OXFORD UNIVERSITY PRESS | Briefings In Bioinformatics

mmeaseR: an R package for Data Integration, Marker Identification, Metabolite Annotation and Enrichment in Large-scale and Long-term Metabolomics

Journal:	*Briefings in Bioinformatics*
Manuscript ID	BIB-22-0984
Manuscript Type:	Problem solving protocol
Date Submitted by the Author:	26-May-2022
Complete List of Authors:	YANG, Qingxia; Nanjing University of Posts and Telecommunications, Department of Bioinformatics Wang, Panpan; Huanghuai University Xie, Jicheng; Nanjing University of Posts and Telecommunications, Department of Bioinformatics Feng, Yuhao; Nanjing University of Posts and Telecommunications, Department of Bioinformatics Liu, Ziqiang; Nanjing University of Posts and Telecommunications, Department of Bioinformatics Zhu, Feng; Zhejiang University Li, Bo; Chongqing Normal University, College of Life Sciences
Keywords:	Large-scale and long-term metabolomics, Data integration, Metabolite annotation, R package

mmeaseR: an R package for Data Integration, Marker Identification, Metabolite Annotation and Enrichment in Large-scale and Long-term Metabolomics

Qingxia Yang[1, *], Bo Li[2], Panpan Wang[3], Jicheng Xie[1], Yuhao Feng[1], Ziqiang Liu[1], Feng Zhu[4, *]

[1] Department of Bioinformatics, Smart Health Big Data Analysis and Location Services Engineering Lab of Jiangsu Province, School of Geographic and Biologic Information, Nanjing University of Posts and Telecommunications, Nanjing, 210023, China.

[2] College of Life Sciences, Chongqing Normal University, Chongqing, Chongqing 401331, China.

[3] College of Chemistry and Pharmaceutical Engineering, Huanghuai University, Zhumadian 463000, China.

[4] College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, Zhejiang 310058, China.

* To whom the correspondence should be addressed: Prof. Feng Zhu, E-mail: zhufeng@zju.edu.cn, and Dr. Qingxia Yang, E-mail: yangqx@njupt.edu.cn.

Running Title: mmeaseR package for large-scale metabolomics

Abstract

Large-scale and long-term metabolomics have attracted widespread attention in biomedical studies focused on identifying biomarkers and interpreting the mechanisms of a variety of complex diseases. In these studies, ineffective methods of data integration and limited capacity for metabolite annotation are serious challenges. The online tool MMEASE (https://idrblab.org/mmease/) was constructed to enable the integration of multiple analytical experiments with enhanced metabolite annotation and enrichment analysis. However, the web interface presents the limitation of reproducible analysis and dealing with large datasets. To address this limitation, a companion R package (mmeaseR) based on R code was developed. This package is unique in that it is capable of (1) integrating multiple analytical experiments to effectively boost the power of statistical analysis, (2) providing enhanced annotation for metabolites, and (3) conducting enrichment analysis based on a metabolite database. The mmeaseR package complements the MMEASE web server to facilitate the flexibility and reproducible analysis of large-scale and long-term metabolomics. The package is freely available from https://github.com/mmeaseR/mmeaseR.

Key words

Large-scale and long-term metabolomics; Data integration; Metabolite annotation; R package

Introduction

Metabolomics profiling of whole sets of metabolites in biological samples has been widely applied for identifying biomarkers in the diagnosis and prediction of disease [1]. A large number of samples are essential in metabolomics to increase the statistical power and address important issues related to heterogeneity in disease biology [2, 3]. Large-scale and long-term metabolomics is a powerful technique that has attracted widespread attention [4]. However, these studies are greatly hampered by the inefficiency of data integration [5] and the limited capacity for metabolite annotation [6].

To cope with these problems of data integration and metabolite annotation, a useful online web-server, MMEASE (https://idrblab.org/mmease/), was developed in 2021 [7]. Data can be integrated from different analytical experiments using this server to effectively boost the power of statistical analysis. Additionally, enriched annotations are provided for >330 thousand metabolites, and enrichment analysis using various categories/subcategories can be constructed. Overall, MMEASE aims to supply a comprehensive service for large-scale and long-term metabolomics to provide valuable guidance for current biomedical studies.

Despite the user-friendliness of MMEASE, the online tool comes with its inherent limitations [8]. For instance, the analytical functions through the web interface make it difficult for users to reproduce their results when reanalyzing the dataset after a long time. Moreover, the web interface also presents significant constraints in terms of analyzing larger datasets [9]. Therefore, it is highly necessary to develop a companion R package with reproducibility and flexibility for large-scale and long-term metabolomic studies.

In this study, the R package mmeaseR based on R code was developed for large-scale and long-term metabolomics. This package is unique in that it is capable of (1) integrating datasets for multiple analytical experiments, (2) separating samples in combined datasets, (3) identifying metabolic markers, (4) providing enhanced annotation for metabolites, and (5) conducting enrichment analysis based on a metabolite database. The mmeaseR package can complement the MMEASE web server to facilitate the flexibility and reproducible analysis of large-scale and long-term metabolomics. The package is freely available from https://github.com/mmeaseR/mmeaseR.

Design and Implementation

The development version of mmeaseR 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 base from the web server, with extensive modifications to ensure functional compatibility across both the web and the R command line. The analytical workflow of the five steps is summarized in **Figure 1**.

Step 1: Dataset Integration

Large-scale and long-term metabolomics could effectively boost the power of statistical analysis and accurately define biological variation. In these studies, hundreds/thousands of samples are in high demand and are frequently broken down into small analytical experiments [10]. In step 1 of the mmeaseR package, the multiple datasets from different analytical experiments can be integrated, and five methods are provided to remove unwanted experimental variations. Herein, a useful strategy is used to integrate multiple analytical experiments applying the permissible tolerance of retention time and the accurate mass of a certain metabolite peak [11]. However, various unwanted variations resulting from different batch effects can emerge during the process of data integration. Therefore, it is essential to remove these unwanted variations after data integration [12]. There are various methods provided in mmeaseR for removing batch effects in different analytical experiments, including batch mean-centering (BMC/PAMR) [13], empirical Bayes method (COMBAT) [14], and global normalization (Z score) [15].

Step 2: Sample Separation

There are four methods of sample separation for visualizing the clustering and separation of different samples in step 2. After data integration and batch effect removal for multiple analytical experiments in large-scale metabolomics, sample separation is used for the visualization of different samples [16]. In the mmeaseR package, four methods are provided for sample separation, including hierarchical clustering analysis (HCA) [17], *k*-means clustering (KMC) [18], self-organizing map (SOM) [19] and principal component analysis (PCA) [20].

Step 3: Marker Identification

There are 13 marker identification strategies provided for discovering metabolic markers for the given datasets in step 3. Because of the great variations of the statistical theories and model assumptions, different strategies could lead to contradictory results even for the same data. The appropriate application of a marker identification strategy is heavily dependent on the innate characteristics of a certain study [21, 22]. To achieve systematic selection, 13 popular strategies are provided in the mmeaseR package, including fold change (FC) [23], partial least squares discrimination analysis (PLS-DA) [24], orthogonal PLS-DA (OPLS-DA) [25], Student's *t*-test [26], Chi-squared test [27], correlation-based feature selection (CFS) [28], entropy-based filter method [28], linear models and empirical Bayes method [29], recursive elimination of features (Relief) [30], random forest-recursive feature elimination (RF-RFE) [31], significance analysis for microarrays (SAM) [32], support vector machine-recursive feature elimination (SVM-RFE) [33], and Wilcoxon rank sum (WRS) [34].

Step 4: Metabolite Annotation

The MS peaks of interest need to be annotated with functions based on their masses, but less than 2% of the detected MS peaks can be annotated in untargeted MS-based metabolomics [6]. Moreover, the way to convert a raw peak feature into a metabolite with biologically interpretable information remains a major challenge [35]. Herein, the metabolites detected by mass spectrometry and tandem mass spectrometry can be annotated in step 4. The metabolite annotation is based on a novel metabolite database, which was constructed by systematic literature reviews and searching from many databases (HMDB, MMCD, LMSD, MoNA, *etc*.). In this novel metabolite database, detailed biological annotation data have been added for these metabolites including the endogenous and exogenous factors (food, plant, drug, cosmetics, microbe, toxin, environmental pollutant and so on). This information was obtained by performing literature reviews and searching different databases (HMDB, T3DB, ECMDB, YMDB, PMDB, Drugbank, CFAM, TCMID, KEGG and so on).

