Bioinformatics, 36(2), 2020, 347–355
doi: 10.1093/bioinformatics/btz565
Advance Access Publication Date: 22 July 2019
Original Paper

OXFORD

Genome analysis

Estimating and testing the microbial causal mediation effect with high-dimensional and compositional microbiome data

Chan Wang[1], Jiyuan Hu[1], Martin J. Blaser[2] and Huilin Li[1,*]

[1]Division of Biostatistics, Department of Population Health, New York University School of Medicine, New York, NY 10016, USA and [2]Department of Medicine and Microbiology, Center for Advanced Biotechnology and Medicine, Rutgers University, Piscataway, NJ 08854-8021, USA

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on January 10, 2019; revised on June 17, 2019; editorial decision on July 12, 2019; accepted on July 16, 2019

Abstract

Motivation: Recent microbiome association studies have revealed important associations between microbiome and disease/health status. Such findings encourage scientists to dive deeper to uncover the causal role of microbiome in the underlying biological mechanism, and have led to applying statistical models to quantify causal microbiome effects and to identify the specific microbial agents. However, there are no existing causal mediation methods specifically designed to handle high dimensional and compositional microbiome data.

Results: We propose a rigorous Sparse Microbial Causal Mediation Model (SparseMCMM) specifically designed for the high dimensional and compositional microbiome data in a typical three-factor (treatment, microbiome and outcome) causal study design. In particular, linear log-contrast regression model and Dirichlet regression model are proposed to estimate the causal direct effect of treatment and the causal mediation effects of microbiome at both the community and individual taxon levels. Regularization techniques are used to perform the variable selection in the proposed model framework to identify signature causal microbes. Two hypothesis tests on the overall mediation effect are proposed and their statistical significance is estimated by permutation procedures. Extensive simulated scenarios show that SparseMCMM has excellent performance in estimation and hypothesis testing. Finally, we showcase the utility of the proposed SparseMCMM method in a study which the murine microbiome has been manipulated by providing a clear and sensible causal path among antibiotic treatment, microbiome composition and mouse weight.

Availability and implementation: https://sites.google.com/site/huilinli09/software and https://github.com/chanw0/SparseMCMM.

Contact: Huilin.Li@nyulangone.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Microbiome research is producing exciting results, with many studies linking specific microbes to particular diseases (Gilbert *et al.*, 2018; Ni *et al.*, 2017; Zheng *et al.*, 2016), physiological properties and environmental parameters (Albenberg and Wu, 2014; Stein *et al.*, 2016; Zeevi *et al.*, 2015). However, knowing the correlation or association between microbiome and another trait (Hu *et al.*, 2018; Koh *et al.*, 2017) is no longer a sufficient research goal, since now the scientific frontier is to understand the causal role of microbiota in the underlying biological mechanism (Fischbach, 2018). For example, early life exposure to antibiotics has been reported to affect human metabolism and cause weight gain and/or

© The Author(s) 2019. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bioinformatics/article/36/2/347/5536874 by University of Michigan user on 28 October 2020

immunological properties by altering the gut microbiota (Livanos *et al.*, 2016; Mahana *et al.*, 2016; Schulfer *et al.*, 2018, 2019). To confirm such causal relationships, researchers conducted experiments randomizing groups of newborn mice control and antibiotic groups, and collecting longitudinal microbiome, weight and immunological measurements through the study period. Within each experiment, researchers sought to understand how the change in microbiome due to the antibiotic exposure caused the change in mouse phenotype. If the altered microbiome played a causal role, which specific microbes were the culprits? To answer these questions, a rigorous causal mediation analytic framework is needed.

In microbiome causal research, there are many possible factors can potentially confound the microbiome effect, such as age, gender and diet (Knight *et al.*, 2018). In order to control or eliminate the possible confounding effects, randomized experiments are usually used to investigate the specific effects of a course of treatment on the microbiome and diseases (Knight *et al.*, 2018). In this paper, we introduce our proposed microbiome causal analytical framework within randomized experimental designs and target understanding the causal pathway among three factors: treatment (T), microbiome (**M**) and outcome (Y) (Fig. 1). Here, microbiome are hypothesized as mediators on the pathway from treatment to outcome. It is important to understand how the causal effect of the treatment on the outcome can be divided into the causal direct effect and the causal indirect effect, acting through the mediator (the latter is also called the causal mediation effect in this paper).

Donald Rubin and Paul Holland developed the potential outcome idea (Neyman, 1923) and established a formal mathematical causal framework for both observational and randomized experimental studies (Holland, 1986; Rubin, 1974, 2005). They defined causation as a hypothetical value through the counterfactual statement: a score difference between the observed outcome of one subject under one treatment condition and the potential outcome if he/she would be under the alternative treatment condition. Later, VanderWeele and his colleagues expanded this framework to the mediation analysis to define causal direct effect and causal mediation effect under the sufficient causal assumptions and allow for the presence of exposure–mediator interactions (Valeri and VanderWeele, 2013; VanderWeele and Vansteelandt, 2009, 2010, 2014; VanderWeele, 2013, 2014). In our motivating example and other microbiome research, it is possible that the treatment can affect microbiome's mediating effect on the outcome, so we follow VanderWeele's counterfactual mediation framework and incorporate the interaction of treatment and micrbiome on the outcome into the proposed method.

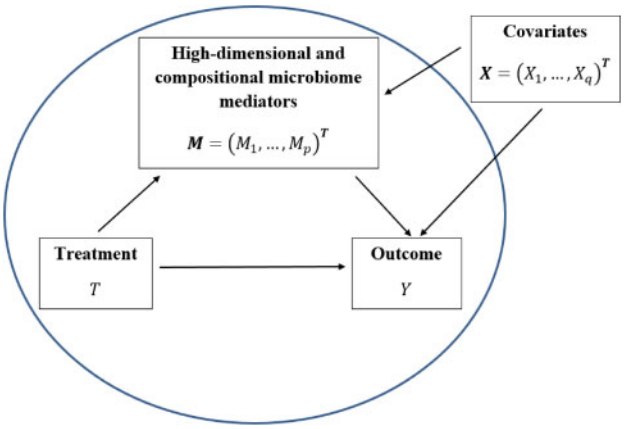

Fig. 1. Relations between treatment *T*, covariates ***X***, microbiome composition (mediators) ***M*** and outcome *Y*. We aim to investigate the effect of treatment directly on the outcome (termed as the direct effect) and the effect of treatment through microbiome composition as mediators (termed the mediation effect), while adjusting for covariates

Recently, with the advent of high-throughput biomedical data, a few causal mediation models have become available to handle high-dimensional mediators by using linear structural equation modeling (LSEM). These methods adopted two ways to reduce the dimensionality of the mediators. One way is through regularization or penalization. For example, by assuming that there is no correlation among mediators, Zhang *et al.* (2016) proposed a joint significance test based on sure independent screening (Fan and Lv, 2008) and minimax concave penalty techniques (Zhang *et al.*, 2010) to evaluate the casual mediating effect of DNA methylation markers. The other way is to transform the correlated high-dimensional mediators into a series of causal mediation models with single continuous mediator. For example, Huang and Pan (2016) used spectral decomposition to transform high-dimensional gene expression mediators into low-dimensional and uncorrelated ones. Chén *et al.* (2018) transformed high-dimensional imaging mediators into orthogonal components and ranked them based on their contributions to the LSEM likelihood.

