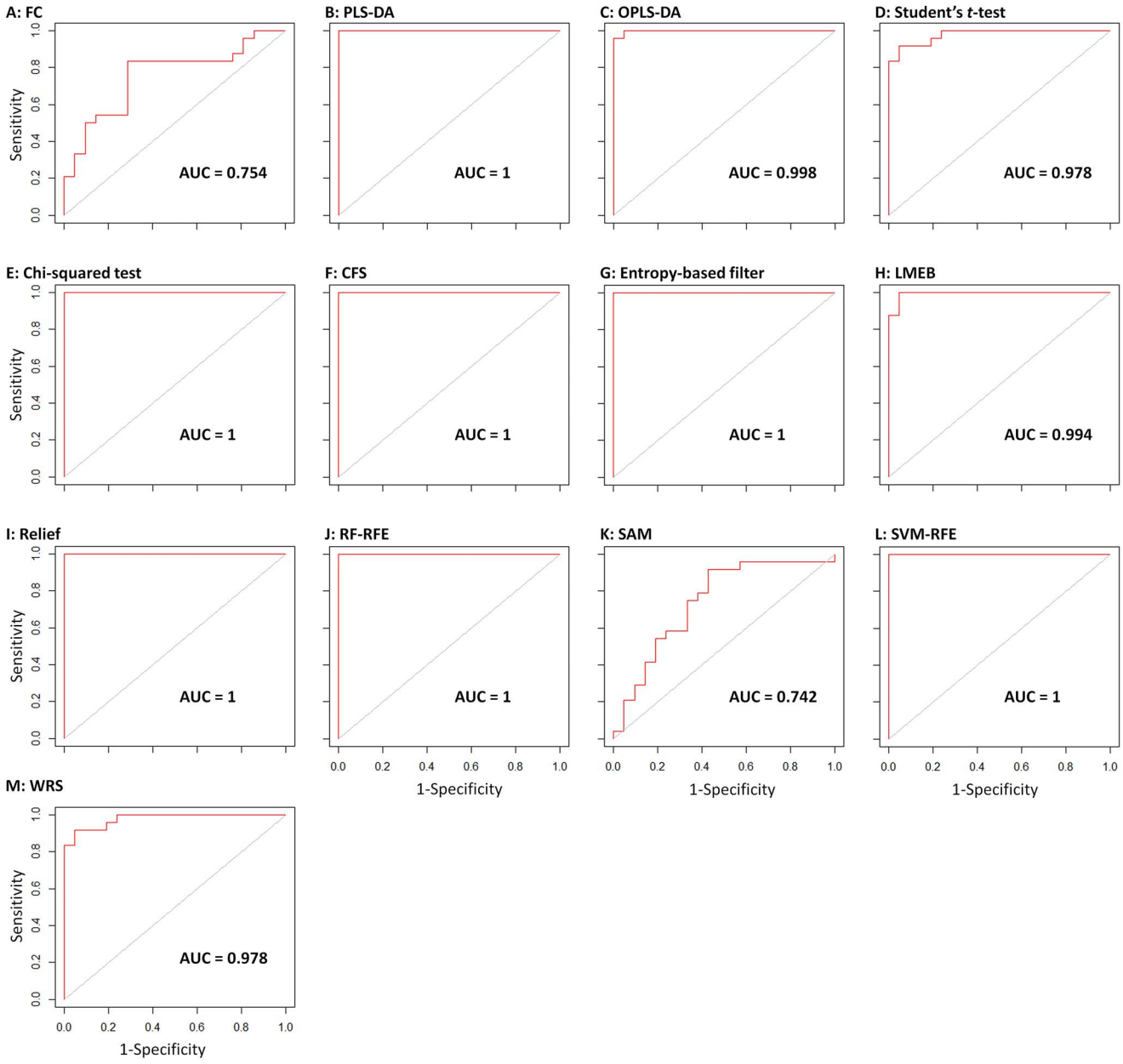

Figure 5. The ROC curves and AUC values of the classification model using the metabolic markers identified by 13 biomarker identification methods. The classification model was constructed using 2-fold cross-validation.

m/z value for the parent ion and the matrix, including m/z values and intensities of MS/MS.

To validate the function of metabolite annotation, the metabolic markers identified by both FC and Student's *t*-test method were applied based on the benchmark dataset (the study of Xiao *et al.*) [91]. The top 10 most upregulated and downregulated metabolites were annotated using the LargeMetabo package. The annotation results of these metabolites, including the name, logFC, adjusted *P*-values and biological interpretation, are shown in Table 3. For example, the most upregulated metabolite, phenylalanyl-threonine, was annotated as an endogenous metabolite. It was reported that phenylalanyl-threonine was one peptide in the analysis of metabolomics [92, 93]. The most downregulated metabolite, UDP-D-glucuronate, was the endogenous substrate of uridine 5′-diphosphate (UDP)-glucuronosyltransferase and was measured in the liver, kidney and placenta [94, 95]. The annotated functions for UDP-D-glucuronate consisted of drug, food, microbial metabolite and plant. Moreover, the results of metabolite annotation using m/z (96.95964) as the enquiry are shown in Supplementary Table S4 available online at http://bib.oxfordjournals.org/. As shown in Supplementary Table S5 and Supplementary Figure S3, available online at http://bib.oxfordjournals.org/, the results of metabolite annotation were created based on the parent ion mass and the MS/MS peak list (m/z and intensity) embedded in the example data.

