Statistics in Biosciences
https://doi.org/10.1007/s12561-024-09450-9

ORIGINAL PAPER

MultiCOP: An Association Analysis of Microbiome-Metabolome Relationships

Zhen Wang[1] · Luyang Fang[1] · Jiazhang Cai[1] · Ping Ma[1] · Wenxuan Zhong[1]

Received: 29 September 2023 / Revised: 6 May 2024 / Accepted: 27 July 2024

Abstract

The connection between human health and the microbiome, including the potential risk of diseases at a metabolic level, is well established. However, comprehending the precise mechanism that underlies this relationship remains unclear due to the analysis challenge caused by the vast amount of data involved and the intricate interactions among them. We propose the multivariate correlation pursuit (MultiCOP) algorithm, which effectively integrates microbiome and metabolome data to uncover microbe-metabolite interactions and find relevant microbes/metabolites by applying correlation pursuit and random projection. The use of correlation search and random projection in the MultiCOP algorithm enables it to surpass the constraints of other methods. Unlike its counterparts, MultiCOP does not rely on assumptions about the relationship, such as linearity, between the two datasets. Additionally, it efficiently handles multivariate data. We conducted extensive simulations to assess the performance of MultiCOP. Additionally, we employed the proposed method to explore microbe-metabolite interactions in patients with inflammatory bowel disease and those with chronic ischemic heart disease, separately. The source code is available at: https://github.com/Luyang8991/MultiCOP.

Keywords Microbiome · Metabolome · Dimension reduction · Feature selection · IBD · CIHD

Zhen Wang and Luyang Fang have contributed equally to this work.

✉ Wenxuan Zhong
wenxuan@uga.edu

[1] Department of Statistics, University of Georgia, Athens, GA 30602, USA

Published online: 22 August 2024

1 Introduction

Recent studies have revealed that the microbiome and metabolome are closely interconnected in human [1–3]. Specifically, the microbiome can produce a wide range of metabolites, including short-chain fatty acids, neurotransmitters, and vitamins, which can impact numerous physiological processes [4]. In addition, the composition of the microbiome can influence the types and levels of metabolites produced by host cells, which can further affect overall health [5, 6]. On the contrary, host-produced metabolites, such as bile acids and glucose, can affect the composition and function of the microbiome [1, 7]. The interactions between the microbiome and the metabolome have been linked to numerous diseases, including obesity [8, 9], diabetes [8], inflammatory bowel disease [3], and cancers [10]. This understanding has led to the development of numerous diseases' preventive and treatment approaches such as probiotics, prebiotics, and dietary modifications aimed at altering the microbiome and metabolite profiles [11–14]. Additionally, interventions targeting specific microbial metabolites or involving fecal microbiome transplantation have shown promise under certain conditions [13]. These strategies have the potential to offer more targeted and effective treatments for a variety of diseases in the future.

An association study of microbiome and metabolome samples is essential to comprehensively understand the interactions between their underlying systems. Popular biotechnologies for processing microbiome samples are high-throughput sequencing technologies. Using these technologies, researchers collect microbiome samples using specialized kits, followed by DNA extraction from the samples. Next-generation sequencing techniques such as 16 S rRNA gene sequencing or shotgun metagenomics are used to analyze microbial DNA. Bioinformatics tools are then utilized to process and analyze the sequencing data, including identifying microbial taxa, assessing diversity, and exploring functional profiles. For metabolome samples, such as blood, urine, or tissues, they are collected and prepared by removing cellular debris. Metabolites are then extracted from the prepared samples using techniques such as liquid-liquid extraction or solid-phase extraction. Metabolomics analysis techniques, such as mass spectrometry (MS) [15] or nuclear magnetic resonance (NMR) [16] spectroscopy, are employed to identify and quantify the metabolites. The resulting data undergoes analysis using bioinformatics tools and statistical methods to identify significant changes or associations in the metabolome.

Various statistical methods have been utilized to study the association of microbiome and metabolome data to comprehensively understand the interactions between these two systems. The simplest ones are correlation-based approaches, such as Spearman correlation [17], canonical correlation analysis (CCA) [18], partial least square (PLS) [19], and their extensions [20, 21]. For instance, Theriot et al. [22] conducted a Spearman correlation analysis to explore potential associations between the mouse gut microbiome and metabolome, specifically focusing on relationships between pairs of metabolites and operational taxonomic units (OTUs). Kostic et al. [18] used a sparse CCA to integrate the gut microbiome and the gut metabolome of infants predisposed to type 1 diabetes (T1D). However, these approaches can only identify the optimal linear combination of the two datasets by maximizing the correlation between components (in CCA) or the covariance between features (in PLS) and are therefore not capable of revealing non-linear relationships. MelonnPan [23] employs Elastic Net linear regression to model the relative abundance of each metabolite using microbial features but has the same limitation as correlation-based methods. To overcome this limitation, Morton et al. [24] proposed a neural network framework to model non-linear relationships by estimating the conditional probability that each molecule is present, given the presence of a specific microorganism. However, this approach is limited to exploring interactions between an individual metabolite and microbes.

To address the aforementioned challenges, we present a novel computational framework, called MultiCOP, which utilizes microbiome and metabolome data to discover microbiome-metabolome association in a data-driven manner. To overcome the limitation of only revealing linear relationships, we utilize the correlation pursuit (COP) algorithm [25], which performs dimension reduction and variable selection without assuming a specific relationship between the two data sets. To handle the multivariate metabolome data, we propose a two-stage estimation procedure. In the first stage, we use the random projection approach to break down the problem of estimating microbe-metabolite interactions into a series of sub-problems, with each subproblem finding the microbes related to the univariate projected metabolome data using the COP algorithm. The final estimate of the relevant microbes is achieved using the majority vote based on the results of all sub-problems. In the second stage, we perform similar procedures to select the metabolites related to the selected microbes from the first stage. One additional challenge in the association analysis of microbiome and metabolome is the high computational cost caused by the vast number of microbes and metabolites. Our MultiCOP algorithm can efficiently reduce the computational cost. For each subproblem, the COP algorithm addresses the high computational cost due to the vast number of predictors by using a stepwise algorithm for selecting variables. Moreover, empirical results show that it is sufficient to take $O(n)$ iterations of random projection to obtain excellent performance, where n represents the sample size.

2 Methods

To investigate the association between the microbiome (denoted as $\mathbf{X}$) and metabolome (denoted as $\mathbf{Y}$) data, we propose the multivariate correlation pursuit (MultiCOP) algorithm for variable selection in the sufficient dimension reduction (SDR) framework. MultiCOP comprises two stages. In the first stage, we focus on dimension reduction and variable selection for the microbiome data. To achieve this, we use random projection to break down the problem into a series of sub-problems, each of which identifies the microbes related to the projected metabolome data $\mathbf{v}'\mathbf{Y}$ along direction $\mathbf{v}$ using the correlation pursuit (COP) algorithm proposed in Zhong et al. [25]. We repeat the above procedure along multiple randomly chosen projection directions and aggregate the selected microbes along each direction by majority vote to obtain the final set of selected microbes. In the second stage, we perform similar procedures to the first stage but on the metabolome data to select the metabolites related to the selected microbes from the first stage.

2.1 Problem Setup

Let $\mathbf{X} = (X_1, \dots, X_p)' \in \mathbb{R}^p$ represent the p dimensional microbiome data with p microbes and $\mathbf{Y} = (Y_1, \dots, Y_p)' \in \mathbb{R}^q$ represent the q-dimensional metabolome data with q metabolites. In practice, the number of microbes p and metabolites q are usually large, which makes it necessary to reduce the dimension of $\mathbf{X}$ and $\mathbf{Y}$ in order to improve the effectiveness of modeling the relationship between metabolite data $\mathbf{Y}$ and microbiome data $\mathbf{X}$.

To reduce the dimension of microbiome data $\mathbf{X}$, we apply the SDR technique, which aims to reduce the dimension of predictors $\mathbf{X}$ while preserving its regression relation with $\mathbf{Y}$. Mathematically, SDR seeks $\mathbf{B} \in \mathbb{R}^{p\times K}$ with the smallest possible column space such that

$$\mathbf{Y} \perp \mathbf{X} \mid \mathbf{B}'\mathbf{X}. \tag{1}$$

The column space of $\mathbf{B}$, or span$\{\mathbf{B}\}$, is known as the central space and is denoted as $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$. Denote $\mathbf{B} = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_K)$ with $\boldsymbol{\beta}_i = (\beta_{1i}, \cdots, \beta_{pi})'$. A regression form of (1) for continuous response $\mathbf{Y}$ is

$$\mathbf{Y} = \mathbf{f}\big(\boldsymbol{\beta}_1'\mathbf{X}, \dots, \boldsymbol{\beta}_K'\mathbf{X}, \epsilon\big), \tag{2}$$

where $\boldsymbol{\epsilon}$ is r-dimensional random error independent of $\mathbf{X}$ (with $r \geq 1$), and $\mathbf{f} : \mathbb{R}^{K+r} \mapsto \mathbb{R}^q$ is an unknown link function. Multivariate response SDR focuses on estimating $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K$ without necessarily estimating the link function $\boldsymbol{f}$. We refer to $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_K$ as the SDR directions. A predictor variable $X_j (1 \leq j \leq p)$ is considered relevant if there is at least one i $(1 \leq i \leq K)$ such that $\beta_{ji} \neq 0$. Let $\mathbf{X}_{\mathcal{A}^*}$ be the set of relevant predictor variables and L be the number of relevant predictor variables. In the case of metabolome data, where p can be quite large, it is reasonable to assume sparsity, implying that only a small subset of the predictors influences $\mathbf{Y}$. Under the SDR model, this assumption is equivalent to the fact that both K and L are much smaller than p.

