3MCor: An Integrative Web Server for Metabolome-Microbiome-Metadata Correlation Analysis

Tao Sun
Shanghai Jiao Tong University Affiliated Six People's Hospital
Mengci Li
Shanghai Jiao Tong University Affiliated Six People's Hospital
Xiangtian Yu
Shanghai Jiao Tong University Affiliated Six People's Hospital
Dandan Liang
Shanghai Sixth Peoples Hospital
Guoxiang Xie
Human Metabolomics Institute
Chao Sang
Shanghai Jiao Tong University Affiliated Six People's Hospital
Wei Jia
Hong Kong Baptist University
Tianlu Chen (chentianlu@sjtu.edu.cn)
Shanghai Jiao Tong University Affiliated Six People's Hospital https://orcid.org/0000-0003-4163-1098

Software article

Keywords: metabolome, microbiome, correlation detection, web server

Posted Date: April 14th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-406175/v1

Version of Record: A version of this preprint was published at Bioinformatics on January 1st, 2021. See the published version at https://doi.org/10.1093/bioinformatics/btab818.

Abstract

Background: Mounting evidences have shown that microbiome and metabolome are closely linked to human health and dual-omics studies expanded our knowledge and understanding of health and life. Here, we designed and developed a full-function and easy-to-use platform, 3MCor (http://3mcor.cn/), for metabolome and microbiome correlation analysis under the instruction of phenotype and with the consideration of confounders.

Results: Many traditional and newly reported correlation analysis methods were integrated for intra- and inter-correlation analysis. Three inter-correlation pipelines are provided for global, hierarchical, and pairwise analysis. Especially, the incorporated network analysis function is conducive to a rapid identification of network clusters and key nodes from a complicated correlation network. Complete numerical results (csv files) and rich figures (pdf files) will be generated in minutes. To our knowledge, 3MCor is the first platform developed specifically for the correlation analysis of metabolome and microbiome. Its functions were compared with corresponding modules of existing omics data analysis platforms. Results from 2 real-world data sets, one from a public library with a continuous phenotype and one from our lab with a categorical phenotype, were used to demonstrate its simple and flexible operation, comprehensive outputs, and distinctive contribution to dual-omics studies. **Conclusions:** 3MCor is powerful with complementary pipelines and comprehensive considerations of phenotypes, confounders, and the interactions among omics features. In addition to the web server, the backend R script is available at https://github.com/chentianlu/3MCorServer.

Background

Over the past 20 years, mounting studies have shown that microbiome and metabolome are closely linked to human health and the compositional and functional abnormalities of their alterations can lead to a variety of diseases, such as obesity, type 2 diabetes mellitus (T2DM), non-alcoholic liver disease, cardio-metabolic disease etc. [1-5]. To better understand the underlying molecular mechanisms of host-microbiota co-metabolism and their impacts on phenotypes, many experiments in humans, animals, and cells were conducted and various correlation analysis methods have been applied on the association analysis of dual-omics data sets.

Currently, the most widely used correlation methods are Pearson, Spearman, Kendall, and generalized linear regression model (GLM). Sparse correlations for compositional data (SparCC) and correlation inference for compositional data through Lasso (CCLasso) were newly developed for the intra-correlation analysis between compositional microbial variables [6-8]. Maximum information coefficient (MIC) was reported for linear/nonlinear correlation pair detection in big omics data sets [9]. More recently, Pedersen et al. introduced a computational strategy for microbiome and metabolome correlation analysis with the consideration of specified phenotype, by using the weighted gene co-expression network analysis (WGCNA) for dimension reduction and using the leave one out method for key variable identification [10, 11]. The core idea of this strategy, dimension reduction, was in line with our previous pipeline used in a

brain-gut axis study [6]. Besides, there was an in-depth comparison of 6 typical correlation methods in the inter-correlation analysis of microbiome and metabolome [12]. An advanced method, generalized correlation analysis for metabolome and microbiome (GRaMM), was designed [13]. We have also explored the co-changes and associations of serum metabolome and gut microbiome in rats across lifespan [6, 14] and in mice under antibiotic intervention [15], high-fat intervention [16], and caloric restriction [6], using traditional and newly developed methods.

The powerful analytical tools and active studies generated tremendous correlation pairs or clusters. However, every method has its own prerequisites and the application of some methods requires programming skills. It is also challenging to identify key metabolites and microbes from huge amounts of results. These enlightened us to develop a multi-function and easy-to-use platform for researchers who are unfamiliar to cross-omics correlation analysis and computational tools. The metabolome-microbiome-metadata correlation analysis platform (3MCor, http://3mcor.cn) is a web server for integrative correlation analysis of metabolome and microbiome under the instruction of phenotype and with the consideration of confounders. 3MCor is independent to measurement platforms. Microbial data sets from amplicon sequencing (16S, 18S, ITS) or shotgun metagenomic sequencing and metabolomic data sets from mass spectrometry or NMR are all acceptable. Intra- and inter- correlation analysis are both supported. Various methods for data pretreatment, module filtering, univariate and multivariate correlation analysis, confounder effect correction, and network and topology analysis were integrated and 4 pipelines for intra-correlation and global, hierarchical, and pairwise inter-correlation correlation analysis were constructed. Comprehensive results will be generated in minutes, including the overall correlation of two omics matrixes, the correlation type, strength and significance between metabolites, microbes, pathways, and/or microbial functions, the topological indices and hub nodes of correlation network and rich visual figures. The resulting data files can be inputted directly into many existing omics data analysis tools.

Implementation

Framework of 3MCor

The 3MCor is composed of four main steps (Figure 1): (1) omics and metadata importation, (2) data preprocessing, (3) inter-correlation and intra-correlation (including global, hierarchical, and pairwise correlation pipelines) analysis, and (4) results visualization.

Data importation and preprocessing

Omics and metadata importation

A total of 4 data files (in .txt, .csv. or .xlsx format), including 2 omics data files and 2 metadata files, are acceptable. At least 1 omics data file is required and the others are optional. Only intra-correlation analysis will be conducted in case there is only 1 omics data file imported. Both intra- and inter-correlation analysis can be selected when two omics data sets are imported. For the metabolome data

set, both profiling and targeted data are supported in which the rows are samples and the columns are annotated metabolites, unknown peaks, or pathway activity score. For the microbiome data set, both amplicon sequencing (16S, 18S, ITS) and metagenome data sets are supported in which the rows are samples and the columns are raw or scaled counts/functions. Both OTUs and ASVs are supported. Two metadata files (optional), one for phenotype and one for confounding factors, are acceptable. The rows are samples and the columns are confounding factors (race, age, BMI, etc.) or the phenotype (grouping variable) of the study.