Based on this enhanced metabolite database, there were eight categories of enrichment analysis for metabolites. Using 302 metabolic markers between triple-negative breast cancer and paired normal breast tissues by both FC and Student's *t*-test methods, the enrichment analysis was validated based on eight categories of metabolite enrichment. The results, including

Downloaded from https://academic.oup.com/bib/article/23/6/bbac455/6768054 by Zhejiang University user on 23 November 2022

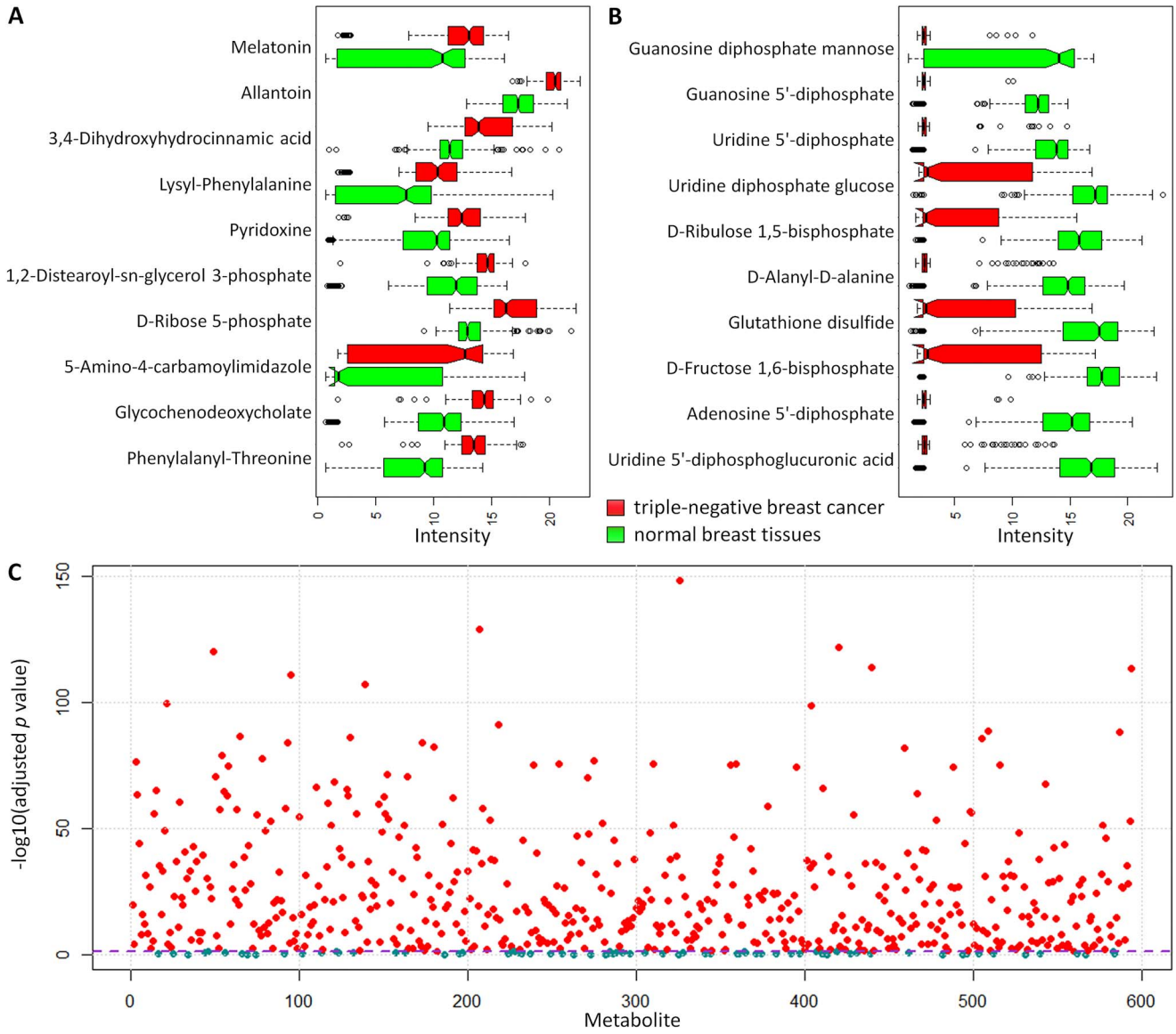

Figure 6. The metabolic markers for triple-negative breast cancer identified using a benchmark dataset (*Cell Res.* 32: 477–90, 2022). Based on the metabolomic data between triple-negative breast cancer patients and the paired normal breast tissues, boxplots of the top 10 most (**A**) upregulated and (**B**) downregulated markers using the FC method were used. (**C**) The plot of logarithmic transformation of the adjusted *P*-values for each metabolite using Student's *t*-test method. The points above dotted line indicate the metabolites with adjusted *P*-values <0.05.

KEGG pathways, metabolic and disease pathways, biological function classes, chemical structure in chemical families, food components and food additives, therapeutic classes of secondary metabolites, species taxonomy and toxins and environmental pollutants, are shown in Figure 7**A**–**H**, respectively. A chord diagram was applied to visualize the enrichment analysis of KEGG pathways, and a bar plot (the number of enrichment terms > 3) or pie chart (the number of enrichment terms ≤ 3) was used to visualize the results of other categories. These pathways and biological functions may be involved in the development of the disease studied or may result from the disease [96, 97]. Moreover, as shown in Supplementary Figure S4 and Supplementary Table S6, available online at http://bib.oxfordjournals.org/, the results of enrichment analysis for KEGG pathways were created using the example data embedded in the LargeMetabo package. Based on the example data and code, the results of enrichment analysis for KEGG pathways were shown in Supplementary Table S7 and Supplementary Figure S5 available online at http://bib.oxfordjournals.org/. And the results of enrichment analysis for the classes of food components and food additives were shown in Supplementary Figure S6 available online at http://bib.oxfordjournals.org/.