To reduce the dimension of metabolome data $\mathbf{Y}$, we again apply the SDR technique that seeks a set of linear combinations of $\mathbf{Y}$, say $\mathbf{H}'\mathbf{Y}$, such that $\mathbf{X}_{\mathcal{A}^*}$ depends on $\mathbf{Y}$ only through $\mathbf{H}'\mathbf{Y}$, i.e., $\mathbf{X}_{\mathcal{A}^*} \perp \mathbf{Y} \mid \mathbf{H}'\mathbf{Y}$, where $\mathbf{H} = (\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_{K'})$ with $\boldsymbol{\gamma}_i = (\gamma_{1i}, \cdots, \gamma_{qi})'$. Let $\mathbf{Y}_{\mathcal{B}^*}$ be the set of relevant predictor variables and L' be the number of relevant predictor variables, we can safely assume that both K' and L' are much smaller than q.

2.2 Random Projection

Since both p and q are large, we need a large number of observations, which is impractical to obtain, to effectively solve the problem in (1). Notice that (1) is equivalent to $\mathbf{v}'\mathbf{Y} \perp \mathbf{X} \mid \mathbf{B}'\mathbf{X}$ for $\forall \mathbf{v} \in \mathbb{R}^q$. Thus, it is natural to transfer the multivariate SDR problem into a bunch of univariate SDR sub-problems with responses $\mathbf{v}_j'\mathbf{Y}$, $j = 1, \cdots m$, and then ensemble the results of all the sub-problems. Given independent observations $\left\{\left(\mathbf{x}_i, \mathbf{y}_i\right)\right\}_{i=1,\dots,n}$ of $(\mathbf{X}, \mathbf{Y})$, where $\mathbf{x}_i = \left(x_{i1}, \dots, x_{ip}\right)'$ and $\mathbf{y}_i = \left(y_{i1}, \dots, y_{iq}\right)'$. We project $\mathbf{Y}$ along randomly sampled directions, say $\mathbf{v}_1, \dots, \mathbf{v}_m \in \mathbb{R}^q$, to obtain m samples of scalar-valued data

$$\left\{\left(\mathbf{x}_1, \mathbf{v}_j'\mathbf{y}_1\right), \dots, \left(\mathbf{x}_n, \mathbf{v}_j'\mathbf{y}_n\right)\right\}, \quad j = 1, \dots, m. \tag{3}$$

That is, observations of $(\mathbf{X}, \mathbf{v}_j'\mathbf{Y})$, $j = 1, \cdots m$. For each subproblem, we aim to solve the univariate SDR problem that seeks $\mathbf{B} \in \mathbb{R}^{p\times K}$ with the smallest possible column space such that

$$Z \perp \mathbf{X} \mid \mathbf{B}'\mathbf{X} \tag{4}$$

with $Z = \mathbf{v}'\mathbf{Y}$, and the corresponding relevant variables.

2.3 Correlation Pursuit for Variable Selection

The estimation of $\mathbf{B} = (\beta_1, \cdots, \beta_K)$, i.e., the SDR directions, in the univariate SDR problem (4) is equivalent to the solutions of the eigenvalue decomposition problem [26]

$$\begin{gathered} Mv_i = \lambda_i \Sigma v_i, \quad v_i'\Sigma v_i = 1, \qquad \text{for } i = 1, 2, \dots, K \\ \lambda_1 \geqslant \lambda_2 \geqslant \dots \geqslant \lambda_K > 0. \end{gathered} \tag{5}$$

where $\Sigma \triangleq \text{var}(\mathbf{X})$, and $M \triangleq \text{var}\{E(\mathbf{X} \mid Z)\}$ is the covariance matrix of the expectation of $\mathbf{X}$ given Z. The SDR directions $\eta_1, \eta_2, \cdots \eta_K$ are the first K eigenvectors of $\Sigma^{-1}M$ and the central space $\mathcal{S}_{Z|\mathbf{X}} = \text{span}\left(\eta_1, \eta_2, \dots, \eta_K\right)$. Here $\lambda_1, \cdots, \lambda_K$ are equivalent to the K squared profile correlations defined as $P^2(\eta_i) = \frac{\eta_i'\text{var}\{E(\mathbf{X}|Z)\}\eta_i}{\eta_i'\Sigma\eta_i} \equiv \frac{\eta_i' M\eta_i}{\eta_i'\Sigma\eta_i}$.

Given independent observations $\left\{\left(\mathbf{x}_i, z_i\right)\right\}_{i=1,\dots,n}$ of $(\mathbf{X}, Z)$, where $\mathbf{x}_i = \left(x_{i1}, \dots, x_{ip}\right)'$, sliced inverse regression (SIR), a popular method proposed by Li [27], applies the following procedure to estimate M. First, the range of $\left\{z_i\right\}_{i=1}^n$ is divided into H disjoint intervals, which are denoted as $S_1, \dots, S_H$. For each interval S_h, the mean vector is calculated by $\bar{\mathbf{x}}_h = n_h^{-1}\Sigma_{z_i \in S_h}\mathbf{x}_i$, where n_h is the sample size in S_h. Using these mean vectors, M is estimated by $\hat{M} = \frac{\sum_{h=1}^{H} n_h(\bar{\mathbf{x}}_h - \bar{\mathbf{x}})(\bar{\mathbf{x}}_h - \bar{\mathbf{x}})'}{n}$. Then, the matrix $\Sigma^{-1}M$ is estimated by $\hat{\Sigma}^{-1}\hat{M}$, where $\hat{\Sigma}$ is the sample covariance matrix. In this way, the principal directions can be estimated accordingly.

However, since only a small number of microbes/metabolites are relevant to the microbiome-metabolome interactions, estimating $p \times p$ covariance matrices Σ, M, and the eigenvalue decomposition of $\hat{\Sigma}^{-1}\hat{M}$ directly will be very inaccurate and computationally heavy due to a large number of irrelevant microbes/metabolites, leading to inaccurate estimates of SDR directions $\hat{\eta}_1, \hat{\eta}_2, \dots, \hat{\eta}_K$ [28]. To address this issue, we apply a stepwise SIR-based method called correlation pursuit (COP) [25]. COP begins with a set of randomly selected predictors and alternates between an addition step, which selects and adds a predictor to the set, and a deletion step, which selects and removes a predictor from the set. The algorithm stops when no new predictor can be added or deleted.

Addition step. Let $\mathcal{A}$ be the set of indices of the currently selected predictors and $X_{\mathcal{A}}$ be the corresponding set of selected variables. First, apply SIR to the data involving only the predictors in $X_{\mathcal{A}}$, we obtain the estimated squared profile correlations $\hat{\lambda}_1^{\mathcal{A}}, \hat{\lambda}_2^{\mathcal{A}}, \ldots, \hat{\lambda}_K^{\mathcal{A}}$. Let X_t be an arbitrary predictor not in $\mathcal{A}$, and define $\mathcal{A} + t = \mathcal{A} \cup \{t\}$. Next, apply SIR to the data containing the predictors in $\mathcal{A} + t$ to obtain the estimated squared profile correlations $\hat{\lambda}_1^{\mathcal{A}+t}, \hat{\lambda}_2^{\mathcal{A}+t}, \ldots, \hat{\lambda}_K^{\mathcal{A}+t}$. Since $\mathcal{A} \subset \mathcal{A} + t$, we have $\hat{\lambda}_1^{\mathcal{A}} \leqslant \hat{\lambda}_1^{\mathcal{A}+t}$. The difference $\hat{\lambda}_i^{\mathcal{A}+t} - \hat{\lambda}_i^{\mathcal{A}}$, $i = 1, \cdots K$ indicates the improvement in the ith profile correlation due to the inclusion of X_t. We standardize this difference and use the resulting test statistic $\mathrm{COP}_i^{\mathcal{A}+t} = \frac{n(\hat{\lambda}_i^{\mathcal{A}+t} - \hat{\lambda}_i^{\mathcal{A}})}{1 - \hat{\lambda}_i^{\mathcal{A}+t}}$, $\quad i = 1, \cdots K$, to assess the significance of adding X_t to $\mathcal{A}$ in improving the ith profile correlation. To assess the overall contribution of adding X_t to the improvement in all K profile correlations, we combine the statistics $\mathrm{COP}_i^{\mathcal{A}+t}$ into a single test statistic $\mathrm{COP}_{1:K}^{\mathcal{A}+t} = \sum_{i=1}^{K} \mathrm{COP}_i^{\mathcal{A}+t}$ and define $\overline{\mathrm{COP}}_{1:K}^{\mathcal{A}} = \max_{t \in \mathcal{A}^{\mathrm{c}}} \left(\mathrm{COP}_{1:K}^{\mathcal{A}+t}\right)$. Let $X_{\bar{t}}$ be a predictor that attains $\overline{\mathrm{COP}}_{1:K}^{\mathcal{A}}$, i.e. $\overline{\mathrm{COP}}_{1:K}^{\mathcal{A}} = \mathrm{COP}_{1:K}^{\mathcal{A}+\bar{t}}$, and let c_{e} be a prespecified threshold. Then, if $\overline{\mathrm{COP}}_{1:K}^{\mathcal{A}} > c_{\mathrm{e}}$, we add $\bar{t}$ to $\mathcal{A}$; otherwise, we do not add any variable.