Unfortunately, none of the methods cited above is designed for the high dimensional, sparse and compositional microbiome data. Compositionality is the crux of the new challenges in causal studies involving the microbiome. Due to varying sequencing read counts across samples, normalization needs to be employed to make the microbial counts comparable before downstream analyses. As a common normalization method, the sequence counts are scaled by the total number of reads or the total number of reads and lengths of reads together. This step converts the count data into the relative abundance (Knight *et al.*, 2018), which is compositional and has the simple unit-sum constraint. Log-ratio analysis (Aitchison, 1982) and Dirichlet regression (Hijazi and Jernigan, 2009) are two available methods to deal with relative abundance. Aitchison's log-ratio methods work on the ratios of the components of the composition. Because those ratios are sensitive to low relative abundance and need additional handling for zero counts, the log-ratio methods have been criticized for its limited interpretability (Hijazi and Jernigan, 2009). As an alternative, Dirichlet regression has drawn attention by analyzing compositional data with multivariate statistical modeling, and has been shown to be useful for compositional data (Campbell and Mosimann, 1987a,b; Hijazi and Jernigan, 2009).

In the paper, we propose a sparse microbial causal mediation framework (SparseMCMM) to clarify the relationship among a binary treatment, a vector of compositional microbial mediators and a continuous outcome. We use Dirichlet regression to model the relationship between treatment and microbiome composition (Hijazi and Jernigan, 2009), and linear log-contrast regression to model treatment effect, log transformed microbiome effects and the effects of their interactions on the outcome (Aitchison and Bacon-Shone, 1984; Lin *et al.*, 2014). The combination of those two regressions is then considered in the counterfactual mediation framework to clarify the causal mediation effect of microbiome on the outcome.

The remainder of this paper is organized as follows: In Section 2, we introduce the proposed SparseMCMM with high-dimensional and compositional mediators under the counterfactual framework and derive the corresponding mediation effect estimators. Then, we describe how to select and estimate non-zero parameters in the high-dimensional causal mediation regressions with regularizations. Moreover, we propose two hypothesis tests for the overall mediation effect and apply permutation procedures to access their

Downloaded from https://academic.oup.com/bioinformatics/article/36/2/347/5536874 by University of Michigan user on 28 October 2020

statistical significance. In Section 3, we conduct extensive simulations to evaluate the performance of the proposed model on both estimation and testing. Subsequently, we implement SparseMCMM to investigate the extent to which the microbiome mediates the causal pathway from antibiotic usage to excess weight gain in a longitudinal murine study. We conclude with discussion of the relevant issues in Section 4.

2 Materials and methods

2.1 Casual mediation model

2.1.1 Notations and two basic regression models

Suppose there are n subjects, p taxa and q covariates and subscripts i, j, k, indicate a subject, a taxon and a covariate, respectively. For the ith subject, let T_i be the treatment status with $T_i = 1$ or 0 for the treatment group or the control group, $\boldsymbol{M}_i = (M_{i1}, \ldots, M_{ip})^T$ be the microbiome relative abundance with the constraint $\sum_{j=1}^{p} M_{ij} = 1$, let $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iq})^T$ represent the covariates such as age, sex and weight, and let Y_i be the continuous outcome. We propose two regression models to model the causal mediation relationships among T, $\boldsymbol{M}$, $\boldsymbol{X}$ and Y as illustrated in Figure 1. The first model depicts that the outcome Y_i is determined by the treatment, compositional mediators, interactions between the treatment and mediators and covariates as:

$$Y_i = \alpha_0 + \boldsymbol{\alpha}_X^T \boldsymbol{X}_i + \alpha_T T_i + \sum_{j=1}^{p-1} \alpha_{Mj} \log(M_{ij}/M_{ip}) + \sum_{j=1}^{p-1} \alpha_{Cj} T_i \log(M_{ij}/M_{ip}) + \epsilon_i.$$

Note that with $\sum_{j=1}^{p} M_{ij} = 1$, the relative abundances of p taxa are built into the model through the log-contrast strategy with the pth taxon as the reference using the log ratio transformation proposed by Aitchison and Bacon-Shone (1984). By defining $\alpha_{Mp} = -\sum_{j=1}^{p-1} \alpha_{Mj}$ and $\alpha_{Cp} = -\sum_{j=1}^{p-1} \alpha_{Cj}$ (Lin *et al.*, 2014), with algebraic equivalents, we can rewrite the above model in the matrix form,

$$Y_i = \alpha_0 + \boldsymbol{\alpha}_X^T \boldsymbol{X}_i + \alpha_T T_i + \boldsymbol{\alpha}_M^T [\log(\boldsymbol{M}_i)] + \boldsymbol{\alpha}_C^T [\log(\boldsymbol{M}_i)] T_i + \epsilon_i, \quad \text{subject to} \quad \boldsymbol{\alpha}_M^T \boldsymbol{1} = 0, \quad \text{and} \quad \boldsymbol{\alpha}_C^T \boldsymbol{1} = 0, \tag{1}$$

where α_0 is the intercept, α_T is the coefficient of treatment, $\boldsymbol{\alpha}_X = (\alpha_{X1}, \ldots, \alpha_{Xq})^T$, $\boldsymbol{\alpha}_M = (\alpha_{M1}, \ldots, \alpha_{Mp})^T$ and $\boldsymbol{\alpha}_C = (\alpha_{C1}, \ldots, \alpha_{Cp})^T$ are the vectors of coefficients of covariates, microbial mediators and interactions between treatment and mediators, respectively, and $\epsilon_i \sim N(0, \sigma^2)$ is the error term.

Secondly, we use the Dirichlet regression (Hijazi and Jernigan, 2009) to model the microbial relative abundance as a function of treatment and covariates. Specifically we assume that $\boldsymbol{M}_i | (T_i, \boldsymbol{X}_i) \sim \text{Dirichlet}\big(\gamma_1(T_i, \boldsymbol{X}_i), \ldots, \gamma_p(T_i, \boldsymbol{X}_i)\big)$, and their microbial relative means are linked with treatment and covariates $(T_i, \boldsymbol{X}_i)$ in the generalized linear model fashion with a log link:

$$E[M_{ij}] = \frac{\gamma_j(T_i, \boldsymbol{X}_i)}{\sum_{m=1}^{p} \gamma_m(T_i, \boldsymbol{X}_i)}, \qquad \log\{\gamma_j(T_i, \boldsymbol{X}_i)\} = \beta_{0j} + \beta_{Tj} T_i + \boldsymbol{\beta}_{Xj}^T \boldsymbol{X}_i \tag{2}$$

where β_{0j} is the intercept and β_{Tj} and $\boldsymbol{\beta}_{Xj}$ are the coefficients of treatment and covariates for the jth taxon, respectively. With this modeling, we quantify the treatment effect on each individual taxon which further allows the quantitation of component-wise (or taxon-wise) mediation effect and the overall (or aggregated) mediation effect of the microbiome community in the next section.