Step 5: Enrichment Analysis

There are eight categories from the metabolite database used for enrichment analysis in step 5. This enrichment analysis is performed to reveal the aggregation degree of functional roles or exogenous sources for the studied metabolite list using a hypergeometric test. The eight categories contain a large collection of metabolites, including (1) 310 metabolic pathways in KEGG pathways; (2) 736 human metabolic and disease pathways of SMPDB; (3) biological function classes reflecting the biological roles of metabolites; (4) structural categories for human metabolites in chemical families; (5) food sources of metabolites in FooDB; (6) therapeutic classes of secondary metabolites from traditional medicine; (7) species taxonomy from traditional medicine; and (8) categories of toxins and environmental pollutants.

Results and Discussion

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

Case 1: Data Integration for Three Analytical Experiments

In the mmeaseR package, multiple analytical experiments can be integrated into a comprehensive dataset. As the input, the *csv* file of the feature-by-sample matrix contains five essential columns providing the information of mass, retention time, intensity, isotope and adduct. The samples must be kept in columns with the sample names in the first row. The labels indicating different sample groups of cases and controls are kept in the second row. The example of input files with the corresponding contents separated by comma is provided in the mmeaseR package. The binary matrix

after data integration for the example datasets is shown in **Supplementary Table S1**. After data integration, three methods are provided to remove the batch effects among different analytical experiments, including batch mean-centering (BMC/PAMR), the empirical Bayes method (ComBat/EB), and global normalization (GlobalNorm). Using an example dataset in mmeaseR, the binary matrix after batch effect removal is shown in **Supplementary Table S2**.

The performance of integrating multiple analytical experiments was evaluated by the dataset from MTBLS17 [36] from MetaboLights [37]. In this benchmark dataset, 78 hepatocellular carcinoma (HCC) patients and 184 cirrhotic (CIR) controls were detected by metabolomic profiling in three analytical experiments. In the first experiment, there were 129 CIR controls and 60 HCC patients. In the second experiment, there were 50 CIR controls and 13 HCC patients included. In the third experiment, there were 5 CIR controls and 5 HCC patients. Using this benchmark dataset, the function of data integration and batch effect removal was validated effectively.

As shown in **Figure 2A** and **Figure 2B**, boxplots were applied for the visualization of raw data for m/z (mass-to-charge ratio) and RT (retention time) values for the three datasets from different analytical experiments. There were certain differences in the m/z and RT values among the different datasets, especially for the third experiment. The boxplots of the intensities in each sample after data integration and batch effect removal are shown in **Figure 2C** and **Figure 2D**, respectively. The intensities for all samples ranged from 4 to 8 after data integration, while the range distribution was from -4 to 0 after removing batch effects. From these boxplots, the range distribution could be significantly reduced by batch effect removal. The principal component analysis (PCA) plots of samples for data integration and batch effect removal are shown in **Figure 2E** and **Figure 2F**, respectively. The PCA plots were applied to visualize the samples in different datasets before and after batch effect removal. From the PCA plots, after data integration, the samples in different datasets were separated from each other, while the samples in different datasets could be clustered together after removing batch effects.

Case 2: Sample Separation & Biomarker Identification for Metabolomic Data

For the benchmark dataset MTBLS17, a combined matrix including 78 HCC patients and 184 CIR controls was obtained by integrating three analytical datasets and removing batch effects. The separation of all samples from this combined dataset could be visualized using four methods. As shown in **Figure 3A**, hierarchical clustering for samples and metabolites was performed using HCA. There were obvious differences in the two clusters for the samples and metabolites. As shown in **Figure 3B**, clustering using k-means for samples in different groups was used to visualize sample separation. The plots of PCA and SOM for samples in different classes are shown in **Figure 3C** and **Figure 3D**, respectively. Applying all metabolites for this dataset, the samples could be separated

distinctly by KMC, while all samples were gathered by PCA. Therefore, the differential metabolites identified using the biomarker identification method could be used for sample separation to realize the better separation of all samples. Moreover, the result of sample separation using example data embedded in the mmeaseR package is shown in **Supplementary Figure S1**, including plots of HCA, KMC, PCA and SOM.

In addition to sample separation, 13 methods of biomarker identification were provided in the mmeaseR package. These methods were applied to discover differential metabolites for the combined dataset. As shown in **Table 1**, the top 20 metabolites with the highest VIP (variable importance in the projection) values using the PLS-DA method were included. Herein, the results of other biomarker identification methods are also provided. There were significant differences for different methods even for the same dataset. Therefore, it was very important to select the most appropriate method for a specific dataset, and was necessary to provide these biomarker identification methods for the convenient choice. The plots of biomarker identification using example data embedded in the mmeaseR package are shown in **Supplementary Figure S2** by six methods, including FC, PLS-DA, *t*-test, CHIS, CFS and LMEB.

Another benchmark dataset was applied to identify the metabolic markers for triple-negative breast cancer, including 330 samples with triple-negative breast cancer and 149 paired normal breast tissues in the study of Xiao *et al*. [38]. In this study, both fold change (FC) and Student's *t*-test methods were used to discover biomarkers from the raw data. The cutoffs (logFC >1 and *p* value <0.05) were set for fold change and Student's *t*-test, respectively. There were 302 differential metabolites at the intersection of the two methods, including 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 4A** and **Figure 4B**, respectively. Taking phenylalanyl-threonine as an example, the value in samples with triple-negative breast cancer was five times that of normal breast tissues. The value of uridine diphosphate glucuronic acid 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 4C**. The points in red indicate the metabolites with adjusted *p* values <0.05 between the two sample groups. These differential metabolites could be used for the downstream analysis in **Case 3**.

Case 3: Metabolite Annotation and Enrichment Analysis for the Specific Metabolite

In the mmeaseR package, the function of metabolite annotation is provided based on an enhanced metabolite database. To validate this function, the differential metabolites identified using both fold change and Student's *t*-test methods were applied. Using the mmeaseR package, the top 10

upregulated and downregulated metabolites were annotated. The results, including the name, logFC, adjusted *p* values and biological function of these metabolites, are shown in **Table 2**. 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 metabolomic profiles [39]. The most downregulated metabolite, uridine 5'-diphosphoglucuronic acid, was the endogenous substrate of UDP-glucuronosyltransferase and could be measured in liver, kidney and placenta [40]. The annotated functions for uridine 5'-diphosphoglucuronic acid consist 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 S3**. As shown in **Supplementary Table S4** and **Supplementary Figure S3**, the results of metabolite annotation for tandem mass spectrometry were created by the parent ion mass (181.04) and the MS/MS peak list (m/z & intensity) embedded in the example data.

Based on this enhanced metabolite database, eight categories were provided for enrichment analysis. Using 302 differential metabolites between triple-negative breast cancer and paired normal breast tissues by both fold change and Student's *t*-test, the enrichment analysis was validated based on five categories of biological function. The results of these five categories, including KEGG pathways, biological function classes, food components & food additives, species taxonomy and toxins & environmental pollutants, are shown in **Figure 4D**, **Figure 5A**, **Figure 5B**, **Figure 5C**, and **Figure 5D**, respectively. These pathways and biological functions might be involved in the development of the disease studied or might result from the disease. Moreover, as shown in **Supplementary Figure S4** and **Supplementary Table S5**, the results of enrichment analysis for KEGG pathways were created using the example data embedded in the mmeaseR package. As shown in **Supplementary Table S6** and **Supplementary Figure S5**, the results of enrichment analysis for the classes of food components and food additives were generated using the example data and code.

Conclusion

The integration of multiple analytical experiments with enhanced metabolite annotation and enrichment analysis is necessary in large-scale and long-term metabolomics. To facilitate flexibility and reproducible analysis, the mmeaseR R package was developed in this study. This package can integrate multiple analytical experiments to enlarge the sample size, provide enhanced annotation for metabolites, and conduct enrichment analysis based on a metabolite database. The package is freely available from https://github.com/mmeaseR/mmeaseR.

