Received: 25 April 2020 | Revised: 31 August 2020 | Accepted: 16 October 2020

DOI: 10.1002/sim.8808

RESEARCH ARTICLE

Statistics in Medicine WILEY

Mediation effect selection in high-dimensional and compositional microbiome data

Haixiang Zhang[1] | **Jun Chen**[2] | **Yang Feng**[3] | **Chan Wang**[4] | **Huilin Li**[4] | **Lei Liu**[5]

[1]Center for Applied Mathematics, Tianjin University, Tianjin, China

[2]Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, Minnesota

[3]Department of Biostatistics, College of Global Public Health, New York University, New York, New York

[4]Division of Biostatistics, Department of Population Health, New York University School of Medicine, New York, New York

[5]Division of Biostatistics, Washington University in St. Louis, St. Louis, Missouri

Correspondence
Lei Liu, Division of Biostatistics, Washington University in St. Louis, St. Louis, MO 63110, USA.
Email: lei.liu@wustl.edu

Funding information
National Center for Advancing Translational Sciences, Grant/Award Number: UL1 TR002345; National Institute of Diabetes and Digestive and Kidney Diseases, Grant/Award Number: R01 DK 110014; National Institute on Aging, Grant/Award Number: R21 AG063370

The microbiome plays an important role in human health by mediating the path from environmental exposures to health outcomes. The relative abundances of the high-dimensional microbiome data have an unit-sum restriction, rendering standard statistical methods in the Euclidean space invalid. To address this problem, we use the isometric log-ratio transformations of the relative abundances as the mediator variables. To select significant mediators, we consider a closed testing-based selection procedure with desirable confidence. Simulations are provided to verify the effectiveness of our method. As an illustrative example, we apply the proposed method to study the mediation effects of murine gut microbiome between subtherapeutic antibiotic treatment and body weight gain, and identify *Coprobacillus* and *Adlercreutzia* as two significant mediators.

KEYWORDS
closed testing, compositional microbiome data, high-dimensional data, isometric log-ratio transformation, mediation analysis

1 | INTRODUCTION

The human microbiome has been linked with complex diseases, such as diabetes, psoriasis, and obesity.[1-4] In particular, gut microbiome plays as a key orchestrator of cancer therapy,[5] especially cancer immunotherapy.[6] Petrosino[7] commented that the microbiome is a key component of precision medicine. More related results on microbiome analysis can be found in the review articles by Li[8] and Xia and Sun.[9]

Mediation analysis has become an important tool in many research fields, for example, behavioral sciences,[10-12] management research,[13] social psychology,[14] psychosomatic medicine,[15] epidemiology,[16] and clinical research.[17] The main goal of mediation analysis is to investigate the role of intermediate variable(s), that is, the mediator(s) that lie in the path between a treatment (or exposure) and an outcome variable. For more related literatures on mediation analysis, we refer to the review articles by Wood et al,[13] MacKinnon et al,[18] Ten Have and Joffe,[19] Preacher,[20] and VanderWeele.[21]

Recently, there have been burgeoning statistical or bioinformatical research devoted to studying the mediation effects of microbiome. For instance, Zhang et al[22] proposed a distance-based approach for testing the mediation effect of the human microbiome. Sohn and Li[23] proposed a sparse compositional mediation model in the simplex space and applied it to a gut microbiome study. Wang et al[24] proposed a rigorous sparse microbial causal mediation model for the high-dimensional and compositional microbiome data. Zhang et al[25] adopted the isometric log-ratio *(ilr)*-transformation and debiased Lasso techniques to develop a joint significance test for the mediation effect of human gut microbiome with a focus on prespecified taxa. In this work, we propose a novel method to select mediating microbial taxa. As existing methods for microbiome mediation analysis are primarily designed to test the overall mediation effect (eg, Sohn and Li[23] and Wang et al[24]), our major contribution is to propose a statistical procedure to select individual taxa that mediate the path between exposure and phenotype.

The remainder of this article is organized as follows: In Section 2, we present the model and method, including *ilr*-transformation formulas for relative abundances, high-dimensional inference for the linear mediation model in the Euclidean space, and a closed testing-based selection procedure for the mediation effects of *ilr*-transformed mediators. In Section 3, we use simulation studies to check the performance of the proposed method. In Section 4, we provide an application to a murine gut microbiome study. Concluding remarks are reported in Section 5.

2 | METHODS

2.1 | Log-ratio transformation for the relative abundances

Suppose that there are a total of d taxa in the microbiome for each sample, where the relative abundances are denoted by a vector $\mathbf{M} = (M_1, \ldots, M_d)'$. The d-part composition $\mathbf{M}$ lies in the "simplex" space,[26] which is given as

$$S^d = \left\{ \mathbf{x} = (x_1, \ldots, x_d)' : \sum_{k=1}^{d} x_k = 1; x_k > 0, k = 1, \ldots, d \right\}.$$

Compositions are subject to two constraints: the components are positive in (0, 1), and sum up to one. Due to the compositional characteristic of relative abundances, many statistical models in the real Euclidean space are inappropriate. To address this issue, additive log-ratio (*alr*) and centered log-ratio (*clr*)-transformations are commonly used.[27,28] However, for linear regression with compositions as covariates, a linear constraint is needed for the estimator of regression coefficients,[23,24] which poses a serious impediment to developing appealing optimization methods and theoretical results.

Alternatively, Egozcue et al[29] suggested the *ilr* transformation by transforming the compositional data from the simplex S^d to the Euclidean space $\mathbb{R}^{d-1}$. The *ilr*-based transformations on the compositional mediators $M_1, \ldots, M_d$ are

$$\tilde{M}_k = \sqrt{\frac{d-k}{d-k+1}} \ln \frac{M_k}{\sqrt[d-k]{\prod_{j=k+1}^{d} M_j}}, k = 1, \ldots, d-1. \tag{1}$$

One key advantage of the *ilr*-based transformation is that we can directly fit these *ilr*-transformed variables with linear models. Hence, many statistical methods for linear models can be used directly to analyze microbiome data.