Data preprocessing

Here, missing value imputation, scaling and transformation are provided for omics data preprocessing. The KNN (k-nearest neighbors, default), mean, minimum, and QRILC (quantile regression imputation of left-censored data) [17] are provided for missing value imputation. QRILC was specifically designed for missing data caused by lower than limit of quantification and is suitable for quantitative data preprocessing. The scaling is to remove unwanted variations caused by samples and/or sequencing. Besides total intensity normalization, rarefaction is also provided for microbial counts data [18, 19]. For transformation, log transformation is provided for metabolome data set. The centered log-ratio transformation (CLR) is the default method for microbiome data, where the data in each sample is transformed by taking the logarithm of the ratio between the count value for each part and the geometric mean count [20, 21]. Users can choose some or all of these steps according to the data characteristics and their study aims.

Metabolomics pathway deregulation matrix generation

Besides metabolite levels, the levels of pathway activity or deregulation may provide more direct insights to function study. Recently, the pathway-based features have been used in some personalized health studies and specific method has been developed for pathway deregulation matrix generation [22-25]. A tool which converts raw metabolite-based data matrix to the pathway-based matrix is implemented in 3MCor. The resulting matrix has the same number of samples as the metabolite-based matrix. The pathway-based features are derived from KEGG pathway database and their levels indicate the activity or deregulation degrees of the pathway features in corresponding samples. It can be inputted into 3MCor directly as a metabolome data set.

Correlation analysis

Intra-correlation analysis (1 pipeline)

Intra-correlation is the pairwise correlation analysis among variables within single omics data set. For metabolome data set, Pearson, Spearman, Kendall, Partial Spearman, Partial Pearson, Partial Kendall, MIC, GLM and GRaMM are provided. For microbiome data set, MIC, SparCC and CCLasso are provided. Among them, MIC is designed for large data sets. It can identify both linear and nonlinear pairs but only has correlation strength without positive or negative information. GLM, all the partial correlation methods,

and GRaMM can handle confounders. SparCC and CCLasso were designed specifically for two microbes of a compositional microbiome data set [26, 27]. It was reported that Pearson and Spearman may achieve unreasonable and even wrong results for intra-correlation analysis within microbiome data set [26, 27].

Inter-correlation analysis (3 pipelines)

There are 3 complementary pipelines for inter-correlation analysis (Figure 1): (1) Global correlation, (2) Hierarchical correlation and (3) Pairwise correlation. Multiple pipelines can be selected to obtain comprehensive results.

- Level 1 - Global correlation pipeline

This is the highest level with the aim to evaluate the overall correlation between two omics data sets. Five methods, Canonical correlation analysis (CCA) [28, 29], Coinertia analysis (CIA) [30], Procrustes analysis (PA) [31], two-way orthogonal partial least squares (O2PLS) [32], sparse partial least squares (sPLS) [33] and Mantel Test [34] are provided. CCA projects multidimensional data and into variables and using linear transformation and then take the correlation coefficient of and as the correlation of and . CCA requires that variables in the dataset are linearly independent and the number of samples should not be less than the number of variables. CIA is a two-table ordination method used to measure the consistency between two datasets and can be considered as a PCA of the joint covariances ofand . PA compares the consistency of two data sets by analyzing the shape distributions of scatter plots derived from other dimension reduction and projection methods [35]. O2PLS decomposes the data sets into five parts: the joint variance of X and Y, the part of X Unique, the part of Y Unique, the residual part of X, and the residual part of Y. The Pearson correlation between the joint variance of X and Y is taken as the global correlation of the two data sets. The sPLS is similar to classical PLS while with an extra Lasso penalization for the computation of loading vectors and the Singular Value Decomposition. Thus, it is more suitable for sparse data sets. The Pearson correlation of the first principal components derived from the Predictor matrixes is taken as the global correlation of the two data sets. The Mantel Test calculates the correlation of two unfolded variables derived from the distance matrix between data sets and .

- Level 2- Hierarchical correlation pipeline

For high throughput omics data, it is time consuming to calculate correlations between every two variables. Also, the core results may be flooded in massive results. In addition, analysis without the consideration of phenotype may deviate from study aim. Hence, some hierarchical strategies with the aid of different dimension reduction methods were proposed by bioinformatician [11] and our group [6]. These strategies were summarized and simplified in 3 steps and incorporated in 3MCor. First, users need to select a dimension reduction method to get the metabolic (Meta) and microbial (Mic) modules. Three methods are provided: WGCNA, principal component analysis (PCA), and principal co-ordinates analysis (PCoA). WGCNA is designed for the genomic data processing and is popular for cooperative affect clustering of omics data sets. The number of modules and the affiliations of variables are automatically

determined. PCA, a traditional method with wide applications, is to find the most important coordinates/components by linear combinations of original variables [36]. The PCoA is similar to PCA and is based on the distance matrix but not raw data set. It is widely used in microbiome data analysis. Second, phenotype (if exist) will be used for modules filtering. For a continuous phenotype, modules that are correlated to the phenotype (Spearman $p<0.05$) will be retained. For a categorical phenotype, Mann-Whitney or Kruskal-Wallis test will be used to screen out differential ($p<0.05$) modules between/among groups. Then, the correlation between the selected metabolic and microbial modules will be calculated taking the eigengene equivalents/effective abundances of modules (for WGCNA) or the first principal component (for PCA and PCOA) as the representative variable of each module. Many univariate correlation methods mentioned above can be selected in this step. Furthermore, the importance of variables in each module are ranked by the within-module connectivity that is determined by summing its connectivity to all other variables in the given module (for WGCNA) or the loadings of each principal components (for PCA/PCoA).

- Level 3- Pairwise correlation pipeline

This is the lowest level and pairwise correlation between individual variables in metabolome and microbiome will be evaluated by using univariate correlation methods mentioned above. Suggestions for method selection are same as that of intra-correlation section. Besides basic pairwise correlation analysis, network and topological analysis were incorporated in this pipeline hoping to present the overall correlation pattern by a network and dozens of characteristic indices and to highlight core nodes or links among the massive correlated pairs. Users may control the number of pairs involved by adjusting the threshold of r (default $|r|>0.3$) and/or p (default $p<0.05$) values. The node size is determined by its centrality of degree (default), betweenness or closeness. Network analysis contains 3 main parts. 1) The entire network can be divided into some clusters (cooccurrence communities) via optimizing the modularity score. 2) a total of 13 topological indices (Table S5), such as average nearest neighbor degree, average path length, degree centrality, degree assortativity, betweenness centrality, closeness centrality, density, transitivity, number of vertices, number of edges, modularity, diameter, and cluster counts will be calculated. 3) A Zi-Pi plot will be generated to show the potential roles of nodes in the network (network hubs, module hubs, connectors, and peripherals) according to their Zi and Pi values. The Zi and Pi values are calculated from the within-cluster and among-cluster connectivities (eq 1 and eq 2 of SI) [37].

Two types of phenotypes are acceptable. For a continuous phenotype (e.g. age, BMI, glucose level), all the samples will be analyzed once. For a categorical phenotype (e.g. case vs. control), the selected analysis will conducted multiple times, using samples of each group and all groups respectively. Further results comparison contains the number of all ($p<0.05$) and strong ($|r|>0.5$ and $p<0.05$) correlated pairs, the distribution of r/p values, the consistency of correlated pairs (a Venn diagram), the network and topological indices, and so on.

