MMCA: a Web-based Server for the Microbiome and Metabolome Correlation Analysis

Yan Ni[1,#], Gang Yu[1], Yongqiong Deng[2], Xiaojiao Zheng[3], Tianlu Chen[3], Junfen Fu[1,#], Wei Jia[3,4,#]

1. Children's Hospital, School of Medicine, Zhejiang University, Hangzhou 310029, P.R. China.

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

3. Shanghai Key Laboratory of Diabetes Mellitus and Center for Translational Medicine, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Shanghai 200233, China.

4. University of Hawaii Cancer Center, Honolulu, HI 96813, USA.

Correspondence should be addressed to Dr. Yan Ni (yanni617@zju.edu.cn), Dr. Junfen Fu (fjf68@zju.edu.cn), and Dr. Wei Jia (wjia@cc.hawaii.edu).

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 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 correlation analysis pipeline (MMCA) to meet the urgent needs for tools that effectively integrate microbiome and metabolome data to derive biological insights.

Results: To make the MMCA pipeline available to a wider research community, we have implemented a web server (http://mmca.met-bioinformatics.cn). MMCA integrates a variety of statistical analysis methods in order to obtain reliable results from multiple analyses, including univariate analysis and multivariate modeling. MMCA also provides KEGG-based functional network analysis in order to investigate their biological interplay between metabolites and microbes. To make it more convenient, an html-based report is available for overview and can be downloaded for later use.

Conclusions: MMCA allows users to upload annotated microbiome and metabolome data, provides a user-friendly interface to analyze and visualize the complex interplay between microbiome and metabolome, and helps users to develop mechanistic hypothesis for nutritional and personalized therapies of diseases.

Keywords: MMCA, 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 the 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 correlation analysis (MMCA). MMCA integrates 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. MMCA 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, MMCA is an automated data analysis and reporting pipeline that can be applied by users with little training in bioinformatics. MMCA is a web-based server with user-friendly interface, and public available to academic users through http://mmca.met-bioinformatics.cn.

Implementation

MMCA 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 written in R 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. Once data analysis is completed, all the data and analysis results will be saved for 72 hours and removed automatically afterwards. MMCA 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 MMCA contains six major steps: (1) data upload and processing, (2) global similarity analysis of microbiome and metabolome data matrix, (3) pairwise metabolite-microbe correlation analysis, (4) multivariate regression-based 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 MMCA. Each step offers a variety of options and procedures to help 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 annotation 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 two datasets may not be exactly the same, but should be paired correctly within the metadata table. Once uploaded successfully, MMCA 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 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 MicrbiomeAnalysis[13], features with very low count (e.g., <2) in a few samples (e.g., <20%), or very stable 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[14].

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[15]. Such missing values may pose general numerical challenges for traditional statistical analysis, thus MMCA provides a variety of methods for missing value imputation, aiming to remove zeros 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)[14]. 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. MMCA 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

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)[16]. CIA can be considered as a PCA model of the joint covariances of two datasets. The RV coefficient is between 0 and 1, and the closer it is to 1 the greater similarity between two datasets (Figure 2A). PA measures the congruence of two-dimensional data distributions from superimposition and scaling of PCA models of two datasets. Spearman correlation analysis was used to measure their similarities (Figure 2B).

Pairwise metabolite-microbe correlation analysis

Correlation analysis methods. MMCA 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 as suggested by our previous work on method comparisons. Since Spearman correlation analysis outperforms other methods due to their overall performances[17], it has been set as a default one in MMCA. 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, MMCA allows users to define the specific criteria of significant microbe-metabolite pairs of interest for subsequent analysis, e.g., correlation coefficients > 0.3 or <-0.03 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 belonging to a specific phylum (e.g., Firmicutes) that have close correlations with a specific group of metabolites (e.g., amino acids) (Figure 2C). The second one is to apply a heat map to illustrate the 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 (Figure 2D). 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 2E).

Multivariate regression–based data integration

Unsupervised multivariate analysis. MMCA provides canonical correlation analysis (CCA)[18, 19] and O2PLS[20] in order to evaluate the inherent correlations between two datasets without considering phenotype information, and to evaluate their relative contributions of variables to their similarities/differences. 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.

