M^2IA: a Web Server for Microbiome and Metabolome Integrative Analysis

Yan Ni[1,#], Gang Yu[1], Huan Chen[2], Yongqiong Deng[3], Philippa M Wells[4], Claire J Steves[4,5], Feng Ju[6,7], Junfen Fu[1,#]

1. National Clinical Research Center for Child Health, The Children's Hospital, Zhejiang University School of Medicine, Hangzhou 310029, P.R. China.

2. Key laboratory of Microbial technology and Bioinformatics of Zhejiang Province, NMPA Key laboratory for Testing and Risk Warning of Pharmaceutical Microbiology, Zhejiang Institute of Microbiology, Hangzhou 310012, P.R. China.

3. Department of Dermatology and STD, the Affiliated Hospital of Southwest Medical University, Luzhou, Sichuan, P.R. China.

4. The Department of Twin Research, Kings College London, South Wing, St Thomas' Hospital, Westminster Bridge Road, London SE1 7EH, UK

5. Department of Ageing and Health, St Thomas' Hospital, 9th floor, North Wing, Westminster Bridge Road, London SE1 7EH, UK

6. School of Engineering, Westlake University, 18 Shilongshan Road, Hangzhou 310024, P.R. China.

7. Institute of Advanced Technology, Westlake Institute for Advanced Study, 18 Shilongshan Road, Hangzhou 310024, P.R. China.

Correspondence should be addressed to Dr. Yan Ni (yanni617@zju.edu.cn) and Dr. Junfen Fu (fjf68@zju.edu.cn).

Abstract

Background: In the last decade, integrative studies of microbiome and metabolome have experienced exponential growth in understanding their impact on human health and diseases. However, analyzing the resulting multi-omics data and their correlations remains a significant challenge in current studies due to the lack of a comprehensive computational tool to facilitate data integration and interpretation. In this study, we have developed a microbiome and metabolome integrative analysis pipeline (M^2IA) to meet the urgent needs for tools that effectively integrate microbiome and metabolome data to derive biological insights.

Results: M^2IA streamlines the integrative data analysis between metabolome and microbiome, including both univariate analysis and multivariate modeling methods. KEGG-based functional analysis of microbiome and metabolome are provided to explore their biological interactions within specific metabolic pathways. The functionality of M^2IA was demonstrated using TwinsUK cohort data with 1116 fecal metabolites and 16s rRNA microbiome from 786 individuals. As a result, we identified two important pathways, i.e., the benzoate degradation and phosphotransferase system, were closely associated with obesity from both metabolome and microbiome.

Conclusions: M^2IA is a public available web server (http://m2ia.met-bioinformatics.cn) with the capability to investigate the complex interplay between microbiome and metabolome, and help users to develop mechanistic hypothesis for nutritional and personalized therapies of diseases.

Keywords: M^2IA, microbiome and metabolome correlation, data integration, network analysis, metabolic function

Background

The study of microbiome in human health has experienced exponential growth over the last decade with the advent of new sequencing technologies for interrogating complex microbial communities[1]. Meanwhile, metabolomics has been an important tool for understanding microbial community functions and their links to health and diseases through the quantitation of dozens to hundreds of small molecules[2]. The gut microbiota is considered a metabolic 'organ' to protect the host against pathogenic microbes, modulate immunity, and regulate metabolic processes, including short chain fatty acid production and bile acid biotransformation[3]. Conversely, these metabolites can modulate gut microbial compositions and functions both directly and indirectly[4]. Thus, the microbiota-metabolites interactions are important to maintain the host health and well-being. Integrative data analysis of gut microbiome and metabolome can offer deep insights on the impact of lifestyle and dietary factors on chronic and acute diseases (e.g., autoimmune diseases, inflammatory bowel disease[5], cancers[6], type 2 diabetes and obesity[7], cardiovascular disorders, and non-alcoholic fatty liver disease[8]), and provide potential diagnostic and therapeutic targets[9].

In the last decade, metabolomics studies in microbiota-related research have increased in a wide range of research areas, such as gastroenterology, biochemistry, endocrinology, microbiology, genetics, to nutrition, food science and pharmacology (Figure S1). However, both metagenomics/16s rRNA-based high-throughput sequencing technologies and mass spectrometry/nuclear magnetic resonance-based metabolomics platforms can produce large and high-dimensional data, posing a major challenge for subsequent data integration[10, 11]. Current integration analysis methods mainly focus on statistical correlations between microbiome and metabolome, such as the spearman correlations and partial least squares discriminant analysis (Table S1). Pedersen et al. recently summarized a step-by-step computational protocol from their previous study that applied WGCNA, a dimension reduction method, to measure the correlations among the host phenotype, gut metagenome, and fasting serum metabolome. However, separate analyses of -omics data through one or two statistical methods only provide fragmented information and do not capture the holistic view of disease mechanisms. Moreover,

although much bioinformatics work has been done to process and analyze the individual omics data, to date, it still lacks a comprehensive strategy or a computational tool to analyze the correlations between microbiome and metabolome[12]. To rapidly advance microbiome and metabolome data integration and understand their roles in diverse diseases, advanced computational methods for multi -omics data integration and interpretation need to be developed[2].

In this study, we have developed a comprehensive computational tool, a standardized workflow for microbiome and metabolome integrative analysis (M^2IA). M^2IA combines a variety of univariate and multivariate methods for correlation analysis and data integration. In addition, functional network analysis implemented with KEGG metabolic pathway database can provide deep insights of their biological correlations, i.e., the possible participation of identified microbes in a specific metabolic reaction or metabolic pathway. M^2IA accepts the microbiome data from 16S rRNA gene or shotgun metagenomic sequencing technologies, and metabolomics data from mass spectrometry or NMR spectroscopy platforms. Other than current bioinformatics tools, M^2IA is an automated data analysis and reporting pipeline that can be applied by users with little training in bioinformatics. M^2IA is a web-based server with user-friendly interface, and public available to academic users through http://m2ia.met-bioinformatics.cn.