2.1.2 Definition of direct and mediation effects in the counterfactual framework

With the above two models, we next determine the average causal direct effect of treatment and the average mediation effect of microbiome on the outcome under the counterfactual framework (VanderWeele and Vansteelandt, 2009, 2014; VanderWeele, 2016). With counterfactual notation, DE refers to the expected difference of the outcome Y between the treatment $T = 1$ and $T = 0$ when the mediators $\boldsymbol{M}$ are set to the value they would have taken had T been set to 0, and is defined as:

$$DE = E[Y_{T=1, \boldsymbol{M}(T=0)} - Y_{T=0, \boldsymbol{M}(T=0)} | \boldsymbol{X}].$$

ME is the indirect effect of treatment on outcome through the compositional microbiome community and refers to the expected difference of the outcome Y between the mediators $\boldsymbol{M}(T = 1)|\boldsymbol{X}$ and $\boldsymbol{M}(T = 0)|\boldsymbol{X}$ when the treatment $T = 1$, where $\boldsymbol{M}(T = t)|\boldsymbol{X}$ represents the microbiome composition we would have observed had T been set to the value t given covariates $\boldsymbol{X}$. Mathematically, ME is defined as:

$$ME = E[Y_{T=1, \boldsymbol{M}(T=1)} - Y_{T=1, \boldsymbol{M}(T=0)} | \boldsymbol{X}].$$

Under four sufficient identifiable assumptions (see Supplementary Materials, Section S1), DE and ME can be further expressed by the parameters in models (1)–(2), respectively, as follows:

$$DE = \alpha_T + \boldsymbol{\alpha}_C^T E[\log(\boldsymbol{M}) | T = 0, \boldsymbol{X}], \tag{3}$$

and

$$\begin{aligned} ME &= (\boldsymbol{\alpha}_M^T + \boldsymbol{\alpha}_C^T)\{E[\log(\boldsymbol{M}) | T = 1, \boldsymbol{X}] - E[\log(\boldsymbol{M}) | T = 0, \boldsymbol{X}]\} \\ &= \sum_{j=1}^{p} (\alpha_{Mj} + \alpha_{Cj})\{E[\log(M_j) | T = 1, \boldsymbol{X}] - E[\log(M_j) | T = 0, \boldsymbol{X}]\} : \\ &= \sum_{j=1}^{p} ME_j \end{aligned} \tag{4}$$

From Equation (4), note that ME is the summation of the individual mediation effects from each taxon ME_j. ME_j is the product of two parts: $(\alpha_{Mj} + \alpha_{Cj})$ which represents the jth microbial effect consisting of the main effect and the interaction effect of this taxon and the treatment on the outcome; and $\{E[\log(M_j) | T = 1, \boldsymbol{X}] - E[\log(M_j) | T = 0, \boldsymbol{X}]\}$ which represents the treatment effect on the jth taxon. Therefore ME_j exists only if when both the jth microbial effect on the outcome and the treatment effect on the jth taxon are not zero. For the expectation part, we have $E[\log(M_j) | T = t, \boldsymbol{X}] = \psi[\gamma_j(T = t, \boldsymbol{X})] - \psi[\sum_{m=1}^{p} \gamma_m \ (T = t, \boldsymbol{X})]$, $\gamma_j(T = t, \boldsymbol{X}) = \exp(\beta_{0j} + \beta_{Tj} t + \boldsymbol{\beta}_{Xj}^T \boldsymbol{X})$, $t = 0$ or 1, and $\psi(\cdot) = \frac{d}{dx} \ln(\Gamma(x))$ is the digamma function.

The total effect of the treatment on the outcome is therefore the summation of DE and ME:

$$TE = DE + ME = E[Y_{T=1, \boldsymbol{M}(T=1)} - Y_{T=0, \boldsymbol{M}(T=0)} | \boldsymbol{X}]. \tag{5}$$

The detailed derivations of DE, ME and TE are provided in the Supplementary Materials, Section S2.

In summary, given covariates $\boldsymbol{X}$, SparseMCMM is able to decompose the treatment effect on the outcome into the direct effect of treatment and the mediation effects through the microbiome.

Downloaded from https://academic.oup.com/bioinformatics/article/36/2/347/5536874 by University of Michigan user on 28 October 2020

It elucidates the mediation role of the microbiome through rigorous statistical modeling, and quantifies the mediation effects for the overall microbiome community and for each specific taxon respectively.

2.2 Parameter estimation

It is challenging to estimate all parameters from the joint log-likelihood function based on models (1)–(2) due to the nonlinearity and constraints. As an alternative approach, we first estimate the regression parameters in models (1)–(2), separately and then include them into Equations (3)–(5) to obtain the estimated DE, ME (ME_j for the individual taxon) and TE, respectively. A similar two-step approach has been used in the genetics field and its computing efficiency has been recognized (Huang and Pan, 2016; Zhang *et al.*, 2016).

Another challenge of estimation in the microbiome setting is the high dimensional mediators. Model (1) has $\sim 2p$ parameters when it considers both main effect and interaction effect and model (2) has $\sim(q+2)p$ parameters. Both models have far greater number of parameters than the sample size n. To deal with the high-dimensional mediators, we propose regularization techniques to simultaneously identify the key taxa with the primary mediation effect and parameter estimation. Specifically, in model (1), we propose a penalized least squares criterion to penalize the main effects and interaction effects of mediators, which has optimal biological explanations; in model (2), we utilize the L_1 norm penalty to select taxa which are altered by the treatment. In the following section, we introduce the estimation procedures for models (1) and (2) sequentially.

2.2.1 Parameter estimation for the linear log-contrast regression

Denote the parameters of regression coefficients in model (1) by $\boldsymbol{\alpha} = (\alpha_0, \alpha_T, \boldsymbol{\alpha}_X^T, \boldsymbol{\alpha}_M^T, \boldsymbol{\alpha}_C^T)^T$. We use the least squares method (Friedman *et al.*, 2001) to estimate $\boldsymbol{\alpha}$. Given the observed data $(T_i, \boldsymbol{X}_i, \boldsymbol{M}_i, Y_i)$, the sum of squared residuals (SSR) which measures the discrepancy between the observed and predicted outcome is:

$$SSR(\boldsymbol{\alpha}) = \sum_{i=1}^{n}\{Y_i - \{\alpha_0 + \boldsymbol{\alpha}_X^T\boldsymbol{X}_i + \alpha_T T_i + \boldsymbol{\alpha}_M^T[\log(\boldsymbol{M}_i)] + \boldsymbol{\alpha}_C^T[\log(\boldsymbol{M}_i)]T_i\}\}^2.$$

It is a well-established variable selection practice in high dimensional linear regression with interaction that the interaction effect exists only if the corresponding main effects are included in the model, which is termed the heredity condition or hierarchy structure (Peixoto, 1987; Radchenko and James, 2010). To apply this constraint, we add two penalties to the SSR and solve the following convex least squares optimization problem:

$$\underset{\alpha}{\operatorname{argmin}}\Big(SSR(\boldsymbol{\alpha}) + \lambda_1\sum_{j=1}^{p}\sqrt{\alpha_{Mj}^2 + \alpha_{Cj}^2} + \lambda_2\sum_{j=1}^{p}|\alpha_{Cj}|\Big), \qquad (6)$$

with the constraints $\boldsymbol{\alpha}_M^T\mathbf{1} = 0$ and $\boldsymbol{\alpha}_C^T\mathbf{1} = 0$. $\lambda_1(\geq 0)$ and $\lambda_2(\geq 0)$ are two tuning parameters. As discussed by Radchenko and James (2010), these penalty functions have desirable properties in both theory and performance in addressing the heredity condition. Specifically, the first penalty, similar to the group-lasso penalty, ensures that the main effect α_{Mj} and the interaction effect α_{Cj} of the jth taxon shrinks on the same scale. The second L_1 norm penalty only works on the interaction terms. In combination, they guarantee that the main effect α_{Mj} from the first penalty could only be shrunk to 0 when the corresponding interaction effect α_{Cj} from the second penalty also is shrunk to 0. When $\alpha_{Cj} = 0$, the first penalty is reduced to a lasso penalty: $\sqrt{\alpha_{Mj}^2 + \alpha_{Cj}^2} = |\alpha_{Mj}|$.