Without loss of generality, we assume the response Y and $\tilde{M}_k$ are centered hereafter. We can fit a high-dimensional linear mediation model in the Euclidean space (Figure 1),

$$\begin{aligned} E(Y|X, Z, \tilde{\mathbf{M}}) &= \gamma X + \beta_1 \tilde{M}_1 + \cdots + \beta_{d-1} \tilde{M}_{d-1} + \mathbf{Z}'\boldsymbol{\theta}, \\ E(\tilde{M}_k|X, Z) &= \alpha_k X + \mathbf{Z}'\boldsymbol{\eta}_k, \quad k = 1, \ldots, d-1, \end{aligned} \tag{2}$$

where $\tilde{\mathbf{M}} = (\tilde{M}_1, \ldots, \tilde{M}_{d-1})'$, X is an exposure, $\mathbf{Z} = (Z_1, \ldots, Z_q)'$ is a vector of covariates; γ represents the *direct effect* of X on Y adjusting for $\{\tilde{M}_k; k = 1, \ldots, d-1\}$ and $\mathbf{Z}$; α_k represents the relation between X and the mediator $\tilde{M}_k$ adjusting for $\mathbf{Z}$; β_k represents the relation between $\tilde{M}_k$ and Y adjusting for the effects of X, $\mathbf{Z}$ and other mediators; $\alpha_k\beta_k$ denotes

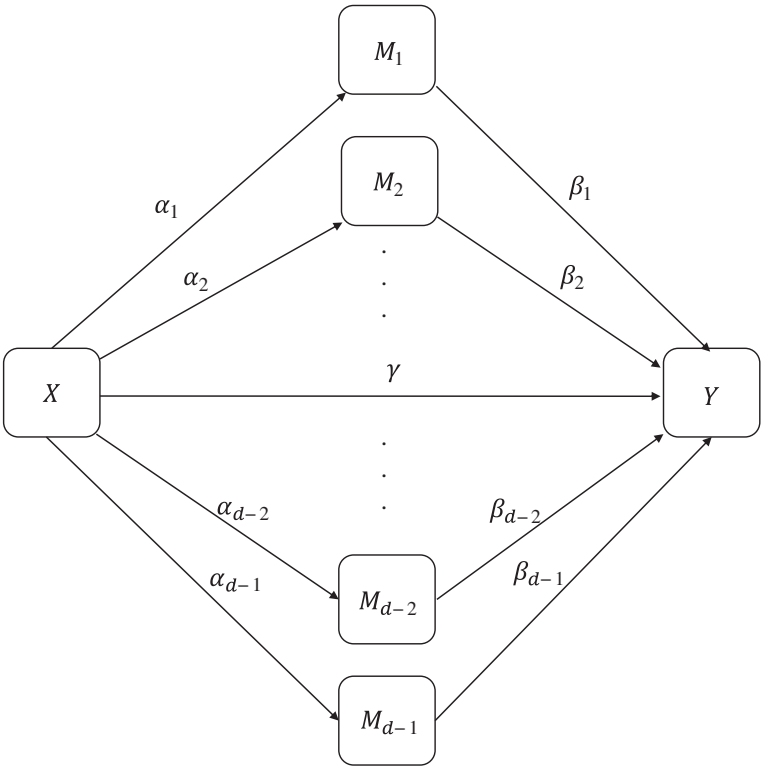

FIGURE 1 A scenario of high-dimensional mediation model with *ilr*-transformed mediators (confounding variables omitted)

the *indirect effect* of X on Y that is transmitted through the mediator $\tilde{M}_k$; $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)'$ and $\boldsymbol{\eta}_k = (\eta_{k1}, \ldots, \eta_{kq})'$ are regression coefficients for $\boldsymbol{Z}$ for $k = 1, \ldots, d-1$.

The transformed mediator $\tilde{M}_1$ is a scaled sum of all log-ratios of the original composition part M_1 and the other parts $M_2, \ldots, M_d$, where the linear relationship is described by

$$\begin{aligned} \tilde{M}_1 &= \frac{1}{\sqrt{d(d-1)}} \left(\ln \frac{M_1}{M_2} + \cdots + \ln \frac{M_1}{M_d} \right) \\ &= \sqrt{\frac{d-1}{d}} \ln \frac{M_1}{\sqrt[d-1]{\prod_{j=2}^{d} M_j}}. \end{aligned} \tag{3}$$

It can be seen that $\tilde{M}_1$ extracts all relative information of M_1 and captures the relative contribution of M_1 with respect to the geometric mean of the remaining parts in the composition.[28] From this point of view, the term $\alpha_1\beta_1$ can describe the "relative effect" rather than "absolute effect" of the original composition M_1. Of note, apart from the coefficient, the *ilr* transformation of the first composition M_1 is equivalent to the *clr* transformation

$$\ln \frac{M_1}{\sqrt[d]{\prod_{j=1}^{d} M_j}} = \frac{d-1}{d} \ln \frac{M_1}{\sqrt[d-1]{\prod_{j=2}^{d} M_j}}.$$

For more interpretations of the "relative effect" of compositions, we refer to section 2.3 of Zhang et al.[25]

2.2 | High-dimensional inference

In Model (2), the interpretation of $\tilde{M}_2, \ldots, \tilde{M}_{d-1}$ are not straightforward, because M_1 is not contained therein. Hence, if we are interested in understanding the mediation effects from taxa M_ℓ, $\ell \in \{2, \ldots, d\}$, we can reorder M_ℓ to play the role of M_1 as $(M_\ell, M_1, \ldots, M_{\ell-1}, M_{\ell+1}, \ldots, M_d)'$ to interpret the effect of M_ℓ. For $\ell = 1, \ldots, d$, we propose the following

ilr-transformation-based linear mediation models as

$$\begin{aligned} E(Y|X,Z,\tilde{\mathbf{M}}^{[\ell]}) &= \gamma^{[\ell]}X+\beta_1^{[\ell]}\tilde{M}_1^{[\ell]}+\cdots+\beta_{d-1}^{[\ell]}\tilde{M}_{d-1}^{[\ell]}+\mathbf{Z}'\boldsymbol{\theta}^{[\ell]},\\ E(\tilde{M}_k^{[\ell]}|X,Z) &= \alpha_k^{[\ell]}X+\mathbf{Z}'\boldsymbol{\eta}_k^{[\ell]},\quad k=1,\ldots,d-1, \end{aligned}$$

where $\tilde{\mathbf{M}}^{[\ell]}=(\tilde{M}_1^{[\ell]},\ldots,\tilde{M}_{d-1}^{[\ell]})'$, and the other notations are defined similarly to those in (2). As mentioned before, $\alpha_1^{[\ell]}\beta_1^{[\ell]}$ is an interpretable mediation effect term related to the ℓth taxon, for $\ell=1,\ldots,d$. Consider the multiple testing problem:

$$H_{0\ell}:\alpha_1^{[\ell]}\beta_1^{[\ell]}=0\quad\text{versus}\quad H_{A\ell}:\alpha_1^{[\ell]}\beta_1^{[\ell]}\neq 0\ \text{for}\ \ell=1,\ldots,d. \tag{4}$$

Assume that $(Y_i,X_i,\mathbf{Z}_i,\mathbf{M}_i)$ are independently and identically distributed (i.i.d.) observations, $i=1,\ldots,n$. For constructing valid P-values, let $P_{\alpha_1^{[\ell]}}=2\{1-\Phi(|\hat{\alpha}_1^{[\ell]}|/\hat{\sigma}_{\alpha_1^{[\ell]}})\}$ and $P_{\beta_1^{[\ell]}}=2\{1-\Phi(|\hat{\beta}_1^{[\ell]}|/\hat{\sigma}_{\beta_1^{[\ell]}})\}$, where $\Phi(x)$ is the cumulative distribution function of $N(0,1)$, $\hat{\alpha}_1^{[\ell]}$ and $\hat{\sigma}_{\alpha_1^{[\ell]}}$ are based on the ordinary least squares method. $\hat{\beta}_1^{[\ell]}$ and $\hat{\sigma}_{\beta_1^{[\ell]}}$ can be obtained through the following debiased Lasso estimate.