Implementation

M^2IA is developed in Java web system, with html, Javascript technology implemented for interactive interface and JFinal and Mysql databases for data management in the backend. All the statistical analyses and visualization were implemented in R programming language. The entire system is deployed on a cloud server with 16 GB of RAM and four virtual CPUs with 2.6 GHz each. Users can register an academic account to manage their own research projects. A demo account has been created for journal reviewers: account is demo and password is demo123 (Please note this sentence will be removed after manuscript review). M^2IA has been tested with major modern browsers such as Google Chrome, Safari, Mozilla Firefox and Microsoft Internet.

Result

Workflow

A standard data analysis workflow in M^2IA contains six major steps: (1) data upload and processing, (2) global similarity analysis between microbiome and metabolome data matrix, (3) pairwise metabolite-microbe correlation analysis, (4) multivariate regression-based data integration analysis, (5) functional network analysis, and (6) an auto-generated html report for overview. Figure 1 summarizes the overall design and the flowchart of M^2IA. Each step offers multiple methods helping users to explore complex correlations between microbiome and metabolome. In this section, each step will be described in detail.

Data upload and processing

Data upload. The first step is to upload three different types of data: (1) microbiome datasets: for 16s rRNA gene sequencing data, an OTU abundance table with taxonomic annotations (.txt format) and a corresponding reference sequence file (.fna format) are required. Similarly, a unigene table with taxonomic annotations and a table with KEGG KO function annotations are required for metagenomic shotgun sequencing study. (2) metabolome dataset: a metabolite abundance table (.csv format) with compound name, HMDB database ID, and chemical class information, and (3) a metadata table containing sample IDs and group information. Sample IDs from microbiome and metabolome datasets may not be exactly the same, but should be paired correctly within the metadata table. Once uploaded successfully, M^2IA system examines the consistency of sample IDs from two datasets and notifies users whether there exist unmatched samples. After that, the metadata table is presented and can be modified by users interactively. For example, instead of uploading a new table, when users want to change grouping information or remove certain samples, they can modify group IDs or uncheck certain samples directly on the table for downstream data analysis.

Missing value processing. This step includes data filtering and missing value imputation. First, unqualified variables can be removed according to the criteria of sample prevalence and variance across samples. For microbiome data, the minimal count number, the

percentage of samples with non-missing values, and the relative standard deviation (RSD) are used for screening. As suggested by MicrobiomeAnalyst[10], features with very low count (e.g., <2) in a few samples (e.g., <20%), or very stable with limited variations across all the samples (e.g., RSD < 30%), are considered difficult to interpret their biological significance in the community and can be removed directly. Similarly, our previous work summarized that those metabolic features with missing values in more than 80% of samples or RSD values smaller than 30% can be filtered at the beginning. The sample prevalence is calculated based on all the samples or samples within each group[13].

Second, both microbiome and metabolome have the characteristics of sparsity seen as the absence of many taxa or metabolites across samples due to biological and/or technical reasons[14]. Such missing values may pose general numerical challenges for traditional statistical analysis, thus M^2IA provides a variety of methods for missing value imputation, aiming to remove zeros or NAs in the data matrix and facilitate subsequent statistical analysis. There are two different approaches to handle missing values: one is to simply replace missing values with a certain value (called pseudo count in microbiome), including half of the minimum or the median value; another is to apply regression models to impute missing values, such as random forest, k-nearest neighbors, Probabilistic PCA, Bayesian PCA, singular value decomposition, and the quantile regression imputation of left-censored data (QRILC)[13]. Users may choose an appropriate method for missing value imputation.

Data normalization. After missing value processing, users can perform normalization method in order to make more meaningful comparisons. M^2IA provides the total sum scaling that calculates the relative percentage of features, or the log transformation when data does not follow normal distribution. To note, scaling or transformation methods are also provided for specific statistical analysis, such as principal component analysis (PCA) and PLS-DA analysis.

Global similarity between two datasets

After data processing, the first step is to apply Coinertia analysis (CIA) and Procrustes analysis (PA) to evaluate the global similarity between metabolome and microbiome dataset (Figure 1)[15]. CIA can be considered as a PCA model of the joint covariances of two datasets. Similarly, PA measures the congruence of two-dimensional data distributions from superimposition and scaling of PCA models of two datasets (Figure 2A). The correlation coefficient R is between 0 and 1, and the closer it is to 1 indicates the greater similarity between two datasets.

Pairwise metabolite-microbe correlation analysis

Correlation analysis methods. M^2IA provides five different types of correlation analysis, including Pearson, Spearman, SparCC, CCLasso, and Maximal Information Coefficient (MIC) analysis. Although each method has it pros and cons in different situations, users can choose an appropriate one. Since Spearman correlation analysis outperforms other methods due to their overall performances[16], it has been set as a default one in M^2IA. Microbes can be analyzed according to their different taxonomic annotations (i.e., phylum, genus, and species) and metabolites at different chemical classifications (e.g., amino acids, sugars, free fatty acids etc.). In addition, M^2IA allows users to define the specific criteria of significance between microbe-metabolite pairs for subsequent analysis, e.g., absolute correlation coefficients > 0.3 and p value <0.05.

Visualization. Three different ways of visualization are provided in order to explore complex correlations between microbes and metabolites. The first one is to apply circos plot that can help users to quickly identify those microbes having close correlations with different classes of metabolites (e.g., amino acids) (Figure 2C). The second one is to apply a heat map to illustrate their relative positive/negative correlations between each microbe and metabolite. Meanwhile, hierarchical clustering is used to analyze the similarities among metabolites or microbes, in which closely correlated metabolites/microbes are usually clustered together. The third one is to provide a microbe-metabolite interaction network using cytoscape technique, which can help users to explore complex relationship between microbes and metabolites (Figure 2D).