Deletion step. Assume that X_t is a predictor in $\mathcal{A}$, and define $\mathcal{A} - t = \mathcal{A} - \{t\}$ as the set obtained by removing X_t from $\mathcal{A}$. Let $\hat{\lambda}_1^{\mathcal{A}-t}, \hat{\lambda}_2^{\mathcal{A}-t}, \ldots, \hat{\lambda}_K^{\mathcal{A}-t}$ be the estimated squared profile correlations based only on the predictors in $\mathcal{A} - t$ only. The effect of deleting X_t from $\mathcal{A}$ on the ith squared profile correlation can be quantified as $\mathrm{COP}_i^{\mathcal{A}-t} = \frac{n(\hat{\lambda}_i^{\mathcal{A}} - \hat{\lambda}_i^{\mathcal{A}-t})}{1 - \hat{\lambda}_i^{\mathcal{A}}}$, $\quad 1 \leqslant i \leqslant K$. To assess the overall effect of removing X_t, we define $\underline{\mathrm{COP}}_{1:K}^{\mathcal{A}} = \min_{t \in \mathcal{A}} \left(\mathrm{COP}_{1:K}^{\mathcal{A}-t}\right)$. Here, $\underline{\mathrm{COP}}_{1:K}^{\mathcal{A}}$ represents the least effect of deleting one predictor from $\mathcal{A}$. Let $X_{\underline{t}}$ be the predictor that achieves $\underline{\mathrm{COP}}_{1:K}^{\mathcal{A}}$, and let c_{d} be a pre-defined threshold for deletion. If $\underline{\mathrm{COP}}_{1:K}^{\mathcal{A}} < c_{\mathrm{d}}$, we delete $X_{\underline{t}}$ from $\mathcal{A}$; otherwise, no deletion happens.

It has been shown [25] that asymptotically, the COP algorithm continues to select variables until all the actual predictors are included. Moreover, once all the true predictors are included, the algorithm will remove all redundant variables from the selected set.

2.4 MultiCOP

We are now ready to present our MultiCOP algorithm. The proposed MultiCOP algorithm consists of two stages. In the first stage, we focus on achieving dimension reduction and variable selection for microbiome data $\mathbf{X}$. We first transfer the multivariate SDR problem into a bunch of univariate SDR sub-problems using random projection. For each set of projected data $(\mathbf{X}, Z)$ with $Z \in \mathbb{R}$, we apply the COP algorithm to estimate the relevant microbes. We then ensemble the selected microbes from each subproblem to form the final pool of the selected microbes using the majority vote. In the second stage, we repeat the above procedure to select the relevant metabolites in $\mathbf{Y}$ based on the selected microbes from the first stage.

First stage. Given independent observations $\left\{\left(\mathbf{x}_i, \mathbf{y}_i\right)\right\}_{i=1,\ldots,n}$ of $(\mathbf{X}, \mathbf{Y})$, where $\mathbf{x}_i = \left(x_{i1}, \ldots, x_{ip}\right)'$ and $\mathbf{y}_i = \left(y_{i1}, \ldots, y_{iq}\right)'$. We first apply the random projection to project $\mathbf{Y}$ along m randomly sampled directions $\mathbf{v}_1, \ldots, \mathbf{v}_m \in \mathbb{R}^q$ to obtain m samples of scalar-valued data

$$\left\{\left(\mathbf{x}_1, \mathbf{v}_j'\mathbf{y}_1\right), \ldots, \left(\mathbf{x}_n, \mathbf{v}_j'\mathbf{y}_n\right)\right\}, \quad j = 1, \ldots, m, \tag{6}$$

i.e., observations of $(\mathbf{X}, \mathbf{v}_j'\mathbf{Y})$, $j = 1, \cdots m$. For each $\mathbf{v}_j$, we use the COP algorithm to solve the univariate SDR problem $Z_j \perp \mathbf{X} \mid \mathbf{B}'\mathbf{X}$ with $Z_j = \mathbf{v}_j'\mathbf{Y}$. Let $\mathcal{A}_j$ be the indices of the selected microbes for each $\mathbf{v}_j$ and $\mathbf{X}_{\mathcal{A}_j}$ be the corresponding set of selected microbes, $j = 1, \cdots m$.

We then ensemble the selected microbes along each projection direction to obtain the final set of the selected microbes by majority vote. Specifically, voting for the most k frequent microbes from all of the m estimated sets $\mathbf{X}_{\mathcal{A}_j}$, we can get the final subset of $\mathbf{X}$ containing the selected microbes, denoted as $\mathbf{X}_{\mathcal{A}^*}$.

Second stage. Given independent observations $\left\{\left(\mathbf{x}_{i,\mathcal{A}^*}, \mathbf{y}_i\right)\right\}_{i=1,\ldots,n}$ of $(\mathbf{X}_{\mathcal{A}^*}, \mathbf{Y})$. We apply the random projection to project $\mathbf{X}_{\mathcal{A}^*}$ along m randomly sampled directions $\mathbf{h}_1, \ldots, \mathbf{h}_m \in \mathbb{R}^k$ to obtain m samples of scalar-valued data

$$\left\{\left(\mathbf{h}_j'\mathbf{x}_{1,\mathcal{A}^*}, \mathbf{y}_1\right), \ldots, \left(\mathbf{h}_j'\mathbf{x}_{n,\mathcal{A}^*}, \mathbf{y}_n\right)\right\}, \quad j = 1, \ldots, m, \tag{7}$$

i.e., observations of $(\mathbf{h}_j'\mathbf{X}_{\mathcal{A}^*}, \mathbf{Y})$, $j = 1, \cdots m$. For each $\mathbf{h}_j$, we use the COP algorithm to solve the univariate SDR problem $W_j \perp \mathbf{Y} \mid \mathbf{B}'\mathbf{Y}$ with $W_j = \mathbf{h}_j'\mathbf{X}_{\mathcal{A}^*}$. Let $\mathcal{B}_j$ be the indices of the selected predictors for each $\mathbf{h}_j$ and $\mathbf{Y}_{\mathcal{B}_j}$ be the corresponding set of selected metabolites, $j = 1, \cdots m$. We then ensemble the results to form the final subset $\mathbf{Y}_{\mathcal{B}^*}$ containing the selected metabolites using the majority vote.

The MultiCOP algorithm is summarized in Algorithm 1.

Algorithm 1 MultiCOP algorithm

```
Input: Microbiome data $\mathbf{X} = (X_1, \cdots, X_p)$, metabolome data $\mathbf{Y} = (Y_1, \cdots, Y_q)$, number of profile correlation directions $K$, thresholds $c_e$ and $c_d$, number of directions $m$.
procedure FIRST STAGE (Selection of microbes)
    Step 1: Generate projected samples $(\mathbf{X}, \mathbf{v}_j'\mathbf{Y})$, where $\mathbf{v}_j,\ j = 1, \cdots m$ are i.i.d randomly selected projection directions.
    Step 2: Apply COP to each of the $m$ projected samples and get $m$ sets of indices of the selected microbes $\mathcal{A}_j$ and the corresponding microbes $\mathbf{X}_{\mathcal{A}_j},\ j = 1, \cdots m$.
        a. With regard to $\mathbf{X} = (X_1, \cdots, X_p)$, randomly select the initial subset of $K$ predictors, and denote $\mathcal{A}$ as the indices of this subset.
        b. Addition step: iterate until no more addition of predictors can be performed.
            (i) find $\bar{t}$ such that $\mathrm{COP}_{1:K}^{\mathcal{A}+\bar{t}} = \overline{\mathrm{COP}}_{1:K}^{\mathcal{A}}$.
            (ii) If $\overline{\mathrm{COP}}_{1:K}^{\mathcal{A}} > c_e$, add $\bar{t}$ to $\mathcal{A}$, i.e. let $\mathcal{A} = \mathcal{A} + \bar{t}$.
        c. Deletion step: iterate until no more deletion of predictors can be performed.
            (i) find $\underline{t}$ such that $\mathrm{COP}_{1:K}^{\mathcal{A}-\underline{t}} = \underline{\mathrm{COP}}_{1:K}^{\mathcal{A}}$.
            (ii) If $\underline{\mathrm{COP}}_{1:K}^{\mathcal{A}} < c_d$, add $\underline{t}$ to $\mathcal{A}$, i.e. let $\mathcal{A} = \mathcal{A} - \underline{t}$.
    Step 3: Majority vote. Combine all the $m$ sets of selected microbes together and vote for the most $k$ frequent microbes. Denote the selected subset of $\mathbf{X}$ as $\mathbf{X}_{\mathcal{A}^*}$.
end procedure
procedure SECOND STAGE (Selection of metabolites)
    Perform the same steps (from Step 1 to Step 3 in the first stage) on $\mathbf{Y} = (Y_1, \cdots, Y_q)$ to get $\mathbf{Y}_{\mathcal{B}^*}$.
end procedure
Output: selected subsets of microbes $\mathbf{X}_{\mathcal{A}^*}$ and metabolites $\mathbf{Y}_{\mathcal{B}^*}$.
```

2.5 Implementation Issues

During the implementation of the MultiCOP algorithm, several tuning parameters need to be defined. In Algorithm 1 *Step 2*, where we apply the COP algorithm to each projected sample, we follow the guidelines set by Zhong et al. [25] to determine the number of profile correlation directions (K) and the thresholds (c_e and c_d) for the addition and deletion steps. Specifically, to determine the threshold c_e, and c_d, we rely on the findings of Zhong et al. [25], who demonstrated that, under mild conditions, the test statistics $\overline{\mathrm{COP}}_{1:K}^{\mathcal{A}}$ converges to a χ^2 distribution in probability. We select a pair of thresholds $c_e = \chi^2_{\alpha,K}$ and $c_d = \chi^2_{\alpha-0.05,K}$. Furthermore, for the selection of K in each COP procedure applied to the projected samples, we employ the G information criterion proposed by Zhu et al. [28]. Following the approach outlined by Zhong et al. [25], we initially define a range for K, typically from 1 to J. Subsequently, we select $K = \arg\min_{1 \leqslant k \leqslant J} G(k)$, where $G(k) = -\log L(k) + \frac{\log(n)}{2} k\left(2p_k - k + 1\right)$. Once the number of profile correlations K is determined for each COP procedure, denoted as $K_1, \ldots, K_m$, we choose $k = \text{mode}\{K_1, \ldots, K_m\}$ as the number of variables in the final selected subset in *Step 3*.