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, MMCA offers PCA score-based differential 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 MMCA, WGCNA algorithm is used to collapse co-abundant metabolites into different clusters, and 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[21] for 16s RNA sequencing data or Diamond 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 (Figure 4A-B).

Metabolic function analysis. The final step aims to interpret the biological correlations between microbes and metabolites through KEGG orthology (KO) and pathway database[23]. First, KO function analysis/prediction in microbiome data provides KO number and their relative expressions (Figure 4C). Meanwhile, MMCA provides metabolic pathway enrichment analysis on differential metabolites (Figure 4D). According to the chemical reactions of metabolites, related enzymes, genes, and their functional KO orthologs in KEGG pathway, MMCA enables to provide microbe-metabolite interaction network indicating the potential involvement of microbes in a specific metabolic pathway (Figure 4E).

Automated html report

Once users upload the data and optimize parameters of data analysis, MMCA provides a simplified user experience through a one-click button to submit a job. It may take a few minutes for MMCA 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, MMCA 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 illustrate the utility of MMCA, we conducted a study to identify characteristic gut microbiome and metabolome of adolescent patients with acne vulgaris, and their correlations. Fecal samples were collected from 15 patients (YAS) and 15 age and gender-matched healthy controls (YCS). Gut microbiome was analyzed by sequencing 16S ribosomal RNA gene (V3-V4 region) and targeted quantitative analysis of 118 fecal metabolites was performed using gas chromatography time-of-flight mass spectrometry. PA and CIA analysis indicated the slight similarity between two data matrix with RV value 0.27 and correlation coefficient 0.34, respectively (Figure 2A-B). Then, spearman correlation analysis was performed to explore specific microbe-metabolite correlations. The circos plot showed four different classes of metabolites (i.e., fatty acids, indoles, organic acids and amino acids) correlated with microbes belonging to Firmicutes, Actinobacteria, Protebacteria, and Verrucomicrol phylum (Figure 2C). The heatmap with specific microbe-metabolite connections indicated that three major metabolites (i.e., aminoadipic acid, 3-indoeacetonitril, p-Hydroxyphenylactic acid) had significant correlations with differential microbes at Genus level (Figure 2D). The network further visualized their significant correlations between microbes and metabolites and their classifications (Figure 2E).

The next step was to build multivariate OPLS-DA and RF model in order to evaluate their relative contributions to the discrimination between YAS and YCS group. A total of 265 variables consisting 147 microbes at Genus level and 118 individual metabolites were integrated together as a data matrix. OPLS-DA model identified 24 microbes and 5 metabolites with VIP >1 and significant correlations $P<0.05$ (Figure 3A-B). The five metabolites were linoleic acid, 4-hydroxybenzoic acid, aminoadipic acid, cis-Aconitic acid, and myristic acid. However, only 10 significant microbes were accepted in the RF model, among which, Bacillus and Lactococcus belonging to Bacilli class were the most important ones (Figure 3C). Thus, it seemed that gut microbia changed more significantly

than metabolome in adolescent patients with acne vulgaris. Finally, we performed functional network analysis to further interpret the biological significance of their correlations. WGCNA identified that there were two major metabolite modules (i.e., fatty acids and organic acids) significantly correlated with microbiotal predicted functions (Figure 4A-B). A total of 40 functions were significant between YAS and YCS group (Figure 4C). Meanwhile, three significant metabolic pathways were identified from metabolic pathway enrichment analysis (p value < 0.05), among which, fatty acid biosynthesis was the most significant pathway (Figure 4D). Finally, MMCA applied KEGG database searching and provided an interaction network for each differential metabolite (e.g., Dodecanoic acid) to indicate all the relevant microbes that may be involved in the same metabolic reaction (Figure 4E). To summarize, gut microbiome changed significantly in adolescent patients with acne vulgaris, particularly for microbes belonging to Bacilli class. In comparison, gut metabolome did not show significant changes, but altered fatty acid metabolism might be associated with the development of acne vulgaris in adolescent patients.