Development environment and configuration

Its back-end is primarily based on PHP and shell and font-end interactive page is based on HTML, CSS and Javascript. All of the statistical analysis and visualizations were implemented using R in a Docker container. Major R packages are: igraph package for network analysis, gramm4R for GRaMM, ggplot2 for figures, WGCNA package for WGCNA algorithm, Lilikoi and Pathifier for pathway deregulation matrix generation [22, 24]. The entire system is deployed on a Google Cloud server with 32GB of RAM and eight virtual CPUs with 2.6 GHz each. The web platform has been tested with Mozilla Firefox, Chrome, Opera and Safari browsers in Windows 7/10, Linux and MacOS systems. To date, 3MCor has been running for over 6 months and was tested by dozens of human and animal data sets derived from different types of samples and detection platforms.

Tutorial documentation

To ensure that the users are able to understand all the available parameters for corresponding pipelines and results interpretation, interactive help information, a detailed user tutorial page and three demo data sets are provided on line. The backend R script is available here .

Results And Discussion

To verify the feasibility and validity of 3MCor, we compared it with existing tools, which cover the correlation analysis in omics data. Furthermore, 2 real-world data sets were used to demonstrate its simple and flexible operation, comprehensive outputs, and distinctive contribution to dual-omics studies.

Function comparisons with existing tools

Here, we compared it with the correlation analysis modules of several widely used omics data analysis platforms (Table 1). 3MCor is more powerful with comprehensive methods and pipelines compared with other tools. As known, MetagenoNets focus on network analysis but supports only microbiome data set, so that it cannot carry out inter-correlation analysis. MaAsLin2 has many data preprocessing and univariate correlation methods, but it does not contain network analysis function and no graphical interface is provided. Especially, these two platforms cannot conduct global and hierarchical correlation analysis and cannot identify nonlinear correlation pairs. 3MCor is comparable to M2IA while with more options in most correlation analysis relevant items. Besides the tools included in Table 1, there are other integrative omics data analysis platforms such as MetaboAnalyst [38, 39], MicrobiomeAnalyst [20, 40], W4M [41, 42], IP4M [43], Metabox [44], and so on. Although correlation analysis is only a small part of their overall framework, some basic correlation analysis (Spearman, Pearson, Kendall, SparCC, partial correlation) methods and/or network drawing functions are provided.

Case study 1— colonization and succession of gut microbiome and metabolome in newborns

Background and data

Initial microbial colonization and later succession in the gut of human infants are linked to health and disease later in life. The gut microbiome and metabolome in 88 African-American newborns were

investigated using fecal samples collected in the first few days (from 2 min to 176 h after birth) of life [45]. The metagenomic sequences (derived from Illumina HiSeq) were downloaded from NCBI (SRP217052). The inhouse pipeline Microbiome Automated Analysis Workflows (MAAWf, http://www.maawf.com/) [46] was used for taxonomic and functional profiles generation. The metabolome data (derived from LC-MS) were provided by the authors. The 276 microbial functions and 86 metabolites of 50 infants, were imported into 3MCor. The sampling time (hour) after birth was taken as a continuous phenotype. The mother's BMI was taken as a confounding factor.

Interpretation of output

We first evaluated the global correlation (Level 1) of the two data sets. Consistently, moderate similarities were revealed by CCA, PA, Mantel test, PA and O2PLS. The joint PC1 scatter plot of O2PLS with highest correlation strength ($r=0.519$, $p<0.05$) was shown as Figure 2a. Then, for the hierarchical pipeline (Level 2), WGCNA was selected for dimension reduction and 11 metabolic and 15 microbial modules were generated. The correlation strengths and significances of modules that are significantly related to phenotype were shown as Figure 2b. The metabolic module Meta 5 (in red) was significantly and positively correlated with most microbial modules (Figures 2b-2c). The importance (loadings) of all the metabolites affiliated to Meta5 were calculated (Figure 2d), and succinate and methylmalonic acid were the top 2 metabolites with highest contributions to Meta 5. The Sankey plot (Figure 2e) was a good summary of hierarchical pipeline with the selected correlation modules, the important metabolites and functions (only top 3 were shown), and their affiliations.

The pairwise pipeline (Level 3) was also conducted. A series networks of different scales were generated by setting different r and p thresholds. Two typical ones with the thresholds of $|r|>0.2$ & $p<0.1$ and $|r|>0.55$ & $p<0.05$ were shown as Figures 2f and 2g respectively. For network 2g, 3 clusters were identified based on its topological structure. The phenotype had high connectivity in cluster 2 and the outstanding metabolites with high connectivity in cluster 3 were succinate and methylmalonic acid. Many topological indices were calculated and as expected, the Zi-Pi analysis (Figure 2h) indicated that these two metabolites were potential keystone nodes. There are many microbial functions closely correlated with them, such as branched amino acid biosynthesis ($r=0.39$, $p<0.01$ with succinate and $r=0.38$, $p<0.05$ with methylmalonic acid), aromatic amino acid biosynthesis ($r=0.44$, $p<0.05$ with succinate and $r=0.44$, $p<0.01$ with methylmalonic acid), fatty acid biosynthesis initiation ($r=0.51$, $p<0.01$ with succinate and $r=0.51$, $p<0.05$ with methylmalonic acid) etc. We further conducted pairwise correlation on microbial species and these two metabolites using the GRaMM method of 3MCor. The most prominent microbe correlated with them is *E. coli* ($r=0.45$, $p<0.05$ with succinate and $r=0.48$, $p<0.05$ with methylmalonic acid). The associations of *E. coli*, succinate, and aforementioned functions were highly consistent with the results of original report and many other studies [45, 47, 48]. The methylmalonic acid was also identified as key node by 3MCor, implying its important roles in very early life. It is well recognized that methylmalonic acid is the upstream metabolite of succinate and they are involved in vitamin homeostasis and TCA cycle [49].

Case study 2—distinct microbiota-bile acid association pattern in type 2 diabetics

Background and data

Bile acid (BA) profiles are closely associated with T2DM and the metabolism of BA are regulated by gut microbiota. Data sets of this case study were from our reports on BA and T2DM [50]. The study was approved by the Ethics Committee of Shanghai Jiao Tong University Affiliated Sixth People's Hospital. Written informed consent was obtained from all participants before recruitment. Please see the original reports for detailed information on diagnose and in/exclusion criteria, sample collection and detection, and data preprocessing. A total of 23 bile acids (quantified metabolome) and 50 high abundance genera (16S rRNA microbiome) derived from the fecal samples of 18 healthy controls (HC) and 29 diabetics (DM) were used here. A binary categorical variable (HC vs. DM) was taken as the phenotype and BMI was taken as a confounding factor. The entire correlation analysis was run three times automatically by 3MCor, using samples of HC, DM, and ALL (HC+DM) respectively. Some of the results were provided here and please see SI for complete parameter setting and results.