Multivariate regression–based data integration

Unsupervised multivariate analysis. M^2IA provides canonical correlation analysis (CCA)[17, 18] and O2PLS[19] in order to evaluate the inherent correlations between two datasets without considering phenotype information. CCA aims to find two new bases (canonical variate) in which the correlation between original parameters of two datasets is maximized. O2PLS is capable of modeling both prediction and systematic variation, and the joint score plot indicates their relationship between two data matrix. Metabolites/microbes with the large canonical coefficients from CCA model or loading values from O2PLS model are considered essential ones for their similarities (Figure 2B).

Supervised multivariate analysis. In comparison, supervised multivariate analysis methods integrate two data matrix initially and identify differential variables (microbes & metabolites) that significantly contribute to the discrimination of different groups. Here, M^2IA offers PCA score-based differential analysis, partial least squares discriminant analysis (PLS-DA), orthogonal partial least squares discriminant analysis (OPLS-DA), and random forest (RF). PCA is originally an unsupervised data mining method; here we examine the differences of the first principal component score values from PCA model, and their correlations with the phenotype information. PLS-DA and OPLS-DA have been commonly applied in the field of metabolomics for data dimension reduction and feature selection. OPLS-DA seeks to maximize the explained variance between groups in a single dimension or the first latent variable (LV), and separate the within group variance (orthogonal to group difference) into orthogonal LVs (Figure 3A). The variable loadings and/or coefficient weights from a validated PLS-DA and OPLS-DA model are used to rank variables with respect to their performance for discriminating between groups (Figure 3B). Boruta algorithm-based RF classifier can identify important features by shuffling samples and adding extra randomness to the system (Figure 3C).

Network analysis

WGCNA-based network analysis. Weighted gene co-expression network analysis (WGCNA) was recently raised to integrate high-throughput '-omics' datasets for identifying the potential mechanistic links[11]. In M^2IA, WGCNA algorithm is used to

collapse co-abundant metabolites into different clusters, and thus metabolites within a cluster are highly correlated. One attractive feature is that both identified and unknown metabolite features can be considered. KEGG microbiome functional modules are annotated using Tax4fun2[20, 21] for 16s RNA sequencing data or HUMAnN pipeline[22] for shotgun sequencing data, respectively. Finally, the correlations between metabolite clusters and microbial functional modules with phenotype information are examined using Spearman correlation analysis and further visualized using heat map and network.

Metabolic function analysis. The final step aims to analyze or predict the biological function profiles using metabolome and microbiome data, respectively. The metabolic pathway activity is analyzed using PAPi package in R software [23]. In parallel, function profiles of microbial communities are predicted using Tax4Fun tool for 16s rRNA data and HUMAnN for shotgun sequencing data. Linear discriminant analysis is further used to identify differential KO identifiers from microbiome and metabolome (Figure 4A-C). Their relationships are further investigated using a microbe-metabolite interaction network by connecting a specific metabolite, its reactions and related enzymes, and the genomes of organism (both host and microbial)(Figure 4D-E).

Automated html report

Once users upload the data and optimize parameters of data analysis, M^2IA provides a simplified user experience through a one-click button to submit a job. It may take a few minutes for M^2IA to perform data analysis and generate interpretable results. The exact processing time depends on the sample size, number of variables, and number of pair-wise group comparisons. Once the analysis is completed, M^2IA will summarize key results and produce an html report for users. In addition, a zip-compressed package containing all the supporting data files (tables and figures) can be downloaded for future use.

Application example

To validate the performance of M^2IA, we analyzed the fecal metabolome and microbiome of 786 individuals from TwinsUK cohort[24]. 16S rRNA was extracted from

fecal samples, PCR amplified, barcoded per sample and sequenced with the Illumina MiSeq platform, as previously described[25]. A total of 1,116 different metabolites were measured using UPLC-MS/MS platforms. First, the Procrustes Analysis revealed moderate similarities between metabolome and microbiome with statistical significance (Figure 2A, $r=0.31$, $p<0.05$). Since most of individuals at the metabolic level surrounded those at the microbiome level, the variations of metabolic profiles among individuals were larger in metabolome. The O2PLS analysis indicated microbial and metabolic data matrixes were highly correlated with correlation coefficient 0.76 ($p<0.05$, Figure 2B), and those variables that contributed to their similarities could be identified according to their weights. Then, Spearman correlation analysis was performed to explore specific microbe-metabolite correlations. The circos plot showed seven different classes of metabolites (i.e., nucleotides, carbohydrates, peptides, cofactors and vitamins, amino acids, xenobiotics, and lipids) were correlated with microbes mainly belonging to Firmicutes, Actinobacteria, Bacteroidetes, and Protebacteria (Figure 2C). Among which, four microbes at genus level belonging to the Firmicutes had the most significant correlations with metabolites, and their specific correlations were further visualized in a network (Figure 2D).