Supplementary Data

The details of the available functions and the usage of example code for mmeaseR package could be found in the **Supplementary Data**.

Competing Interests

The authors declare that they have no competing interests.

Funding

This work was funded by the National Natural Science Foundation of Jiangsu (BK20210597), and the NUPTSF (Grant No. NY220169).

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. 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.
6. da Silva RR, Dorrestein PC, Quinn RA. Illuminating the dark matter in metabolomics. *Proc Natl Acad Sci U S A* 2015;**112**:12549-50.
7. 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.
8. Chong J, Xia J. MetaboAnalystR: an R package for flexible and reproducible analysis of metabolomics data. *Bioinformatics* 2018;**34**:4313-4.
9. Zhang YW, Tamba CL, Wen YJ *et al*. mrMLM v4.0.2: An R Platform for Multi-locus Genome-wide Association Studies. *Genomics Proteomics Bioinformatics* 2020;**18**:481-7.
10. 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.
11. 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.

12. De Livera AM, Dias DA, De Souza D *et al*. Normalizing and integrating metabolomics data. *Anal. Chem* 2012;**84**:10768-76.

13. 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.

14. 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.

15. 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.

16. 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**:436-48.

17. 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.

18. Ren S, Hinzman AA, Kang EL *et al*. Computational and statistical analysis of metabolomics data. *Metabolomics* 2015;**11**:1492-513.

19. 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.

20. Want EJ, Wilson ID, Gika H *et al*. Global metabolic profiling procedures for urine using UPLC-MS. *Nat Protoc* 2010;**5**:1005-18.

21. 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.

22. Li F, Zhou Y, Zhang Y *et al*. POSREG: proteomic signature discovered by simultaneously optimizing its reproducibility and generalizability. *Brief Bioinform* 2022;**23**:bbac040.

23. 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.

24. 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.

25. Bylesjo M, Rantalainen M, Cloarec O *et al*. OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. *Journal of Chemometrics* 2006;**20**:341-51.

26. Begun A. Power estimation of the t test for detecting differential gene expression. *Funct Integr Genomics* 2008;**8**:109-13.

27. 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.

28. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics* 2007;**23**:2507-17.

29. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 2004;**3**:Article3.

30. 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.

31. 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.

32. 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.

33. 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 Analyt Technol Biomed Life Sci* 2012;**910**:149-55.

34. 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.

35. Zhang S, Amahong K, Sun X *et al*. The miRNA: a small but powerful RNA for COVID-19. *Brief Bioinform* 2021;**22**:1137-49.

36. 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.

37. 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**:440-4.

38. Xiao Y, Ma D, Yang YS *et al*. Comprehensive metabolomics expands precision medicine for triple-negative breast cancer. *Cell Res* 2022;**32**:477-90.

39. Zhao Q, Shen H, Su KJ *et al*. A joint analysis of metabolomic profiles associated with muscle mass and strength in Caucasian women. *Aging (Albany NY)* 2018;**10**:2624-35.

40. 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.

Table 1. Thirteen methods of biomarker identification were applied to discover metabolic markers. The top 20 metabolites with the highest VIP (variable importance in the projection) values using the PLS-DA method were included. Herein, the results of other biomarker identification methods are also provided. PLS-DA (partial least squares discrimination analysis), FC (fold change), OPLS-DA (orthogonal PLS-DA), CFS (correlation-based feature selection), Cor (correlation coefficient), RELIEF (recursive elimination of features), RF-RFE (random forest-recursive feature elimination), ACC (mean decrease accurary), WRS (Wilcoxon rank sum), SAM (significance analysis for microarrays), LMEB (linear models and empirical Bayes), CHIS (Chi-squared test), Import (Importance), ENTROPY (entropy-based filter method), InfGain (InfGain Order), SVM-RFE (support vector machine-recursive feature elimination).

MZ/RT	PLS-DA VIP	FC FC	OPLS-DA VIP	*t* test *p* value	CFS Cor	RELIEF VIP	RF-RFE ACC	WRS *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

Table 2. The functions annotated for the top 10 upregulated and downregulated metabolites between triple-negative breast cancer patients and the paired normal breast tissues.

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 (AICA)	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 (GDP)	-8.68	4.49E-149	Endogenous; Food; Microbial Metabolite; Plant
13	Uridine 5'-diphosphate (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 (D-Ala-D-Ala)	-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 (ADP)	-10.80	5.22E-121	Endogenous; Food; Microbial Metabolite; Plant; TCM Ingredient
20	Uridine 5'-diphosphoglucuronic acid (UDP-D-glucuronate)	-12.26	7.80E-130	Drug; Endogenous; Food; Microbial Metabolite; Plant

Figure 1. The analytical workflow of the mmeaseR package. There are five major functions: (A) Data Integration, (B) Sample Separation, (C) Marker Identification, (D) Metabolite Annotation, and (E) Enrichment Analysis.

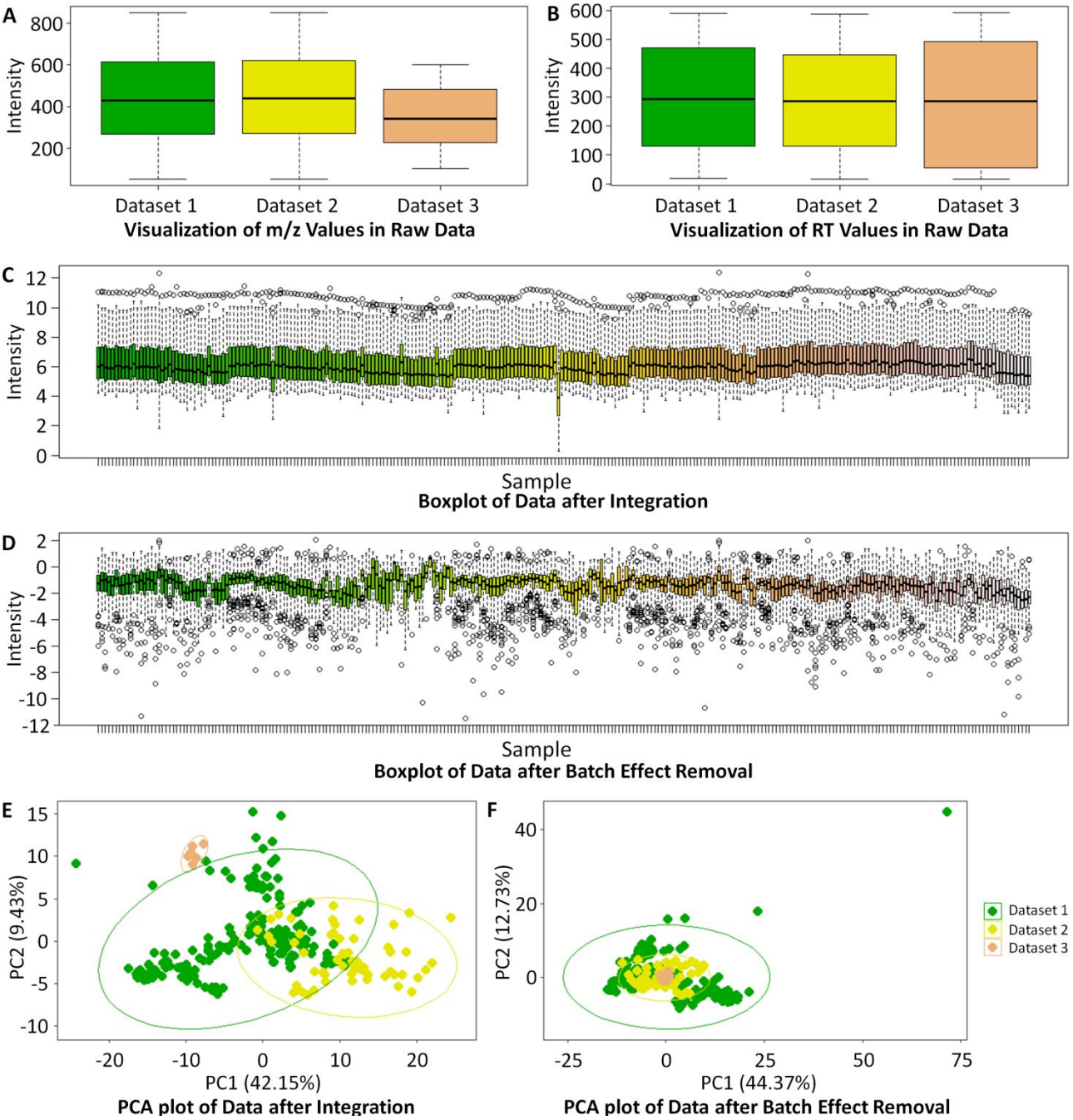


Figure 2. The data integration for three analytical experiments in MTBLS17. The boxplots were applied for the visualization of raw data for (A) m/z (mass-to-charge ratio) and (B) RT (retention time) values for three analytical datasets. Boxplots of intensities in each sample were shown for data integration (C) and batch effect removal (D). The PCA plots were applied to visualize the distribution of samples in different analytical datasets before (E) and after (F) batch effect removal.