Interpretation of output

First, high overall correlations between metabolome and microbiome were observed and the correlation strength of HC (Figure 3a, $r=0.68$, $p<0.05$) is weaker than that of ALL (Figure 3b, $r=0.89$, $p<0.05$) and DM (Figure 3c, $r=0.86$, $p<0.05$) groups in PA analysis. Second, in the hierarchical pipeline, Mic 5 of HC group, Mic 6 of ALL group, and Mic 9 of DM group were significantly correlated with most Meta modules. Meta 2 of HC group, Meta 1 of ALL group, and Meta 3 of DM group were significantly correlated with most Mic modules (Figures 3d-3f). The top 2 metabolites with great contributions to each selected Meta modules included 7_ketoDCA, 7_ketoLCA, LCA, isoLCA, TBA, and CDCA (Figures 3g-3i). For microbes, *g_Bacillus*, *g_Ruminococcus*, *g_Oscillospira*, *f_Ruminococcaceae,* and *f_Rikenellaceae* were ranked as the tops (Figures 3j-3l). Using the pairwise inter-correlation pipeline, 3 networks and 3 Zi-Pi plots (Figures 3m-3r) were constructed using the same thresholds ($p<0.1$). Consistent to the results of global and hierarchical pipelines, the metabolome-microbiome correlation patterns and key nodes of HC, ALL, and DM groups were different and higher correlation strength of were observed in ALL and DM groups, compared with HC, as the numbers of nodes involved in the networks and the numbers of identified key nodes are both higher in ALL and DM groups. We counted the numbers of all ($p<0.05$) and strongly ($|r|>0.5$ and $p<0.05$) correlated pairs derived from HC, DM, and ALL groups and found that there were more strong correlation pairs in DM group although the number of total correlation pairs of DM was lower than that of ALL groups (Figure 3s). Notably, HCA (Hyocholic acid), 12-ketoCDCA, *g_Streptococcus* (SI Figure 24P), *f_Mogibacteriaceae* (SI Figure 24Q) and *g_Ruminococcus* were identified as high-contribution variables and hub nodes for DM group, in the hierarchical and pairwise pipeline respectively. Their levels in HC and DM groups were shown as Figures 3t. nor_DCA, *f_Ruminococcaceae*, *f_Mogibacteriaceae*, *g_Ruminococcus*, *g_Anaerostipes* were highlighted in both pipelines for ALL group. There was no overlapped node for HC group. Therefore, diabetes has a discernible impact on human metabolome-microbiome correlation pattern and HCA, 12-ketoCDCA, *g_Streptococcus, f_Mogibacteriaceae* and *g_Ruminococcus* are identified to be key nodes associated with diabetes. Our human, animal, and cell experiments have elucidated that HCA, one of the non-12α-hydroxylated secondary BAs, can upregulated

glucagon-like peptide-1 (GLP-1) production and secretion in enteroendocrine cells to effectively regulate blood glucose homeostasis, via simultaneously activating Takeda G-protein-coupled receptor 5 (TGR5) and inhibiting farnesoid X receptor (FXR) [51]. The 12-ketoCDCA is also a non-12α-hydroxylated BA which is proved to be involved in the occurrence and development of DM [52]. For *g_Streptococcus*, it has been reported that diabetes is the most frequently identified risk factors in invasive *Streptococcus* agalactiae (GBS) infections [53]. Recent machine learning studies have found that *f_Mogibacteriaceae* is one of the markers that can effectively predict the risk of DM [54]. *Ruminococcus*.spp was positively associated with DM in many studies [55-59].

Conclusions

To our knowledge, 3MCor is the first web server specifically for metabolome-microbiome correlation analysis with comprehensive considerations of phenotype, confounders, and the complicated relationships of omics features. It integrates more than 20 methods relevant to metabolome/microbiome correlation analysis, including 1 pipeline for intra-correlation and 3 pipelines for inter-correlation analysis. To assess the advantage of 3MCor, function comparison has been carried out for 3MCor with other typical tools. Both linear and non-linear correlations can be detected. Continuous and categorial phenotypes are acceptable. It is independent to sample type and detection platform. In addition to basic analysis on multiple levels and/or groups, network analysis may greatly facilitate the quick identification of correlation patterns, sub-nets, and hub nodes from tremendous and/or inconsistent results. The jointly usage of multiple pipelines is also a good way to get more reliable results. In the two real-world applications, 3MCor captured many metabolites and microbes associated to very early life and type 2 diabetes using the hierarchical and pairwise pipelines. Some of them are highly consistent to previous reports and some of them are brand new ones with extra insights to human health and disease. These illustrated the practical values of 3MCor. Furthermore, as a user-friendly and extensible tool, the interface and operation of 3MCor is simple, and sufficient guidance for method and parameter setting and results interpretation are provided, where the inputs and outputs of 3MCor are compatible to existing omics data analysis tools.

There are lots of work remains to be done. First, the current methods and pipelines are all data-driven. Some biological and chemical associations between two features, such as the minimum reaction steps, the structure similarity, the affiliations to the same or related pathway/class/phylum, should also be considered. We are devoting to integrate diverse biological information from public databases (e.g. KEGG, SMPDB, HMDB, Biocyc) and literatures to the pipelines which may further improve the reliability of integrative (correlation) analysis and relieve the burden of subsequent experiment validation. This is a challenging but promising task. Second, omics data sets are of similarity. 3MCor is of potential to be used for other omics (e.g. genomics, transcriptomics, and proteomics) or more than 3 omics data sets correlation analysis. Substantial function extension and performance evaluation are needed.

Availability and requirements

Project name: 3MCor

Project home page: http://3mcor.cn/

Operating system(s): Platform independent

Programming language: PHP, HTML, CSS, JS, R, Shell

Other requirements: Apache2, Docker

License: GNU GLP >2

Any restrictions to use by non-academics: licence needed

Abbreviations

T2DM: type 2 diabetes mellitus; GLM: generalized linear regression model; SparCC: Sparse correlations for compositional data; CCLasso: correlation inference for compositional data through Lasso; MIC: Maximum information coefficient; WGCNA: the weighted gene co-expression network analysis; GRaMM: generalized correlation analysis for metabolome and microbiome; KNN: k-nearest neighbors; QRILC: quantile regression imputation of left-censored data; CLR: The centered log-ratio transformation; CCA: Canonical correlation analysis; CIA: Coinertia analysis; PA: Procrustes analysis; O2PLS: two-way orthogonal partial least squares; sPLS: sparse partial least squares; Meta: metabolic; Mic: microbial; PCA: principal component analysis; PCoA: principal co-ordinates analysis; BA: Bile acid; HC: healthy controls; DM: diabetics; HCA: Hyocholic acid; GLP-1: glucagon-like peptide-1; TGR5: Takeda G-protein-coupled receptor 5; FXR: farnesoid X receptor; GBS: invasive Streptococcus agalactiae.

Declarations

1. **Ethics approval and consent to participate**

The data of case study 1 were download from public database and this study was approved by the Committee for the Protection of Human Subjects (Internal Review Board) of the Children's Hospital of Philadelphia. The case study 2 was approved by the Ethics Committee of Shanghai Jiao Tong University Affiliated Sixth People's Hospital. Written informed consent was obtained from all participants before recruitment.