First, we calculate the Lasso estimate by

$$\left(\tilde{\gamma}^{[\ell]},\tilde{\boldsymbol{\beta}}^{[\ell]},\tilde{\boldsymbol{\theta}}^{[\ell]}\right)=\arg\min_{\gamma,\boldsymbol{\beta},\boldsymbol{\theta}}\left\{\frac{1}{2n}\sum_{i=1}^{n}\left(Y_i-\gamma^{[\ell]}X_i-\sum_{j=1}^{d-1}\beta_j^{[\ell]}\tilde{M}_{ij}^{[\ell]}-\sum_{j=1}^{q}\theta_j^{[\ell]}Z_{ij}\right)^2+\lambda\sum_{j=1}^{d-1}|\beta_j^{[\ell]}|\right\}, \tag{5}$$

where $\lambda>0$ is a penalty parameter, which can be determined using 10-fold cross-validation. By Zhang and Zhang,[30] the debiased Lasso estimator of $\beta_1^{[\ell]}$ is

$$\hat{\beta}_1^{[\ell]}=\tilde{\beta}_1^{[\ell]}+\frac{\sum_{i=1}^{n}R_i^{[\ell]}(Y_i-\tilde{\gamma}^{[\ell]}X_i-\sum_{j=1}^{d-1}\tilde{\beta}_j^{[\ell]}\tilde{M}_{ij}^{[\ell]}-\sum_{j=1}^{q}\tilde{\eta}_j^{[\ell]}Z_{ij})}{\sum_{i=1}^{n}R_i^{[\ell]}\tilde{M}_{i1}^{[\ell]}},$$

where $\tilde{\gamma}^{[\ell]}$, $\tilde{\boldsymbol{\beta}}^{[\ell]}$ and $\tilde{\boldsymbol{\theta}}^{[\ell]}$ are defined in (5); $R_i^{[\ell]}=\tilde{M}_{i1}^{[\ell]}-\hat{\phi}_1^{[\ell]}X_i-\sum_{j=2}^{d-1}\hat{\phi}_j^{[\ell]}\tilde{M}_{ij}^{[\ell]}-\sum_{j=1}^{q}\hat{\phi}_{d-1+j}^{[\ell]}Z_{ij}$ is the residual from a Lasso regression of $\tilde{M}_{i1}^{[\ell]}$ versus X_i, $\mathbf{Z}_i$ and $\tilde{\mathbf{M}}_{ij}^{[\ell]}$, $i=1,\ldots,n, j=2,\ldots,d-1$; and $\hat{\boldsymbol{\phi}}^{[\ell]}=(\hat{\phi}_1^{[\ell]},\ldots,\hat{\phi}_{d+q-1}^{[\ell]})'$ is the Lasso solution from

$$\hat{\boldsymbol{\phi}}^{[\ell]}=\arg\min_{\boldsymbol{\phi}}\left\{\frac{1}{2n}\sum_{i=1}^{n}\left(\tilde{M}_{i1}^{[\ell]}-\phi_1^{[\ell]}X_i-\sum_{j=2}^{d-1}\phi_j^{[\ell]}\tilde{M}_{ij}^{[\ell]}-\sum_{j=1}^{q}\phi_{d-1+j}^{[\ell]}Z_{ij}\right)^2+\lambda^*\sum_{j=1}^{d+q-1}|\phi_i^{[\ell]}|\right\},$$

where $\lambda^*>0$ is a penalty parameter determined by 10-fold cross-validation as for λ.

It follows from Zhang and Zhang[30] that the standard error for $\hat{\beta}_1^{[\ell]}$ is

$$\hat{\sigma}_{\beta_1^{[\ell]}}=n^{-1/2}\frac{\hat{\sigma}_\epsilon^{[\ell]}\sqrt{\sum_{i=1}^{n}(R_i^{[\ell]})^2/n}}{|\sum_{i=1}^{n}R_i^{[\ell]}\tilde{M}_{i1}^{[\ell]}/n|},$$

where $(\hat{\sigma}_\epsilon^{[\ell]})^2=\sum_{i=1}^{n}(Y_i-\tilde{\gamma}^{[\ell]}X_i-\sum_{j=1}^{d-1}\tilde{\beta}_j^{[\ell]}\tilde{M}_{ij}^{[\ell]}-\sum_{j=1}^{q}\tilde{\theta}_j^{[\ell]}Z_{ij})^2/(n-\hat{s})$, and $\hat{s}$ is the number of nonzero coefficients in the Lasso estimator $\tilde{\boldsymbol{\beta}}^{[\ell]}$.

Based on the idea of *joint significance test*,[11,31] the raw (unadjusted) P-values for (4) are given as follows,

$$P_{\text{raw},\ell}=\max\{P_{\alpha_1^{[\ell]}},P_{\beta_1^{[\ell]}}\},\quad \ell=1,\ldots,d. \tag{6}$$

To adjust for multiple testing, one possible solution is the Bonferroni criterion, that is, $P_{\text{raw},\ell}<0.05/d$. However, this Bonferroni-based method suffers from vanishing power in the case of large-scale multiple hypothesis testing.[32] In the

following section, we will propose a closed testing-based selection procedure adapting the method of Goeman et al,[32] which can effectively control false discovery proportions (FDPs) when the number of hypotheses goes to infinity.

2.3 | Closed testing-based mediation effect selection procedure

In this section, we focus on developing a selection procedure for mediation effects. Denote by $\{H_{0\ell}\}_{\ell=1}^{d}$ the collection of hypotheses of interest (*elementary hypotheses*) in Equation (4), $T_0 \subseteq \{1, \dots, d\}$ is the index set of true null hypotheses. The closed testing methods[33,34] consider not only the elementary hypotheses, but also all intersection hypotheses of the form $H_A = \bigcap_{i\in A} H_{0i}$, where $A \subseteq \{1, \dots, d\}$ and $A \neq \emptyset$. An intersection hypothesis H_A is true if and only if H_i is true for all $i \in A$. Let $\mathcal{A}$ be the collection of all subsets of $\{1, \dots, d\}$. We define $\mathcal{U}_{0.05}$ as the collection of all $A \in \mathcal{A}$ such that H_A is rejected, where $P(T_0 \notin \mathcal{U}_{0.05}) \geq 0.95$ (or equivalently $P(T_0 \in \mathcal{U}_{0.05}) \leq 0.05$). For any $S \subseteq \{1, \dots, d\}$ of selected hypotheses to reject (referred to as discoveries), Goeman et al[32] provided a simultaneous 0.95-confidence lower-bound, $LB_{0.05}(S)$, for the number of true discoveries in S. Below we summarize the procedure of Goeman et al[32] in Algorithm 1.