To validate the enrichment analysis of the LargeMetabo package, a benchmark dataset [98] of the human adult urinary metabolic variations with age was applied in this study. Based on the metabolic markers identified by the study of a large cohort of 183 adults with age, the KEGG pathways were enriched. As shown in Table 4, there were 10 KEGG pathways in this study, including caffeine metabolism, alanine, aspartate and glutamate metabolism, phenylalanine metabolism, tryptophan metabolism, central carbon metabolism in cancer, arginine biosynthesis, pantothenate and CoA biosynthesis, beta-alanine metabolism, phenylalanine, tyrosine and tryptophan biosynthesis and protein digestion and absorption. To validate the results of enrichment analysis based on the enhanced metabolite database, the relationship between pathways and variations with age was

Downloaded from https://academic.oup.com/bib/article/23/6/bbac455/6768054 by Zhejiang University user on 23 November 2022

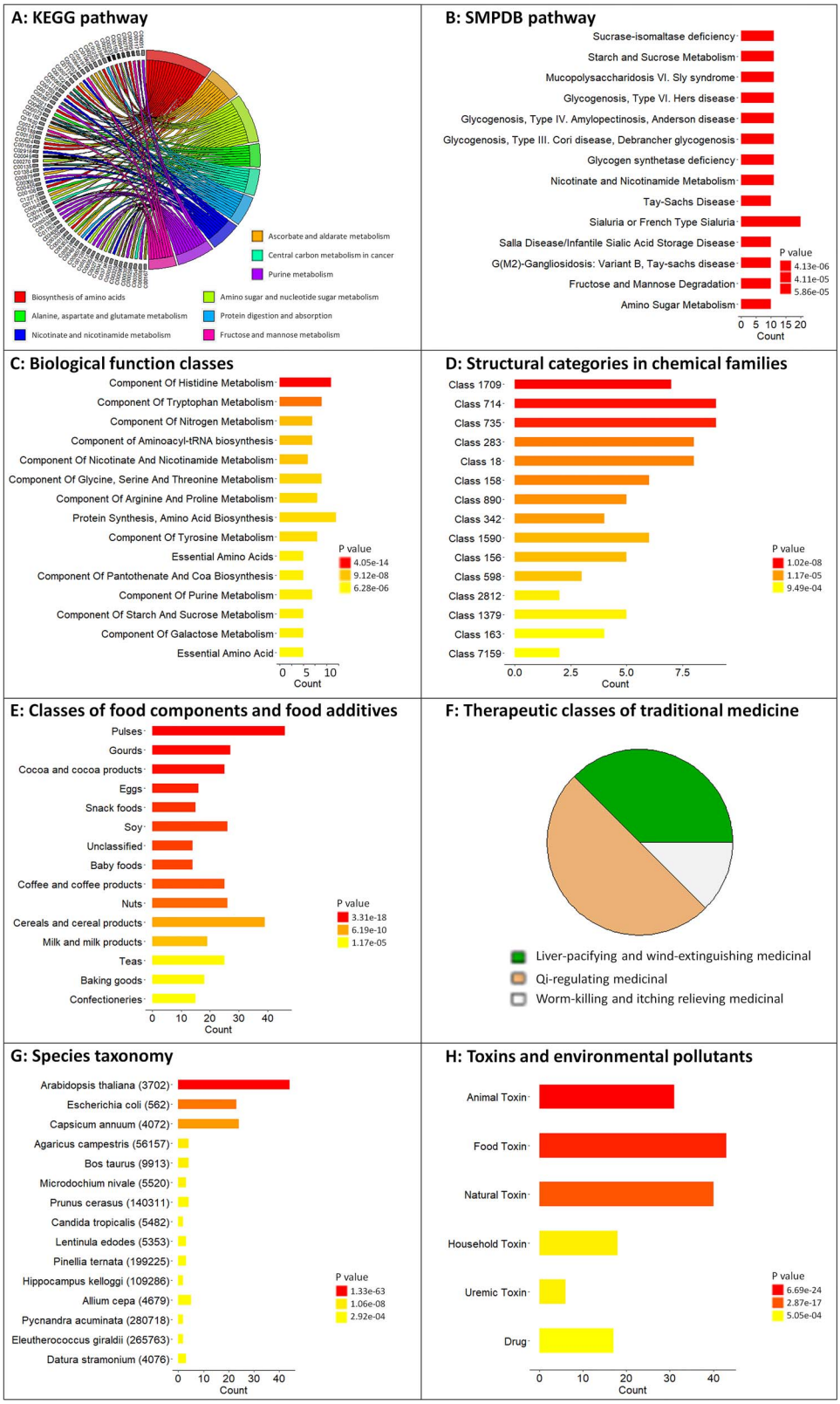

Figure 7. Enrichment analysis was performed based on the metabolite database using these markers by both FC and Student's *t*-test for a benchmark dataset (*Cell Res*. 32: 477–90, 2022). The plots of biological functions included (**A**) KEGG pathway, (**B**) SMPDB pathway, (**C**) biological function classes, (**D**) structural categories in chemical families, (**E**) classes of food components and food additives, (**F**) therapeutic classes of traditional medicine, (**G**) species taxonomy and (**H**) toxins and environmental pollutants.

studied based on a comprehensive literature review. Except for central carbon metabolism in cancer, these nine pathways have been reported to be related to variations with age. Herein, the functions of enrichment analysis have been validated by a literature review, but more benchmark datasets are required for comprehensive validation in the future.