2. **Consent for publication**

Not applicable.

3. **Availability of data and material**

The backend R script is available at . The raw data of case study 1 can be downloaded from NCBI (https://www.ncbi.nlm.nih.gov/sra/?term=SRP217052) and the raw data of case study 2 have been deposited and available at Metabolights repository with accession code MTBLS2343.

4. **Competing interests**

The authors declare no competing interests.

5. **Funding**

This work was supported by the National Key R&D Program of China (2019YFA0802300) and National Natural Science Foundation of China (31972935).

6. **Authors' contributions**

C. and W.J. conceptualized the study and designed the research. T.S., M.L. and D.L. coordinated the web server design, development, and test. X.Y., G.X. and C.S. performed the tool evaluation. T.S. and T.C. drafted the manuscript. All the authors critically revised the manuscript.

References

1. Jaacks L, Vandevijvere S, A P, Cj M, Imamura CW, Swinburn FDM. B, M E: **The obesity transition: stages of the global epidemic**. *Yearbook of Paediatric Endocrinology* 2019.
2. Lakka HM, Laaksonen DE, Lakka TA, Niskanen LK, Kumpusalo E, Tuomilehto J, Salonen JT. The metabolic syndrome and total and cardiovascular disease mortality in middle-aged men. Jama. 2002;288:2709–16.
3. Reddy KS, Yusuf S. Emerging epidemic of cardiovascular disease in developing countries. Circulation. 1998;97:596–601.
4. Zheng Y, Ley SH, Hu FB. Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. Nat Rev Endocrinol. 2018;14:88–98.
5. Younossi ZM, Koenig AB, Abdelatif D, Fazel Y, Henry L, Wymer M. Global epidemiology of nonalcoholic fatty liver disease-Meta-analytic assessment of prevalence, incidence, and outcomes. Hepatology. 2016;64:73–84.
6. Chen T, You Y, Xie G, Zheng X, Zhao A, Liu J, Zhao Q, Wang S, Huang F, Rajani C, et al. Strategy for an Association Study of the Intestinal Microbiome and Brain Metabolome Across the Lifespan of Rats. Anal Chem. 2018;90:2475–83.
7. Chong J, Xia J. **Computational Approaches for Integrative Analysis of the Metabolome and Microbiome**. *Metabolites* 2017, 7.

8. Dhakan DB, Maji A, Sharma AK, Saxena R, Pulikkan J, Grace T, Gomez A, Scaria J, Amato KR, Sharma VK: **The unique composition of Indian gut microbiome, gene catalogue, and associated fecal metabolome deciphered using multi-omics approaches**. *Gigascience* 2019, **8**.
9. Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, Lander ES, Mitzenmacher M, Sabeti PC. Detecting novel associations in large data sets. Science. 2011;334:1518–24.
10. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.
11. Pedersen HK, Forslund SK, Gudmundsdottir V, Petersen A, Hildebrand F, Hyötyläinen T, Nielsen T, Hansen T, Bork P, Ehrlich SD, et al. A computational framework to integrate high-throughput '-omics' datasets for the identification of potential mechanistic links. Nat Protoc. 2018;13:2781–800.
12. You Y, Liang D, Wei R, Li M, Li Y, Wang J, Wang X, Zheng X, Jia W, Chen T. Evaluation of metabolite-microbe correlation detection methods. Anal Biochem. 2019;567:106–11.
13. Liang D, Li M, Wei R, Wang J, Li Y, Jia W, Chen T. Strategy for Intercorrelation Identification between Metabolome and Microbiome. Anal Chem. 2019;91:14424–32.
14. Zhang X, Yang Y, Su J, Zheng X, Wang C, Chen S, Liu J, Lv Y, Fan S, Zhao A, et al: **Age-related compositional changes and correlations of gut microbiome, serum metabolome, and immune factor in rats**. *Geroscience* 2020.
15. Zheng X, Xie G, Zhao A, Zhao L, Yao C, Chiu NH, Zhou Z, Bao Y, Jia W, Nicholson JK, Jia W. The footprints of gut microbial-mammalian co-metabolism. J Proteome Res. 2011;10:5512–22.
16. Zheng X, Huang F, Zhao A, Lei S, Zhang Y, Xie G, Chen T, Qu C, Rajani C, Dong B, et al. Bile acid is a significant host factor shaping the gut microbiome of diet-induced obese mice. BMC Biol. 2017;15:120.
17. Wei R, Wang J, Su M, Jia E, Chen S, Chen T, Ni Y. Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. Sci Rep. 2018;8:663.
18. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome Datasets Are Compositional: And This Is Not Optional. Front Microbiol. 2017;8:2224.
19. McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput Biol. 2014;10:e1003531.
20. 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:W180-w188.
21. Nagpal S, Singh R, Yadav D, Mande SS. MetagenoNets: comprehensive inference and meta-insights for microbial correlation networks. Nucleic Acids Res. 2020;48:W572-w579.
22. Drier Y, Sheffer M, Domany E. Pathway-based personalized analysis of cancer. Proc Natl Acad Sci U S A. 2013;110:6388–93.

23. Huang S, Chong N, Lewis NE, Jia W, Xie G, Garmire LX. Novel personalized pathway-based metabolomics models reveal key metabolic pathways for breast cancer diagnosis. Genome Med. 2016;8:34.
24. Al-Akwaa FM, Yunits B, Huang S, Alhajaji H, Garmire LX. **Lilikoi: an R package for personalized pathway-based classification modeling using metabolomics data**. *Gigascience* 2018, 7.
25. Fang X, Liu Y, Ren Z, Du Y, Huang Q, Garmire LX. **Lilikoi V2.0: a deep learning-enabled, personalized pathway-based R package for diagnosis and prognosis predictions using metabolomics data**. *Gigascience* 2021, 10.
26. Fang H, Huang C, Zhao H, Deng M. CCLasso: correlation inference for compositional data through Lasso. Bioinformatics. 2015;31:3172–80.
27. Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. PLoS Comput Biol. 2012;8:e1002687.
28. Kostic AD, Gevers D, Siljander H, Vatanen T, Hyötyläinen T, Hämäläinen AM, Peet A, Tillmann V, Pöhö P, Mattila I, et al. The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. Cell Host Microbe. 2015;17:260–73.
29. 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.
30. Dolédec S, Chessel D. Co-inertia analysis: an alternative method for studying species-environment relationships. Freshw Biol. 2006;31:277–94.
31. McHardy IH, Goudarzi M, Tong M, Ruegger PM, Schwager E, Weger JR, Graeber TG, Sonnenburg JL, Horvath S, Huttenhower C, et al. Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships. Microbiome. 2013;1:17.
32. Bylesjö 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;52:1181–91.
33. KA LC, Rossouw D, Robert-Granié C, Besse P. A sparse PLS for variable selection when integrating omics data. Stat Appl Genet Mol Biol. 2008;7:Article 35.
34. Mantel N. The detection of disease clustering and a generalized regression approach. Cancer Res. 1967;27:209–20.
35. Quinn RA, Navas-Molina JA, Hyde ER, Song SJ, Vázquez-Baeza Y, Humphrey G, Gaffney J, Minich JJ, Melnik AV, Herschend J, et al: **From Sample to Multi-Omics Conclusions in under 48 Hours**. *mSystems* 2016, 1.
36. Wold S, Esbensen K, Geladi P. Principal Component Analysis. Chemometr Intell Lab Syst. 1987;2:37–52.
37. Cumbo F, Paci P, Santoni D, Di Paola L, Giuliani A. GIANT: a cytoscape plugin for modular networks. PLoS One. 2014;9:e105001.