Since the procedure in Algorithm 1 is asymmetric between the two datasets, we need to specify the priority of one dataset over the other. In this study, we select variables from the microbiome data first, followed by the metabolome data, based on the biological understanding of the causal relationship between the microbiome and metabolome. The microbiome, which consists of the microbial community in a particular environment (e.g., gut, skin), is known to influence the metabolome [2, 29, 30]. Microbes engage in various metabolic processes and produce metabolites that shape the metabolomic profile. Consequently, as the microbiome can be considered a driver or cause of the metabolome, we prioritize the selection of features from the microbiome data.

3 Results

3.1 Simulation Study

We perform comprehensive simulation studies to compare MultiCOP with other four established variable selection methods through four distinct simulation scenarios. These scenarios indicate four different ways in which the microbiome ($\mathbf{X}$) and metabolome data (Y) are associated with each other, including a linear association, a non-linear association, a non-linear association with a large number of microbes (p) and metabolites (q), and a heteroscedastic association.

For each scenario, we set $\mathbf{X} = \left(X_1, X_2, \ldots, X_p\right)'$ follows a p-variate normal distribution with mean 0 and covariance $\text{cov}\left(X_i, X_j\right) = \rho^{|i-j|}$ for $1 \leqslant i, j \leqslant p$, and $\epsilon = (\epsilon_1, \epsilon_2, \cdots, \epsilon_q)'$ is independent of X. Each ϵ_i independently follows $N(0, \sigma^2)$.

Scenario 1. Consider the linear model $\mathbf{Y} = \mathbf{BX} + \epsilon$, with $p = 5$ and $q = 6$. We set,

$$\begin{cases} Y_i = 3X_1 + 1.5X_2 + 2X_3 + \epsilon_i \quad i = 1, 2 \\ Y_i = \epsilon_i, \quad i = 3, 4, 5, 6 \end{cases} \tag{8}$$

Scenario 2. Consider the non-linear model with $p = 5$ and $q = 6$,

$$\begin{cases} Y_i = \frac{(X_1 + X_2)}{0.5 + (1.5 + X_1)^2} + \epsilon_i, \quad i = 1, 2 \\ Y_i = \epsilon_i, \quad i = 3, 4, \cdots, q \end{cases} \tag{9}$$

Scenario 3. Consider the non-linear model

$$\begin{cases} Y_i = \frac{(X_{i+1} + \ldots + X_{i+3})}{0.5 + (1.5 + 0.5(X_1 + \ldots + X_6))^2} + \epsilon_i, \quad i = 1, 2, 3 \\ Y_i = \epsilon_i, \quad i = 4, \ldots, q \end{cases} \tag{10}$$

Scenario 4. Consider the heteroscedastic model with $p = 10$ and $q = 10$,

$$\begin{cases} Y_i = \frac{\epsilon_i}{0.2+0.5\sum_{j=1}^{4}\beta_j^{(i)}X_j} & i = 1, 2, 3 \\ Y_i = \epsilon_i, \quad i = 4, \cdots, 10, \end{cases} \tag{11}$$

where $\beta^{(1)} = (\beta_1^{(1)}, \beta_2^{(1)}, \beta_3^{(1)}, \beta_4^{(1)})^T = (1, 1, 1, 0)^T$, $\beta^{(2)} = (\beta_1^{(2)}, \beta_2^{(2)}, \beta_3^{(2)}, \beta_4^{(2)})^T = (1, -1, 1, 0)^T$, and $\beta^{(3)} = (\beta_1^{(3)}, \beta_2^{(3)}, \beta_3^{(3)}, \beta_4^{(3)})^T = (1, 0, 2, 1)^T$.

For each scenario, we further create three different situations by specifying different values of σ, n, p, and q to test the stability of the algorithm. In scenario 3, we test the performance of MultiCOP when p and q are large.

Under each situation, we generate 30 replicate data sets and compare the performance between MultiCOP and four other variable selection methods: canonical correlation analysis (CCA) [31], partial least square (PLS) [1], multivariate LASSO [32], and reduced rank regression (RRR) [33] on each synthetic data set. We measure the performance of each method in variable selection using two metrics: the false positive rate (FPR) and the false negative rate (FNR) for both **X** and **Y**.

When implementing MultiCOP, the pool of possible c_e is set as $\left\{\chi^2_{0.90,K}, \chi^2_{0.95,K}, \chi^2_{0.99,K}\right\}$, and the associated pool of c_d is $\left\{\chi^2_{0.85,K}, \chi^2_{0.90,K}, \chi^2_{0.95,K}\right\}$. The possible values of K are from 1 to $\min\{20, \max\{p, q\}\}$. The number of random projection directions m is set to be $10n$. For sparse CCA, we run the R function PMA::CCA [31], which performs sparse canonical correlation analysis using the penalized matrix decomposition. For PLS, we employ the R function plsVarSel::VIP [1], which filters variables by variable importance on projections. For multivariate LASSO, we use the famous R package glmnet [32], which is widely used for LASSO regression. We fit the multivariate LASSO model treating **Y** as the response to filter out the **X**'s with a coefficient equal to 0, and treat **X** as the response to filter **Y**'s. For RRR, we implement the command rrpack::rrr [33] to calculate the coefficients for each variable. Since RRR cannot be used to select variables, we pick out the top variables with the highest absolute value. The number of variables to pick is the same as the number of selected variables in MultiCOP. During the comparison, all the parameters in the compared method are set as the default value.

We report the results for each scenario in Tables 1, 2, 3, 4 separately. The results demonstrate that the performance of MultiCOP is competitive compared to the other methods, as indicated by the relatively low values of FNR and FPR. Notably, the other methods have a slightly better performance than MultiCOP in some settings in the first three scenarios, due to their inherent frameworks being designed to detect the linear associations. In contrast, MultiCOP does not hinge on any pre-defined association type. As such, it is comprehensible for our methods to slightly underperform in comparison in some settings. However, when dealing with non-linear associations between multivariate datasets, MultiCOP exhibits significantly improved performance in terms of FNR and FPR. Results in simulation show that MultiCOP can efficiently explore the association between multivariate and multivariate data, which is consistent with our theoretical analysis.

Table 1 Results for Scenario 1. We consider three different settings and compare the proposed MultiCOP algorithm with CCA, PLS, multivariate LASSO, and RRR. We report the mean value and standard deviation (in parentheses) of 30 repetitions

	FPR in x	FNR in x	FPR in y	FNR in y
Results for ρ=0.5, σ=0.5, n=100 (rep=30)				
CCA	0.00 (0.000)	0.67 (0.000)	0.00 (0.000)	0.50 (0.000)
PLS	0.00 (0.000)	0.08 (0.143)	0.00 (0.000)	0.00 (0.000)
LASSO	0.75 (0.315)	0.23 (0.202)	0.46 (0.348)	0.11 (0.253)
RRR	0.00 (0.000)	0.16 (0.169)	0.01 (0.046)	0.00 (0.000)
MultiCOP (ours)	0.00 (0.000)	0.16 (0.169)	0.01 (0.046)	0.00 (0.000)
Results for ρ=0.5, σ=0.5, n=1000 (rep=30)				
CCA	0.00 (0.000)	0.67 (0.000)	0.00 (0.000)	0.50 (0.000)
PLS	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)
LASSO	0.80 (0.311)	0.27 (0.192)	0.41 (0.290)	0.04 (0.169)
RRR	0.02 (0.091)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)
MultiCOP (ours)	0.02 (0.091)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)
Results for ρ=0.5, σ=2, n=100 (rep=30)				
CCA	0.00 (0.000)	0.67 (0.000)	0.00 (0.000)	0.50 (0.000)
PLS	0.00 (0.000)	0.08 (0.143)	0.00 (0.000)	0.00 (0.000)
LASSO	0.75 (0.341)	0.24 (0.199)	0.53 (0.339)	0.13 (0.271)
RRR	0.00 (0.000)	0.00 (0.000)	0.01 (0.046)	0.00 (0.000)
MultiCOP (ours)	0.00 (0.000)	0.00 (0.000)	0.01 (0.046)	0.00 (0.000)

3.2 Real Data 1: Inflammatory Bowel Disease (IBD)

Inflammatory bowel disease (IBD) is a chronic condition that affects the digestive tract, causing inflammation and damage to intestinal tissue. Research has shown that patients with IBD often exhibit a dysbiotic gut microbiome, characterized by an imbalance in the composition of bacterial species and reduced diversity [34]. This dysbiosis can potentially stimulate an abnormal immune response, leading to chronic inflammation and subsequent intestinal tissue damage [34, 35]. In particular, in addition to changes in microbial composition, alterations in metabolites produced by the gut microbiome have also been observed in patients with IBD [36].