Discussion

We have implemented MMCA pipeline as a user-friendly web server that can provide microbiome and metabolome data integration 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 network 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. MMCA also implements KEGG database in order to link orthologous gene groups to reactions and annotated compounds. So that researchers will have better understanding of microbial metabolism, i.e., the participation and relative contributions of gut microbiota to certain metabolic reactions or pathway. To make it more convenient, MMCA automatically produces a comprehensive report that summarizes and interprets the key results.

Users can apply MMCA to their own microbiome and metabolome data. In terms of analytical platforms, MMCA 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 integration 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. MMCA 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[24, 25] for 16s RNA sequencing, and XCMS for MS-based metabolic profiling[26]. MMCA 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 MMCA and more

independent knowledge databases (e.g., enzyme database and drug metabolism database) will be integrated within MMCA for deeper data interpretation.

Conclusions

In summary, MMCA 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 implementation. YD and KJ conducted the application study of adolescent acne vulgaris. TC and XZ provided valuable suggestions on the method development of -omics data integration. FF and WJ supported this project and contributed to the review and modification of the manuscript before submission for publication.

Availability of data and materials

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

Funding

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

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. Chong J, Xia J. Computational Approaches for Integrative Analysis of the Metabolome and Microbiome. Metabolites 2017, **7**(4).

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. 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.
14. 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.
15. Tsilimigras MC, Fodor AA. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. Ann Epidemiol 2016, **26**(5)**:** 330-335.
16. 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.
17. 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.
18. 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.
19. 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.
20. 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.
21. Tax4Fun2. https://sourceforgenet/projects/tax4fun2/.
22. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods 2015, **12**(1)**:** 59-60.
23. KEGG database. https://www.kegg.jp.
24. Qimme 2. https://qiime2.org.
25. 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.
26. 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 MMCA.

Figure 2. Illustration of similarity analysis and spearman correlation analysis results. (A) CIA. (B) PA. The length of lines connecting two points indicates the agreement of samples between two datasets. (C-E) Circos plot, heat map, and network analysis of spearman correlations between microbes and metabolites.

Figure 3. (A-B) Score and variable importance plot of OPLS-DA model. (C) Feature selection of RF analysis.

Figure 4. Functional network analysis. (A-B) WGCNA-based heatmap and network plot between significant metabolite modules and microbial KO functions. (C) Bar plot of microbial KO functions. (D) Scatter plot of metabolic pathway enrichment analysis results. (E) Metabolite-microbe interaction network.