Algorithm 1. Closed Testing-Based Algorithm

Step 1: Obtain the raw (unadjusted) P-values $P_1, \cdots, P_d$ for the elementary hypotheses $H_{01}, \cdots, H_{0d}$. Sort these P-values in the increasing order as $P_{r_1} \leq P_{r_2} \leq \cdots \leq P_{r_d}$, where $r_i \in \{1, \cdots, d\}$. Let $R_i = \{r_1, \cdots, r_i\}$ be the set of the smallest i P-values. Similarly define $R_{d-i} = \{r_1, \cdots, r_{d-i}\}$ to be the set of the smallest $d - i$ P-values. Denote by $L_i = \{1, \cdots, d\} \backslash R_{d-i}$ the set of the largest i P-values, as in eq. (8) of Goeman et al[32]

Step 2: Calculate the 0.95-confidence lower-bound for the number of true discoveries in S,

$$LB_{0.05}(S) = \max_{1\leq k\leq |S|} 1 - k + |\{i \in S : h_{0.05} P_i \leq 0.05k\}|, \tag{7}$$

where $|S|$ denotes the number of elements in S, and $h_{0.05} = \max\{1 \leq i \leq d : L_i \notin \mathcal{U}_{0.05}\}$ denotes the size of the set of largest P-values not rejected.

Of note, the term $LB_{0.05}(S)$ is given in theorem 1 of Goeman et al,[32] which can provide a 0.95-confidence lower-bound for the number of true discoveries in any set S. For example, if $S = \{1, 2, 3, 4, 5\}$ and $LB_{0.05}(S) = 3$, we can say that there are at least 3 true discoveries in S with 0.95-confidence. Here, we point out that the lower-bound (7) is exact and applicable to all S, and we can easily obtain $LB_{0.05}(S)$ using the R package `hommel` for practical applications.

Let $\mathcal{S}_0 = \{\ell : \alpha_1^{[\ell]} \beta_1^{[\ell]} \neq 0,\ \ell = 1, \dots, d\}$ be the index set of significant mediators. Below, we propose a novel mediation effect selection procedure. Our basic idea is to sort the P-values $\{P_{\text{raw},\ell}\}_{\ell=1}^{d}$ in (6) as $P_{\text{raw},r_1} \leq P_{\text{raw},r_2} \leq \cdots P_{\text{raw},r_d}$. Let $K = \max\{i : P_{\text{raw},r_i} \leq 0.05,\ i = 1, \dots, d\}$, and $S_1 = \{r_1\}, S_2 = \{r_1, r_2\} \cdots S_K = \{r_1, r_2, \dots, r_K\}$. Based on (7), we know that $LB_{0.05}(S_1) \leq LB_{0.05}(S_2) \leq \cdots \leq LB_{0.05}(S_K)$. It is reasonable to regard M_{r_k} as a significant mediator if there is a jump of size one from $LB_{0.05}(S_{k-1})$ to $LB_{0.05}(S_k)$, that is, the corresponding mediator M_{r_k} leads to an increase of the 0.95-confidence lower-bound of true discoveries. More specifically, we show our proposed mediation effect selection procedure in Algorithm 2.

Algorithm 2. Mediation Effect Selection Algorithm

Step 1: Sort the P-values $\{P_{\text{raw},\ell}\}_{\ell=1}^{d}$ in (6) as $P_{\text{raw},r_1} \leq P_{\text{raw},r_2} \leq \cdots \leq P_{\text{raw},r_d}$. Let $S_1 = \{r_1\}, S_2 = \{r_1, r_2\}, \cdots, S_K = \{r_1, r_2, \cdots, r_K\}$, and $K = \max\{i : P_{\text{raw},r_i} \leq 0.05, i = 1, \cdots, d\}$.

Step 2: Run Algorithm 1 to obtain the values of $LB_{0.05}(S_k)$ in (7), for $k = 1, \cdots, K$.

Step 3: Define $J_1 = LB_{0.05}(S_1)$ and $J_k = LB_{0.05}(S_k) - LB_{0.05}(S_{k-1})$, where $k = 2, \cdots, K$. The estimated index set of significant mediators is

$$\hat{S} = \{r_k : J_k = 1, \text{ for } k = 1, \cdots, K\}. \tag{8}$$

Remark 1. In Algorithm 2, sorting the P-values $\{P_{\text{raw},\ell}\}_{\ell=1}^{d}$ can save computation time as only K values of $LB_{0.05}(S)$ are calculated. Otherwise, a total of d lower-bounds $LB_{0.05}(S)$ need to be calculated. Hence, our method has a computational advantage, especially when the value of K/d is small.

In summary, we first impose the *ilr*-transformation on the relative abundances, then we refit the *ilr*-transformed variables in the linear mediation models. Because only the first element of the *ilr*-transformed variables is interpretable, we permute the orders of the original d mediators in turn to ensure that each taxon should play the role of the first element. In the structural equation modeling (SEM) framework, we obtain the raw P-values by joint significant tests for the component-wise mediation effects. Furthermore, we apply a novel closed testing-based selection method for the *ilr*-transformed high-dimensional mediators.

3 | NUMERICAL SIMULATION

In this section, we conduct some simulations to check the performance of our proposed method. For simplicity and concentration, we do not consider covariates in the simulation, although our `microHIMA` package has the capacity for covariate adjustment.

3.1 | Simulation study 1

To begin with, we introduce some compositional operators as follows. For two compositions $\boldsymbol{\eta}, \boldsymbol{\zeta} \in \mathbb{S}^d$, the perturbation operator is defined by

$$\boldsymbol{\eta} \oplus \boldsymbol{\zeta} = \left(\frac{\eta_1 \zeta_1}{\sum_{j=1}^{d} \eta_j \zeta_j}, \ \cdots \ , \frac{\eta_d \zeta_d}{\sum_{j=1}^{d} \eta_j \zeta_j} \right)'. \tag{9}$$

Of note, for three compositions $\boldsymbol{\eta}$, $\boldsymbol{\omega}$ and $\boldsymbol{\zeta} \in \mathbb{S}^d$, $\boldsymbol{\eta} \oplus \boldsymbol{\omega} \oplus \boldsymbol{\zeta} = (\boldsymbol{\eta} \oplus \boldsymbol{\omega}) \oplus \boldsymbol{\zeta}$. Moreover, the power transformation for a composition $\boldsymbol{\eta}$ by a scalar v is given as

$$\boldsymbol{\eta}^v = \left(\frac{\eta_1^v}{\sum_{j=1}^{d} \eta_j^v}, \ \cdots \ , \frac{\eta_d^v}{\sum_{j=1}^{d} \eta_j^v} \right)'. \tag{10}$$

First we generate data from the compositional mediation model in Sohn and Li,[23]

$$\mathbf{M} = \mathbf{m}_0 \oplus \mathbf{a}^X \oplus \mathbf{e}, \tag{11}$$

$$Y = c_0 + cX + (\log \mathbf{M})' \mathbf{b} + \epsilon, \tag{12}$$

where $\sum_{i=1}^{d} b_i = 0$, the baseline composition $\mathbf{m}_0$ is from the standard uniform distribution, Unif(0,1), under the unit-sum constraint. We randomly generate the exposure X from $N(0,1)$, and $\mathbf{a}^X$ is defined in (10). The compositional distribution $\mathbf{e}$ is generated from a multivariate logistic normal distribution with mean zero and variance Σ_e, $c_0 = 1$ and $c = 0.5$. We consider the following three settings:

Case I: We set $\mathbf{a} = (1/3, 1/4, 1/5, a_4, \ \cdots \ , a_d)'$, and $a_i = \frac{13u_i}{60 \sum_{i=4}^{d} u_i}$ for $i = 4, \ \cdots \ , d$, where u_i follows from $U(0,1)$; and $\mathbf{b} = (1.3, -0.7, -0.6, 0, \ \cdots \ , 0)'$. Following Sohn and Li,[23] let $\Sigma_e = 2\boldsymbol{\mathcal{N}}$ with $\boldsymbol{\mathcal{N}} = \boldsymbol{\mathcal{I}}_{d-1} + \mathbf{1}_{d-1}\mathbf{1}_{d-1}'$, where $\boldsymbol{\mathcal{I}}_{d-1}$ is the $(d-1) \times (d-1)$ identity matrix, and $\mathbf{1}_{d-1} = (1, \ \cdots \ , 1)'$. The compositional regression disturbance ϵ follows from a normal distribution with mean 0 and variance 2. Let $S_0 = \{1, 2, 3\}$ denote the index set of significant mediators.

Case II: The setting is the same as Case I, except that $\Sigma_e = \boldsymbol{\mathcal{N}}$ and ϵ is generated from the t_3 distribution, that is, t distribution with 3 df.

TABLE 1 Accuracy of mediation effect selection in simulation study 1[†]

		$d = 50$				$d = 100$			
	Methods	**MS**	**CMR**	**FPR**	**FDP**	**MS**	**CMR**	**FPR**	**FDP**
Case I	Proposed	3.004	0.996	8.5×10^{-5}	0.0010	3.002	0.998	2.1×10^{-5}	5×10^{-4}
	B-H	3.110	0.906	0.0023	0.0256	3.102	0.906	0.0011	0.0246
	CMM^a	6.22	0.05	0.0685	0.4784	11.08	0	0.0833	0.7121
	CMM^b	3.36	0.70	0.0077	0.0835	4.76	0.21	0.0181	0.3131
Case II	Proposed	2.984	0.974	0.0001	0.0015	2.968	0.972	2.1×10^{-5}	5×10^{-4}
	B-H	3.078	0.914	0.0019	0.0210	3.086	0.896	0.0011	0.0268
	CMM^a	6.66	0.02	0.0779	0.5129	11.33	0	0.0859	0.7213
	CMM^b	3.96	0.39	0.0204	0.1982	6.29	0.04	0.0339	0.4778
Case III	Proposed	0.002	0.998	4×10^{-5}	0.0020	0.002	0.998	2×10^{-5}	0.0020
	B-H	0.032	0.968	0.0006	0.0320	0.016	0.984	0.0002	0.0160
	CMM^a	4.59	0.01	0.0918	0.99	9.05	0	0.0905	1
	CMM^b	0.80	0.40	0.0160	0.60	2.37	0.10	0.0237	0.9

[†]"Proposed" denotes our method in Algorithm 2; "B-H" denotes the classical B-H algorithm in Benjamini and Hochberg,[35] adopted by Zhang et al;[25] "CMM^a" and "CMM^b" denote the methods in Sohn and Li;[23] "MS" denotes the mean model size; "CMR" denotes the proportion of times selecting the correct model $\mathcal{S}_0$; "FPR" and " FDP" denote the false positive rate, and the false discovery proportion.

Case III: The setting is the same as Case I, except that $\mathbf{b} = (0, \ldots, 0)'$ with $\mathcal{S}_0 = \emptyset$. There are no significant mediators in this case.

Let $\hat{\boldsymbol{S}}$ be the estimated index set of the significant mediators ($\mathcal{S}_1$) based on Algorithm 2. To evaluate the performance of mediation effect selection, we record: the model size (MS), $|\hat{\boldsymbol{S}}|$; the rate that the correct model is selected (CMR), $I(\hat{\boldsymbol{S}} = \mathcal{S}_1)$; the false positive rate (FPR), $|\hat{\boldsymbol{S}} \setminus \mathcal{S}_1|/(d - |\boldsymbol{S}_1|)$, where $\hat{\boldsymbol{S}} \setminus \mathcal{S}_1$ denotes the set difference of $\hat{\boldsymbol{S}}$ and $\mathcal{S}_1$; the FDP, $|\hat{\boldsymbol{S}} \setminus \mathcal{S}_1|/|\hat{\boldsymbol{S}}|$.

For comparison, we also consider the classical B-H algorithm[35] which adjusts the P-values in Equation (6) directly, as in Zhang et al.[25] Moreover, let CI_j be the $100(1-\alpha)\%$ bootstrap confidence interval (CI; 1000 bootstrap samples) for the jth component-wise mediation effect in Sohn and Li,[23] $j = 1, \ldots, d$. The selected index set of the significant mediators by Sohn and Li[23] is

$$\hat{\boldsymbol{S}} = \{j : 0 \notin CI_j, \text{for } j = 1, \ldots, d\}. \tag{13}$$

For fair competition, we set $\alpha = 0.05$ (denoted as CMM^a) and $\alpha = 0.05/d$ (denoted as CMM^b) for the $100(1-\alpha)\%$ CI proposed by Sohn and Li,[23] where the computation procedure is available from the R package `ccmm` (https://CRAN.R-project.org/package=ccmm). For the "Proposed" and "B-H" methods, the computations are based on 500 replicates, while results from CMM^a and CMM^b are based on 100 replicates in view of their computational burden. For all cases, the sample size is $n = 200$. Table 1 indicates that our method tends to select a smaller model, and has notable advantages over B-H algorithm[25,35] in the performance of CMR, FPR and FDP. Compared with Sohn and Li,[23] our method has substantial advantages in all four criteria under consideration.