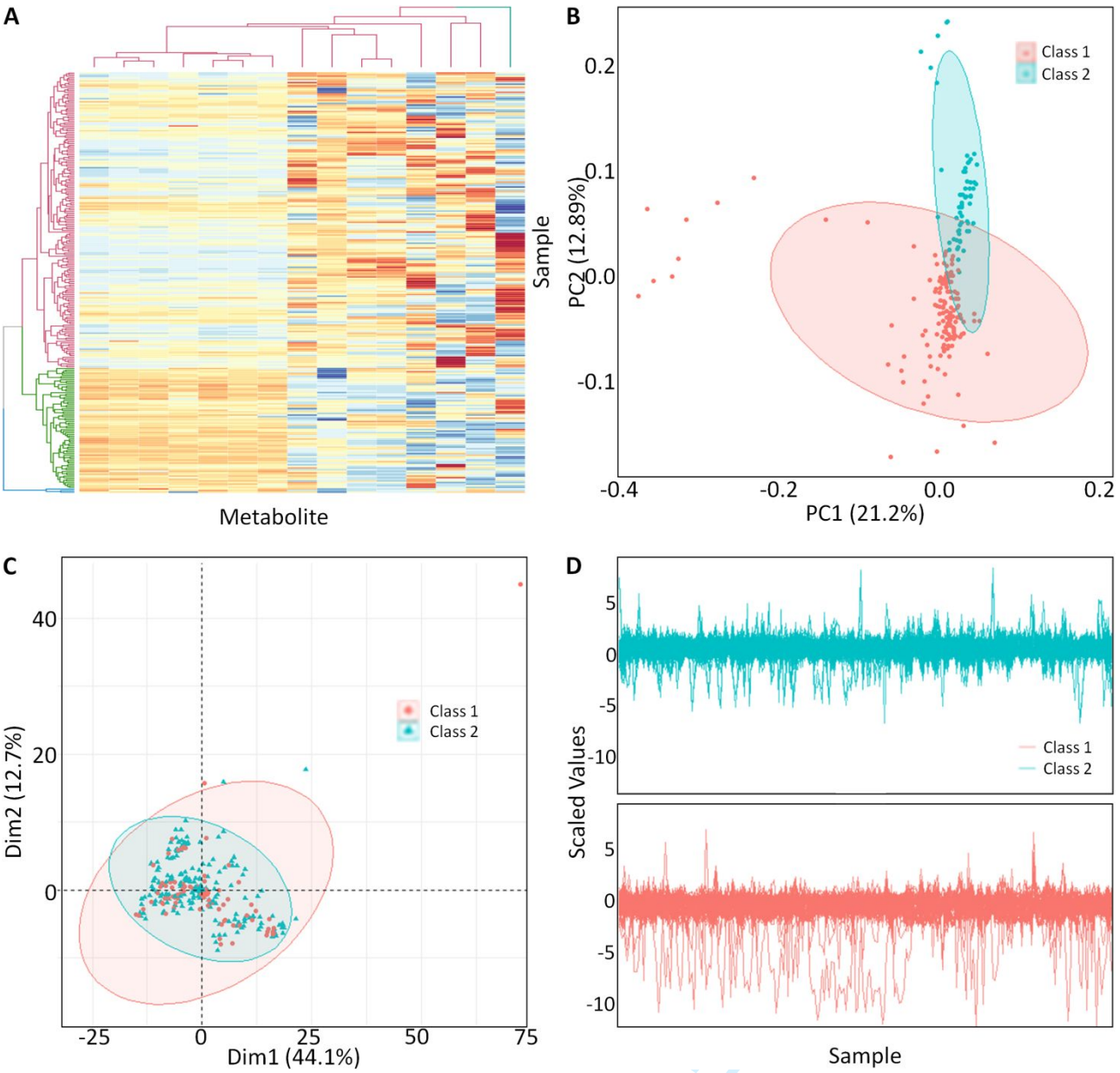


Figure 3. Four methods were applied for sample separation for the dataset after integration and batch effect removal, including (A) hierarchical clustering analysis (HCA), (B) k-means clustering (KMC), (C) principal component analysis (PCA) and (D) self-organizing map (SOM).

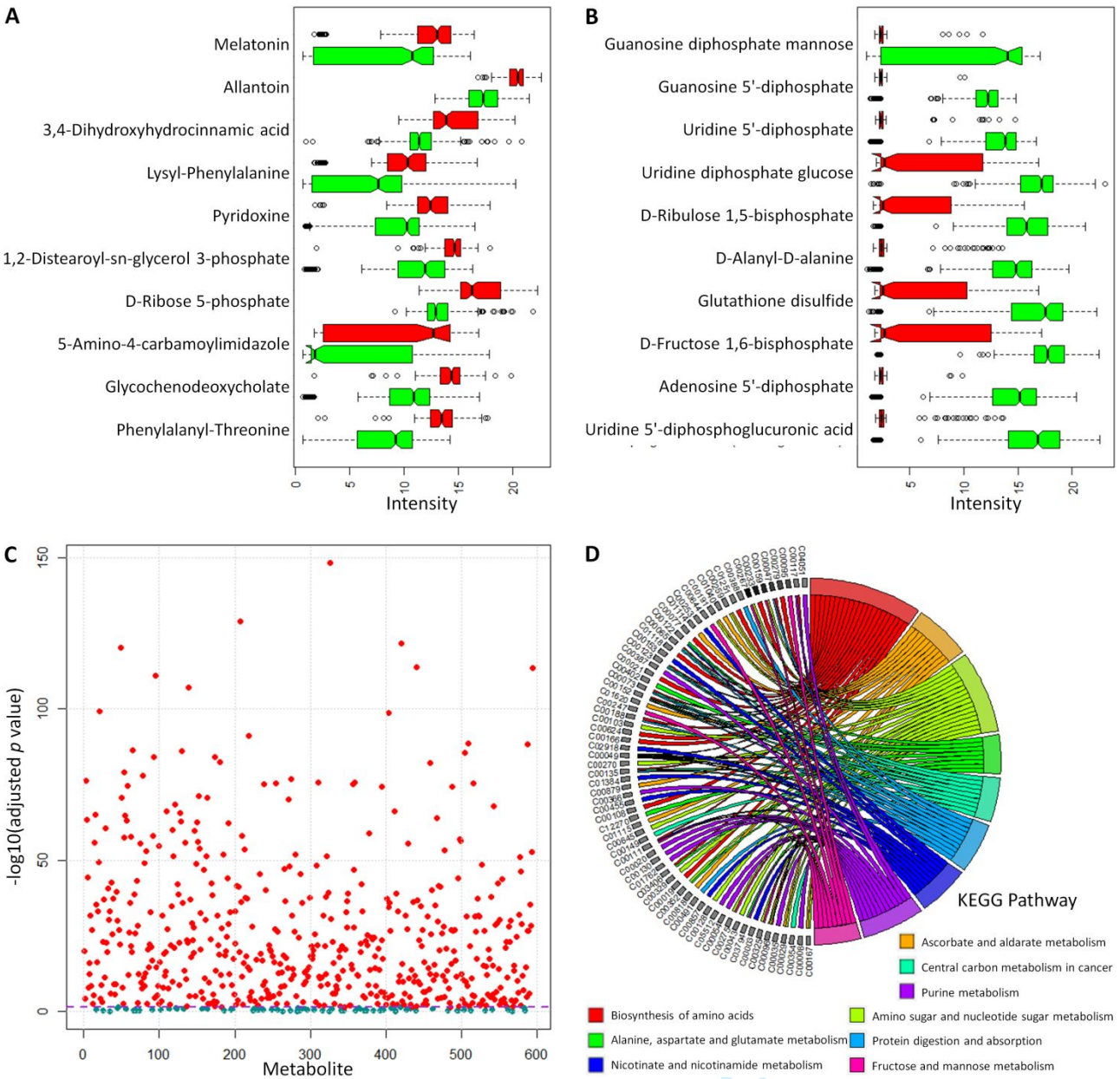


Figure 4. Based on the metabolomic data between triple-negative breast cancer patients and the paired normal breast tissues, (A) boxplots of the top 10 most upregulated, and (B) top 10 most downregulated biomarkers using the fold change method are shown. (C) The plot of logarithmic transformation of the adjusted *p* values for each metabolite using Student's *t*-test. The points in red indicate the metabolites with adjusted p values <0.05 between the two groups. (D) Chord diagrams for the KEGG pathways enriched using the differential metabolites based on the metabolite database.