Figure 1

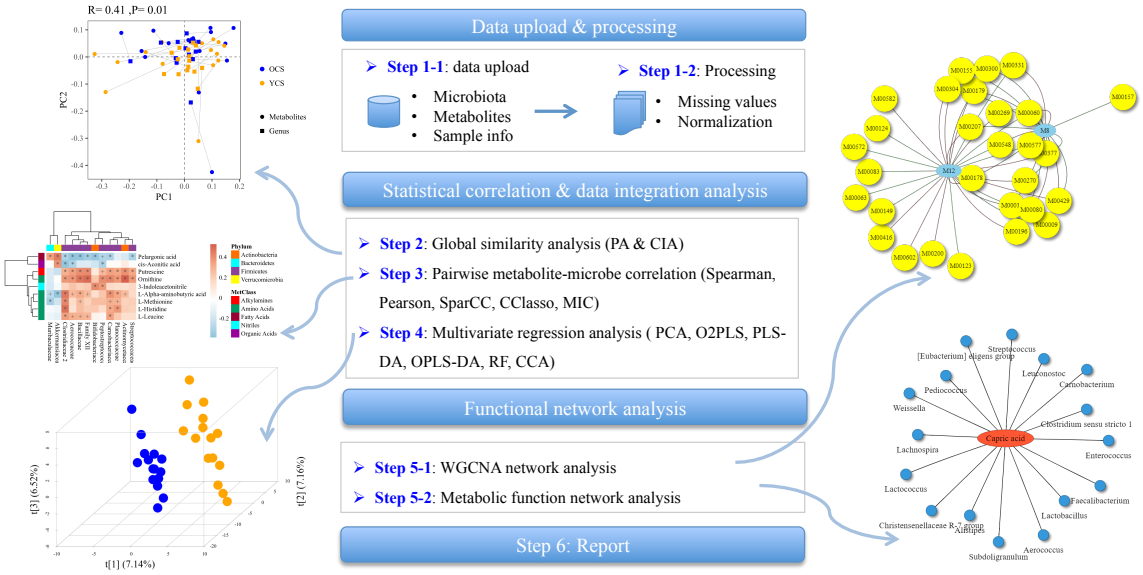

R= 0.41 ,P= 0.01
PC2
PC1
OCS
YCS
Metabolites
Genus
Data upload & processing
Step 1-1: data upload
Microbiota
Metabolites
Sample info
Step 1-2: Processing
Missing values
Normalization
Statistical correlation & data integration analysis
Step 2: Global similarity analysis (PA & CIA)
Step 3: Pairwise metabolite-microbe 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
t[1] (7.14%)
t[2] (7.16%)
t[3] (6.52%)
Capric acid
Streptococcus
[Eubacterium] eligens group
Leuconostoc
Carnobacterium
Pediococcus
Weissella
Clostridium sensu stricto 1
Lachnospira
Enterococcus
Lactococcus
Faecalibacterium
Lactobacillus
Christensenellaceae R-7 group
Alistipes
Aerococcus
Subdoligranulum