Downloaded from https://academic.oup.com/bib/article/23/6/bbac455/6768054 by Zhejiang University user on 23 November 2022

Table 3. The results of metabolite annotation for the top 10 most upregulated and downregulated metabolites between triple-negative breast cancer patients and the paired normal breast tissues using the LargeMetabo package

No.	Metabolite	logFC	adj *P*-value	Annotation
1	Phenylalanyl-threonine	5.45	2.39E-63	Endogenous
2	Glycochenodeoxycholate	4.22	1.81E-41	Endogenous; food; microbial metabolite
3	5-Amino-4-carbamoylimidazole	4.12	3.83E-13	Endogenous; food; microbial metabolite; TCM ingredient
4	D-ribose 5-phosphate	3.65	6.88E-38	Endogenous; food; microbial metabolite; plant; TCM ingredient
5	1,2-Distearoyl-sn-glycerol 3-phosphate	3.47	9.58E-36	.
6	Pyridoxine	3.35	5.40E-21	Cosmetic; drug; endogenous; food; microbial metabolite; plant; TCM ingredient; toxins/pollutant
7	Lysyl-phenylalanine	3.19	8.78E-15	Endogenous
8	3,4-Dihydroxyhydrocinnamic acid	3.10	2.60E-27	Food; microbial metabolite; plant; TCM ingredient
9	Allantoin	3.04	2.57E-76	Carcinogenic potency; cosmetic; drug; endogenous; food; microbial metabolite; plant; TCM ingredient; toxins/pollutant
10	Melatonin	3.02	2.04E-12	Cosmetic; drug; endogenous; food; microbial metabolite; TCM ingredient; toxins/pollutant
11	Guanosine diphosphate mannose	−8.58	8.24E-85	Endogenous; food; microbial metabolite; plant; TCM ingredient
12	Guanosine 5′-diphosphate	−8.68	4.49E-149	Endogenous; food; microbial metabolite; plant
13	UDP	−9.32	9.78E-112	Endogenous; food; microbial metabolite; plant; TCM ingredient
14	Uridine diphosphate glucose (UDP-D-glucose)	−9.75	3.32E-56	Endogenous; food; microbial metabolite; plant; TCM ingredient
15	D-ribulose 1,5-bisphosphate	−9.78	1.53E-60	.
16	D-alanyl-D-alanine	−10.00	3.79E-100	Endogenous; food; microbial metabolite; plant; TCM ingredient
17	Glutathione disulfide	−10.66	9.85E-64	Cosmetic; drug; endogenous; food; microbial metabolite; plant; TCM ingredient
18	D-fructose 1,6-bisphosphate	−10.70	1.01E-56	Endogenous; food; microbial metabolite; TCM ingredient
19	Adenosine 5′-diphosphate	−10.80	5.22E-121	Endogenous; food; microbial metabolite; plant; TCM ingredient
20	UDP-D-glucuronate	−12.26	7.80E-130	Drug; endogenous; food; microbial metabolite; plant

Table 4. Validation on the enrichment results of KEGG pathways using metabolic makers identified by the study of 183 adults with age (*J Proteome Res*. 14: 3322–35, 2015); the relationship between KEGG pathways and variations with age is shown according to a comprehensive literature review

No.	KEGG pathways	The relationship between enrichment term and variations with age
1	Caffeine metabolism	Critical changes in adenosinergic neurotransmission occur with aging, and caffeine is an adenosine receptor antagonist (*Neurobiol Aging*. 26: 957–64, 2005). The higher concentrations of caffeine and its metabolites suggest an increased consumption and/or a decreased metabolic activity with age (*J Proteome Res*. 14: 3322–35, 2015).
2	Alanine, aspartate and glutamate metabolism	Aspartic acid and glutamate show significant concentration changes with age (*Proc Natl Acad Sci U S A*. 108: 6181–6, 2011).
3	Phenylalanine metabolism	The rate of assimilation and hydroxylation of phenylalanine diminishes with age, which defines the age-related characteristic of phenylalanine metabolism (*Vopr Pitan*. 1982. 4: 52–5, 1982). Phenylalanine shows significant concentration changes with age (*Proc Natl Acad Sci U S A*. 108: 6181–6, 2011).
4	Tryptophan metabolism	Studies in multiple organisms have implicated tryptophan metabolism as a powerful regulator of lifespan (*Trends Mol Med*. 19: 336–44, 2013).
5	Central carbon metabolism in cancer	–
6	Arginine biosynthesis	The demonstrated anti-aging benefits of L-arginine show greater potential than any pharmaceutical or nutraceutical agent ever previously discovered (*J Adv Res*. 1: 169–77, 2010).
7	Pantothenate and CoA biosynthesis	Extracellular precursors, especially vitamin B5 (calcium pantothenate), is essential for eukaryotic cells to obtain CoA, which plays key roles in aging-related neurodegeneration (*Nat Chem Biol*. 11: 784–92, 2015).
8	Beta-alanine metabolism	The anti-aging effect of beta-alanine is shown by significantly increasing the skeletal muscle carnosine (*J Int Soc Sports Nutr*. 5: 21, 2008).
9	Phenylalanine, tyrosine and tryptophan biosynthesis	The metabolism of tyrosine and phenylalanine is significantly affected by pyridoxine which is an essential co-factor in the body's fight against elevated homocysteine levels. A high homocysteine level is associated with aging-related alzheimer's disease (*Int J Pharma Bio Sci*. 1: 1–17, 2010).
10	Protein digestion and absorption	The rate of protein digestion affects protein gain differently during aging in humans (*J Physiol*. 549: 635–44, 2003).