A recent study called IBDMDB [37], as part of the Integrative Human Microbiome Project (HMP2 or iHMP), was conducted to generate integrated molecular profiles of both host and microbial activity during disease (IBD) using numerous individual specimens. This study resulted in the establishment of a publicly accessible database that encompassed multiple types of measurements, including metagenomes, metatranscriptomes, proteomes, metabolites, and other related data, available at https://ibdmdb.org/. Specifically, metagenome sequences and metabolite profiles were primarily derived from stool specimens. We utilized the MultiCOP algorithm to analyze the metagenome sequence data and the corresponding metabolite data of

Table 2 Results for Scenario 2. We consider three different settings and compare the proposed MultiCOP algorithm with CCA, PLS, multivariate LASSO, and RRR. We report the mean value and standard deviation (in parentheses) of 30 repetitions

	FPR in x	FNR in x	FPR in y	FNR in y
Results for ρ=0.5, σ=0.5, n=100 (rep=30)				
CCA	0.00 (0.000)	0.50 (0.000)	0.00 (0.000)	0.50 (0.000)
PLS	0.02 (0.085)	0.02 (0.091)	0.00 (0.00)	0.00 (0.000)
LASSO	0.53 (0.407)	0.18 (0.280)	0.34 (0.311)	0.08 (0.218)
RRR	0.03 (0.102)	0.05 (0.153)	0.23 (0.101)	0.42 (0.190)
MultiCOP (ours)	0.00 (0.000)	0.00 (0.000)	0.04 (0.115)	0.05 (0.153)
Results for ρ=0.5, σ=3, n=100 (rep=30)				
CCA	0.01 (0.061)	0.52 (0.091)	0.00 (0.000)	0.50 (0.000)
PLS	0.10 (0.178)	0.08 (0.190)	0.05 (0.102)	0.07 (0.173)
LASSO	0.36 (0.371)	0.17 (0.260)	0.43 (0.371)	0.17 (0.284)
RRR	0.13 (0.166)	0.20 (0.249)	0.07 (0.117)	0.15 (0.233)
MultiCOP (ours)	0.00 (0.000)	0.00 (0.000)	0.05 (0.102)	0.10 (0.203)
Results for ρ=0.5, σ=3, n=200 (rep=30)				
CCA	0.00 (0.000)	0.50 (0.000)	0.00 (0.000)	0.50 (0.000)
PLS	0.04 (0.115)	0.05 (0.153)	0.02 (0.063)	0.00 (0.000)
LASSO	0.49 (0.324)	0.13 (0.247)	0.57 (0.380)	0.26 (0.347)
RRR	0.07 (0.136)	0.10 (0.203)	0.03 (0.086)	0.07 (0.173)
MultiCOP (ours)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)

IBD patients to explore the interactions between the microbiome and metabolites during the progression of IBD.

For the microbiome dataset, we initiated with a differential abundance analysis to pinpoint microbial features exhibiting distinct distributions between the datasets. We specifically compared our dataset with the control group detailed in Lloyd-Price et al. [37], where data for the control group are provided. We utilized the DESeq2 [38] package to analyze the count data. The differences in taxa were evaluated using the Wald test based on the negative binomial model, which is consistent with the approach proposed in Love et al. [39]. In the subsequent phase, we employed a filtering process, retaining only those taxa that appeared in a minimum of 10% of the samples, as the procedure described in Morton et al. [24]. Then, we implemented a pseudo-count of one to all data points.

We followed the procedures detailed in Morton et al. [24] for processing the metagenomics data. Specifically, microbes present in fewer than 10 samples were excluded because they lacked enough data to confidently infer their interactions with metabolites. The metabolite data was profiled using a combination of four LC-MS methods including HILIC-pos, HILIC-neg, C18-neg, and C8-pos [18]. The raw LC-MS data was preprocessed, and peak identification was conducted by matching peaks based on attributes such as retention time, m/z ratio, and intensity. This process enabled the annotation of detected peaks, leading to the identification of 589 distinct metabolites. We restricted our analysis to only the annotated metabolites.

Table 3 Results for Scenario 3. We consider three different settings and compare the proposed MultiCOP algorithm with CCA, PLS, multivariate LASSO, and RRR. We report the mean value and standard deviation (in parentheses) of 30 repetitions

	FPR in x	FNR in x	FPR in y	FNR in y
Results for ρ=0.5, σ=0.1, n=200, p=q=10 (rep=30)				
CCA	1.00 (0.000)	0.17 (0.000)	1.00 (0.000)	0.33 (0.000)
PLS	0.01 (0.456)	0.27 (0.102)	0.00 (0.000)	0.00 (0.000)
LASSO	0.54 (0.329)	0.12 (0.157)	0.92 (0.134)	0.47 (0.336)
RRR	0.08 (0.137)	0.30 (0.111)	0.37 (0.072)	0.98 (0.08)
MultiCOP (ours)	0.00 (0.000)	0.24 (0.085)	0.00 (0.026)	0.13 (0.166)
Results for ρ=0.5, σ=0.1, n=300, p=q=10 (rep=30)				
CCA	1.00 (0.000)	0.17 (0.000)	1.00 (0.000)	0.33 (0.000)
PLS	0.00 (0.000)	0.25 (0.105)	0.00 (0.000)	0.00 (0.000)
LASSO	0.61 (0.276)	0.10 (0.162)	0.97 (0.132)	0.65 (0.178)
RRR	0.04 (0.095)	0.20 (0.102)	0.35 (0.082)	0.92 (0.143)
MultiCOP (ours)	0.00 (0.000)	0.17 (0.120)	0.00 (0.000)	0.10 (0.155)
Results for ρ=0.5, σ=0.1, n=500, p=q=50 (rep=30)				
CCA	0.05 (0.032)	0.12 (0.075)	0.38 (0.059)	0.00 (0.000)
PLS	0.02 (0.009)	0.01 (0.030)	0.05 (0.027)	0.00 (0.000)
LASSO	0.20 (0.106)	0.04 (0.072)	0.33 (0.161)	0.01 (0.061)
RRR	0.02 (0.013)	0.13 (0.092)	0.10 (0.008)	1.00 (0.000)
MultiCOP (ours)	0.00 (0.000)	0.01 (0.042)	0.04 (0.007)	0.01 (0.061)

This was to mitigate the potential for significant variations in the metabolite profiles between consecutive samples from the same subjects, which might have been introduced by unidentified compounds [37].

We subsequently applied the MultiCOP algorithm to the dataset, which, after filtering, comprised $n = 387$ samples with $p = 562$ microbes and $q = 589$ metabolites. The thresholds were established as $c_e = \chi^2_{0.95,K}$ and $c_d = \chi^2_{0.90,K}$. We selected the value for K based on the G information criterion, in alignment with the consistency theorems presented in Zhong et al. [25]. Furthermore, we set the number of random projection directions m to be $10n$.

The relevant microbes and metabolites identified in the microbiome-metabolome association are presented in Fig. 1a. The solid lines represent microbiome-metabolome associations that align with findings from previous literature, whereas the dashed lines indicate highly potential associations that biologists can explore in the future. Our findings are consistent with previous studies, underscoring certain microbial and metabolic features pivotal to the association between the microbiome and metabolome in IBD patients. Key microbial entities include members of the Roseburia genus, namely *Roseburia hominis* and *Roseburia intestinalis*, and the Klebsiella genus, such as *Klebsiella oxytoca* and *Klebsiella pneumoniae*. On the metabolic front, crucial features encompass carnitines (*NH4*), bile acids (*ketodeoxycholate* and *alpha-muricholate*), and Short-chain

Table 4 Results for Scenario 4. We consider three different settings and compare the proposed MultiCOP algorithm with CCA, PLS, multivariate LASSO, and RRR. We report the mean value and standard deviation (in parentheses) of 30 repetitions

	FPR in x	FNR in x	FPR in y	FNR in y
Results for ρ=0, σ=0.5, n=100 (rep=30)				
CCA	0.11 (0.082)	0.91 (0.123)	0.12 (0.054)	0.94 (0.126)
PLS	0.38 (0.163)	0.72 (0.165)	0.48 (0.143)	0.60 (0.296)
LASSO	0.32 (0.155)	0.95 (0.121)	0.24 (0.153)	0.89 (0.237)
RRR	0.39 (0.126)	0.72 (0.205)	0.50 (0.082)	0.99 (0.061)
MultiCOP (ours)	0.02 (0.051)	0.16 (0.154)	0.09 (0.079)	0.03 (0.102)
Results for *ρ=0.5, σ=0.5, n=100 (rep=30)*				
CCA	0.09 (0.084)	0.89 (0.126)	0.11 (0.058)	0.93 (0.136)
PLS	0.47 (0.177)	0.68 (0.219)	0.49 (0.122)	0.60 (0.238)
LASSO	0.33 (0.133)	0.98 (0.076)	0.30 (0.170)	0.88 (0.223)
RRR	0.32 (0.141)	0.65 (0.233)	0.49 (0.080)	1.00 (0.000)
MultiCOP (ours)	0.01 (0.042)	0.19 (0.157)	0.06 (0.089)	0.01 (0.061)
Results for ρ=0, σ=1, n=100 (rep=30)				
CCA	0.13 (0.068)	0.95 (0.102)	0.11 (0.061)	0.92 (0.143)
PLS	0.36 (0.150)	0.66 (0.250)	0.51 (0.134)	0.62 (0.300)
LASSO	0.32 (0.123)	0.99 (0.046)	0.30 (0.200)	0.90 (0.217)
RRR	0.34 (0.123)	0.71 (0.208)	0.47 (0.076)	0.98 (0.085)
MultiCOP (ours)	0.02 (0.058)	0.22 (0.101)	0.06 (0.072)	0.02 (0.085)

fatty acids (SCFA) like *butyrate*. These observations are in harmony with the original research by Lloyd-Price et al. [37]. Furthermore, our findings reinforce the documented presence of the *Propionibacterium* genus as outlined in Morton et al. [24]. Notably, certain species within this genus produce *1,4-dihydroxy-2-naphthoic acid* (DHNA), promoting growth in bacteria like *Bifidobacterium*, which is known for its potential role in alleviating IBD symptoms.