According to BMI, participants were divided into obese (OB, BMI >28) and healthy control group (HC, BMI <24). The multivariate OPLS-DA model based on the integrated data matrix of metabolome and microbiome indicated the obvious difference between OB and HC group (Figure 3A). Differential metabolites and microbes contributing to their separations were identified using both VIP values (VIP>1) and correlation coefficients ($p<0.05$), as indicated by the V-plot (Figure 3B). Alternative multivariate models e.g., RF model, could be further used to validate the relative contributions of microbes and metabolites to the discrimination between two groups (Figure 3C). For example, arginine was identified as a significant marker from both OPLS-DA and RF model. Then, we performed functional network analysis to further interpret the biological significance of their correlations. A total of 39 and 16 significant biological functions (KOs) were identified in microbiome and metabolome between OB and HC group, respectively ($p < 0.05$, Figure 4A,B). Two important pathways (i.e., the benzoate degradation and

phosphotransferase system) were overlapped (Figure 4C), among which, the benzoate degradation was the most significant metabolic pathway in both microbial and metabolic function analysis. Then, each differential metabolite and relevant microbes that may participate in its metabolic reactions were further identified through searching the KEGG database. For example, 3,4-dihydroxybezoate (also called protocatechuic acid) in the benzoate degradation was a major metabolite of antioxidant polyphenols formed by gut microbiota, and might have beneficial effects against inflammation and insulin resistance in obesity[26]. Our results indicated that 3,4-dihydroxybezoate was significantly reduced in obesity, and eight different gut microbes might participate in its metabolism (Figure 4D). Similarly, a total of 28 different microbes identified in this study were involved in the glycerol metabolism of phosphotransferase system (Figure 4E).

Discussion

We have implemented M^2IA pipeline as a user-friendly web server that can provide microbiome and metabolome data integration analysis to understand the important roles of microbial metabolism in diverse disease contexts. This is a bioinformatics workflow that integrates a wide range of univariate and multivariate methods, including PA/CIA-based data matrix similarity analysis, univariate-based correlation analysis, multivariate regression-based analysis, and knowledge-based function analysis. The advantage of applying these methods simultaneously is to understand the inherent characteristics of the high-dimensional omics data in different ways, and obtain reliable results from multiple analyses. M^2IA also implements KEGG database in order to link orthologous gene groups to metabolic reactions and annotated compounds. So that researchers will have better understanding of their origins of metabolites, from host or microbial, and what microbes participate in specific metabolic reactions. To make it more convenient, M^2IA automatically produces a comprehensive report that summarizes and interprets the key results.

Users can apply M^2IA to their own microbiome and metabolome data. In terms of analytical platforms, M^2IA accepts the microbiome data from 16S rRNA gene sequencing or shotgun metagenomic sequencing technologies, and metabolomics data from mass

spectrometry or NMR analytical platforms. For sample types, many human studies on gut microbiota collect fecal samples, due to its non-invasive characteristics. The fecal metabolome provides a functional readout of the gut microbiome, and the integrative analysis can provide better understanding of correlations between gut microbiome and metabolome. We recommend applying the same fecal samples of human subjects or fecal/colon contents of animals for analysis. However, samples types can be from any site, i.e., saliva, buccal mucosa, and colon tissue samples. Simultaneously, metabolomics studies can analyze feces, plasma/serum, urine, saliva, exhaled breaths, cerebrospinal fluid, and tissues of target organs. Thus, sample types may differ, but they need to require originating from a same subject for both microbiome and metabolome analysis.

Finally, the limitations and future directions of this study disserve to be mentioned. M^2IA accepts annotated microbioa and metabolites as input data matrix. However, the raw data preprocessing steps are not included in this pipeline, e.g., quality control and microbial annotation for microbiome data, and peak detection and metabolite annotation for metabolomics data. Currently, many sophisticated software tools or pipelines have been developed for original data preprocessing, such as Qimme2[27, 28] for 16s RNA sequencing, and XCMS for MS-based metabolic profiling[29]. M^2IA focuses on providing a standardized pipeline and a comprehensive workflow for the integrative analysis of metabolome and microbiome data, not only exploring the statistically significant correlations but also investigating the biological significance of their interaction network. Moreover, KEGG database based functional network analysis can help to explain their biological correlations and develop mechanistic hypothesis potentially applicable to the development of nutritional and personalized therapies. In the future, more advanced integration methods will be introduced in M^2IA and more independent knowledge databases (e.g., enzyme database and drug metabolism database) will be integrated within M^2IA for deeper biological interpretation.

Conclusions

In summary, M^2IA is an effective and efficient computational tool for experimental biologists to comprehensively analyze and interpret the important interactions between microbiome and metabolome in the big data era.

Authors' contributions

YN contributed to develop the methodology, algorithms, and write the manuscript. YG contributed to the software development and implementation. YD provided valuable testing datasets during method development of our pipeline. PW and CS contributed to TwinsUK data preparation and processing for pipeline validation. HC and FJ provided valuable suggestions on software development and manuscript modification. FF supported this project and contributed to the review and modification of the manuscript.

Availability of data and materials

Data used in this study can be accessed at http://m2ia.met-bioinformatics.cn.

Funding

This work was funded by National Natural Science Foundation of China (81900510), the National Key Research and Development Program of China (No. 2016YFC1305301), and the startup funding from the Children's Hospital, Zhejiang University School of Medicine.