Downloaded from https://academic.oup.com/bib/article/23/6/bbac455/6768054 by Zhejiang University user on 23 November 2022

There were still some limitations for the LargeMetabo package. For large-scale metabolomics, multiple datasets from different analytical experiments can be combined into an integrated dataset based on the mass and RT using this package. However, only when the sample preparation and experimental conditions in different analytical experiments are similar, these datasets can fulfill the requests for data integration in MS-based metabolomics. In the future, the applications of this data integration strategy can be extended to a larger scope of metabolomics. Moreover, more experimental datasets should be applied in the LargeMetabo package to validate the functions of processing and analyzing large-scale metabolomic data.

Conclusion

Currently, data integration from multiple analytical experiments and enhanced metabolite annotation are urgently needed for processing and analyzing large-scale metabolomic data. To facilitate the flexibility and reproducibility of data processing and analysis, the LargeMetabo R package was developed in this study. This package can (1) integrate multiple analytical experiments into a combined dataset; (2) identify metabolic markers by selecting the most appropriate method by performance assessment and (3) conduct enhanced metabolite annotation and enrichment analysis based on a new metabolite database. The LargeMetabo package is freely available from https://github.com/LargeMetabo/LargeMetabo.

Key Points

- The LargeMetabo package was developed based on R code for processing and analyzing large-scale metabolomic data.
- The LargeMetabo package can integrate multiple datasets from different analytical experiments to effectively boost the power of statistical analysis in large-scale metabolomics.
- The appropriate method can be selected by assessment among various biomarker identification methods for large-scale metabolic data.
- The functions of metabolite annotation and enrichment analysis based on an enhanced metabolite database are provided for interpretation and biological mechanism of disease.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Data availability

The implemented code and experimental dataset are available online at https://github.com/LargeMetabo/LargeMetabo.

Funding

National Natural Science Foundation of China (62201289); Natural Science Foundation of Jiangsu Province (BK20210597); NUPTSF (NY220169); Science and Technology Research Program of Chongqing Municipal Education Commission (KJQN202100538).

References

1. Han S, Van Treuren W, Fischer CR, *et al.* A metabolomics pipeline for the mechanistic interrogation of the gut microbiome. *Nature* 2021;**595**:415–20.
2. Wishart DS. Emerging applications of metabolomics in drug discovery and precision medicine. *Nat Rev Drug Discov* 2016;**15**:473–84.
3. Fu J, Zhang Y, Wang Y, *et al.* Optimization of metabolomic data processing using NOREVA. *Nat Protoc* 2022;**17**:129–51.
4. Shanmuganathan M, Kroezen Z, Gill B, *et al.* The maternal serum metabolome by multisegment injection-capillary electrophoresis-mass spectrometry: a high-throughput platform and standardized data workflow for large-scale epidemiological studies. *Nat Protoc* 2021;**16**:1966–94.
5. Kim T, Tang O, Vernon ST, *et al.* A hierarchical approach to removal of unwanted variation for large-scale metabolomics data. *Nat Commun* 2021;**12**:4992.
6. Li C, Hou L, Sharma BY, *et al.* Developing a new intelligent system for the diagnosis of tuberculous pleural effusion. *Comput Methods Programs Biomed* 2018;**153**:211–25.
7. Hu L, Hong G, Ma J, *et al.* An efficient machine learning approach for diagnosis of paraquat-poisoned patients. *Comput Biol Med* 2015;**59**:116–24.
8. Yu B, Zanetti KA, Temprosa M, *et al.* The consortium of metabolomics studies (COMETS): metabolomics in 47 prospective cohort studies. *Am J Epidemiol* 2019;**188**:991–1012.
9. Teo G, Chew WS, Burla BJ, *et al.* MRMkit: automated data processing for large-scale targeted metabolomics analysis. *Anal Chem* 2020;**92**:13677–82.
10. Fu J, Tang J, Wang Y, *et al.* Discovery of the consistently well-performed analysis chain for SWATH-MS based pharmacoproteomic quantification. *Front Pharmacol* 2018;**9**:681.
11. Dunn WB, Broadhurst D, Begley P, *et al.* Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nat Protoc* 2011;**6**:1060–83.
12. Li B, Tang J, Yang Q, *et al.* NOREVA: normalization and evaluation of MS-based metabolomics data. *Nucleic Acids Res* 2017;**45**:W162–70.
13. Cambiaghi A, Ferrario M, Masseroli M. Analysis of metabolomic data: tools, current strategies and future challenges for omics data integration. *Brief Bioinform* 2017;**18**:498–510.
14. da Silva RR, Dorrestein PC, Quinn RA. Illuminating the dark matter in metabolomics. *Proc Natl Acad Sci U S A* 2015;**112**:12549–50.
15. Zhang S, Amahong K, Sun X, *et al.* The miRNA: a small but powerful RNA for COVID-19. *Brief Bioinform* 2021;**22**:1137–49.
16. Fu J, Zhang Y, Liu J, *et al.* Pharmacometabonomics: data processing and statistical analysis. *Brief Bioinform* 2021;**22**:bbab138.
17. Wen B, Mei Z, Zeng C, *et al.* metaX: a flexible and comprehensive software for processing metabolomics data. *BMC Bioinformat* 2017;**18**:183.
18. Yang Q, Li B, Chen S, *et al.* MMEASE: online meta-analysis of metabolomic data by enhanced metabolite annotation, marker selection and enrichment analysis. *J Proteomics* 2021;**232**:104023.
19. Pang Z, Chong J, Zhou G, *et al.* MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights. *Nucleic Acids Res* 2021;**49**:W388–96.
20. Ding X, Yang F, Chen Y, *et al.* Norm ISWSVR: a data integration and normalization approach for large-scale metabolomics. *Anal Chem* 2022;**94**:7500–9.