We utilize the sequential quadratic programming (SQP) method, a popular method for solving constrained nonlinear optimization problems, with R package nloptr (Kraft, 1988; Ypma, 2014) to optimize Equation (6) and obtain the estimate $\hat{\boldsymbol{\alpha}}$. Particularly, SQP seeks a numerical solution by solving a sequence of quadratic subproblems, each of which optimizes a quadratic objective function subject to the linear constraints. Tuning parameters λ_1 and λ_2 are determined by Bayesian information criterion (BIC) (Chen and Chen, 2008).

2.2.2 Parameter estimation for the Dirichlet regression

Denote the parameters of regression coefficients in model (2) by $\boldsymbol{\beta} = (\beta_0, \beta_{T1}, \ldots, \beta_{Tp}, \boldsymbol{\beta}_{X1}^T, \ldots, \boldsymbol{\beta}_{Xp}^T)^T$. Given the observed data $(T_i, \boldsymbol{X}_i, \boldsymbol{M}_i)$, the log-likelihood function for n observations is given by

$$\begin{aligned} l(\boldsymbol{\beta}; \boldsymbol{M}, \boldsymbol{X}, T) = & \\ \sum_{i=1}^{n}\left\{\log\left[\Gamma(\sum_{j=1}^{p}\gamma_{ij})\right] - \sum_{j=1}^{p}\log\left[\Gamma(\gamma_{ij})\right] + \sum_{j=1}^{p}(\gamma_{ij}-1)\log(M_{ij})\right\} & \\ = \sum_{i=1}^{n}\tilde{\Gamma}[\sum_{j=1}^{p}\exp(\beta_{0j} + \beta_{Tj}T_i + \boldsymbol{\beta}_{Xj}^T\boldsymbol{X}_i)] + \sum_{i=1}^{n}\sum_{j=1}^{p}\{\log(M_{ij}) & \\ [\exp(\beta_{0j} + \beta_{Tj}T_i + \boldsymbol{\beta}_{Xj}^T\boldsymbol{X}_i) - 1] - \tilde{\Gamma}[\exp(\beta_{0j} + \beta_{Tj}T_i + \boldsymbol{\beta}_{Xj}^T\boldsymbol{X}_i)]\}, & \end{aligned}$$

where $\gamma_{ij} = \gamma_j(T_i, \boldsymbol{X}_i)$ as defined in model (2), and $\tilde{\Gamma}(\cdot)$ is the log gamma function. In order to select the taxa whose relative abundances are altered by treatment, we minimize the following penalized version of the log-likelihood with L_1 penalty:

$$\underset{\beta}{\operatorname{argmin}}\Big\{-l(\boldsymbol{\beta}; \boldsymbol{M}, \boldsymbol{X}, T) + \lambda_3[\sum_{j=1}^{p}|\beta_{Tj}| + \sum_{j=1}^{p}|\boldsymbol{\beta}_{Xj}|_1]\Big\}, \qquad (7)$$

where $\lambda_3 \geq 0$ is the tuning parameter and is determined by BIC as those in model (6). The Newton-Raphson algorithm in nloptr R package (Bonnans *et al.*, 2006) is used to find the numerical estimate $\hat{\boldsymbol{\beta}}$.

Inserting the estimates $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ into Equations (3)–(5), we can obtain the estimates $\hat{DE}$, $\hat{ME}$, $\hat{ME}_j$ and $\hat{TE}$ respectively.

2.3 Hypothesis tests for mediation effects

We propose two tests to examine whether the microbiome has any mediation effect on the outcome or not, at the community and taxon levels, denoted as OME and CME, respectively, in Equation (4).

Since the null hypothesis of no overall mediation effect at the community level can be expressed $H_0 : ME = 0$, the first test is defined as

$$\text{OME} = (\hat{\boldsymbol{\alpha}}_M^T + \hat{\boldsymbol{\alpha}}_C^T)\{\hat{E}[\log(\boldsymbol{M})|T=1, \boldsymbol{X}] - \hat{E}[\log(\boldsymbol{M})|T=0, \boldsymbol{X}]\} \equiv f(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}). \qquad (8)$$

From Equation (4), we can see that the overall mediation effect of the microbiome community is defined as the summation of all component-wise ME_js and it can be counteracted when both positive and negative component-wise ME_js are present. Thus, in the second test, we consider the following null hypothesis which tests whether at least one component-wise ME_j is significantly non-zero:

Downloaded from https://academic.oup.com/bioinformatics/article/36/2/347/5536874 by University of Michigan user on 28 October 2020

$$H_0 : ME_j = 0, \forall j \in \{1, \ldots, p\}. \quad (9)$$

To tackle this problem with the high dimensionality of the mediators, we propose an equivalent null hypothesis $H_0 : \sum_{j=1}^{p} ME_j^2 = 0$, as indicated by Huang and Pan (2016). Then CME test is formulated as

$$\text{CME} = \sum_{j=1}^{p} \hat{ME}_j^2 \equiv g(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}). \quad (10)$$

It is not trivial to derive the asymptotic distributions of the test statistics OME and CME. As an alternative, we apply the following permutation procedure to estimate *P*-values (Boca *et al.*, 2014; Taylor and MacKinnon, 2012; Zhang *et al.*, 2018).

First, we randomly shuffle treatment T and outcome Y separately to obtain the permuted treatment $T^{(b)}$ and outcome $Y^{(b)}$, $b = 1, \ldots, B$. Second, we get estimates $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\alpha}}^{(b)}$ from model (1) based on data $(T, \boldsymbol{X}, \boldsymbol{M}, Y)$ and $(T, \boldsymbol{X}, \boldsymbol{M}, Y^{(b)})$ respectively, and $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}^{(b)}$ from model (2) based on data $(T, \boldsymbol{X}, \boldsymbol{M})$ and $(T^{(b)}, \boldsymbol{X}, \boldsymbol{M})$ respectively. Third, we calculate three sets of permuted statistics: 1. $\text{OME}_1^{(b)} = f\left(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}^{(b)}\right)$, $\text{CME}_1^{(b)} = g\left(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}^{(b)}\right)$; 2. $\text{OME}_2^{(b)} = f\left(\hat{\boldsymbol{\alpha}}^{(b)}, \hat{\boldsymbol{\beta}}\right)$, $\text{CME}_2^{(b)} = g\left(\hat{\boldsymbol{\alpha}}^{(b)}, \hat{\boldsymbol{\beta}}\right)$; and 3. $\text{OME}_3^{(b)} = f\left(\hat{\boldsymbol{\alpha}}^{(b)}, \hat{\boldsymbol{\beta}}^{(b)}\right)$, $\text{CME}_3^{(b)} = g\left(\hat{\boldsymbol{\alpha}}^{(b)}, \hat{\boldsymbol{\beta}}^{(b)}\right)$ to cover three different types of null hypotheses: 1. $T \not\Rightarrow \mathbf{M} \Rightarrow Y$; 2. $T \Rightarrow \mathbf{M} \not\Rightarrow Y$; and 3. $T \not\Rightarrow \mathbf{M} \not\Rightarrow Y$ in testing the mediation effect, respectively. And the final test statistics for the b^{th} permutation are defined as

$$|\text{OME}^{(b)}| = \max\{|\text{OME}_1^{(b)}|, |\text{OME}_2^{(b)}|, |\text{OME}_3^{(b)}|\} \text{ and}$$

$$\text{CME}^{(b)} = \max\{\text{CME}_1^{(b)}, \text{CME}_2^{(b)}, \text{CME}_3^{(b)}\}$$

respectively. Therefore, the testing *P*-values for OME and CME are

$$p^{\text{OME}} = \frac{\sum_{b=1}^{B} I(|\text{OME}^{(b)}| \geq |\text{OME}|) + 1}{B + 1} \text{ and}$$

$$p^{\text{CME}} = \frac{\sum_{b=1}^{B} I(\text{CME}^{(b)} \geq \text{CME}) + 1}{B + 1}$$

respectively, where $I(\cdot)$ is the indicator function.