Acknowledgements

Not applicable

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Reference

1. Cho I, Blaser MJ. APPLICATIONS OF NEXT-GENERATION SEQUENCING The human microbiome: at the interface of health and disease. Nat Rev Genet 2012, **13**(4)**:** 260-270.
2. Shaffer M, Armstrong AJS, Phelan VV, Reisdorph N, Lozupone CA. Microbiome and metabolome data integration provides insight into health and disease. Transl Res 2017, **189:** 51-64.
3. Wang WL, Xu SY, Ren ZG, Tao L, Jiang JW, Zheng SS. Application of metagenomics in the human gut microbiome. World J Gastroentero 2015, **21**(3)**:** 803-814.
4. Wahlstrom A, Sayin SI, Marschall HU, Backhed F. Intestinal Crosstalk between Bile Acids and Microbiota and Its Impact on Host Metabolism. Cell Metab 2016, **24**(1)**:** 41-50.
5. Franzosa EA, Sirota-Madi A, Avila-Pacheco J, Fornelos N, Haiser H, Reinker S, Vatanen T, Hall AB, Mallick H, Mclver LJ, Sauk JS, Wilson RG, Stevens BW, Scott JM, Pierce K, Deik AA, Bullock K, Imhann F, Porter JA, Zhernakova A, Fu JY, Weersma RK, Wijmenga C, Clish CB, Vlamakis H, Huttenhower C, Xavier RJ. Gut microbiome structure and metabolic activity in inflammatory bowel disease. Nat Microbiol 2019, **4**(2)**:** 293-305.
6. Louis P, Hold GL, Flint HJ. The gut microbiota, bacterial metabolites and colorectal cancer. Nat Rev Microbiol 2014, **12**(10)**:** 661-672.
7. Gu YY, Wang XK, Li JH, Zhang YF, Zhong HZ, Liu RX, Zhang DY, Feng Q, Xie XY, Hong J, Ren HH, Liu W, Ma J, Su Q, Zhang HM, Yang JL, Wang XL, Zhao XJ, Gu WQ, Bi YF, Peng YD, Xu XQ, Xia HH, Li F, Xu X, Yang HM, Xu GW, Madsen L, Kristiansen K, Ning G, Wang WQ. Analyses of gut microbiota and plasma bile acids enable stratification of patients for antidiabetic treatment. Nat Commun 2017, **8**.
8. Caussy C, Hsu C, Lo MT, Liu A, Bettencourt R, Ajmera VH, Bassirian S, Hooker J, Sy E, Richards L, Schork N, Schnabl B, Brenner DA, Sirlin CB, Chen CH, Loomba R, Consortium GNT. Link between gut-microbiome derived metabolite and shared gene-effects with hepatic steatosis and fibrosis in NAFLD. Hepatology 2018, **68**(3)**:** 918-932.
9. Vernocchi P, Del Chierico F, Putignani L. Gut Microbiota Profiling: Metabolomics Based Approach to Unravel Compounds Affecting Human Health. Front Microbiol 2016, **7:** 1144.
10. Dhariwal A, Chong J, Habib S, King IL, Agellon LB, Xia J. MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. Nucleic Acids Res 2017, **45**(W1)**:** W180-W188.
11. Pedersen HK, Forslund SK, Gudmundsdottir V, Petersen AO, Hildebrand F, Hyotylainen T, Nielsen T, Hansen T, Bork P, Ehrlich SD, Brunak S, Oresic M, Pedersen O, Nielsen HB. A computational framework to integrate high-

throughput '-omics' datasets for the identification of potential mechanistic links. Nat Protoc 2018, **13**(12)**:** 2781-2800.
12. Chong J, Xia JG. Computational Approaches for Integrative Analysis of the Metabolome and Microbiome. Metabolites 2017, **7**(4).
13. Wei R, Wang J, Su M, Jia E, Chen S, Chen T, Ni Y. Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. Scientific reports 2018, **8**(1)**:** 663.
14. Tsilimigras MC, Fodor AA. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. Ann Epidemiol 2016, **26**(5)**:** 330-335.
15. McHardy IH, Goudarzi M, Tong M, Ruegger PM, Schwager E, Weger JR, Graeber TG, Sonnenburg JL, Horvath S, Huttenhower C, McGovern DP, Fornace AJ, Jr., Borneman J, Braun J. Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships. Microbiome 2013, **1**(1)**:** 17.
16. You YJ, Liang DD, Wei RM, Li MC, Li YT, Wang JY, Wang XY, Zheng XJ, Jia W, Chen TL. Evaluation of metabolite-microbe correlation detection methods. Anal Biochem 2019, **567:** 106-111.
17. Smolinska A, Tedjo DI, Blanchet L, Bodelier A, Pierik MJ, Masclee AAM, Dallinga J, Savelkoul PHM, Jonkers D, Penders J, van Schooten FJ. Volatile metabolites in breath strongly correlate with gut microbiome in CD patients. Anal Chim Acta 2018, **1025:** 1-11.
18. Kostic AD, Gevers D, Siljander H, Vatanen T, Hyotylainen T, Hamalainen AM, Peet A, Tillmann V, Poho P, Mattila I, Lahdesmaki H, Franzosa EA, Vaarala O, de Goffau M, Harmsen H, Ilonen J, Virtanen SM, Clish CB, Oresic M, Huttenhower C, Knip M, Group DS, Xavier RJ. The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. Cell Host Microbe 2015, **17**(2)**:** 260-273.
19. Bylesjo M, Eriksson D, Kusano M, Moritz T, Trygg J. Data integration in plant biology: the O2PLS method for combined modeling of transcript and metabolite data. Plant J 2007, **52**(6)**:** 1181-1191.
20. Tax4Fun2. https://sourceforgenet/projects/tax4fun2/.
21. Asshauer KP, Wemheuer B, Daniel R, Meinicke P. Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. Bioinformatics 2015, **31**(17)**:** 2882-2884.
22. Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL, Rodriguez-Mueller B, Zucker J, Thiagarajan M, Henrissat B, White O, Kelley ST, Methe B, Schloss PD, Gevers D, Mitreva M, Huttenhower C. Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comput Biol 2012, **8**(6)**:** e1002358.
23. KEGG database. https://www.kegg.jp.
24. Zierer J, Jackson MA, Kastenmüller G, Mangino M, Long T, Telenti A, Mohney RP, Small KS, Bell JT, Steves CJ, Valdes AM, Spector TD, Menni C. The fecal metabolome as a functional readout of the gut microbiome. 2018, **- 50**(- 6)**:** - 795.