Downloaded from https://academic.oup.com/bib/article/23/6/bbac455/6768054 by Zhejiang University user on 23 November 2022

21. Guitton Y, Tremblay-Franco M, Le Corguille G, *et al.* Create, run, share, publish, and reference your LC-MS, FIA-MS, GC-MS, and NMR data analysis workflows with the Workflow4Metabolomics 3.0 Galaxy online infrastructure for metabolomics. *Int J Biochem Cell Biol* 2017;**93**:89–101.
22. Tautenhahn R, Patti GJ, Rinehart D, *et al.* XCMS Online: a web-based platform to process untargeted metabolomic data. *Anal Chem* 2012;**84**:5035–9.
23. Gowda H, Ivanisevic J, Johnson CH, *et al.* Interactive XCMS Online: simplifying advanced metabolomic data processing and subsequent statistical analyses. *Anal Chem* 2014;**86**: 6931–9.
24. Zhang Y, Tamba C, Wen Y, *et al.* mrMLM v4.0.2: an R platform for multi-locus genome-wide association studies. *Genom Proteom Bioinform* 2020;**18**:481–7.
25. Chong J, Xia J. MetaboAnalystR: an R package for flexible and reproducible analysis of metabolomics data. *Bioinformatics* 2018;**34**:4313–4.
26. Liu X, Xu Y, Wang R, *et al.* A network-based algorithm for the identification of moonlighting noncoding RNAs and its application in sepsis. *Brief Bioinform* 2021;**22**:581–8.
27. Fu T, Zheng G, Tu G, *et al.* Exploring the binding mechanism of metabotropic glutamate receptor 5 negative allosteric modulators in clinical trials by molecular dynamics simulations. *ACS Chem Nerosci* 2018;**9**:1492–502.
28. Kastenmuller G, Romisch-Margl W, Wagele B, *et al.* metaP-server: a web-based metabolomics data analysis tool. *J Biomed Biotechnol* 2011;**2011**:839862.
29. Zhang W, Lei Z, Huhman D, *et al.* MET-XAlign: a metabolite cross-alignment tool for LC/MS-based comparative metabolomics. *Anal Chem* 2015;**87**:9114–9.
30. De Livera AM, Dias DA, De Souza D, *et al.* Normalizing and integrating metabolomics data. *Anal Chem* 2012;**84**: 10768–76.
31. Irshad O, Khan MUG. A comparative analysis of biological data integration systems famous for data exploitation and knowledge discovery. *Curr Bioinform* 2021;**16**:662–81.
32. Kuligowski J, Perez-Guaita D, Lliso I, *et al.* Detection of batch effects in liquid chromatography-mass spectrometry metabolomic data using guided principal component analysis. *Talanta* 2014;**130**:442–8.
33. Sanchez-Illana A, Pineiro-Ramos JD, Sanjuan-Herraez JD, *et al.* Evaluation of batch effect elimination using quality control replicates in LC-MS metabolite profiling. *Anal Chim Acta* 2018;**1019**:38–48.
34. Lazar C, Meganck S, Taminau J, *et al.* Batch effect removal methods for microarray gene expression data integration: a survey. *Brief Bioinform* 2013;**14**:469–90.
35. Yang Q, Wang Y, Zhang Y, *et al.* NOREVA: enhanced normalization and evaluation of time-course and multi-class metabolomic data. *Nucleic Acids Res* 2020;**48**:W436–48.
36. Beckonert O, Keun HC, Ebbels TM, *et al.* Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nat Protoc* 2007;**2**:2692–703.
37. Ren S, Hinzman AA, Kang EL, *et al.* Computational and statistical analysis of metabolomics data. *Metabolomics* 2015;**11**: 1492–513.
38. Zou Q, Lin G, Jiang X, *et al.* Sequence clustering in bioinformatics: an empirical study. *Brief Bioinform* 2020;**21**:1–10.
39. Goodwin CR, Covington BC, Derewacz DK, *et al.* Structuring microbial metabolic responses to multiplexed stimuli via self-organizing metabolomics maps. *Chem Biol* 2015;**22**:661–70.
40. Want EJ, Wilson ID, Gika H, *et al.* Global metabolic profiling procedures for urine using UPLC-MS. *Nat Protoc* 2010;**5**: 1005–18.
41. Jia C, Zuo Y, Zou Q. O-GlcNAcPRED-II: an integrated classification algorithm for identifying O-GlcNAcylation sites based on fuzzy undersampling and a K-means PCA oversampling technique. *Bioinformatics* 2018;**34**:2029–36.
42. Denkert C, Budczies J, Kind T, *et al.* Mass spectrometry-based metabolic profiling reveals different metabolite patterns in invasive ovarian carcinomas and ovarian borderline tumors. *Cancer Res* 2006;**66**:10795–804.
43. Gromski PS, Muhamadali H, Ellis DI, *et al.* A tutorial review: metabolomics and partial least squares-discriminant analysis—a marriage of convenience or a shotgun wedding. *Anal Chim Acta* 2015;**879**:10–23.
44. Bylesjo M, Rantalainen M, Cloarec O, *et al.* OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. *J Chemometr* 2006;**20**:341–51.
45. Begun A. Power estimation of the t test for detecting differential gene expression. *Funct Integr Genomics* 2008;**8**:109–13.
46. Lee IH, Lushington GH, Visvanathan M. A filter-based feature selection approach for identifying potential biomarkers for lung cancer. *J Clin Bioinforma* 2011;**1**:11.
47. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics* 2007;**23**:2507–17.
48. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 2004;**3**: Article 3.
49. Baumgartner C, Bohm C, Baumgartner D, *et al.* Supervised machine learning techniques for the classification of metabolic disorders in newborns. *Bioinformatics* 2004;**20**:2985–96.
50. Darst BF, Malecki KC, Engelman CD. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC Genet* 2018;**19**:65.
51. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* 2001;**98**:5116–21.
52. Lin X, Yang F, Zhou L, *et al.* A support vector machine-recursive feature elimination feature selection method based on artificial contrast variables and mutual information. *J Chromatogr B* 2012;**910**:149–55.
53. Rosner B, Glynn RJ, Lee ML. Incorporation of clustering effects for the Wilcoxon rank sum test: a large-sample approach. *Biometrics* 2003;**59**:1089–98.
54. Hu J, Chen H, Heidari AA, *et al.* Orthogonal learning covariance matrix for defects of grey wolf optimizer: Insights, balance, diversity, and feature selection. *Knowl Based Syst* 2021;**213**:106684.
55. Hu J, Gui W, Heidari AA, *et al.* Dispersed foraging slime mould algorithm: Continuous and binary variants for global optimization and wrapper-based feature selection. *Knowl Based Syst* 2022;**237**:107761.
56. Zhang Y, Liu R, Wang X, *et al.* Boosted binary Harris hawks optimizer and feature selection. *Eng Comput* 2021;**37**:3741–70.
57. Christin C, Hoefsloot HC, Smilde AK, *et al.* A critical assessment of feature selection methods for biomarker discovery in clinical proteomics. *Mol Cell Proteomics* 2013;**12**:263–76.
58. Li F, Zhou Y, Zhang Y, *et al.* POSREG: proteomic signature discovered by simultaneously optimizing its reproducibility and generalizability. *Brief Bioinform* 2022;**23**:bbac040.
59. Wang R, Zheng X, Wang J, *et al.* Improving bulk RNA-seq classification by transferring gene signature from single cells in acute myeloid leukemia. *Brief Bioinform* 2022;**23**:bbac002.