38. Chong J, Soufan O, Li C, Caraus I, Li S, Bourque G, Wishart DS, Xia J. MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. Nucleic Acids Res. 2018;46:W486-w494.
39. Chong J, Wishart DS, Xia J. Using MetaboAnalyst 4.0 for Comprehensive and Integrative Metabolomics Data Analysis. Curr Protoc Bioinformatics. 2019;68:e86.
40. Chong J, Liu P, Zhou G, Xia J. Using MicrobiomeAnalyst for comprehensive statistical, functional, and meta-analysis of microbiome data. Nat Protoc. 2020;15:799–821.
41. Giacomoni F, Le Corguillé G, Monsoor M, Landi M, Pericard P, Pétéra M, Duperier C, Tremblay-Franco M, Martin JF, Jacob D, et al. Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics. Bioinformatics. 2015;31:1493–5.
42. Guitton Y, Tremblay-Franco M, Le Corguillé G, Martin JF, Pétéra M, Roger-Mele P, Delabrière A, Goulitquer S, Monsoor M, Duperier C, et al. Create, run, share, publish, and reference your LC-MS, FIA-MS, GC-MS, and NMR data analysis workflows with the Workflow4Metabolomics 3.0 Galaxy online infrastructure for metabolomics. Int J Biochem Cell Biol. 2017;93:89–101.
43. Liang D, Liu Q, Zhou K, Jia W, Xie G, Chen T. IP4M: an integrated platform for mass spectrometry-based metabolomics data mining. BMC Bioinformatics. 2020;21:444.
44. Wanichthanarak K, Fan S, Grapov D, Barupal DK, Fiehn O. Metabox: A Toolbox for Metabolomic Data Analysis, Interpretation and Integrative Exploration. PLoS One. 2017;12:e0171046.
45. Bittinger K, Zhao C, Li Y, Ford E, Friedman ES, Ni J, Kulkarni CV, Cai J, Tian Y, Liu Q, et al. Bacterial colonization reprograms the neonatal gut metabolome. Nat Microbiol. 2020;5:838–47.
46. Zhu S, Sun T, Zhu C, Qing T, Jiang Y, Ding R, Su H, Sun Y, Xu X, Xu K, et al: *MAAWf: An Integrated and Visual Tool for Microbiome Data Analyses*. 2020.
47. Sánchez AM, Bennett GN, San KY. Novel pathway engineering design of the anaerobic central metabolic pathway in Escherichia coli to increase succinate yield and productivity. Metab Eng. 2005;7:229–39.
48. Thakker C, Martínez I, San KY, Bennett GN. Succinate production in Escherichia coli. Biotechnol J. 2012;7:213–24.
49. Elin RJ, Winter WE. Methylmalonic acid: a test whose time has come? Arch Pathol Lab Med. 2001;125:824–7.
50. Zheng X, Chen T, Zhao A, Ning Z, Kuang J, Wang S, You Y, Bao Y, Ma X, Yu H, et al. Hyocholic acid species as novel biomarkers for metabolic disorders. Nat Commun. 2021;12:1487.
51. Zheng X, Chen T, Jiang R, Zhao A, Wu Q, Kuang J, Sun D, Ren Z, Li M, Zhao M, et al: **Hyocholic acid species improve glucose homeostasis through a distinct TGR5 and FXR signaling mechanism**. *Cell Metab* 2020.
52. Jia W, Wei M, Rajani C, Zheng X. **Targeting the alternative bile acid synthetic pathway for metabolic diseases**. *Protein Cell* 2020.
53. Graux E, Hites M, Martiny D, Maillart E, Delforge M, Melin P, Dauby N. **Invasive group B Streptococcus among non-pregnant adults in Brussels-Capital Region, 2005–2019**. *Eur J Clin Microbiol Infect Dis*

2020:1–9.

54. Gou W, Ling CW, He Y, Jiang Z, Fu Y, Xu F, Miao Z, Sun TY, Lin JS, Zhu HL, et al: **Interpretable Machine Learning Framework Reveals Robust Gut Microbiome Features Associated With Type 2 Diabetes**. *Diabetes Care* 2020.
55. Allin KH, Tremaroli V, Caesar R, Jensen BAH, Damgaard MTF, Bahl MI, Licht TR, Hansen TH, Nielsen T, Dantoft TM, et al. Aberrant intestinal microbiota in individuals with prediabetes. Diabetologia. 2018;61:810–20.
56. Candela M, Biagi E, Soverini M, Consolandi C, Quercia S, Severgnini M, Peano C, Turroni S, Rampelli S, Pozzilli P, et al. Modulation of gut microbiota dysbioses in type 2 diabetic patients by macrobiotic Ma-Pi 2 diet. Br J Nutr. 2016;116:80–93.
57. Patrone V, Vajana E, Minuti A, Callegari ML, Federico A, Loguercio C, Dallio M, Tolone S, Docimo L, Morelli L. Postoperative Changes in Fecal Bacterial Communities and Fermentation Products in Obese Patients Undergoing Bilio-Intestinal Bypass. Front Microbiol. 2016;7:200.
58. Salamon D, Sroka-Oleksiak A, Kapusta P, Szopa M, Mrozińska S, Ludwig-Słomczyńska AH, Wołkow PP, Bulanda M, Klupa T, Małecki MT, Gosiewski T. Characteristics of gut microbiota in adult patients with type 1 and type 2 diabetes based on next–generation sequencing of the 16S rRNA gene fragment. Pol Arch Intern Med. 2018;128:336–43.
59. Zhang X, Shen D, Fang Z, Jie Z, Qiu X, Zhang C, Chen Y, Ji L. Human gut microbiota changes reveal the progression of glucose intolerance. PLoS One. 2013;8:e71108.