Figure 2

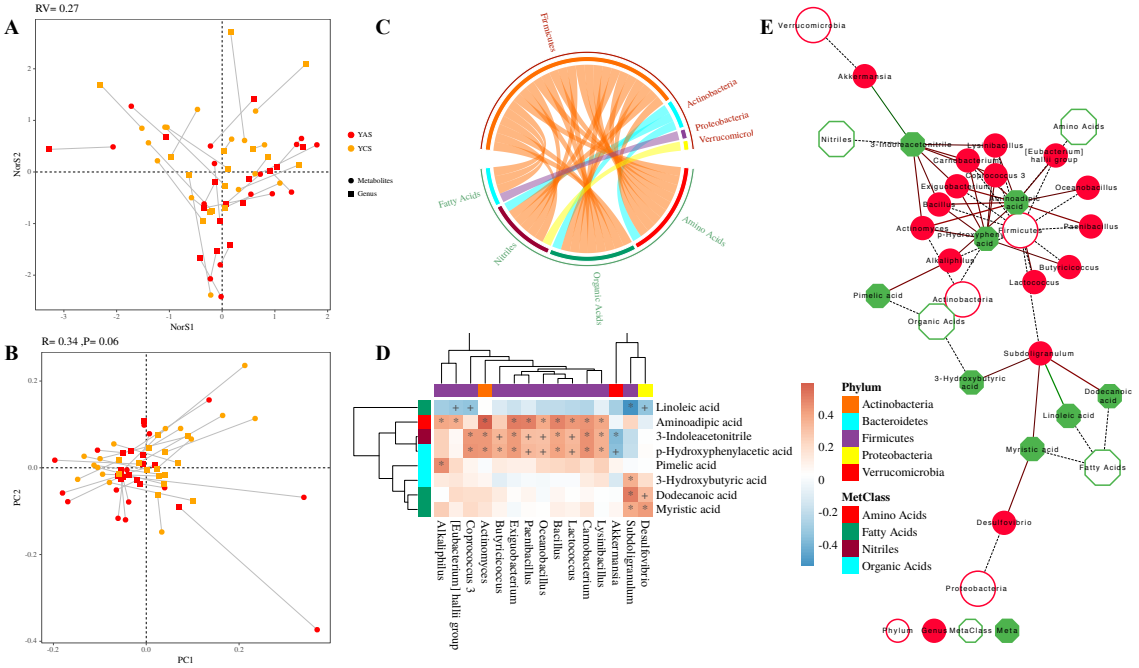

A
RV= 0.27
NorS2
NorS1
YAS
YCS
Metabolites
Genus
B
R= 0.34 ,P= 0.06
PC2
PC1
C
Firmicutes
Actinobacteria
Proteobacteria
Verrucomicrol
Amino Acids
Organic Acids
Nitriles
Fatty Acids
D
Linoleic acid
Aminoadipic acid
3-Indoleacetonitrile
p-Hydroxyphenylacetic acid
Pimelic acid
3-Hydroxybutyric acid
Dodecanoic acid
Myristic acid
Alkaliphilus
[Eubacterium] hallii group
Coprococcus 3
Actinomyces
Butyricicoccus
Exiguobacterium
Paenibacillus
Oceanobacillus
Bacillus
Lactococcus
Carnobacterium
Lysinibacillus
Akkermansia
Subdoligranulum
Desulfovibrio
E
Verrucomicrobia
Akkermansia
Nitriles
3-Indoleacetonitrile
Lysinibacillus
Carnobacterium
[Eubacterium] hallii group
Amino Acids
Coprococcus 3
Exiguobacterium
Oceanobacillus
Bacillus
Aminoadipic acid
Actinomyces
p-Hydroxyphenylacetic acid
Firmicutes
Paenibacillus
Alkaliphilus
Butyricicoccus
Lactococcus
Pimelic acid
Actinobacteria
Organic Acids
Subdoligranulum
3-Hydroxybutyric acid
Dodecanoic acid
Linoleic acid
Myristic acid
Fatty Acids
Desulfovibrio
Proteobacteria
Phylum
Genus
MetaClass
Meta
Phylum
Actinobacteria
Bacteroidetes
Firmicutes
Proteobacteria
Verrucomicrobia
MetClass
Amino Acids
Fatty Acids
Nitriles
Organic Acids