Downloaded from https://academic.oup.com/bib/article/23/6/bbac455/6768054 by Zhejiang University user on 23 November 2022

60. Li F, Yin J, Lu M, *et al.* ConSIG: consistent discovery of molecular signature from OMIC data. *Brief Bioinform* 2022;**23**:bbac253.
61. Yang Q, Li B, Tang J, *et al.* Consistent gene signature of schizophrenia identified by a novel feature selection strategy from comprehensive sets of transcriptomic data. *Brief Bioinform* 2020;**21**:1058–68.
62. Yang Q, Li Y, Li B, *et al.* A novel multi-class classification model for schizophrenia, bipolar disorder and healthy controls using comprehensive transcriptomic data. *Comput Biol Med* 2022;**148**:105956.
63. Wishart DS, Guo A, Oler E, *et al.* HMDB 5.0: the human metabolome database for 2022. *Nucleic Acids Res* 2022;**50**: D622–31.
64. Cui Q, Lewis IA, Hegeman AD, *et al.* Metabolite identification via the madison metabolomics consortium database. *Nat Biotechnol* 2008;**26**:162–4.
65. Smith CA, O'Maille G, Want EJ, *et al.* METLIN: a metabolite mass spectral database. *Ther Drug Monit* 2005;**27**:747–51.
66. Sud M, Fahy E, Cotter D, *et al.* LMSD: LIPID MAPS structure database. *Nucleic Acids Res* 2007;**35**:D527–32.
67. Horai H, Arita M, Kanaya S, *et al.* MassBank: a public repository for sharing mass spectral data for life sciences. *J Mass Spectrom* 2010;**45**:703–14.
68. Zhang Y, Zheng Q. In silico analysis revealed a unique binding but ineffective mode of amantadine to influenza virus B M2 channel. *J Phys Chem Lett* 2021;**12**:1169–74.
69. Zhang Y, Zhang H, Zheng Q. In silico study of membrane lipid composition regulating conformation and hydration of influenza virus B M2 channel. *J Chem Inf Model* 2020;**60**:3603–15.
70. Lin B, Zhang H, Zheng Q. How do mutations affect the structural characteristics and substrate binding of CYP21A2? An investigation by molecular dynamics simulations. *Phys Chem Chem Phys* 2020;**22**:8870–7.
71. Zhang Y, Zheng Q. What are the effects of the serine triad on proton conduction of an influenza B M2 channel? An investigation by molecular dynamics simulations. *Phys Chem Chem Phys* 2019;**21**:8820–6.
72. Zhang C, Tao L, Qin C, *et al.* CFam: a chemical families database based on iterative selection of functional seeds and seed-directed compound clustering. *Nucleic Acids Res* 2015;**43**: D558–65.
73. Wishart DS, Feunang YD, Guo AC, *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 2018;**46**:D1074–82.
74. Sajed T, Marcu A, Ramirez M, *et al.* ECMDB 2.0: A richer resource for understanding the biochemistry of E. coli. *Nucleic Acids Res* 2016;**44**:D495–501.
75. Kanehisa M, Furumichi M, Sato Y, *et al.* KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res* 2021;**49**: D545–51.
76. Castrignano T, De Meo PD, Cozzetto D, *et al.* The PMDB: protein model database. *Nucleic Acids Res* 2006;**34**:D306–9.
77. Wishart D, Arndt D, Pon A, *et al.* T3DB: the toxic exposome database. *Nucleic Acids Res* 2015;**43**:D928–34.
78. Huang L, Xie D, Yu Y, *et al.* TCMID 2.0: a comprehensive resource for TCM. *Nucleic Acids Res* 2018;**46**:D1117–20.
79. Jewison T, Knox C, Neveu V, *et al.* YMDB: the yeast metabolome database. *Nucleic Acids Res* 2012;**40**:D815–20.