Figures

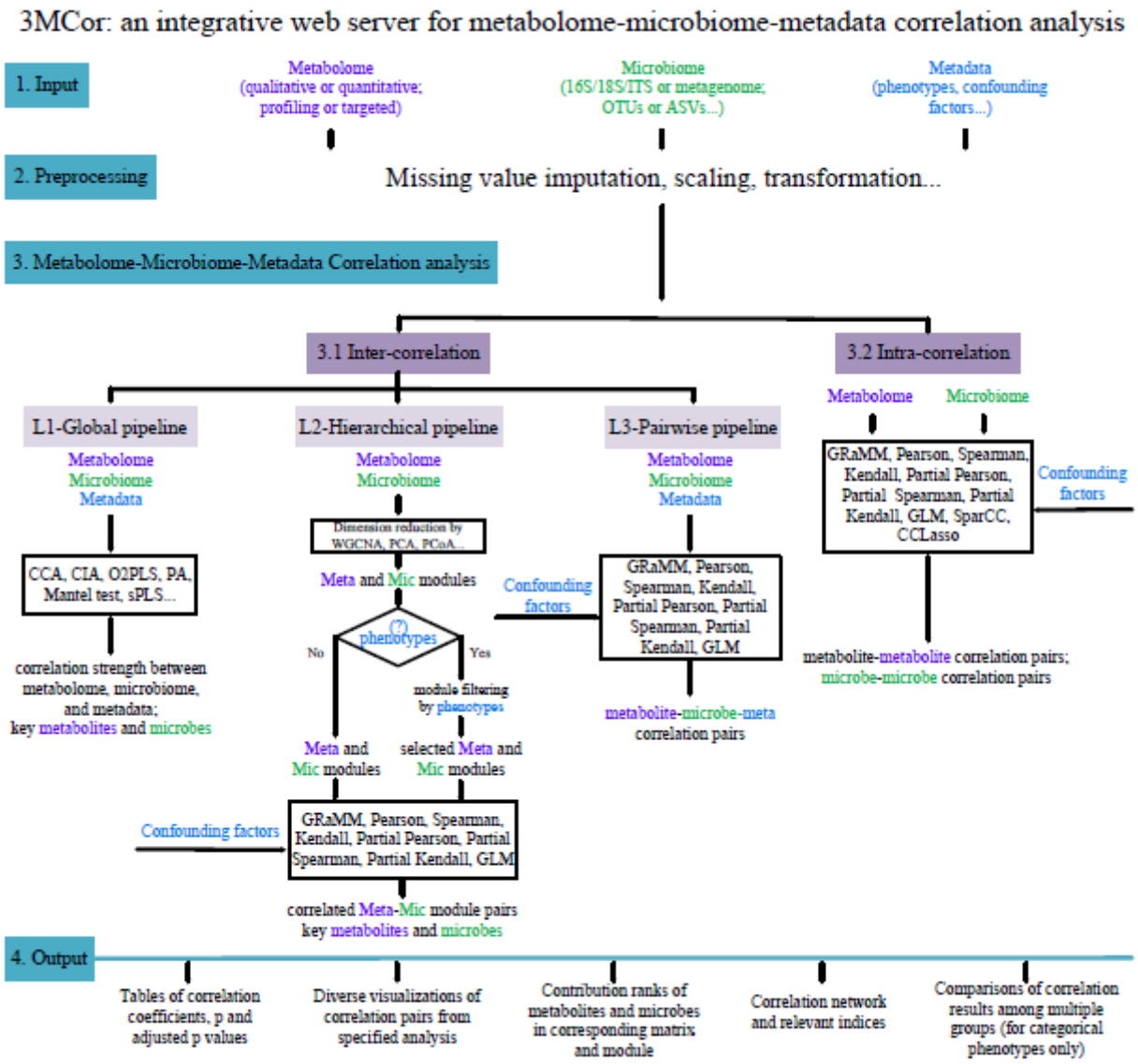

Figure 1

The flowchart of 3MCor, including the data input, data preprocessing, correlation analysis and results output step