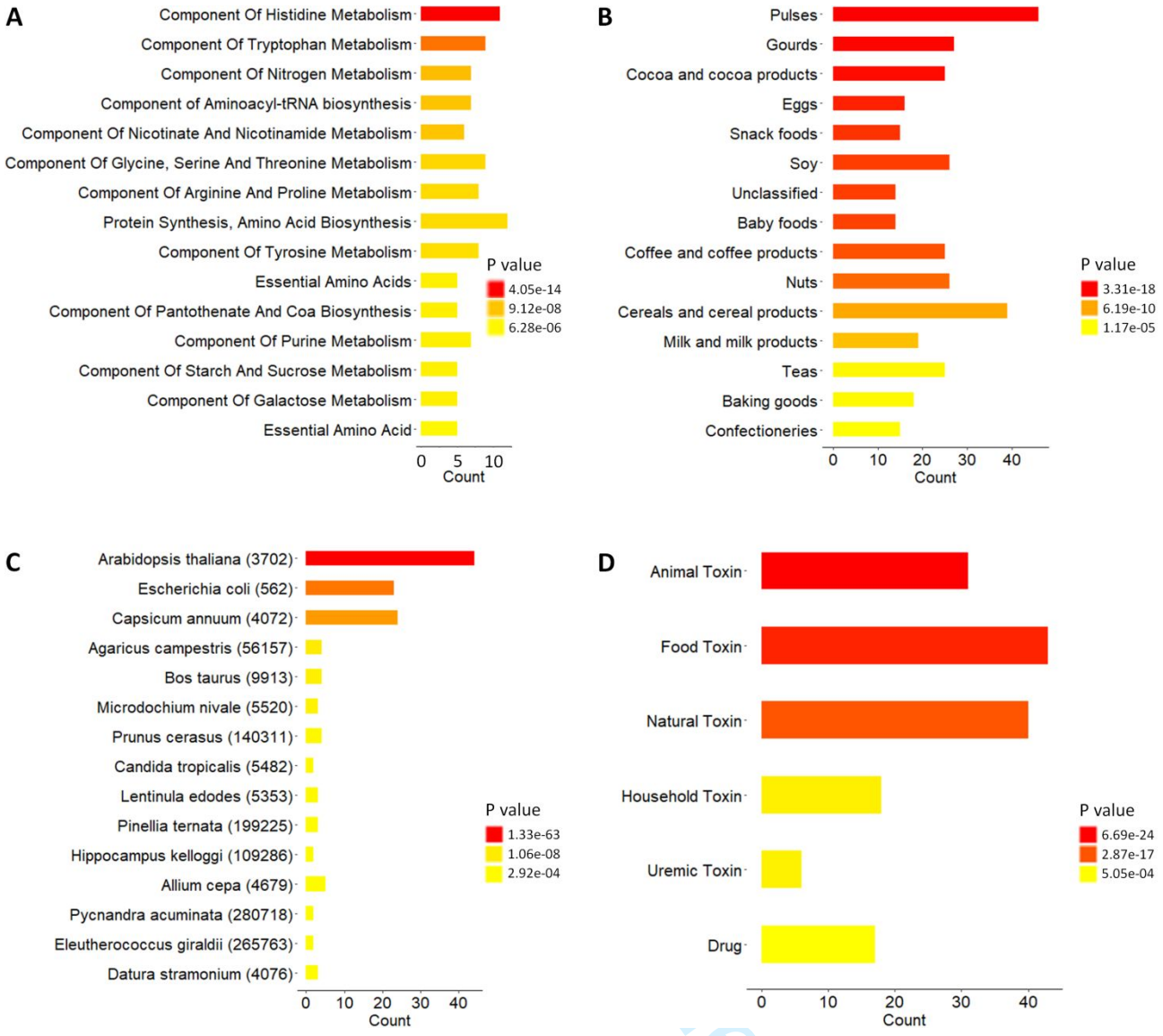


Figure 5. Enrichment analysis was performed using these differential metabolites by both fold change and Student's *t*-test. The biological functions include (A) biological function classes, (B) food components & food additives, (C) species taxonomy and (D) toxins & environmental pollutants.

Supplementary Information

mmeaseR: an R package for Data Integration, Marker Identification, Metabolite Annotation and Enrichment in Large-scale and Long-term Metabolomics

Qingxia Yang[1, *], Bo Li[2], Panpan Wang[3], Jicheng Xie[1], Yuhao Feng[1], Ziqiang Liu[1], Feng Zhu[4, *]

[1] Department of Bioinformatics, Smart Health Big Data Analysis and Location Services Engineering Lab of Jiangsu Province, School of Geographic and Biologic Information, Nanjing University of Posts and Telecommunications, Nanjing, 210023, China.

[2] College of Life Sciences, Chongqing Normal University, Chongqing, Chongqing 401331, China.

[3] College of Chemistry and Pharmaceutical Engineering, Huanghuai University, Zhumadian 463000, China.

[4] College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, Zhejiang 310058, China.

* To whom the correspondence should be addressed: Prof. Feng Zhu, E-mail: zhufeng@zju.edu.cn, and Dr. Qingxia Yang, E-mail: yangqx@njupt.edu.cn.

Running Title: mmeaseR package for large-scale metabolomics

Install mmeaseR Package

The mmeaseR package is provided through GitHub. In order to install it, devtools package available in CRAN (https://cran.r-project.org/) is required. To install devtools, the user must type the following commands in an R session:

> install.packages("devtools")

> library(devtools)

Once devtools package has been installed, the user can install mmeaseR package by typing the following commands in an R session:

> install_github("mmeaseR/mmeaseR", force = TRUE)

> library(mmeaseR)

1. Tutorial for the Data Integration Step

1.1 Data Integration for the three analytical experiments

For data integration, multiple datasets from different analytical experiments can be used as the input of the mmeaseR package. Before data integration, the *csv* files containing a feature-by-sample matrix should be prepared in advance. Each dataset (*csv* file) contains five essential columns providing the information of mass, retention time, intensity, isotope and adduct. The first two columns provide the mass and retention time, and samples must be kept in columns with the sample names in the first row. The group label in the second row indicates distinct sample groups, such as case and control. Input data values (mass, retention time, intensity) should be numeric, and a blank cell or "NA" should be adopted to indicate any missing values.

An example input file with the corresponding contents separated by comma (mutile_Group) is provided in mmeaseR. The example code for data integration from the package is shown below. The binary matrix after data integration for the example datasets is shown in **Supplementary Table S1**. In this table, each metabolite was named according to the order of integration of all metabolites, such as Align_1 and Align_2. The sample names from different batches, m/z (mass-to-charge ratio) and rt (retention time) values are shown in the column.