25. Goodrich JK, Davenport ER, Beaumont M, Jackson MA, Knight R, Ober C, Spector TD, Bell JT, Clark AG, Ley RE. Genetic Determinants of the Gut Microbiome in UK Twins. Cell Host Microbe 2016, **19**(5)**:** 731-743.
26. Ormazabal P, Scazzocchio B, Vari R, Santangelo C, D'Archivio M, Silecchia G, Iacovelli A, Giovannini C, Masella R. Effect of protocatechuic acid on insulin responsiveness and inflammation in visceral adipose tissue from obese individuals: possible role for PTP1B. Int J Obes (Lond) 2018, **42**(12)**:** 2012-2021.
27. Qimme 2. https://qiime2.org.
28. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 2010, **7**(5)**:** 335-336.
29. Forsberg EM, Huan T, Rinehart D, Benton HP, Warth B, Hilmers B, Siuzdak G. Data processing, multi-omic pathway mapping, and metabolite activity analysis using XCMS Online. Nat Protoc 2018, **13**(4)**:** 633-651.

Legends

Figure 1. The flowchart of M^2IA.

Figure 2. Illustration of similarity analysis and spearman correlation analysis results. (A) PA. The length of lines connecting two points indicates the agreement of samples between two datasets. (B) O2PLS model. (C) Circos plot of spearman correlations between microbes and metabolites. (D) Network of spearman correlations between differential metabolites and microbes belonging to the Firmicutes.

Figure 3. (A) Score plot of an OPLS-DA model between OB and HC group. (B) Variable importance plot (V-plot) of an OPLS-DA model between OB and HC group. (C) Feature selection of Random Forest modeling.

Figure 4. (A-B) Bar plot of differential KO functions from metabolome and microbiome data. (C) Venn diagram of differential KOs. (D-E) Interaction network of each metabolite (3,4-dihydroxybenzoate and glycerol) with microbes that may participate its metabolism.

Figure 1

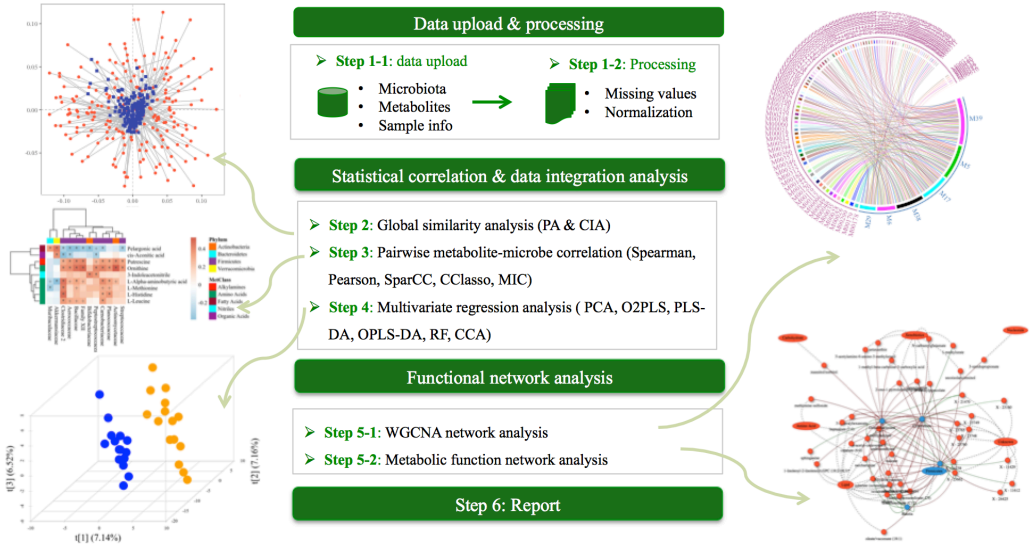
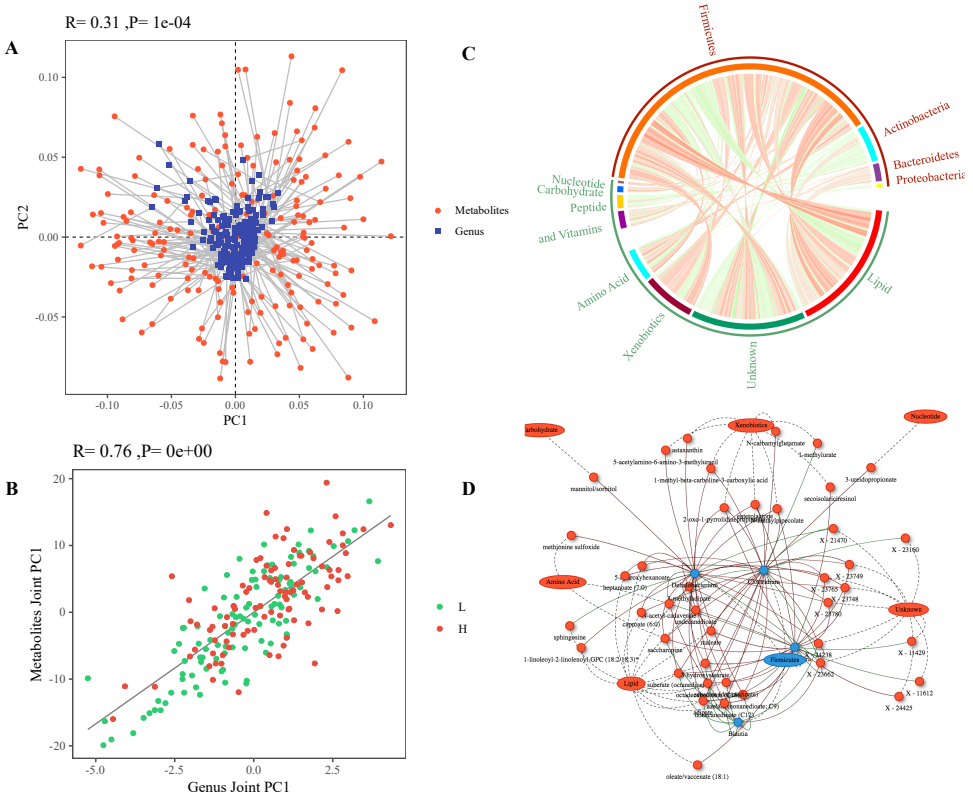
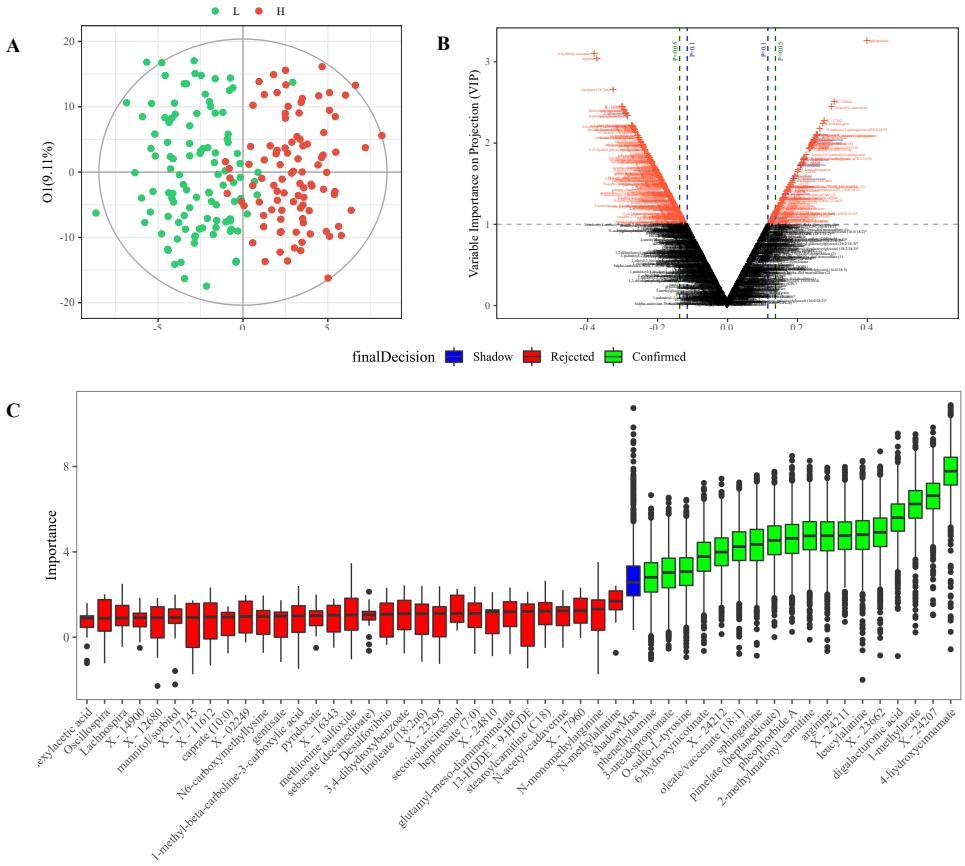