80. Zhang Y, Li J, Zhang X, *et al.* Advances of mechanisms-related metabolomics in Parkinson's disease. *Front Neurosci* 2021;**15**:614251.
81. Hong J, Luo Y, Zhang Y, *et al.* Protein functional annotation of simultaneously improved stability, accuracy and false discovery rate achieved by a sequence-based deep learning. *Brief Bioinform* 2020;**21**:1437–47.
82. Xia W, Zheng L, Fang J, *et al.* PFmulDL: a novel strategy enabling multi-class and multi-label protein function annotation by integrating diverse deep learning methods. *Comput Biol Med* 2022;**145**:105465.
83. Ressom HW, Xiao JF, Tuli L, *et al.* Utilization of metabolomics to identify serum biomarkers for hepatocellular carcinoma in patients with liver cirrhosis. *Anal Chim Acta* 2012;**743**: 90–100.
84. Haug K, Cochrane K, Nainala VC, *et al.* MetaboLights: a resource evolving in response to the needs of its scientific community. *Nucleic Acids Res* 2020;**48**:D440–4.
85. Wehrens R, Franceschi P, Vrhovsek U, *et al.* Stability-based biomarker selection. *Anal Chim Acta* 2011;**705**: 15–23.
86. Liu X, Zheng X, Wang J, *et al.* A long non-coding RNA signature for diagnostic prediction of sepsis upon ICU admission. *Clin Transl Med* 2020;**10**:e123.
87. Zhang Y, Liu R, Heidari AA, *et al.* Towards augmented kernel extreme learning models for bankruptcy prediction: algorithmic behavior and comprehensive analysis. *Neurocomputing* 2021;**430**: 185–212.
88. Chen H, Wang G, Ma C, *et al.* An efficient hybrid kernel extreme learning machine approach for early diagnosis of Parkinson's disease. *Neurocomputing* 2016;**184**:131–44.
89. Tang J, Mou M, Wang Y, *et al.* MetaFS: performance assessment of biomarker discovery in metaproteomics. *Brief Bioinform* 2021;**22**:bbaa105.
90. Hong J, Luo Y, Mou M, *et al.* Convolutional neural network-based annotation of bacterial type IV secretion system effectors with enhanced accuracy and reduced false discovery. *Brief Bioinform* 2020;**21**:1825–36.
91. Xiao Y, Ma D, Yang YS, *et al.* Comprehensive metabolomics expands precision medicine for triple-negative breast cancer. *Cell Res* 2022;**32**:477–90.
92. Zhao Q, Shen H, Su KJ, *et al.* A joint analysis of metabolomic profiles associated with muscle mass and strength in Caucasian women. *Aging* 2018;**10**:2624–35.
93. Yin J, Li F, Zhou Y, *et al.* INTEDE: interactome of drug-metabolizing enzymes. *Nucleic Acids Res* 2021;**49**:D1233–43.
94. Cappiello M, Giuliani L, Rane A, *et al.* Uridine 5′-diphosphoglucuronic acid (UDPGLcUA) in the human fetal liver, kidney and placenta. *Eur J Drug Metab Pharmacokinet* 2000;**25**:161–3.
95. Fu T, Li F, Zhang Y, *et al.* VARIDT 2.0: structural variability of drug transporter. *Nucleic Acids Res* 2022;**50**:D1417–31.
96. Zhu F, Li X, Yang S, *et al.* Clinical success of drug targets prospectively predicted by in silico study. *Trends Pharmacol Sci* 2018;**39**: 229–31.
97. Xue W, Fu T, Deng S, *et al.* Molecular mechanism for the allosteric inhibition of the human serotonin transporter by antidepressant escitalopram. *ACS Chem Neurosci* 2022;**13**: 340–51.
98. Thevenot EA, Roux A, Xu Y, *et al.* 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* 2015;**14**: 3322–35.

Downloaded from https://academic.oup.com/bib/article/23/6/bbac455/6768054 by Zhejiang University user on 23 November 2022