3.2 | Simulation study 2

We generate microbial relative abundances by the method of Wang et al,[24] using the Dirichlet regression to model the microbial relative abundance as a function of exposure. For $i = 1, \ldots, n$, we assume that $\mathbf{M}_i|X_i \sim$ Dirichlet$(\omega_1(X_i), \ldots, \omega_d(X_i))$, and their microbial relative means are linked with exposure in the generalized linear model fashion:

TABLE 2 Accuracy of mediation effect selection in simulation study 2 with $d = 36$ [†]

	Methods	MS	CMR	FPR	FDP
Case A	Proposed	3	1	0	0
	B-H	3.004	0.998	0.0001	0.0008
	SparseMCMM^a	20.54	0	0.53	0.85
	SparseMCMM^b	15.10	0	0.37	0.79
Case B	Proposed	0	1	0	0
	B-H	0.008	0.992	0.0002	0.0080
	SparseMCMM^a	19.30	0	0.54	1
	SparseMCMM^b	13.88	0	0.39	1

[†]"Proposed" denotes our method in Algorithm 2; "B-H" denotes the classical B-H algorithm in Benjamini and Hochberg,[35] adopted by Zhang et al;[25] "SparseMCMM^a" and "SparseMCMM^b" denote the methods in Wang et al;[24] "MS" denotes the mean model size; "CMR" denotes the proportion of times selecting the correct model S_0; "FPR" and " FDP" denote the false positive rate, and the false discovery proportion.

$$E(M_{ij}) = \frac{\omega_j(X_i)}{\sum_{m=1}^{d} \omega_m(X_i)},$$
$$\log\{\omega_j(X_i)\} = \phi_{0j} + \phi_{X_j} X_i, \quad \text{for } j = 1, \ldots, d,$$

where ϕ_{0j} represents the log-transformed baseline relative abundance for the jth taxon, and ϕ_{X_j} is the coefficient of exposure for the jth taxon. Let $\boldsymbol{\phi}_0 = (\phi_{01}, \ldots, \phi_{0d})'$. Following Wang et al,[24] we set $\boldsymbol{\phi}_0$ as the corresponding estimates from the murine microbiome experiment data in Section 4, that is, $\boldsymbol{\phi}_0 = (-1.771, -1.682, -1.587, -1.130, -1.747, -1.861, -1.873, -1.642, -1.543, -1.265, -1.788, 2.278, -1.618, -1.137, -1.559, -1.344, -1.462, -1.568, -0.522, -1.078, -1.584, -1.471, -1.680, -1.711, -1.757, -0.135, -1.325, -1.541, -0.961, -0.565, -1.612, -1.702, -0.578, -1.883, 0.072, -1.684)'$. We set $\boldsymbol{\phi}_X = (1, 1.2, 1.5, 0, \ldots, 0)'$ with $d = 36$ (the number of taxa in Section 4). We consider the following two situations:

Case A: The exposure X_i follows from binomial distribution $B(1, 0.5)$. The response Y is generated from Model (12), where $c_0 = 0$, $c = 1$, and $\mathbf{b} = (3, -1.5, -1.5, 0, \ldots, 0)'$; ϵ is generated from $N(0, 1)$.

Case B: As suggested by a reviewer, we consider a setting with $\mathbf{b} = (1.3, -0.7, -0.6, 0, \ldots, 0)'$, and $\boldsymbol{\phi}_X = (0, 0, 0, \phi_{X_4}, \ldots, \phi_{X_d})'$, where $\phi_{X_j} = \frac{u_j}{\sum_{j=4}^{d} u_j}$, and u_j follows from $U(0, 1)$. Other parameters are set the same as in Case A. In Case B there are no significant mediators, that is, $S_0 = \emptyset$, representing a "null hypothesis" setting for our method.

Similar to Simulation study 1, we run the method in Zhang et al[25] using the classical B-H algorithm[35] to adjust the P-values in Equation (6). Furthermore, we consider the sparse MCMM methods in Wang et al.[24] Let CI_j be the $100(1-\alpha)\%$ bootstrap CI (100 bootstrap samples) for the jth component-wise mediation effect in Wang et al,[24] where the resulting $\hat{S}$ is similarly given as (13). Denote SparseMCMM^a and SparseMCMM^b for the situations with $\alpha = 0.05$ and $\alpha = 0.05/d$, respectively. Of note, the results for Wang et al[24] are obtained via the R package `SparseMCMM` (https://github.com/chanw0/SparseMCMM).

For the "Proposed" and "B-H" methods, the computations are based on 500 replicates. In view of the computation burden of Wang et al,[24] the simulations of SparseMCMM^a and SparseMCMM^b are based on 50 replicates. For all methods, the sample size is chosen as $n = 300$. We run Algorithm 2 to get the estimated index set $\hat{S}$ of significant mediators. We present the MS, CMR, FPR, and FDP for our method, B-H algorithm and Wang et al[24] in Table 2.

From the results in Table 2, we can see that among all 500 replicates, our method correctly identifies all 3 mediators in Case A, and 0 mediators under the "null hypothesis" in Case B, demonstrating the excellent performance of our method. By contrast, the B-H algorithm[35] implemented in Zhang et al[25] performs slightly worse. In addition, our method is superior over SparseMCMM^a and SparseMCMM^b in terms of selecting significant mediators accurately.

TABLE 3 Summary results of potential mediating taxa[†]

Genus	Proposed	CMM^b	SparseMCMM^b
Coprobacillus	P-value $= 9 \times 10^{-7}$	CI = [0.3956, 2.6055]	*
Adlercreutzia	P-value $= 6 \times 10^{-6}$	*	CI = [−0.0893, −0.1940]

[†]"Proposed" denotes our method in Algorithm 2; CMM^b and SparseMCMM^b are defined in Tables 1 and 2, respectively; P-value is given in (6); CI is the $100(1 - 0.05/d)\%$ confidence interval for the component-wise mediation effect; * denotes that the corresponding causal taxon is not selected out.

4 | APPLICATION TO MICROBIOME DATA

We apply our method to a murine microbiome experiment,[36] where the DNAs were extracted from fecal samples using the 96-well MO BIO PowerSoil DNA Isolation Kit by targeting the V4 region of the bacterial 16S rRNA gene. We focus on 36 male mice, where the taxa table was constructed using the QIIME pipeline[37] on day 28. There were originally 149 genera. The number of taxa in the microbiome dataset is high-dimensional, with high absence rates of many taxa across samples. We remove the taxa that appear in less than 10% of the mice with mean relative abundance less than 10^{-4}, leaving 36 taxa for each sample in our analysis. Since the number of sequencing reads varies greatly across samples, the count data are transformed into compositions after zero counts are replaced by the maximum rounding error 0.5.[24,38] The observed body weight (in grams) prior to sacrifice, that is, on day 116 for the male mice, is taken as the outcome. We consider subtherapeutic antibiotic treatment ($X = 1$) and control group ($X = 0$) as the exposure. Moreover, we assume that all potential confounders were well controlled in the randomized experiment. Our interest is to select significant gut microbial taxa that play the mediating role between treatment and body weight gain.

In Table 3, we report the summary results of two potential mediating taxa, which are survived by our closed testing-based method in Algorithm 2. For the identified taxa, *Coprobacillus* and *Adlercreutzia*, we also give the corresponding CIs based on Sohn and Li[23] and Wang et al,[24] respectively. By our method, for *Coprobacillus*, the estimated pathway effect (α) on $X \to M$ is −3.04 (SE 0.44); the estimated pathway effect (β) on $M \to Y$ is −34.92 (SE 7.11); for *Adlercreutzia*, the pathway effects on $X \to M$ and $M \to Y$ are 0.96 (SE 0.21) and −15.56 (SE 3.18), respectively. Hence, *Coprobacillus* has a positive mediation effect, while *Adlercreutzia* has a negative mediation effect. These conclusions are consistent with the directions of mediation effects estimated by CMM^b and SparseMCMM^b (Table 3).