Data upload & processing
Step 1-1: data upload
Microbiota
Metabolites
Sample info
Step 1-2: Processing
Missing values
Normalization
Statistical correlation & data integration analysis
Step 2: Global similarity analysis (PA & CIA)
Step 3: Pairwise metabolite-microbe correlation (Spearman, Pearson, SparCC, CClasso, MIC)
Step 4: Multivariate regression analysis (PCA, O2PLS, PLS-DA, OPLS-DA, RF, CCA)
Functional network analysis
Step 5-1: WGCNA network analysis
Step 5-2: Metabolic function network analysis
Step 6: Report

Figure 2



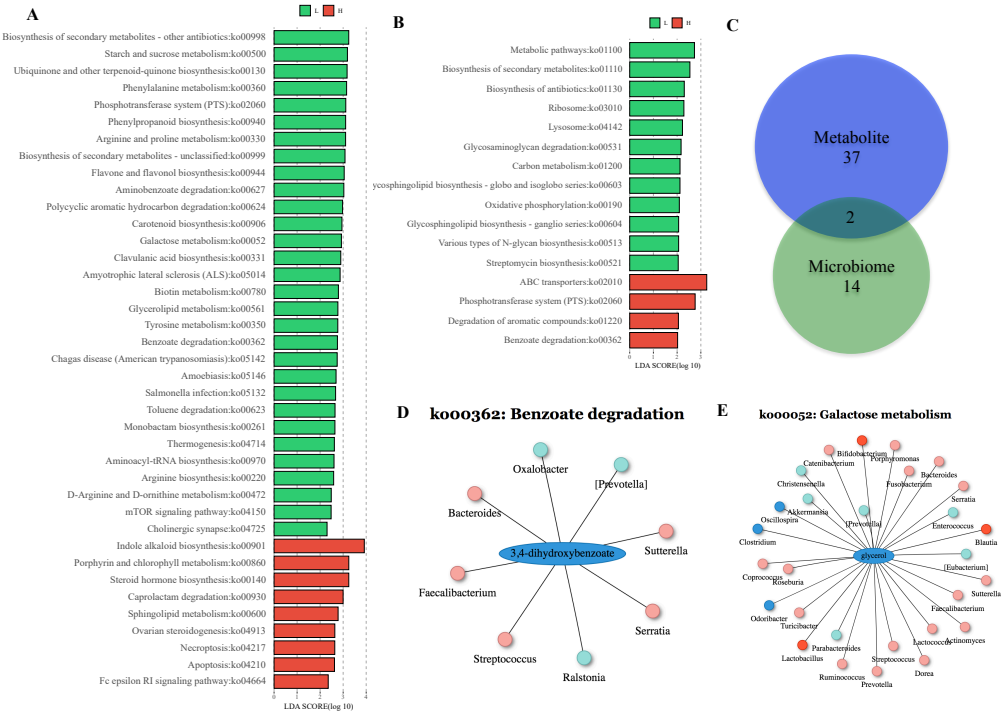
A
R= 0.31 ,P= 1e-04
PC2
PC1
Metabolites
Genus
B
R= 0.76 ,P= 0e+00
Metabolites Joint PC1
Genus Joint PC1
L
H
C
Firmicutes
Actinobacteria
Bacteroidetes
Proteobacteria
Nucleotide
Carbohydrate
Peptide
and Vitamins
Amino Acid
Xenobiotics
Unknown
Lipid
D
Xenobiotics
Nucleotide
Amino Acid
Unknown
Firmicutes
Lipid