Figure 3

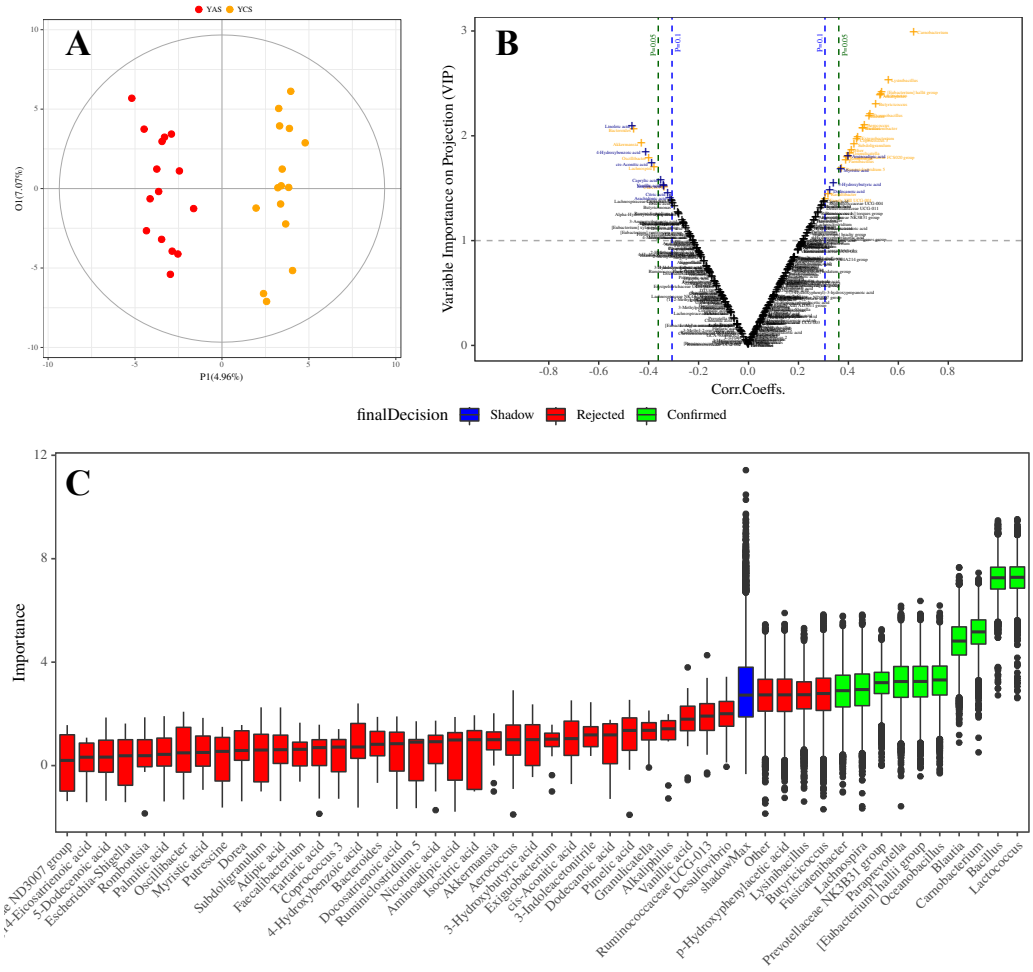
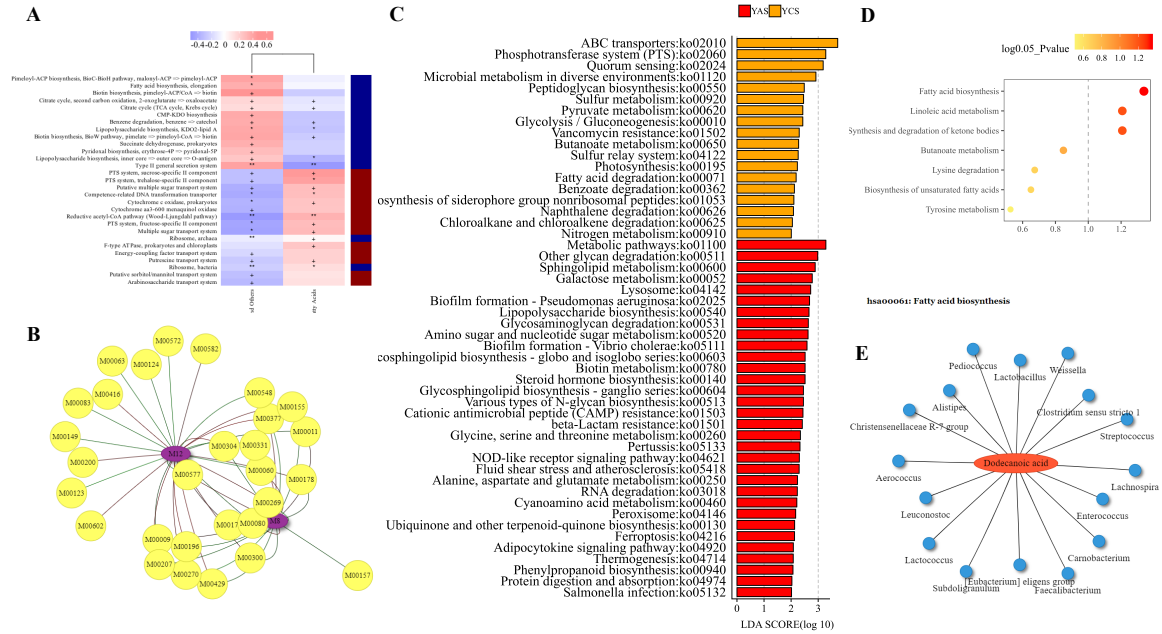

A
YAS
YCS
P1(4.96%)
O1(7.07%)
B
Variable Importance on Projection (VIP)
Corr.Coeffs.
finalDecision
Shadow
Rejected
Confirmed
C
Importance
Shadow/Max
Other