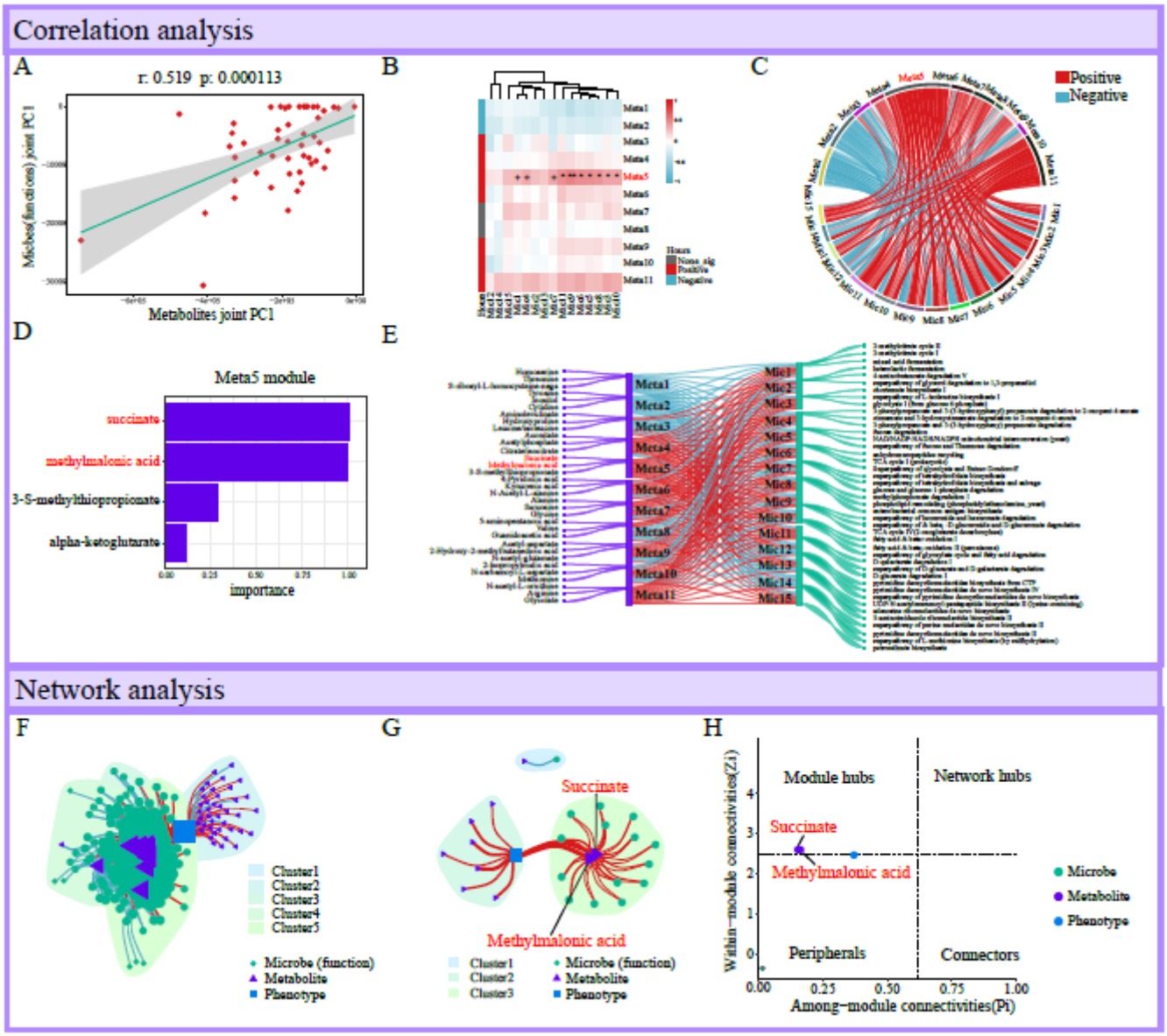

Figure 2

Visual output of case study 1 (A) The O2PLS scatter diagram. The x and y axis is the first principal component (PC1) of the joint part of metabolome and microbiome respectively. r and p is the correlation coefficient and significance between the 2 principal components. (B) Heatmap of the correlation coefficients between metabolic and microbial modules and the confounder (hours to sample collection) by partial spearman (red: positive; blue: negative; gray: not significant; +: p<0.1; *: p<0.05; **: p<0.01). (C) Circos diagram of the correlation between metabolic and microbial modules by partial spearman. red: positive; blue: negative. (D) The rank of importance/contribution of metabolites in Meta 5 to the module. (E) Sankey plot with the selected modules, top 3 metabolic and microbial features, and their affiliations (red: positive; blue: negative). (F) A network with pairs selected by |r|>0.2 and p<0.1. Each node represents a metabolite, a microbial function, or the phenotype. Its size is determined by the centrality. The edge size is determined by the absolute value of spearman correlation coefficient. Edge color in red: positive, in blue: negative; The entire network is divided into several sub-nets/clusters. (G) A network with correlated

pairs selected by |r|>0.55 and p<0.05. (H) A Zi-Pi plot. Nodes are divided into four types as peripherals, connectors, module hubs, and network hubs based on the Zi (default=2.5) and Pi (default=0.62) values.

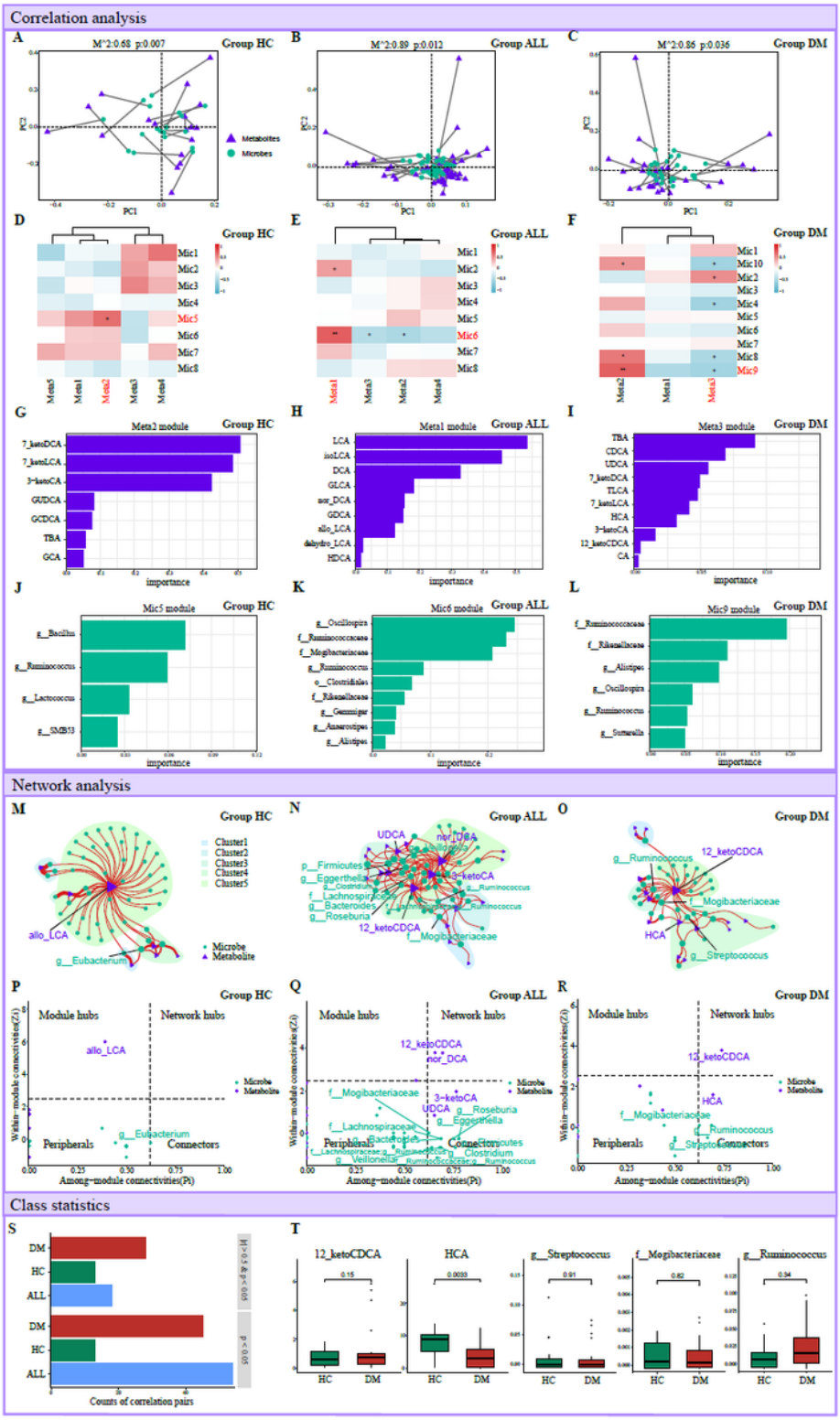

Figure 3

Visual output of case study 2 (A-C) PA scatter diagram of healthy control (HC), diabetes (DM) and all samples (ALL). The x and y axis is the first two principal components (PC1 and PC2) derived from the metabolome (purple) and microbiome (green). The same samples are connected by lines. The line

distance represents the degree of similarity. M^2 and p represents the correlation coefficient and significance of metabolome and microbiome. (D-F) Three Heatmaps of the partial spearman (BMI is the confounder) correlation coefficients between metabolic and microbial modules in HC, ALL, and DM groups respectively (red: positive; blue: negative; gray: not significant; +: p<0.1; *: p< 0.05; **: p<0.01). (G-I) The rank of importance/contribution of metabolites in Meta 2 module of HC, Meta 1 module of DM, and Meta 3 module of ALL. (J-L) The rank of importance/contribution of microbes in Mic 5 module of HC, Mic 6 module of DM and Mic 9 module of ALL. (M-O) Three networks with pairs selected by p<0.1. Each node represents a metabolite and a microbe. Its size is determined by the centrality. The edge size is determined by the absolute value of GRaMM correlation coefficient. The entire network is divided into several sub-nets/clusters. (P-R) Three Zi-Pi plots of HC, ALL, and DM groups. Nodes are divided into four types as peripherals, connectors, module hubs, and network hubs based on the Zi (default=2.5) and Pi (default=0.62) values. (S) A bar plot of counts of correlation pairs in HC, ALL, and DM groups within p<0.05 and |r|>0.5& p<0.05. (T) Box plots of levels of HCA, 12-ketoCDCA, g_Streptococcus, f_Mogibacteriaceae and g_Ruminococcus in in HC and DM groups.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table1.pdf
- 3MCorSI.pdf