Utilizing the MultiCOP algorithm, we've validated not only known associations and discovered novel but also previously unrecognized connections. Primarily, our selected metabolic subset revealed additional SCFAs: *caproate*, *valerate*/*isovalerate*, and *caprate*. These SCFAs hold significant potential relevance to IBD patients and may contribute to treatment avenues. A study indicated that *caproate* promotes colonic wound repair and modulates the expression of PCNA and cyclin D in the colonic mucosa of rats with TNBS colitis [40]. Moreover, such SCFAs have been reported to have beneficial effects in reducing intestinal inflammation and preventing disruptions in the intestinal microflora [41]. A meta-analysis also observed reduced valerate levels in the fecal matter of IBD patients [42]. Yet, the exact role of valerate in IBD warrants deeper investigation [43]. Our findings seem to resonate with insights from previous research, reinforcing their validity. Secondly, our microbial subset of interest identifies species from the Eubacterium genus, specifically *Eubacterium dolichum* and *Eubacterium biforme*. This insight refines the prevailing understanding among

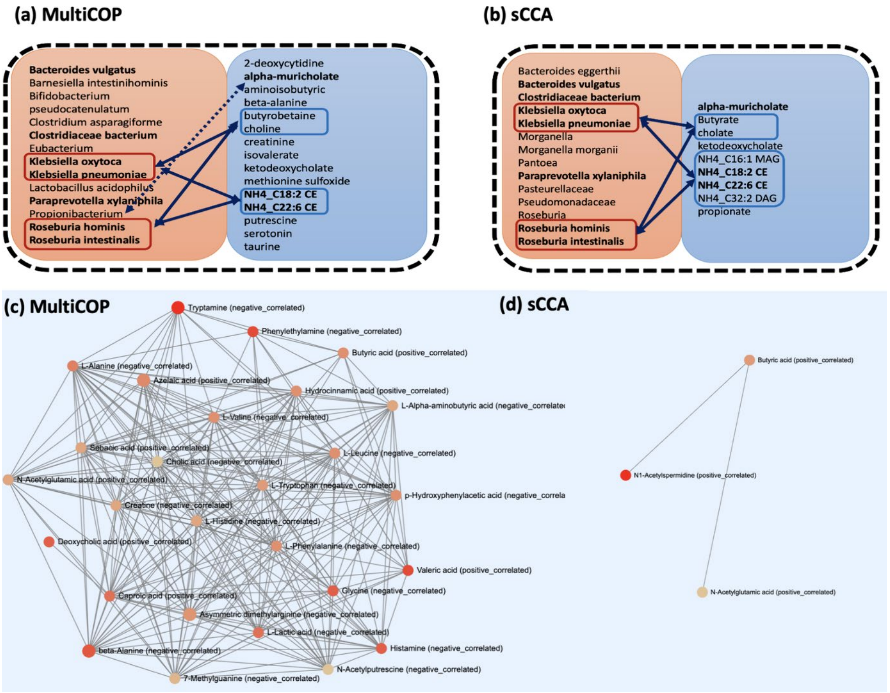
(a) MultiCOP
Bacteroides vulgatus
Barnesiella intestinihominis
Bifidobacterium
pseudocatenulatum
Clostridium asparagiforme
Clostridiaceae bacterium
Eubacterium
Klebsiella oxytoca
Klebsiella pneumoniae
Lactobacillus acidophilus
Paraprevotella xylaniphila
Propionibacterium
Roseburia hominis
Roseburia intestinalis
2-deoxycytidine
alpha-muricholate
aminoisobutyric
beta-alanine
butyrobetaine
choline
creatinine
isovalerate
ketodeoxycholate
methionine sulfoxide
NH4_C18:2 CE
NH4_C22:6 CE
putrescine
serotonin
taurine
(b) sCCA
Bacteroides eggerthii
Bacteroides vulgatus
Clostridiaceae bacterium
Klebsiella oxytoca
Klebsiella pneumoniae
Morganella
Morganella morganii
Pantoea
Paraprevotella xylaniphila
Pasteurellaceae
Pseudomonadaceae
Roseburia
Roseburia hominis
Roseburia intestinalis
alpha-muricholate
Butyrate
cholate
ketodeoxycholate
NH4_C16:1 MAG
NH4_C18:2 CE
NH4_C22:6 CE
NH4_C32:2 DAG
propionate
(c) MultiCOP
Tryptamine (negative_correlated)
Phenylethylamine (negative_correlated)
Butyric acid (positive_correlated)
L-Alanine (negative_correlated)
Azelaic acid (positive_correlated)
Hydrocinnamic acid (positive_correlated)
L-Alpha-aminobutyric acid (negative_correlatec
L-Valine (negative_correlated)
Sebacic acid (positive_correlated)
L-Leucine (negative_correlated)
Cholic acid (negative_correlated)
N-Acetylglutamic acid (positive_correlated)
L-Tryptophan (negative_correlated)
p-Hydroxyphenylacetic acid (negative_correla
Creatine (negative_correlated)
L-Histidine (negative_correlated)
Deoxycholic acid (positive_correlated)
L-Phenylalanine (negative_correlated)
Valeric acid (positive_correlated)
Glycine (negative_correlated)
Caproic acid (positive_correlated)
Asymmetric dimethylarginine (negative_correlated)
L-Lactic acid (negative_correlated)
Histamine (negative_correlated)
beta-Alanine (negative_correlated)
N-Acetylputrescine (negative_correlated)
7-Methylguanine (negative_correlated)
(d) sCCA
Butyric acid (positive_correlated)
N1-Acetylspermidine (positive_correlated)
N-Acetylglutamic acid (positive_correlated)

Fig. 1 In **a**, the outcomes derived from the proposed MultiCOP method are displayed, while **b** showcases the results of the sparse CCA. Solid lines in both **a** and **b** depict microbiome-metabolome relationships affirmed by prior research. In contrast, dashed lines suggest potential connections that merit further investigation by biologists. **c** and **d** present the enrichment network based on TSEA analysis, corresponding to the findings in (**a** and **b**), respectively. Each node in **c** and **d** symbolizes a taxon set: its color signifies the p-value, and its size indicates the count of correspondences with the identified microbes. Nodes interconnect when the shared taxa constitute more than 20% of their combined total

gut microbiologists: particular strains of butyrate-producing microbes, especially from the Eubacterium, Roseburia, and Faecalibacterium genera, are considered advantageous for human health [44].

We then applied Taxon Set Enrichment Analysis (TSEA) [45] to directly investigate whether a specified list of microbes of interest showcases enrichments within taxon sets functionally related to the microbiome-metabolite interaction. The TSEA employs the hypergeometric test as part of the Over Representation Analysis (ORA) to determine if a specific Taxon set is overrepresented compared to what would be expected by chance within a given list of microbes of interest. The enrichment network is depicted in Fig. 1c, where taxon sets manifest as nodes. The node color signifies its p-value, while its size correlates with the count of selected microbial matches to that specific taxon set (node). Nodes are linked if the number of shared taxa exceeds 20% of the total combined taxa between them. The results reveal that most of the selected microbes are included in the 306 manually curated taxon sets, which is consistent with our findings.

For comparison purposes, we also employed sparse CCA analysis, described in [46], and selected the number of components same with the number of microbial and metabolite features with MultiCOP. Figure 1b displays the chosen features and Fig. 1d illustrates the corresponding enrichment graph generated through TSEA. Comparing two enrichment graphs, the graph generated based on MultiCOP appears to be more densely packed and compact compared to the one based on sCCA. This more complicated network indicates that more information about the microbiome-metabolome association is revealed by the proposed MultiCOP method. Additionally, two of the three nodes, *Butyric* and *N-Acetyglutamic*, selected by sCCA as presented in Fig. 1c, are also featured in Fig. 1a. It has been shown that *Butyric Acid (Butyric)* serves multifaceted roles, functioning as an anti-inflammatory agent, an energy source for *colonocytes*, and an immunomodulatory. While *Butyric Acid* is recognized for its direct involvement in IBD processes, the specific implications of *N-Acetyglutamic Acid* and *N1-Acetylspermidine* in the context of IBD may be more indirect.

3.3 Real Data 2: Chronic Ischemic Heart Disease (CIHD)

Chronic ischemic heart disease (CIHD) is characterized by reduced blood flow to the heart muscle due to narrowed or blocked coronary arteries. This condition can lead to various symptoms, such as chest pain (angina) and shortness of breath. Increasing evidence points to a pivotal role of the human microbiome in CIHD's onset and progression [47]. The microbiome, consisting of microorganisms inhabiting the human body, appears intricately linked to the metabolome—the ensemble of the body's small molecule metabolites. Studies have shown that the gut microbiome can influence the development of atherosclerosis, the underlying condition of CIHD, by producing metabolites that impact inflammation, lipid metabolism, and other crucial processes. A case in point is certain gut bacteria's ability to produce *trimethylamine-N-oxide* (TMAO), a metabolite associated with an amplified heart disease risk [48]. Exploring the association between the microbiome and the metabolome in CIHD has the potential to uncover novel biomarkers and therapeutic targets for the disease. By investigating this association, we may uncover indicators that can aid in the diagnosis, prognosis, and treatment of CIHD. Furthermore, targeting interventions to modulate the gut microbiome to mitigate the production of detrimental metabolites shows promise as an effective strategy to prevent or treat CIHD.