Figure 4


A
-0.4 -0.2 0 0.2 0.4 0.6
Pimeloyl-ACP biosynthesis, BioC-BioH pathway, malonyl-ACP => pimeloyl-ACP
Fatty acid biosynthesis, elongation
Biotin biosynthesis, pimeloyl-ACP/CoA => biotin
Citrate cycle, second carbon oxidation, 2-oxoglutarate => oxaloacetate
Citrate cycle (TCA cycle, Krebs cycle)
CMP-KDO biosynthesis
Benzene degradation, benzene => catechol
Lipopolysaccharide biosynthesis, KDO2-lipid A
Biotin biosynthesis, BioW pathway, pimelate => pimeloyl-CoA => biotin
Succinate dehydrogenase, prokaryotes
Pyridoxal biosynthesis, erythrose-4P => pyridoxal-5P
Lipopolysaccharide biosynthesis, inner core => outer core => O-antigen
Type II general secretion system
PTS system, sucrose-specific II component
PTS system, trehalose-specific II component
Putative multiple sugar transport system
Competence-related DNA transformation transporter
Cytochrome c oxidase, prokaryotes
Cytochrome aa3-600 menaquinol oxidase
Reductive acetyl-CoA pathway (Wood-Ljungdahl pathway)
PTS system, fructose-specific II component
Multiple sugar transport system
Ribosome, archaea
F-type ATPase, prokaryotes and chloroplasts
Energy-coupling factor transport system
Putrescine transport system
Ribosome, bacteria
Putative sorbitol/mannitol transport system
Arabinosaccharide transport system
B
M00572
M00582
M00063
M00124
M00416
M00083
M00548
M00155
M00377
M00011
M00149
M00304
M00331
M00200
M00577
M00060
M00178
M00123
M00269
M00602
M00017
M00080
M00009
M00196
M00300
M00207
M00270
M00429
M00157
C
YAS
YCS
ABC transporters:ko02010
Phosphotransferase system (PTS):ko02060
Quorum sensing:ko02024
Microbial metabolism in diverse environments:ko01120
Peptidoglycan biosynthesis:ko00550
Sulfur metabolism:ko00920
Pyruvate metabolism:ko00620
Glycolysis / Gluconeogenesis:ko00010
Vancomycin resistance:ko01502
Butanoate metabolism:ko00650
Sulfur relay system:ko04122
Photosynthesis:ko00195
Fatty acid degradation:ko00071
Benzoate degradation:ko00362
osynthesis of siderophore group nonribosomal peptides:ko01053
Naphthalene degradation:ko00626
Chloroalkane and chloroalkene degradation:ko00625
Nitrogen metabolism:ko00910
Metabolic pathways:ko01100
Other glycan degradation:ko00511
Sphingolipid metabolism:ko00600
Galactose metabolism:ko00052
Lysosome:ko04142
Biofilm formation - Pseudomonas aeruginosa:ko02025
Lipopolysaccharide biosynthesis:ko00540
Glycosaminoglycan degradation:ko00531
Amino sugar and nucleotide sugar metabolism:ko00520
Biofilm formation - Vibrio cholerae:ko05111
cosphingolipid biosynthesis - globo and isoglobo series:ko00603
Biotin metabolism:ko00780
Steroid hormone biosynthesis:ko00140
Glycosphingolipid biosynthesis - ganglio series:ko00604
Various types of N-glycan biosynthesis:ko00513
Cationic antimicrobial peptide (CAMP) resistance:ko01503
beta-Lactam resistance:ko01501
Glycine, serine and threonine metabolism:ko00260
Pertussis:ko05133
NOD-like receptor signaling pathway:ko04621
Fluid shear stress and atherosclerosis:ko05418
Alanine, aspartate and glutamate metabolism:ko00250
RNA degradation:ko03018
Cyanoamino acid metabolism:ko00460
Peroxisome:ko04146
Ubiquinone and other terpenoid-quinone biosynthesis:ko00130
Ferroptosis:ko04216
Adipocytokine signaling pathway:ko04920
Thermogenesis:ko04714
Phenylpropanoid biosynthesis:ko00940
Protein digestion and absorption:ko04974
Salmonella infection:ko05132
0 1 2 3
LDA SCORE(log 10)
D
log0.05_Pvalue
0.6 0.8 1.0 1.2
Fatty acid biosynthesis
Linoleic acid metabolism
Synthesis and degradation of ketone bodies
Butanoate metabolism
Lysine degradation
Biosynthesis of unsaturated fatty acids
Tyrosine metabolism
0.6 0.8 1.0 1.2
hsa00061: Fatty acid biosynthesis
E
Pediococcus
Lactobacillus
Weissella
Alistipes
Clostridium sensu stricto 1
Christensenellaceae R-7 group
Streptococcus
Dodecanoic acid
Aerococcus
Lachnospira
Leuconostoc
Enterococcus
Lactococcus
Carnobacterium
[Eubacterium] eligens group
Subdoligranulum
Faecalibacterium