3 Results

3.1 Simulation studies

We conducted extensive simulation studies to evaluate: (i) the estimation performance of SparseMCMM in terms of bias and mean squared error (MSE) for DE and ME, respectively; and (ii) the testing performance of SparseMCMM in terms of the empirical type I error rate and power, compared with tests Delta.T and tau.T proposed in Huang and Pan (2016) and HIMA proposed in Zhang *et al.* (2016), representing regularization-based and transformation-based tests to handle high-dimensional mediators respectively. Since these competing tests are designed to deal with continuous mediators, rather than the compositional microbiome data, we log transformed the relative abundances to make them more normal. Detailed introduction for these three competing tests is provided in the Supplementary Materials, Section S3.

3.1.1 Simulation design

We designed our simulation settings based on the experimental design of the murine microbiome study we analyzed as the real data example in Section 3.2 (Schulfer *et al.*, 2019). To make the simulation simple and focused, we did not consider any covariates in the simulation, although our SparseMCMM package has the full capacity to handle any covariate adjustment. We generated the simulation data in three steps: (i) generate the treatment T; (ii) generate the microbiome/mediators $\boldsymbol{M}$ based on treatment T; and (iii) generate the outcome Y based on $(T, \boldsymbol{M})$. The detailed simulation design is provided below.

Generate the treatment T: for n total subjects, we randomly assigned 50% to treatment group ($T = 1$) and the others to control group ($T = 0$). $n = 50, 100, 300$ and 500 for evaluating estimation and $n = 100$ for evaluating testing.

Generate the microbiome M: we simulated microbiome data based on the Dirichlet distribution reflecting the real microbial composition in our murine example. The mean relative abundances of p taxa for the treatment and control groups follow the simplified model (2), as below:

$$E[M_{ij}] = \frac{\gamma_j(T_i)}{\sum_{m=1}^{p} \gamma_m(T_i)}, \quad \log\{\gamma_j(T_i)\} = \beta_{0j} + \beta_{Tj} T_i,$$

where $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p})^T$ represents the log-transformed baseline relative abundances for p taxa and they were set as the corresponding estimates from the male control mice at day 28 (using R package dirmult; Tvedebrink, 2009). The real data include 149 genera and we randomly divided them into p taxa. The specific values of $\boldsymbol{\beta}_0$ with $p = 10$, 25 and 50 used in the simulations are listed in Supplementary Table S1. For the treatment group, we randomly chose p_r out of p taxa as the causal ones ($p_r < p$) and denote the set of their indices as $\wedge$ and their treatment coefficients as $\boldsymbol{\beta}_T^* = (\beta_{T1}^*, \ldots, \beta_{Tp_r}^*)^T$. For those non-causal taxa, $\beta_{Tj} = 0$, $j \notin \wedge$. We considered various sets of $\boldsymbol{\beta}_T^*$ to represent different simulation settings. With the specific $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_T$ values, microbiome composition data $\boldsymbol{M}_i$ could be generated from the Dirichlet distribution.

Generate the outcome Y: the outcome Y was generated based on model (1) with the simulated treatment T, and microbiome composition $\boldsymbol{M}$. We set the regression coefficients of intercept, treatment and non-causal taxa as $\alpha_0 = 0$, $\alpha_T = 1$ and $\alpha_{Mj} = \alpha_{Cj} = 0$, $j \notin \wedge$, respectively. For p_r causal taxa, we denoted their regression coefficients for the main effect and interaction effect by $\boldsymbol{\alpha}_M^* = (\alpha_{M1}^*, \ldots, \alpha_{Mp_r}^*)^T$ and $\boldsymbol{\alpha}_C^* = (\alpha_{C1}^*, \ldots, \alpha_{Cp_r}^*)^T$, respectively, and their values were set differently to represent various simulation scenarios.

Specify $\boldsymbol{\beta}_T^*$, $\boldsymbol{\alpha}_M^*$ and $\boldsymbol{\alpha}_C^*$: First, the number of causal taxa was set as $p_r = 2, 3$ and 5 for $p = 10$, 25 and 50 respectively.

1. To evaluate the performance of the DE and ME estimators, parameter values (Supplementary Table S2) were set up as the corresponding estimates of male mice at day 28 in the real data analysis, so that the true TE and ME were around 1.8 and 0.6 respectively (Supplementary Table S3). Finally, 1000 independent replications were conducted to calculate bias and MSE for the DE and ME estimators. The corresponding computational time is given in Supplementary Table S4.
2. To evaluate the performance of OME and CME tests, we considered two simulation scenarios: (i) all individual ME_js of the causal taxa are positive; and (ii) both positive and negative ME_js are present. Further, four strength levels of ME: null, small, medium and large were considered in each scenario. Note that null ME was used to evaluate the empirical type I error rate and the other three strength were used to evaluate the empirical power. Please see the parameter setting for $\boldsymbol{\beta}_T^*$, $\boldsymbol{\alpha}_M^*$ and $\boldsymbol{\alpha}_C^*$ in

Downloaded from https://academic.oup.com/bioinformatics/article/36/2/347/5536874 by University of Michigan user on 28 October 2020

Table 1. Bias and MSE of the proposed causal direct effect (DE) and mediation effect (ME) estimators for various sample sizes and dimensions of the compositional mediators

		DE		ME	
p	n	Bias	MSE	Bias	MSE
10	50	0.0016	0.0093	−0.0047	0.0091
	100	−0.0025	0.0043	0.0019	0.0047
	300	−0.0025	0.0013	−0.0013	0.0017
	500	−0.0016	0.0009	0.00021	0.0011
25	50	0.062	0.092	0.020	0.35
	100	0.026	0.053	0.019	0.22
	300	0.021	0.051	−0.010	0.16
	500	0.012	0.031	−0.0075	0.090
50	50	0.21	1.70	−0.43	1.00
	100	0.22	0.23	−0.081	0.30
	300	0.038	0.048	−0.026	0.13
	500	0.017	0.037	−0.0039	0.12

Supplementary Table S5. To ease our computation, we restricted this part of simulation to the sample size of 100. The P-values were estimated based on 500 permutations (Supplementary Table S6 reports the computational time), and then the empirical type I error rate and power were calculated by the proportion of P-values less than the given significance level (usually 0.05) with 1000 independent replications.

3.1.2 Estimations of causal direct effect and mediation effect

Table 1 shows the bias and MSE for the DE and ME estimators respectively. First, as the dimension of the mediators p increases from 10 to 50, both bias and MSE of DE and ME estimators increase. In contrast, with each fixed p, as the sample size n goes up, the bias and MSE go down. When $n = 500$, the bias of both DE and ME estimators approach zero. This indicates that the DE and ME estimators are approximately unbiased. The proposed method, SparseMCMM, therefore has good performance in direct effect and mediation effect estimation.