To explore the association between the microbiome and the metabolome in CIHD, we applied our MultiCOP method to a publicly available dataset [49] that includes 158 patients of CIHD with 45 annotated urine metabolites, 1212 annotated serum metabolites, and 729 metagenomics species. In our implementation, we set c_e equal to $\chi^2_{0.90,K}$, c_d equal to $\chi^2_{0.85,K}$, and determined the optimal number of profile correlation directions (K) using the G information criterion. Through our analysis, we unveiled valuable insights into the intricate relationship between the microbiome and metabolome in the context of CIHD. The analysis procedure resulted in the identification of two selected subsets consisting of 10 metabolites and 21 microbes, as depicted in Fig. 2a. Within Fig. 2a, solid lines represent microbiome-metabolome

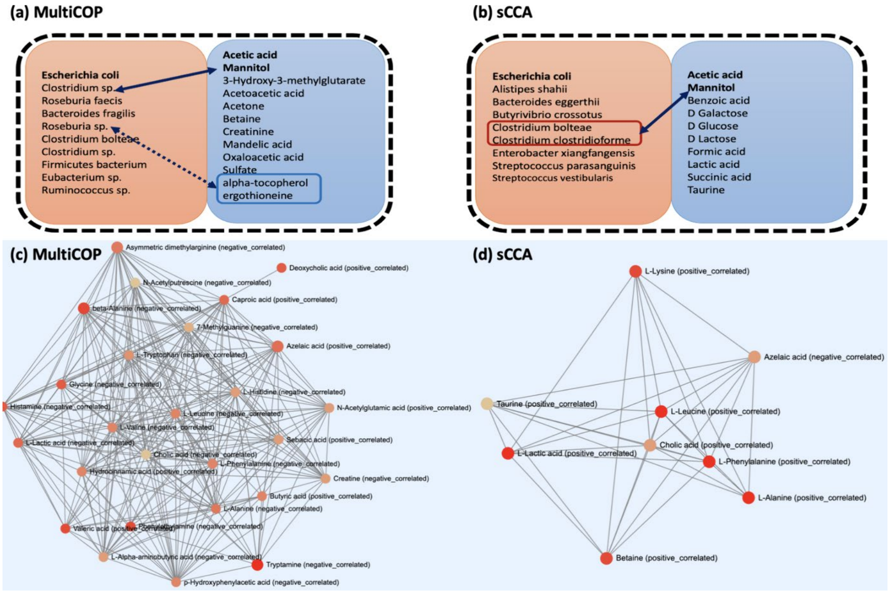
(a) MultiCOP
Escherichia coli
Clostridium sp.
Roseburia faecis
Bacteroides fragilis
Roseburia sp.
Clostridium bolteae
Clostridium sp.
Firmicutes bacterium
Eubacterium sp.
Ruminococcus sp.
Acetic acid
Mannitol
3-Hydroxy-3-methylglutarate
Acetoacetic acid
Acetone
Betaine
Creatinine
Mandelic acid
Oxaloacetic acid
Sulfate
alpha-tocopherol
ergothioneine
(b) sCCA
Escherichia coli
Alistipes shahii
Bacteroides eggerthii
Butyrivibrio crossotus
Clostridium bolteae
Clostridium clostridioforme
Enterobacter xiangfangensis
Streptococcus parasanguinis
Streptococcus vestibularis
Acetic acid
Mannitol
Benzoic acid
D Galactose
D Glucose
D Lactose
Formic acid
Lactic acid
Succinic acid
Taurine
(c) MultiCOP
Asymmetric dimethylarginine (negative_correlated)
Deoxycholic acid (positive_correlated)
N-Acetylputrescine (negative_correlated)
Caproic acid (positive_correlated)
beta-Alanine (negative_correlated)
7-Methylguanine (negative_correlated)
Azelaic acid (positive_correlated)
L-Tryptophan (negative_correlated)
Glycine (negative_correlated)
L-Histidine (negative_correlated)
Histamine (negative_correlated)
N-Acetylglutamic acid (positive_correlated)
L-Leucine (negative_correlated)
L-Valine (negative_correlated)
L-Lactic acid (negative_correlated)
Sebacic acid (positive_correlated)
Cholic acid (negative_correlated)
L-Phenylalanine (negative_correlated)
Hydrocinnamic acid (positive_correlated)
Creatine (negative_correlated)
Butyric acid (positive_correlated)
L-Alanine (negative_correlated)
Valeric acid (positive_correlated)
Phenylethylamine (negative_correlated)
L-Alpha-aminobutyric acid (negative_correlated)
Tryptamine (negative_correlated)
p-Hydroxyphenylacetic acid (negative_correlated)
(d) sCCA
L-Lysine (positive_correlated)
Azelaic acid (negative_correlated)
Taurine (positive_correlated)
L-Leucine (positive_correlated)
Cholic acid (positive_correlated)
L-Lactic acid (positive_correlated)
L-Phenylalanine (positive_correlated)
L-Alanine (positive_correlated)
Betaine (positive_correlated)

Fig. 2 In **a**, the outcomes derived from the proposed MultiCOP method are displayed, while **b** showcases the results of the sparse CCA. Solid lines in both **a** and **b** depict microbiome-metabolome relationships affirmed by prior research. In contrast, dashed lines suggest potential connections that merit further investigation by biologists. **c** and **d** present the enrichment network based on TSEA analysis, corresponding to the findings in (**a** and **b**), respectively. Each node in **c** and **d** symbolizes a taxon set: its color signifies the p-value, and its size indicates the count of correspondences with the identified microbes. Nodes interconnect when the shared taxa constitute more than 20% of their combined total

relationships that have been established and validated in the previous literature. On the contrary, dashed lines indicate potential relationships that warrant further investigation and exploration by biologists. In our study, we observed specific bacterial species within our selected subset, *Clostridium* sp., *Roseburia* sp., and *Firmicutes bacterium*, that exhibited notable changes in individuals with CIHD, as supported by a previous study [44]. Literature suggests that various species of *Clostridium* possess the ability to ferment mannitol [50, 51]. While mannitol has potential applications in certain medical contexts, such as fluid balance management or reduction of edema, its precise role or impact in chronic ischemic heart disease remains unclear. However, further exploration of its potential significance in this condition holds promise. However, further exploration of its potential role in this condition is worthwhile. Furthermore, we observed the presence of certain metabolites in our selected subset that are associated with CIHD, including *ergothioneine* and *alpha-tocopherol*, both of which are antioxidants [52, 53]. Öhrvall et al. [53] suggested a possible protective role of *alpha-tocopherol* in reducing the risk of cardiovascular diseases. These findings highlight the relevance of investigating the involvement of these metabolites in the context of CIHD and their potential implications for disease management.

By comparison, we applied sCCA analysis to select an equal number of microbial features. Figure 2b showcases the selected features, while Fig. 2d visualizes the enrichment graph generated using TSEA. A noticeable distinction arises when comparing our enrichment graph to the alternative counterparts; ours exhibits a denser and more compact structure. This denser structure implies a more intricate network that captures a richer set of interrelationships between microbes and metabolites.

4 Discussion

In this paper, we propose the MultiCOP algorithm, which effectively detects the association between the microbiome and metabolome data to identify microbe-metabolite interactions. The MultiCOP algorithm addresses the multivariate SDR problem by decomposing it into a set of univariate SDR problems through random projection. We then employ the COP algorithm to solve each univariate SDR problem and identify the relevant variables (microbes/metabolites). The outcomes of each subproblem are subsequently ensembled through the majority vote, giving the final set of associated microbes and metabolites that elucidate the microbiome-metabolome interaction. To evaluate the efficacy of MultiCOP, we conducted extensive experiments using simulated data, as well as real data from patients with inflammatory bowel disease and chronic ischemic heart disease. We compared the performance of our algorithm against other established methods, and the results demonstrated the superior performance of MultiCOP in terms of FPR and FNR. These findings strongly suggest that the proposed MultiCOP algorithm holds great promise as a tool for exploring microbiome-metabolome associations and identifying relevant microbes and metabolites. While we empirically showed that taking $O(n)$ iterations of random projection yields excellent performance, we aim to theoretically prove this in our future study. Another future direction is to explore more efficient methods for selecting projection directions, moving beyond the use of random projection.

Acknowledgements This work was partially supported by U.S. National Science Foundation [DMS-1925066, DMS-1903226, DMS-2124493, DMS-2311297, DMS-2319279, DMS-2318809] and National Institutes of Health [NIH R01GM152814].

Data Availability The simulation data is available at https://github.com/Luyang8991/MultiCOP. The inflammatory bowel disease data is available at https://ibdmdb.org/tunnel/public/summary.html. The chronic ischemic heart disease data is available in [49].

Declarations

Conflict of interest The authors declare no conflicts of interest that are relevant to the content of this article.