Generate the binary matrix for data integration

> Integrate_Data(mutile_Group, RT_Tolerance_1 = 10, mz_Tolerance_1 = 0.1, RT_Tolerance_2 = 10, mz_Tolerance_2 = 0.1)

Supplementary Table S1. The binary matrix after data integration

Name	Batch-A-01	…	Batch-B-01	…	Batch-C-01	…	Batch-C-m/z	Batch-C-rt
Align_1	0.00	…	67508.99	…	29095.84	…	105.03	167.92
Align_2	0.00	…	99253.94	…	174527.97	…	114.07	30.44
Align_3	277.10	…	466.13	…	4544.05	…	451.21	212.90
Align_4	0.00	…	44936.18	…	13801.50	…	167.06	91.05
Align_5	0.00	…	70034.85	…	17360.19	…	181.07	142.19
Align_6	10035.51	…	19820.23	…	41515.82	…	314.23	329.48
Align_7	0.00	…	24467.62	…	35807.34	…	188.07	143.72
Align_8	0.00	…	48555.35	…	17464.66	…	202.05	167.96
Align_9	90972.12	…	79414.68	…	116558.55	…	437.20	559.70
Align_10	0.00	…	3390.46	…	7250.90	…	220.12	121.91

1.2 Batch Effects Removal after Data Integration

After data integration, it was essential to remove the unwanted variations among different batches. Various methods are provided in mmeaseR for removing batch effects in different analytical experiments, including batch mean-centering (BMC/PAMR), the empirical Bayes method (ComBat/EB), and global normalization (Z score). An example input file for batch effect removal is provided in mmeaseR. **Supplementary Table S2** shows the binary matrix for batch effect removal from the package created by the example code below.

```
# Generate the binary matrix for batch effects removal
> Removal_Batch(mutile_align, n = 3, algorithm = "BMC/PAMR")
```

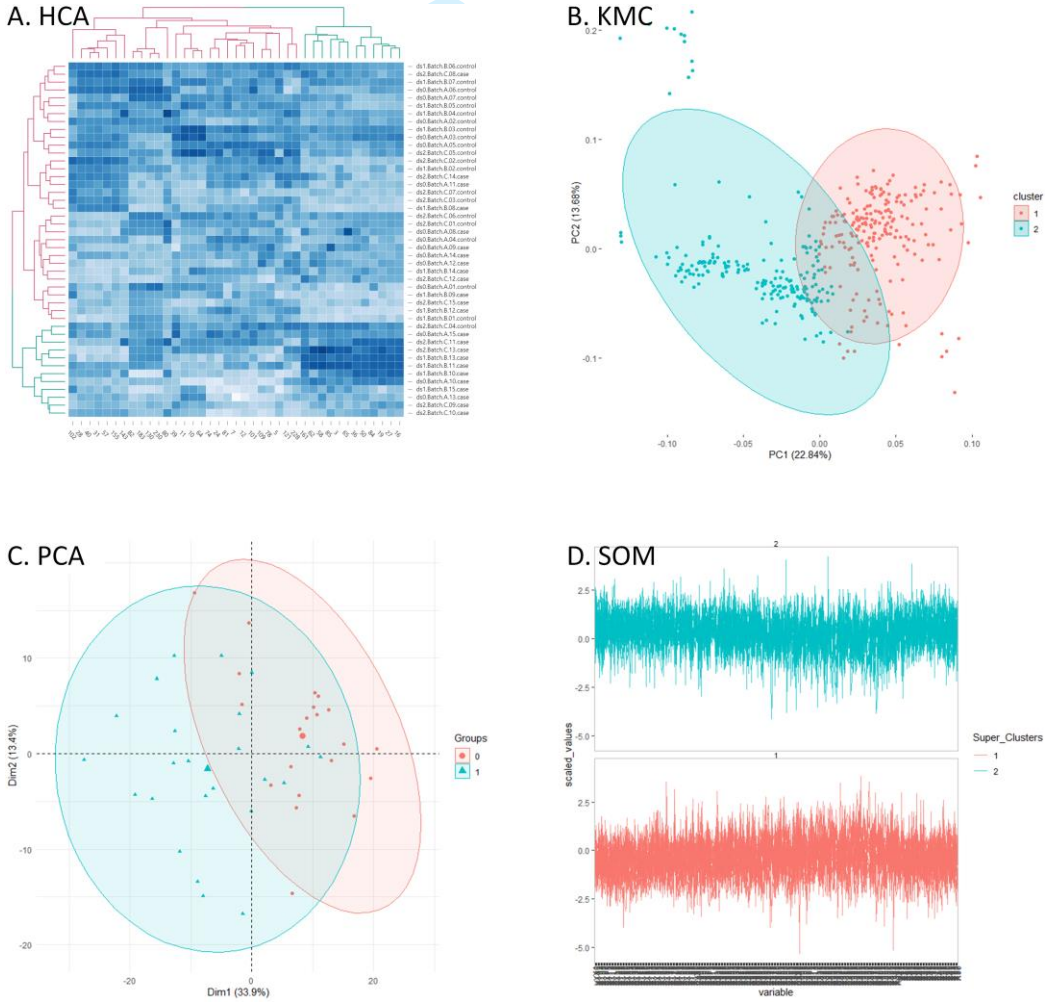
Supplementary Table S2. The binary matrix after batch effects removal

Name	Batch-A-01	Batch-A-02	…	Batch-B-01	Batch-B-02	…	Batch-C-01	…
Align_1	0	0	…	-0.40	0.78	…	-1.30	…
Align_2	0	0	…	-0.64	-0.09	…	0.61	…
Align_3	-3.92	-0.74	…	-4.18	0.78	…	-0.45	…
Align_4	0	0	…	0.67	0.79	…	-1.19	…
Align_5	0	0	…	1.15	1.37	…	-0.45	…

Align_6	-0.79	-1.11	…	-0.40	0.30	…	1.00	…
Align_7	0	0	…	-0.57	1.62	…	0.51	…
Align_8	0	0	…	0.13	0.57	…	-1.01	…
Align_9	0.06	-0.33	…	-0.22	-0.08	…	0.29	…
Align_10	0	0	…	-1.10	-0.87	…	0.45	…

2. Tutorial for the Sample Separation Step

There are four sample separation methods for visualizing the clustering and separation of different samples. In the mmeaseR package, the four methods are provided for sample separation after data integration and batch effect removal, including hierarchical clustering analysis (HCA), k-means clustering (KMC), self-organizing map (SOM) and principal component analysis (PCA). An example input file for sample separation is provided in mmeaseR. **Supplementary Figure S1** shows the plots of batch effects removal from the package created by the example code below.