Supplementary Figure S2 exhibits the boxplot of 1000 estimated component-wise ME_js for $j = 1, \ldots, p$, with $p = 10$, 25, 50 and $n = 50, 100, 300, 500$ respectively. The bias and MSE of component-wise ME_js clearly decrease, as the sample size increases. The figures show that for the non-causal taxa, their ME_j estimates all are around zeros, while for the causal taxa, their ME_j estimates all stand out and are away from zero except for taxon 22 when $p = 50$ (its true ME_{22}=0.002). Overall, SparseMCMM presents good performance in both causal taxa selection and its estimations.

3.1.3 Power and type I error rate

Supplementary Table S7 reports the empirical type I error rates of OME, CME, Delta.T, tau.T and HIMA. They all are below the nominal significance levels 5%. OME and CME have conservative type I errors, which is consistent with the results in Boca *et al.* (2014) and Zhang *et al.* (2018). Delta.T and tau.T have relatively more conservative type I error rates than OME and CME do, which agrees with their conservative performance in the power section.

Figure 2 ($p = 10$, 50) and Supplementary Figure S3 ($p = 25$) present the power estimations of OME, CME, Delta.T, tau.T and HIMA for the same effect direction (scenario 1) and mixed effect directions (scenario 2) with small, medium and large overall MEs. Compared to Delta.T, tau.T and HIMA, the proposed tests OME and CME have superior performances in both scenarios with $p = 10$, 25 and 50, and they exhibit an increasing power trend as the overall ME increases. For the comparison between OME and CME, as expected, OME gains more power than CME in scenario 1, when the component-wise ME_js for the causal taxa all are positive, but it is the reverse for scenario 2 when the directions of the individual causal effects are mixed. Among those three competing tests, HIMA has the best performance in both scenarios. However, none has comparable performance to OME or CME. This implies that one needs to be cautious when directly applying the existing high dimensional mediation methods, which are not designed for taking care of the unit sum constraint, to the compositional microbiome data.

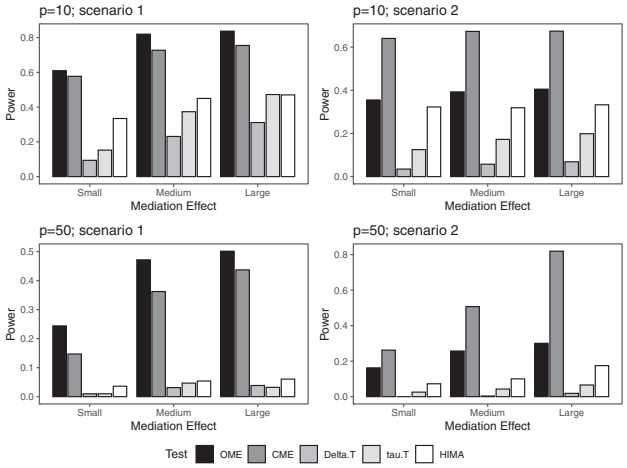

Fig. 2. Empirical power for testing mediation effect with $p = 10$ and 50 in scenarios 1–2 (significance level = 5%). Note that the magnitudes of mediation effect are not comparable across different ps and different scenarios. The detailed setting is given in Supplementary Table S5

3.2 Real data analysis

Schulfer *et al.* (2019) conducted a murine microbiome experiment to explore whether STAT (sub-therapeutic antibiotic treatment) would alter gut microbiome composition and whether this shift would change the body weight gain later in life. Since this study adopts the typical treatment-microbiome (mediator)-outcome design, we use SparseMCMM to re-examine the mediating role that the gut microbiome played in body weight gain. To be specific, we first use OME test to determine whether the overall mediation effect of microbiome is significant, and then use CME test to determine whether at least one individual taxon have significant mediation effects on the body weight gain. Subsequently, SparseMCMM gives the overall ME and individual ME_j estimates for microbiome.

In this study, 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, as described in Caporaso *et al.* (2010). Samples with less than 1800 reads were excluded from the analysis. The OTU table for 21 female (12 STAT and 9 controls) and 37 male (24 STAT and 13 controls) mice was constructed using the QIIME pipeline (Caporaso *et al.*, 2010) at day 21 and 28. Originally there were 149 genera. After filtering those genera that appeared in <10% of mice and with mean proportions $< 10^{-4}$ at each time point separately, there were 38 and 37 genera retained at days 21 and 28 respectively. The observed body weight (in grams) prior to sacrifice, i.e. at day 145 for the female and at day 116 for the male mice was regarded as the outcome. No additional covariates were included in the model, assuming that all potential confounders had been well-controlled in the randomized experiment.

Downloaded from https://academic.oup.com/bioinformatics/article/36/2/347/5536874 by University of Michigan user on 28 October 2020

Table 2. The estimated *P*-values for the microbial mediation effect (at genus rank), the mediation effect estimates and the proportion of the total causal effect on the body weight gain at days 21 and 28 for female and male respectively

	P-value of		Mediation effect	
Day	OME	CME	Estimate	Proportion[a]
Female				
21	0.154	0.438	0.102	3.0%
28	0.048	0.040	0.634	18.8%
Male				
21	0.018	0.002	0.069	3.8%
28	0.004	0.004	0.622	32.2%

[a]Proportion= ME estimate/TE estimate $\times 100\%$.

Table 3. Component-wise point and CI estimates of ME_j for the causal genera at day 28 on body weight gain for female mice

				95% $\hat{CI}$[a]	
Genus	Term1[b]	Term2[c]	$\hat{ME}_j$[d]	Lower	Upper
Akkermansia	0.118	1.466	0.173	0.126	0.220
Lactobacillus	−0.301	−1.289	0.388	0.324	0.452
Eubacterium	0.117	1.462	0.171	0.147	0.199
Rikenellaceae_Other	0.065	−1.508	−0.098	−0.119	−0.077

[a]95% $\hat{CI}$ was calculated by bootstrapping procedure, and the number of bootstrapping is 100.

[b]Term1 represents $(\alpha_{\hat{M}j} + \alpha_{\hat{C}j})$.

[c]Term2 represents $\{\hat{E}[\log(M_j)|T=1] - \hat{E}[\log(M_j)|T=0]\}$.

[d]$\hat{ME}_j = (\alpha_{\hat{M}j} + \alpha_{\hat{C}j})\{\hat{E}[\log(M_j)|T=1] - \hat{E}[\log(M_j)|T=0]\}$.

Supplementary Figure S4 illustrates the distribution of the body weight prior to sacrifice in the control and STAT groups, respectively, for female (left panel) and male mice (right panel). Male mice are known to be heavier than the female mice, and within each gender, the STAT group was heavier than the control group. Considering these different weight distributions between genders and the sexually dimorphic effect of STAT in mice, we explored the mediation effect of gut microbiome on the acceleration of weight gain for female and male mice separately. Since that Delta.T and tau.T tests proposed in Huang and Pan (2016) have no significant results (Supplementary Table S8), and HIMA proposed in Zhang *et al.* (2016) gives testing results only at the genera level (Supplementary Table S9), we only discuss the results for the proposed causal mediation method SparseMCMM next.