References

1. Tremaroli V, Bäckhed F (2012) Functional interactions between the gut microbiota and host metabolism. Nature 489(7415):242–249

2. Zierer J, Jackson MA, Kastenmüller G, Mangino M, Long T, Telenti A, Mohney RP, Small KS, Bell JT, Steves CJ et al (2018) The fecal metabolome as a functional readout of the gut microbiome. Nat Genet 50(6):790–795
3. Franzosa EA, Sirota-Madi A, Avila-Pacheco J, Fornelos N, Haiser HJ, Reinker S, Vatanen T, Hall AB, Mallick H, McIver LJ et al (2019) Gut microbiome structure and metabolic activity in inflammatory bowel disease. Nat Microbiol 4(2):293–305
4. Levy M, Blacher E, Elinav E (2017) Microbiome, metabolites and host immunity. Curr Opin Microbiol 35:8–15
5. Rooks MG, Garrett WS (2016) Gut microbiota, metabolites and host immunity. Nat Rev Immunol 16(6):341–352
6. Tang WW, Li DY, Hazen SL (2019) Dietary metabolism, the gut microbiome, and heart failure. Nat Rev Cardiol 16(3):137–154
7. Wahlström A, Sayin SI, Marschall H-U, Bäckhed F (2016) Intestinal crosstalk between bile acids and microbiota and its impact on host metabolism. Cell Metab 24(1):41–50
8. Devaraj S, Hemarajata P, Versalovic J (2013) The human gut microbiome and body metabolism: implications for obesity and diabetes. Clin Chem 59(4):617–628
9. Canfora EE, Meex RC, Venema K, Blaak EE (2019) Gut microbial metabolites in obesity, nafld and t2dm. Nat Rev Endocrinol 15(5):261–273
10. Yachida S, Mizutani S, Shiroma H, Shiba S, Nakajima T, Sakamoto T, Watanabe H, Masuda K, Nishimoto Y, Kubo M et al (2019) Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. Nat Med 25(6):968–976
11. Yap IK, Li JV, Saric J, Martin F-P, Davies H, Wang Y, Wilson ID, Nicholson JK, Utzinger J, Marchesi JR et al (2008) Metabonomic and microbiological analysis of the dynamic effect of vancomycin-induced gut microbiota modification in the mouse. J Proteome Res 7(9):3718–3728
12. Brunkwall L, Orho-Melander M (2017) The gut microbiome as a target for prevention and treatment of hyperglycaemia in type 2 diabetes: from current human evidence to future possibilities. Diabetologia 60(6):943–951
13. Suez J, Elinav E (2017) The path towards microbiome-based metabolite treatment. Nat Microbiol 2(6):1–5
14. Routy B, Le Chatelier E, Derosa L, Duong CP, Alou MT, Daillère R, Fluckiger A, Messaoudene M, Rauber C, Roberti MP et al (2018) Gut microbiome influences efficacy of pd-1-based immunotherapy against epithelial tumors. Science 359(6371):91–97
15. Dettmer K, Aronov PA, Hammock BD (2007) Mass spectrometry-based metabolomics. Mass Spectrom Rev 26(1):51–78
16. Soininen P, Kangas AJ, Würtz P, Suna T, Ala-Korpela M (2015) Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. Circ: Cardiovasc Genet 8(1):192–206
17. Shankar V, Homer D, Rigsbee L, Khamis HJ, Michail S, Raymer M, Reo NV, Paliy O (2015) The networks of human gut microbe-metabolite associations are different between health and irritable bowel syndrome. ISME J 9(8):1899–1903
18. Kostic AD, Gevers D, Siljander H, Vatanen T, Hyötyläinen T, Hämäläinen A-M, Peet A, Tillmann V, Pöhö P, Mattila I et al (2015) The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. Cell Host Microbe 17(2):260–273
19. El Aidy S, Derrien M, Merrifield CA, Levenez F, Doré J, Boekschoten MV, Dekker J, Holmes E, Zoetendal EG, Van Baarlen P et al (2013) Gut bacteria-host metabolic interplay during conventionalisation of the mouse germfree colon. ISME J 7(4):743–755
20. Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami AM, Culhane AC (2016) Dimension reduction techniques for the integrative analysis of multi-omics data. Brief Bioinform 17(4):628–641
21. Chong J, Xia J (2017) Computational approaches for integrative analysis of the metabolome and microbiome. Metabolites 7(4):62
22. Theriot CM, Koenigsknecht MJ, Carlson PE Jr, Hatton GE, Nelson AM, Li B, Huffnagle GB, Li Z, Young VB (2014) Antibiotic-induced shifts in the mouse gut microbiome and metabolome increase susceptibility to clostridium difficile infection. Nat Commun 5(1):3114
23. Mallick H, Franzosa EA, Mclver LJ, Banerjee S, Sirota-Madi A, Kostic AD, Clish CB, Vlamakis H, Xavier RJ, Huttenhower C (2019) Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. Nat Commun 10(1):3136

24. Morton JT, Aksenov AA, Nothias LF, Foulds JR, Quinn RA, Badri MH, Swenson TL, Van Goethem MW, Northen TR, Vazquez-Baeza Y et al (2019) Learning representations of microbe-metabolite interactions. Nat Methods 16(12):1306–1314
25. Zhong W, Zhang T, Zhu Y, Liu JS (2012) Correlation pursuit: forward stepwise variable selection for index models. J R Stat Soc: Ser B (Stat Methodol) 74(5):849–870
26. Chen C-H, Li K-C (1998) Can sir be as popular as multiple linear regression? Stat Sin 8:289–316
27. Li K-C (1991) Sliced inverse regression for dimension reduction. J Am Stat Assoc 86(414):316–327
28. Zhu L, Miao B, Peng H (2006) On sliced inverse regression with high-dimensional covariates. J Am Stat Assoc 101(474):630–643
29. Nicholson JK, Holmes E, Kinross J, Burcelin R, Gibson G, Jia W, Pettersson S (2012) Host-gut microbiota metabolic interactions. Science 336(6086):1262–1267
30. Grice EA, Segre JA (2011) The skin microbiome. Nat Rev Microbiol 9(4):244–253
31. Witten DM, Tibshirani R, Hastie T (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 10(3):515–534
32. Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33(1):1
33. Chen L, Huang JZ (2012) Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. J Am Stat Assoc 107(500):1533–1545
34. Manichanh C, Borruel N, Casellas F, Guarner F (2012) The gut microbiota in ibd. Nat Rev Gastroenterol Hepatol 9(10):599–608
35. Ni J, Wu GD, Albenberg L, Tomov VT (2017) Gut microbiota and ibd: causation or correlation? Nat Rev Gastroenterol Hepatol 14(10):573–584
36. Lavelle A, Sokol H (2020) Gut microbiota-derived metabolites as key actors in inflammatory bowel disease. Nat Rev Gastroenterol Hepatol 17(4):223–237
37. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, Andrews E, Ajami NJ, Bonham KS, Brislawn CJ et al (2019) Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature 569(7758):655–662
38. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome Biol 15:1–21
39. Love M, Anders S, Huber W (2014) Differential analysis of count data-the deseq2 package. Genome Biol 15(550):10–1186
40. Xing J, Blottière H, Jarry A, Segain J, Cherbut C, Laboisse C, Galmiche J (1998) Butyrate and caproate enhance colonic wound repair and modulate the expression of pcna and cyclin d in colonic mucosa of rats with tnbs colitis. Gastroenterology 114:1117
41. Ferreira CM, Vieira AT, Vinolo MAR, Oliveira FA, Curi R, Martins FdS (2014) The central role of the gut microbiota in chronic inflammatory diseases. J Immunol Res 2014:689492
42. Zhuang X, Li T, Li M, Huang S, Qiu Y, Feng R, Zhang S, Chen M, Xiong L, Zeng Z (2019) Systematic review and meta-analysis: short-chain fatty acid characterization in patients with inflammatory bowel disease. Inflamm Bowel Dis 25(11):1751–1763
43. Yu C, Chen Y, Ahmadi S, Wu D, Wu J, Ding T, Liu D, Ye X, Chen S, Pan H (2023) Goji berry leaf exerts a comparable effect against colitis and microbiota dysbiosis to its fruit in dextran-sulfate-sodium-treated mice. Food Funct 14(7):3026–3037
44. Gibson GR, Hutkins R, Sanders ME, Prescott SL, Reimer RA, Salminen SJ, Scott K, Stanton C, Swanson KS, Cani PD et al (2017) Expert consensus document: the international scientific association for probiotics and prebiotics (isapp) consensus statement on the definition and scope of prebiotics. Nat Rev Gastroenterol Hepatol 14(8):491–502
45. Chong J, Liu P, Zhou G, Xia J (2020) Using microbiomeanalyst for comprehensive statistical, functional, and meta-analysis of microbiome data. Nat Protoc 15(3):799–821
46. Rohart F, Gautier B, Singh A, Lê Cao K-A (2017) mixomics: an r package for omics feature selection and multiple data integration. PLoS Comput Biol 13(11):1005752
47. Wang Z, Klipfell E, Bennett BJ, Koeth R, Levison BS, DuGar B, Feldstein AE, Britt EB, Fu X, Chung Y-M et al (2011) Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. Nature 472(7341):57–63
48. Koeth RA, Wang Z, Levison BS, Buffa JA, Org E, Sheehy BT, Britt EB, Fu X, Wu Y, Li L et al (2013) Intestinal microbiota metabolism of l-carnitine, a nutrient in red meat, promotes atherosclerosis. Nat Med 19(5):576–585

49. Fromentin S, Forslund SK, Chechi K, Aron-Wisnewsky J, Chakaroun R, Nielsen T, Tremaroli V, Ji B, Prifti E, Myridakis A et al (2022) Microbiome and metabolome features of the cardiometabolic disease spectrum. Nat Med 28(2):303–314
50. Trøseid M, Andersen GØ, Broch K, Hov JR (2020) The gut microbiome in coronary artery disease and heart failure: current knowledge and future directions. EBioMedicine 52:102649
51. Smits WK, Lyras D, Lacy DB, Wilcox MH, Kuijper EJ (2016) Clostridium difficile infection. Nat Rev Dis Prim 2(1):1–20
52. Cheah IK (1822) Halliwell B (2012) Ergothioneine; antioxidant potential, physiological function and role in disease. Biochim et Biophys Acta Mol Basis Dis 5:784–793
53. Öhrvall M, Vessby B, Sundlöf G (1996) Gamma, but not alpha, tocopherol levels in serum are reduced in coronary heart disease patients. J Intern Med 239(2):111–117