5 | CONCLUDING REMARKS

We have proposed a novel closed testing-based selection method for the *ilr*-transformed high-dimensional mediators. Simulations and a real data example are provided to illustrate the validity and applicability of the method. Specifically, numerical studies indicated that our proposed method is more accurate than the FDR-based procedure toward mediator selection. One possible explanation for this phenomenon is that the well-known Benjamini-Hochberg procedure[35] controls the FDR under independence or positive-dependence structure.[39] However, the P-values may have more complicated dependence structure in microbiome studies. We also note that our method has some computationally advantages since we do not need to rely on bootstrapping for inference.

The proposed method focuses on the high-dimensional microbiome sequencing data. It will take all the available microbes' relative abundances into the modeling and use the penalization technique to select the important microbes. We are thus able to conduct the microbiome-wide quantification to identify microbes that play important mediating roles in linking the treatment/exposure and human phenotype/response. By contrast, a confirmative hypothesis is needed if only a few candidate taxa are of interest. For such analysis, the conventional mediation methods can be used directly, which is not subject to the high-dimensional and compositional data challenges.

Note that the aim of the *ilr*-transformation in our method is to remove the compositional structure of the original d mediators. However, one drawback of the *ilr*-transformed mediators is how to reasonably interpret the corresponding meditation effect. Numerically, our proposed method can correctly select those significant mediators if they have true mediation effects in some popular compositional mediation models. For example, when data are generated from the models in Sohn and Li[23] and Wang et al,[24] our method can select significant mediators satisfactorily, demonstrating the robustness of our method.

Our proposed procedure can be extended in several directions. First, we have replaced zero values by 0.5 to deal with the zero-inflated problem in the microbiome data. With more rigor, we can adopt the two part models (eg, Chen and Li,[40] Chai et al,[41] and Liu et al[42]) to separately model the odds of the presence of zero values and the amount of positive values. Second, as a reviewer pointed out, it is interesting to incorporate the interaction between the exposure and the microbiome in model (2). Third, the bacterial taxa are related to each other by a phylogenetic tree.[43,44] We can propose a novel tree-guided procedure similar to the tree-guided fused Lasso in Wang and Zhao.[44] Fourth, longitudinal measurements of microbial communities can be obtained in many microbiome studies.[40,45] Mediation analysis for longitudinal microbiome data is a topic for future research. Finally, there are two frameworks of mediation analysis: the SEM framework, for example, Zhang et al,[31] Boca et al,[46] Sampson et al;[47] and the counterfactual or potential outcome approach, for example, Huang and Pan,[48] Cheng et al,[49] and Derkach et al.[50] Our current approach falls into the SEM framework. Note that one benefit of counterfactual-based mediation models is how nonlinear methods are handled. It is of great interest to adapt our method in the counterfactual or potential outcome framework in further research.

ACKNOWLEDGEMENTS

The authors would like to thank the Editor, the Associate Editor and three reviewers for their constructive and insightful comments that greatly improved the article. Research reported in this publication was supported by NIH R21 AG063370, UL1 TR002345, and R01 DK 110014. The content is solely the responsibility of the authors and does not necessarily represent the official view of the NIH.

DATA AVAILABILITY STATEMENT

An R package microHIMA is available to implement the proposed method at https://github.com/joyfulstones/microbiome-mediation-SIS. The real data that support the findings of this study were published in Schulfer et al, to which data request should be addressed.

ORCID

Haixiang Zhang https://orcid.org/0000-0002-7311-5605
Lei Liu https://orcid.org/0000-0003-1844-338X

REFERENCES

1. Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. *Nat Rev Genet*. 2012;13(4):260-270. https://doi.org/10.1038/nrg3182.
2. Cho I, Yamanishi S, Cox L, et al. Antibiotics in early life alter the murine colonic microbiome and adiposity. *Nature*. 2012;488(7413):621-626. https://doi.org/10.1038/nature11400.
3. Ahn J, Sinha R, Pei Z, et al. Human gut microbiome and risk for colorectal cancer. *JNCI J Nat Cancer Inst*. 2013;105(24):1907-1911. https://doi.org/10.1093/jnci/djt300.
4. Alekseyenko AV, Perez-Perez GI, Souza AD, et al. Community differentiation of the cutaneous microbiota in psoriasis. *Microbiome*. 2013;1(1):31. https://doi.org/10.1186/2049-2618-1-31.
5. Roy S, Trinchieri G. Microbiota: a key orchestrator of cancer therapy. *Nat Rev Cancer*. 2017;17(5):271-285. https://doi.org/10.1038/nrc.2017.13.
6. Zitvogel L, Ma Y, Raoult D, Kroemer G, Gajewski TF. The microbiome in cancer immunotherapy: diagnostic tools and therapeutic strategies. *Science*. 2018;359(6382):1366-1370. https://doi.org/10.1126/science.aar6918.
7. Petrosino JF. The microbiome in precision medicine: the way forward. *Genome Med*. 2018;10(1):12. https://doi.org/10.1186/s13073-018-0525-6.
8. Li H. Microbiome, metagenomics, and high-dimensional compositional data analysis. *Ann Rev Stat Appl*. 2015;2(1):73-94. https://doi.org/10.1146/annurev-statistics-010814-020351.
9. Xia Y, Sun J. Hypothesis testing and statistical analysis of microbiome. *Genes and Diseases*. 2017;4(3):138-148. https://doi.org/10.1016/j.gendis.2017.06.001.
10. Baron RM, Kenny DA. The moderator–mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J Pers Soc Psychol*. 1986;51(6):1173-1182. https://doi.org/10.1037/0022-3514.51.6.1173.
11. MacKinnon DP. *Introduction to Statistical Mediation Analysis*. New York, NY: Erlbaum and Taylor Francis Group; 2008.
12. Saunders CT, Blume JD. A regression framework for causal mediation analysis with applications to behavioral science. *Multivar Behav Res*. 2019;54(4):555-577. https://doi.org/10.1080/00273171.2018.1552109.
13. Wood RE, Goodman JS, Beckmann N, Cook A. Mediation testing in management research: a review and proposals. *Organ Res Methods*. 2008;11(2):270-295. https://doi.org/10.1177/1094428106297811.
14. Rucker DD, Preacher KJ, Tormala ZL, Petty RE. Mediation analysis in social psychology: current practices and new recommendations. *Soc Personal Psychol Compass*. 2011;5(6):359-371. https://doi.org/10.1111/j.1751-9004.2011.00355.x.