Table 2 reports the testing results for OME and CME based on 500 permutations. For the females, OME test shows that the overall mediation effect of microbiome is significant (*P*-value = $0.048 < 0.05$) at day 28, but not significant at day 21 (*P*-value = 0.154). The estimated overall ME increases from 0.102 (3% of total causal effect on the weight gain) at day 21 to 0.634 (18.8%) at day 28. On the other hand, CME test is also significant at day 28 with *P*-value 0.04, which shows that there is at least one genus playing a mediation role on the weight gain at day 28. With SparseMCMM, we further identify four candidate genera, reported in Table 3 with the point and 95% confidence interval (CI) estimates for their mediation effects. Table 3 also reports the breakdown of mediation effect estimates. Column 2 indicates the microbiome effect on the weight, column 3 indicates the treatment effect on the microbiome, and the multiplication of those two columns equals the mediation effect estimate for each genus in column 4. Among the identified genera, three genera (*Akkermansia*, *Eubacterium* and *Lactobacillus*) had positive mediation effects. Genera *Akkermansia* and *Eubacterium* were positively associated with weight gain and STAT increased the weight gain by increasing their relative abundances, while *Lactobacillus* was negatively associated with weight gain and STAT increased the weight gain by decreasing its relative abundance; an unclassified genus from family *Rikenellaceae* had a negative mediation effect: it was positively associated with weight gain, however STAT decreased the weight gain by decreasing its relative abundance. The observation about the effect of *Lactobacillus* on weight gain is consistent with several prior studies (Armougom *et al.*, 2009; Clarke *et al.*, 2012; Turnbaugh *et al.*, 2008).

Similar analyses have been done for male mice. Please see them in the Supplementary Materials, Section S5. In summary, we provide evidence that microbiome plays a significant mediating role in the relationship between antibiotic usage and weight gain for both female and male mice, especially at day 28 in this study. We also observe differences in causal genera between genders, which has been well-demonstrated in Schulfer *et al.* (2019). Among the identified causal genera, the effect directions are mixed and CME test has more significant results than OME, which suggests that CME test is more powerful when both positive and negative component-wise mediation effects are present.

4 Discussion

In this paper, we proposed a rigorous causal mediation analytic framework SparseMCMM to investigate the causal mediating role of the high-dimensional and compositional microbiome in the relationship between a treatment and a continuous outcome (see workflow of SparseMCMM in the Supplementary Materials, Section S7). We quantified the causal direct effect of treatment, the overall mediation effect of microbiome community and the component-wise mediation effect for each individual microbe under the counterfactual framework. We developed regularization strategies to handle the high-dimensional mediators and to select the signature causal microbes. We further proposed two tests to examine whether the overall mediation effect of microbiome community on the outcome is significant (test OME) or at least one of component-wise ME_js is significantly non-zero (test CME), respectively. Through extensive simulations, we demonstrated that SparseMCMM provided asymptotically unbiased DE and ME estimates with small MSEs and the proposed tests OME and CME controlled their type I error rate around the significance level even under the constraint of data sparsity. Compared with the competing methods (Delta.T, tau.T and HIMA), SparseMCMM uniformly achieved higher statistical power under almost all scenarios. Finally, we applied SparseMCMM to analyze a STAT murine microbiome study and to detect unambiguous causal paths among STAT (antibiotic treatment), gut microbiome and body weight (outcome) for both female and male mice respectively.

Recently, Sohn and Li (2019) proposed a compositional causal mediation model (CMM) to describe the relationships among treatment, microbiome composition and continuous outcome. Although sharing a similar framework with CMM, our proposed method SparseMCMM has several essential differences. First, we use Dirichlet regression to characterize the relationship between treatment and microbiome composition, while Sohn and Li utilize the algebraic structure of a composition under the simplex space. Secondly, with Dirichlet regression, SparseMCMM can handle the interaction between treatment and microbiome with relation to the outcome in a more flexible manner through the proposed

Downloaded from https://academic.oup.com/bioinformatics/article/36/2/347/5536874 by University of Michigan user on 28 October 2020

regularization strategy, which addresses concerns regarding the potential bias caused by neglecting the presence of interaction effects (Richiardi *et al.*, 2013; Valeri and VanderWeele, 2013). Moreover, SparseMCMM can automatically drop the interaction terms when the data suggest that they are absent with the proposed penalized least squares criterion; CMM lacks such flexibility. Thirdly, SparseMCMM selects casual taxa with regularization techniques, while CMM identifies the key taxa based on confidence interval estimates.

Considering the completely distinct model assumptions, we did not include CMM in our simulation study. However, we applied CMM (through its R package) to our real data example (only at day 28) to compare its results with ours. For female mice, CMM only identified that genus *Lactobacillus* and an unclassified genus from family *Rikenellaceae* had significant mediation effects on the weight gain. Both of these genera were on the causal list from SparseMCMM. However, no causal overall or component-wise mediation effect was identified by CMM for male mice. Overall, SparseMCMM captured more causal signals than CMM in this real data example, which further demonstrates the sound performance of SparseMCMM.

The phylogenetic tree describes the taxonomical and evolutionary relationships among taxa and provides the possibility to further interpret the casual path among treatment, microbiome and outcome (Knight *et al.*, 2018; Silverman *et al.*, 2017). If the causal taxa are phylogenetically related, utilizing the phylogenetic tree information will increase efficiency of SparseMCMM as in the microbiome association tests (Hu *et al.*, 2018; Koh *et al.*, 2017). Our SparseMCMM R package provides a tree option which takes the prior structure information into account by incorporating a graph Laplacian penalty induced by the phylogenetic tree (Chen *et al.*, 2013; Li and Li, 2008). However, please note that since this option require an additional tuning parameter estimation, it increases the computational complexity. Due to the unknown true nature of the state and the tradeoff of the computational time, we defer the choice of utilization of the phylogenetic tree in SparseMCMM to the user.

In addition, the recent increase in the number of microbial longitudinal studies offers new opportunities to investigate the dynamics of microbial communities. Recent investigations show that the dynamics and stability of the microbial community could be a strong predictor of disease activity (Halfvarson *et al.*, 2017; Novakova *et al.*, 2017), but it is challenging to accommodate the high-dimensional longitudinal mediators in the causal mediation model. One limitation of SparseMCMM is that it can only take a single time point of microbiome data into the proposed framework. As a future research area, we will aim to incorporate microbial dynamic system modelling (Zhang and Davis, 2013) into SparseMCMM to provide a causal path relating treatment, longitudinal microbiome composition and outcome.

The proposed causal mediation models in this study has been developed into the SparseMCMM R package, and can be installed from https://sites.google.com/site/huilinli09/software and https://github.com/chanw0/SparseMCMM.

Funding

This work was supported in part by National Institutes of Health grants R01DK090989, R01DK110014 and U01AI22285, the Fondation Leducq Transatlantic Network, and the Zlinkoff and C&D Funds.

Conflict of Interest: none declared.

References

Aitchison,J. (1982) The statistical analysis of compositional data. *J. R. Stat. Soc. Ser. B (Methodological)*, **44**, 139–177.

Aitchison,J. and Bacon-Shone,J. (1984) Log contrast models for experiments with mixtures. *Biometrika*, **71**, 323–330.

Albenberg,L.G. and Wu,G.D. (2014) Diet and the intestinal microbiome: associations, functions, and implications for health and disease. *Gastroenterology*, **146**, 1564–1572.

Armougom,F. *et al.* (2009) Monitoring bacterial community of human gut microbiota reveals an increase in lactobacillus in obese patients and methanogens in anorexic patients. *PLoS One*, **4**, e7125.

Boca,S.M. *et al.* (2014) Testing multiple biological mediators simultaneously. *Bioinformatics*, **30**, 214–220.

Bonnans,J.-F. *et al.* (2006) *Numerical Optimization: Theoretical and Practical Aspects*. Springer Science & Business Media.

Campbell,G. and Mosimann,J. (1987a) *Modelling Continuous Proportional Data with the Dirichlet Distribution*. Unpublished manuscript.

Campbell,G. and Mosimann,J. (1987b) Multivariate methods for proportional shape. In: *ASA Proceedings of the Section on Statistical Graphics*, vol. 1, pp. 10–17. Washington.