Figure 3



A
L
H
O1(9.11%)
B
Variable Importance on Projection (VIP)
finalDecision
Shadow
Rejected
Confirmed
C
Importance
exylacetic acid
Oscillospira
Lachnospira
X - 14900
X - 12680
mannitol/sorbitol
X - 17145
X - 11612
caprate (10:0)
X - 02249
N6-carboxymethyllysine
gemisate
1-methyl-beta-carboline-3-carboxylic acid
pyridoxate
X - 16343
methionine sulfoxide
sebacate (decanedioate)
Desulfovibrio
3,4-dihydroxybenzoate
linoleate (18:2n6)
X - 23295
secoisolariciresinol
heptanoate (7:0)
X - 24810
glutamyl-meso-diaminopimelate
13-HODE + 9-HODE
stearoylcarnitine (C18)
N-acetyl-cadaverine
X - 17960
N-monomethylarginine
N-methylalanine
shadowMax
phenethylamine
3-ureidopropionate
O-sulfo-L-tyrosine
6-hydroxynicotinate
X - 24212
oleate/vaccenate (18:1)
sphinganine
pimelate (heptanedioate)
pheophorbide A
2-methylmalonyl carnitine
arginine
X - 24211
leucylalanine
X - 23662
digalacturonic acid
1-methylurate
X - 24207
4-hydroxycinnamate

Figure 4


A
L
H
Biosynthesis of secondary metabolites - other antibiotics:ko00998
Starch and sucrose metabolism:ko00500
Ubiquinone and other terpenoid-quinone biosynthesis:ko00130
Phenylalanine metabolism:ko00360
Phosphotransferase system (PTS):ko02060
Phenylpropanoid biosynthesis:ko00940
Arginine and proline metabolism:ko00330
Biosynthesis of secondary metabolites - unclassified:ko00999
Flavone and flavonol biosynthesis:ko00944
Aminobenzoate degradation:ko00627
Polycyclic aromatic hydrocarbon degradation:ko00624
Carotenoid biosynthesis:ko00906
Galactose metabolism:ko00052
Clavulanic acid biosynthesis:ko00331
Amyotrophic lateral sclerosis (ALS):ko05014
Biotin metabolism:ko00780
Glycerolipid metabolism:ko00561
Tyrosine metabolism:ko00350
Benzoate degradation:ko00362
Chagas disease (American trypanosomiasis):ko05142
Amoebiasis:ko05146
Salmonella infection:ko05132
Toluene degradation:ko00623
Monobactam biosynthesis:ko00261
Thermogenesis:ko04714
Aminoacyl-tRNA biosynthesis:ko00970
Arginine biosynthesis:ko00220
D-Arginine and D-ornithine metabolism:ko00472
mTOR signaling pathway:ko04150
Cholinergic synapse:ko04725
Indole alkaloid biosynthesis:ko00901
Porphyrin and chlorophyll metabolism:ko00860
Steroid hormone biosynthesis:ko00140
Caprolactam degradation:ko00930
Sphingolipid metabolism:ko00600
Ovarian steroidogenesis:ko04913
Necroptosis:ko04217
Apoptosis:ko04210
Fc epsilon RI signaling pathway:ko04664
0
1
2
3
4
LDA SCORE(log 10)
B
L
H
Metabolic pathways:ko01100
Biosynthesis of secondary metabolites:ko01110
Biosynthesis of antibiotics:ko01130
Ribosome:ko03010
Lysosome:ko04142
Glycosaminoglycan degradation:ko00531
Carbon metabolism:ko01200
ycosphingolipid biosynthesis - globo and isoglobo series:ko00603
Oxidative phosphorylation:ko00190
Glycosphingolipid biosynthesis - ganglio series:ko00604
Various types of N-glycan biosynthesis:ko00513
Streptomycin biosynthesis:ko00521
ABC transporters:ko02010
Phosphotransferase system (PTS):ko02060
Degradation of aromatic compounds:ko01220
Benzoate degradation:ko00362
0
1
2
3
LDA SCORE(log 10)
C
Metabolite
37
2
Microbiome
14
D
ko00362: Benzoate degradation
Oxalobacter
[Prevotella]
Bacteroides
3,4-dihydroxybenzoate
Sutterella
Faecalibacterium
Serratia
Streptococcus
Ralstonia
E
ko00052: Galactose metabolism
Bifidobacterium
Porphyromonas
Catenibacterium
Bacteroides
Christensenella
Fusobacterium
Serratia
Akkermansia
Oscillospira
[Prevotella]
Enterococcus
Clostridium
Blautia
glucose
[Eubacterium]
Coprococcus
Roseburia
Sutterella
Faecalibacterium
Odoribacter
Turicibacter
Actinomyces
Lactococcus
Parabacteroides
Lactobacillus
Streptococcus
Dorea
Ruminococcus
Prevotella