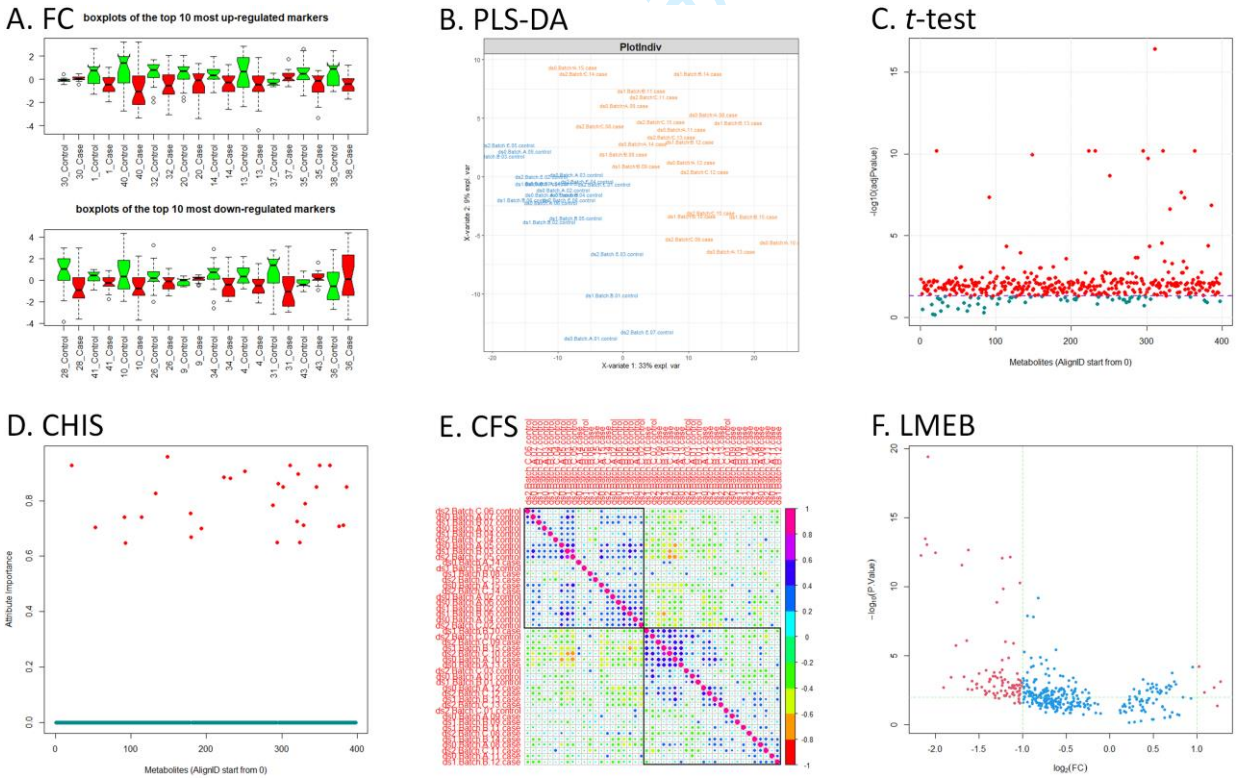


Supplementary Figure S1. The results for sample separation using four methods

```
# The binary matrix for sample separation

> finalData <- MarkerData$finalData

# The label for sample separation

> finalLabel <- MarkerData$finalLabel

# Create the plots

> Sample_Separation(finalData, finalLabel, clusters = 2, method = "HCA")
```

3. Tutorial for the Marker Identification Step

In the marker identification step, there are 13 popular strategies to identify metabolic markers for the given datasets. These strategies include fold change (FC), partial least squares discrimination analysis (PLS-DA), orthogonal PLS-DA (OPLS-DA), Student's *t*-test, Chi-squared test, correlation-based feature selection (CFS), entropy-based filter method, linear models and empirical Bayes method, recursive elimination of features (Relief), random forest-recursive feature elimination (RF-RFE), significance analysis for microarrays (SAM), support vector machine-recursive feature elimination (SVM-RFE), and Wilcoxon rank sum (WRS). An example input file for marker identification is provided in mmeaseR. The plots of marker identification from the package created by the example code below are shown in **Supplementary Figure S2**.



Supplementary Figure S2. The results for marker identification using the different methods

```
# The binary matrix for marker identification

> finalData <- MarkerData$finalData

# The label for marker identification

> finalLabel <- MarkerData$finalLabel

# Create the plots

> Marker_Identify(finalData, finalLabel, method = "FC")
```

4. Tutorial for the Metabolite Annotation Step

4.1 Metabolite Annotation for Primary Mass Spectrometry

In the mmeaseR package, a database for enhanced annotation is provided for metabolite annotation. When performing metabolite annotation for primary mass spectrometry, a compound list containing the studied m/z features should be properly provided. An example input for metabolite annotation for primary mass spectrometry is provided in the mmeaseR package. **Supplementary Table S3** shows the result of metabolite annotation from the package created by the example code below.

```
# The compound list (m/z) of primary mass spectrometry to be annotated

> AnnotaMS <- AnnotaData$AnnotaMS

# Metabolite annotation for primary mass spectrometry

> MetaboA_res <- MetaboAnnotation(AnnotaMS)

# The result of metabolite annotation for primary mass spectrometry

> MetaboA_res$`M+H-2H2O`
```

Supplementary Table S3. Metabolite annotation results (query m/z: 96.95964) using primary mass spectrometry

MMEASE ID	Mass	MID	Common Name	Annotation
MMEASE0000018	132.05	294	3-Ureidopropionic acid	Endogenous; Food; Microbial Metabolites
MMEASE0000024	135.05	85	Adenine	Cosmetic; Drug; Endogenous; Food; Microbial Metabolites; Plant; TCM Ingredients; Toxins/Pollutant

MMEASE0000046	131.07	7	Creatine	Cosmetic; Drug; Endogenous; Food; TCM Ingredients; Toxins/Pollutant
MMEASE0000049	129.08	50	DL-pipecolic acid	Endogenous; Food; Microbial Metabolites; Plant; TCM Ingredients
MMEASE0000056	128.06	291	Hydrouracil (Dihydrothymine)	Endogenous; Food; Microbial Metabolites; Toxins/Pollutant

4.2 Metabolite Annotation for Tandem Mass Spectrometry

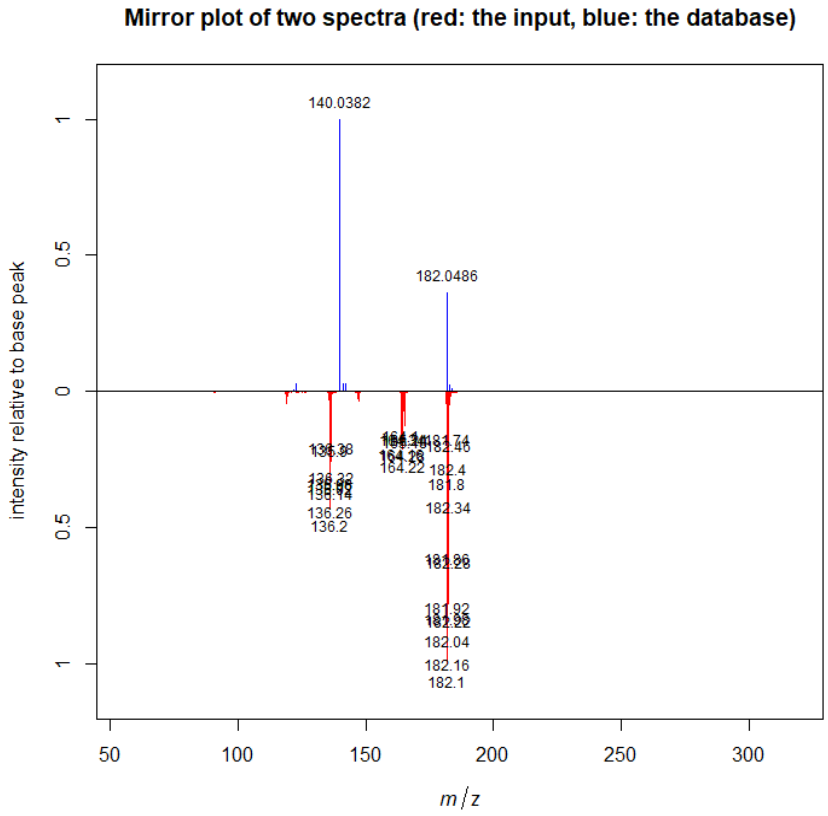
In the mmeaseR package, when performing metabolite annotation for tandem mass spectrometry, the information containing parent ion mass and MS/MS peak list (the first column is m/z and the second column is intensity) should be properly provided. An example input for metabolite annotation for tandem mass spectrometry is provided in the mmeaseR package. These example data embedded in the mmeaseR package include the parent ion mass (181.04) and MS/MS peak list (m/z & intensity). In this MS/MS peak list, the intensities is 0.588541, 2.974737, 100.000000, 2.710494, 2.722505, 36.241342, 2.382192 and 1.165072 for the peaks 122.0278, 123.0119, 140.0382, 141.0409, 142.0337, 182.0486, 183.0518 and 184.0452, respectively. As shown in **Supplementary Table S4** and **Supplementary Figure S3**, the results of metabolite annotation for tandem mass spectrometry were created by the example code below.