Caporaso,J.G. *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, **7**, 335.

Chen,J. and Chen,Z. (2008) Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, **95**, 759–771.

Chen,J. *et al.* (2013) Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics*, **14**, 244–258.

Chén,O.Y. *et al.* (2018) High-dimensional multivariate mediation with application to neuroimaging data. *Biostatistics*, **19**, 121–136.

Clarke,S.F. *et al.* (2012) The gut microbiota and its relationship to diet and obesity: new insights. *Gut Microbes*, **3**, 186–202.

Fan,J. and Lv,J. (2008) Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)*, **70**, 849–911.

Fischbach,M.A. (2018) Microbiome: focus on causation and mechanism. *Cell*, **174**, 785–790.

Friedman,J. *et al.* (2001) *The Elements of Statistical Learning*. Springer Series in Statistics, New York.

Gilbert,J.A. *et al.* (2018) Current understanding of the human microbiome. *Nat. Med.*, **24**, 392.

Halfvarson,J. *et al.* (2017) Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat. Microbiol.*, **2**, 17004.

Hijazi,R.H. and Jernigan,R.W. (2009) Modelling compositional data using Dirichlet regression models. *J. Appl. Probab. Stat.*, **4**, 77–91.

Holland,P.W. (1986) Statistics and causal inference. *J. Am. Stat. Assoc.*, **81**, 945–960.

Hu,J. *et al.* (2018) A two-stage microbial association mapping framework with advanced FDR control. *Microbiome*, **6**, 131.

Huang,Y.-T. and Pan,W.-C. (2016) Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. *Biometrics*, **72**, 402–413.

Knight,R. *et al.* (2018) Best practices for analysing microbiomes. *Nat. Rev. Microbiol.*, **16**, 410.

Koh,H. *et al.* (2017) A powerful microbiome-based association test and a microbial taxa discovery framework for comprehensive association mapping. *Microbiome*, **5**, 45.

Kraft,D. (1988) A software package for sequential quadratic programming. *Forschungsbericht- Deutsche Forschungs- und Versuchsanstalt fur Luft- und Raumfahrt*.

Li,C. and Li,H. (2008) Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, **24**, 1175–1182.

Lin,W. *et al.* (2014) Variable selection in regression with compositional covariates. *Biometrika*, **101**, 785–797.

Livanos,A.E. *et al.* (2016) Antibiotic-mediated gut microbiome perturbation accelerates development of type 1 diabetes in mice. *Nat. Microbiol.*, **1**, 16140.

Downloaded from https://academic.oup.com/bioinformatics/article/36/2/347/5536874 by University of Michigan user on 28 October 2020

Mahana,D. *et al*. (2016) Antibiotic perturbation of the murine gut microbiome enhances the adiposity, insulin resistance, and liver disease associated with high-fat diet. *Genome Med*., **8**, 48.

Neyman,J. (1923) Sur les applications de la théorie des probabilités aux experiences agricoles: essai des principes. *Roczniki Nauk Rolniczych*, **10**, 1–51.

Ni,J. *et al*. (2017) A role for bacterial urease in gut dysbiosis and Crohn's disease. *Sci. Transl. Med*., **9**, eaah6888.

Novakova,E. *et al*. (2017) Mosquito microbiome dynamics, a background for prevalence and seasonality of west Nile virus. *Front. Microbiol*., **8**, 526.

Peixoto,J.L. (1987) Hierarchical variable selection in polynomial regression models. *Am. Stat*., **41**, 311–313.

Radchenko,P. and James,G.M. (2010) Variable selection using adaptive nonlinear interaction structures in high dimensions. *J. Am. Stat. Assoc*., **105**, 1541–1553.

Richiardi,L. *et al*. (2013) Mediation analysis in epidemiology: methods, interpretation and bias. *Int. J. Epidemiol*., **42**, 1511–1519.

Rubin,D.B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol*., **66**, 688.

Rubin,D.B. (2005) Causal inference using potential outcomes: design, modeling, decisions. *J. Am. Stat. Assoc*., **100**, 322–331.

Schulfer,A.F. *et al*. (2018) Intergenerational transfer of antibiotic-perturbed microbiota enhances colitis in susceptible mice. *Nat. Microbiol*., **3**, 234.

Schulfer,A.F. *et al*. (2019) The impact of early-life sub-therapeutic antibiotic treatment (stat) on excessive weight is robust despite transfer of intestinal microbes. *ISME J*., **13**, 1280.

Silverman,J.D. *et al*. (2019) A phylogenetic transform enhances analysis of compositional microbiota data. *Elife*, **6**, e21887.

Sohn,M.B. and Li,H. (2019) Compositional mediation analysis for microbiome studies. *Ann. Appl. Stat*., **13**, 661–681.

Stein,M.M. *et al*. (2016) Innate immunity and asthma risk in Amish and Hutterite farm children. *N. Engl. J. Med*., **375**, 411–421.

Taylor,A.B. and MacKinnon,D.P. (2012) Four applications of permutation methods to testing a single-mediator model. *Behav. Res. Methods*, **44**, 806–844.

Turnbaugh,P.J. *et al*. (2008) Marked alterations in the distal gut microbiome linked to diet-induced obesity. *Cell Host Microbe*, **3**, 213.

Tvedebrink,T. (2009) dirmult: Estimation in Dirichlet-multinomial distribution. *R Package Version 0.1*, **3**.

Valeri,L. and VanderWeele,T.J. (2013) Mediation analysis allowing for exposure–mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. *Psychol. Methods*, **18**, 137.

VanderWeele,T. and Vansteelandt,S. (2014) Mediation analysis with multiple mediators. *Epidemiol. Methods*, **2**, 95–115.

VanderWeele,T.J. (2013) A three-way decomposition of a total effect into direct, indirect, and interactive effects. *Epidemiology (Cambridge, Mass.)*, **24**, 224.

VanderWeele,T.J. (2014) A unification of mediation and interaction: a four-way decomposition. *Epidemiology (Cambridge, Mass.)*, **25**, 749.

VanderWeele,T.J. (2016) Mediation analysis: a practitioner's guide. *Annu. Rev. Public Health*, **37**, 17–32.

VanderWeele,T.J. and Vansteelandt,S. (2009) Conceptual issues concerning mediation, interventions and composition. *Stat. Interface*, **2**, 457–468.

VanderWeele,T.J. and Vansteelandt,S. (2010) Odds ratios for mediation analysis for a dichotomous outcome. *Am. J. Epidemiol*., **172**, 1339–1348.

Ypma,J. (2014) Introduction to nloptr: an R interface to NLopt. Technical Report.

Zeevi,D. *et al*. (2015) Personalized nutrition by prediction of glycemic responses. *Cell*, **163**, 1079–1094.

Zhang,C.-H. *et al*. (2010) Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat*., **38**, 894–942.

Zhang,H. *et al*. (2016) Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics*, **32**, 3150–3154.

Zhang,J. *et al*. (2018) A distance-based approach for testing the mediation effect of the human microbiome. *Bioinformatics*, **34**, 1875–1883.

Zhang,Y. and Davis,R. (2013) Principal trend analysis for time-course data with applications in genomic medicine. *Ann. Appl. Stat*., **7**, 2205–2228.

Zheng,P. *et al*. (2016) Gut microbiome remodeling induces depressive-like behaviors through a pathway mediated by the host's metabolism. *Mol. Psychiatry*, **21**, 786.

Downloaded from https://academic.oup.com/bioinformatics/article/36/2/347/5536874 by University of Michigan user on 28 October 2020