15. Lockhart G, MacKinnon DP, Ohlrich V. Mediation analysis in psychosomatic medicine research. *Psychosom Med*. 2011;73(1):29-43. https://doi.org/10.1097/psy.0b013e318200a54b.
16. Valeri L, Reese SL, Zhao S, et al. Misclassified exposure in epigenetic mediation analyses. does DNA methylation mediate effects of smoking on birthweight? *Epigenomics*. 2017;9(3):253-265. https://doi.org/10.2217/epi-2016-0145.
17. Hayes AF, Rockwood NJ. Regression-based statistical mediation and moderation analysis in clinical research: observations, recommendations, and implementation. *Behav Res Ther*. 2017;98:39-57. https://doi.org/10.1016/j.brat.2016.11.001.
18. MacKinnon D, Fairchild A, Fritz M. Mediation analysis. *Annu Rev Psychol*. 2007;58:593-614.
19. Ten Have TR, Joffe MM. A review of causal estimation of effects in mediation analyses. *Stat Methods Med Res*. 2012;21(1):77-107. https://doi.org/10.1177/0962280210391076.
20. Preacher KJ. Advances in mediation analysis: a survey and synthesis of new developments. *Annu Rev Psychol*. 2015;66(1):825-852. https://doi.org/10.1146/annurev-psych-010814-015258.
21. VanderWeele TJ. Mediation analysis: a practitioner's guide. *Annu Rev Public Health*. 2016;37(1):17-32. https://doi.org/10.1146/annurev-publhealth-032315-021402.
22. Zhang J, Wei Z, Chen J. A distance-based approach for testing the mediation effect of the human microbiome. *Bioinformatics*. 2018;34(11):1875-1883. https://doi.org/10.1093/bioinformatics/bty014.
23. Sohn MB, Li H. Compositional mediation analysis for microbiome studies. *Ann Appl Stat*. 2019;13(1):661-681. https://doi.org/10.1214/18-aoas1210.
24. Wang C, Hu J, Blaser MJ, Li H. Estimating and testing the microbial causal mediation effect with high-dimensional and compositional microbiome data. *Bioinformatics*. 2020;36:347-355. https://doi.org/10.1093/bioinformatics/btz565.
25. Zhang H, Chen J, Li Z, Liu L. Testing for mediation effect with application to human microbiome data. *Stat Biosci*. 2019. https://doi.org/10.1007/s12561-019-09253-3.
26. Aitchison J. *The Statistical Analysis of Compositional Data*. London: Chapman & Hall; 1986.
27. Aitchison J. Logratios and natural laws in compositional data analysis. *Math Geol*. 1999;31(5):563-580. https://doi.org/10.1023/a:1007568008032.
28. Hron K, Filzmoser P, Thompson K. Linear regression with compositional explanatory variables. *J Appl Stat*. 2012;39(5):1115-1128. https://doi.org/10.1080/02664763.2011.644268.
29. Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barcelo-Vidal C. Isometric logratio transformations for compositional data analysis. *Math Geol*. 2003;35(3):279-300. https://doi.org/10.1023/a:1023818214614.
30. Zhang C-H, Zhang SS. Confidence intervals for low dimensional parameters in high dimensional linear models. *J Royal Stat Soc Ser B (Stat Methodol)*. 2014;76(1):217-242. https://doi.org/10.1111/rssb.12026.
31. Zhang H, Zheng Y, Zhang Z, et al. Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics*. 2016;32(20):3150-3154. https://doi.org/10.1093/bioinformatics/btw351.
32. Goeman JJ, Meijer RJ, Krebs TJP, Solari A. Simultaneous control of all false discovery proportions in large-scale multiple hypothesis testing. *Biometrika*. 2019;106(4):841-856. https://doi.org/10.1093/biomet/asz041.
33. Marcus R, Eric P, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*. 1976;63(3):655-660. https://doi.org/10.1093/biomet/63.3.655.
34. Goeman JJ, Solari A. Multiple testing for exploratory research. *Stat Sci*. 2011;26(4):584-597. https://doi.org/10.1214/11-sts356.
35. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Stat Soc Ser B (Methodol)*. 1995;57(1):289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
36. Schulfer AF, Schluter J, Zhang Y, et al. The impact of early-life sub-therapeutic antibiotic treatment (STAT) on excessive weight is robust despite transfer of intestinal microbes. *ISME J*. 2019;13(5):1280-1292. https://doi.org/10.1038/s41396-019-0349-4.
37. Caporaso JG, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods*. 2010;7(5):335-336. https://doi.org/10.1038/nmeth.f.303.
38. Lin W, Shi P, Feng R, Li H. Variable selection in regression with compositional covariates. *Biometrika*. 2014;101(4):785-797. https://doi.org/10.1093/biomet/asu031.
39. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat*. 2001;29(4):1165-1188.
40. Chen EZ, Li H. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics*. 2016;32(17):2611-2617. https://doi.org/10.1093/bioinformatics/btw308.
41. Chai H, Jiang H, Lin L, Liu L. A marginalized two-part Beta regression model for microbiome compositional data. *PLoS Comput Biol*. 2018;14(7):e1006329. https://doi.org/10.1371/journal.pcbi.1006329.
42. Liu L, Shih Y-CT, Strawderman RL, Zhang D, Johnson BA, Chai H. Statistical analysis of zero-inflated continuous data: a review. *Stat Sci*. 2019;34(2):253-279. https://doi.org/10.1214/18-sts681.
43. Tang Z-Z, Chen G, Alekseyenko AV, Li H. A general framework for association analysis of microbial communities on a taxonomic tree. *Bioinformatics*. 2016;btw804:33. https://doi.org/10.1093/bioinformatics/btw804.
44. Wang T, Zhao H. Constructing predictive microbial signatures at multiple taxonomic levels. *J Am Stat Assoc*. 2017;112(519):1022-1031. https://doi.org/10.1080/01621459.2016.1270213.
45. Lugo-Martinez J, Ruiz-Perez D, Narasimhan G, Bar-Joseph Z. Dynamic interaction network inference from longitudinal microbiome data. *Microbiome*. 2019;7(1):54. https://doi.org/10.1186/s40168-019-0660-3.
46. Boca SM, Sinha R, Cross AJ, Moore SC, Sampson JN. Testing multiple biological mediators simultaneously. *Bioinformatics*. 2014;30(2):214-220. https://doi.org/10.1093/bioinformatics/btt633.

47. Sampson JN, Boca SM, Moore SC, Heller R. FWER and FDR control when testing multiple mediators. *Bioinformatics*. 2018;34(14):2418-2424. https://doi.org/10.1093/bioinformatics/bty064.
48. Huang Y-T, Pan W-C. Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. *Biometrics*. 2016;72(2):402-413. https://doi.org/10.1111/biom.12421.
49. Cheng J, Cheng NF, Guo Z, Gregorich S, Ismail AI, Gansky SA. Mediation analysis for count and zero-inflated count data. *Stat Methods Med Res*. 2018;27(9):2756-2774. https://doi.org/10.1177/0962280216686131.
50. Derkach A, Pfeiffer RM, Chen T-H, Sampson JN. High dimensional mediation analysis with latent variables. *Biometrics*. 2019;75(3):745-756. https://doi.org/10.1111/biom.13053.

How to cite this article: Zhang H, Chen J, Feng Y, Wang C, Li H, Liu L. Mediation effect selection in high-dimensional and compositional microbiome data. *Statistics in Medicine*. 2021;40:885–896. https://doi.org/10.1002/sim.8808