Supplementary Table S4. The table of metabolite annotation for tandem mass spectrometry

MMEASE ID	Mass	Name	Annotation	Fit
MMEASE0009095	181.04	Acamprosate	Drug; Toxins/Pollutant	1
MMEASE0171400	181.07	Tyrosine	Cosmetic; Drug; TCM Ingredients	0.03
MMEASE0174293	181.01	Thiocyclam	Agricultural Chemicals	0
MMEASE0173964	181.00	2-(Methylthio)-benzothiazole	Food; Microbial Metabolites	0
MMEASE0171400	181.07	Tyrosine	Cosmetic; Drug; TCM Ingredients	0

```
# The parent ion mass of metabolite

> Parent_mass <- AnnotaData$Parent_mass

# The MS/MS peak list (the first column is m/z and the second column is intensity)

> TandomData <- AnnotaData$TandomData

# The parameters for metabolite annotation of tandem mass spectrometry
```

```
> AnnotaParamTandom < - AnnotaTandom(Parent_mass, TandomData)

# The table of metabolite annotation of tandem mass spectrometry

> annotaDataTandom(AnnotaParamTandom)

# The plot of metabolite annotation of tandem mass spectrum

> annotaTandom_plot(AnnotaParamTandom, TandomData)
```



Supplementary Figure S3. The plot of metabolite annotation for tandem mass spectrometry

5. Tutorial for the Enrichment Analysis Step

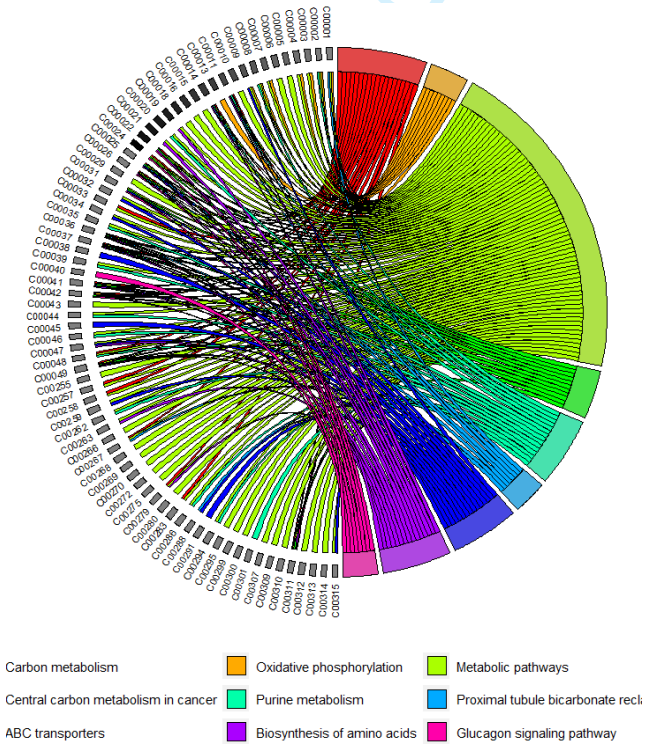
There are eight categories from the metabolite database used for enrichment analysis in the mmeaseR package, including (1) KEGG pathway, (2) SMPDB pathway, (3) Chemical family, (4) Classes of food components and food additives, (5) Biological function classes, (6) Therapeutic classes of secondary metabolites of traditional medicine, (7) Species taxonomy, and (8) Categories of toxins and environmental pollutants.

5.1 Enrichment Analysis for KEGG Pathways

In the mmeaseR package, when performing enrichment analysis for KEGG pathways, a compound

list should be properly provided. An example input for enrichment analysis for the KEGG pathways is provided in mmeaseR. **Supplementary Figure S4** and **Supplementary Table S5** show the results of enrichment analysis for the KEGG pathway from the package created by the example code below.

```
# The compound list for metabolite enrichment

> sampleDatakegg <- EnrichData$sampleDatakegg

# Enrichment analysis for the KEGG pathways

> EnrichParam <- KEGGEnrichPlotPanel(sampleDatakegg, enrichDB = "kegg", pvalcutoff = 0.05, IDtype = 1, cateIdx = 1)

# The table for metabolite enrichment

> EnrichResultList <- Enrichment(EnrichParam)

# The parameters in the plot of metabolite enrichment using KEGG database

> EnrichFC <- seq(from = -2,to = 2, length.out = 24)

# The plot of metabolite enrichment using KEGG database

> KEGGEnrichPlot(EnrichResultList = EnrichResultList, cpdID = sampleDatakegg, cpdFC = EnrichFC)
```



Supplementary Figure S4. The plot of enrichment analysis for the KEGG pathways

Supplementary Table S5. The table of enrichment analysis for the KEGG pathways

Pathway	Name	Count	Ratio	Adj *p*-value	Compound ID
hsa00630	Glyoxylate and dicarboxylate metabolism	16	16/88	2.75E-12	C00007; C00011; C00014; C00022; C00024; C00025; C00026; C00027; C00036; C00037; C00042; C00048; C00258; C00266; C00311; C00313
hsa01200	Carbon metabolism	19	19/88	5.62E-11	C00011; C00014; C00022; C00024; C00025; C00026; C00033; C00036; C00037; C00041; C00042; C00048; C00049; C00257; C00258; C00267; C00279; C00283; C00311
hsa00190	Oxidative phosphorylation	9	9/88	6.77E-10	C00001; C00002; C00003; C00004; C00007; C00008; C00009; C00013; C00042

5.2 Enrichment Analysis for other Databases

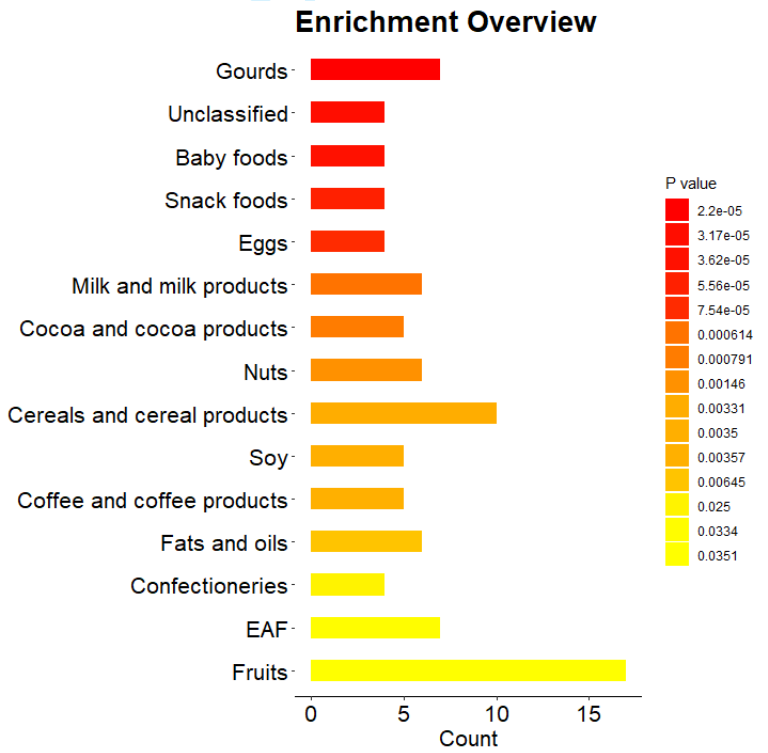
In the mmeaseR package, when performing enrichment analysis for databases other than KEGG pathways, a compound list should also be properly provided. An example input for enrichment analysis for the classes of food components and food additives is provided in the mmeaseR package. As shown in **Supplementary Table S6** and **Supplementary Figure S5**, the results of enrichment analysis for the classes of food components and food additives from the package are creted by the example code below.

```
# The compound list for metabolite enrichment

> sampleDatacas <- EnrichData$sampleDatacas

# The database name for metabolite enrichment

> enrichDB <- EnrichData$enrichDB

# The parameters in the plot of metabolite enrichment using other Databases

> EnrichParam <- KEGGEnrichPlotPanel(sampleDatacas, enrichDB = enrichDB, pvalcutoff = 0.05, IDtype = 2, cateIdx = 1)

# The table of enrichment analysis for other databases

> EnrichResultList <- Enrichment(EnrichParam)

# The database name for metabolite enrichment
```

```
> dbChoice < - enrichDB

# plot of metabolite enrichment for other databases

> EnrichPlot(dbChoice, EnrichResultList)
```

Supplementary Table S6. The table of enrichment analysis for other pathways

Name	Count	Ratio	Adj *p*-value	HMDB ID
Gourds	7	7/31	5.49E-04	MSFA04594; MSFA06356; MSFA12135; MSFA12497; MSFA12578; MSFA12582; MSFA14723
Unclassified	4	4/31	7.94E-04	MSFA12135; MSFA12497; MSFA12578; MSFA14723
Baby foods	4	4/31	9.06E-04	MSFA12135; MSFA12497; MSFA12578; MSFA14723
Snack foods	4	4/31	1.39E-03	MSFA12135; MSFA12497; MSFA12578; MSFA14723
Eggs	4	4/31	1.88E-03	MSFA12135; MSFA12497; MSFA12578; MSFA14723



Supplementary Figure S5. The plot of enrichment analysis for other pathways