Bioinformatics, 34(11), 2018, 1875–1883
doi: 10.1093/bioinformatics/bty014
Advance Access Publication Date: 15 January 2018
Original Paper

Genetics and population analysis

A distance-based approach for testing the mediation effect of the human microbiome

Jie Zhang[1], Zhi Wei[2,*] and Jun Chen[3,*]

[1]Adobe Systems Incorporated, San Jose, CA 95110, USA, [2]Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102, USA and [3]Division of Biomedical Statistics and Informatics and Center for Individualized Medicine, Mayo Clinic, Rochester, MN 55905, USA

*To whom correspondence should be addressed.

BMJ ref
① what is the microbiome? → 1 slide
② what is mediation analysis?

Associate Editor: Oliver Stegle

Received on May 12, 2017; revised on November 30, 2017; editorial decision on December 31, 2017; accepted on January 12, 2018

Abstract

Motivation: Recent studies have revealed a complex interplay between environment, the human microbiome and health and disease. Mediation analysis of the human microbiome in these complex relationships could potentially provide insights into the role of the microbiome in the etiology of disease and, more importantly, lead to novel clinical interventions by modulating the microbiome. However, due to the high dimensionality, sparsity, non-normality and phylogenetic structure of microbiome data, none of the existing methods are suitable for testing such clinically important mediation effect.

Results: We propose a distance-based approach for testing the mediation effect of the human microbiome. In the framework, the nonlinear relationship between the human microbiome and independent/dependent variables is captured implicitly through the use of sample-wise ecological distances, and the phylogenetic tree information is conveniently incorporated by using phylogeny-based distance metrics. Multiple distance metrics are utilized to maximize the power to detect various types of mediation effect. Simulation studies demonstrate that our method has correct Type I error control, and is robust and powerful under various mediation models. Application to a real gut microbiome dataset revealed that the association between the dietary fiber intake and body mass index was mediated by the gut microbiome.

Availability and implementation: An R package 'MedTest' is freely available at https://github.com/jchen1981/MedTest.

Contact: zhiwei@njit.edu or chen.jun2@mayo.edu

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

1 Introduction

With the advancement of the next generation sequencing techniques (Zhang and Wei, 2016; Zhang *et al.*, 2017), numerous microbiome studies have been conducted at different body sites, such as skin, gut and respiratory tract (Arumugam *et al.*, 2011; Charlson *et al.*, 2010; Cotillard *et al.*, 2013; David *et al.*, 2014; Grice *et al.*, 2009; Nunes-Alves, 2016; Qin *et al.*, 2010; Wu *et al.*, 2011; Yatsunenko *et al.*, 2012), with the aim to understand the genetic and environmental forces shaping the human microbiome, and the relationship between the microbiome compositional variation and biological or clinical outcomes (Chen *et al.*, 2012). With an in-depth understanding of the biological mechanisms underlying the complex interplay between environmental and genetic factors, microbiome compositions and diseases, we could develop clinical interventions to treat diseases by modulating the related microbiota (Faith *et al.*, 2013; Le Chatelier *et al.*, 2013; Lozupone *et al.*, 2012).

Quite a few statistical methods have been proposed to test the association between the microbiome compositions and covariates of interest (e.g. environmental factors or clinical outcomes) based on 16S data, where 16S rRNA gene sequence tags are clustered into

OTUs (operational taxonomic units) based on sequence divergence (Chen *et al.*, 2012; Tang *et al.*, 2016; Zhao *et al.*, 2015). They usually utilize distance metrics, measuring the pairwise dissimilarity in the microbiome profiles, to compute test statistics, and employ permutation tests to calculate the *P*-value. The performance of these distance-based methods depends on the choice of the distance metric (Chen *et al.*, 2012). Numerous distance measures, with different properties and capabilities, have been proposed to detect diverse patterns in microbiome data (Bray and Curtis, 1957; Chen *et al.*, 2012; Lozupone and Knight, 2005; Lozupone *et al.*, 2007). However, it is usually difficult to choose a proper distance in advance for a particular dataset (Zhao *et al.*, 2015). Recently, methods that accommodate multiple distances have been proposed (Tang *et al.*, 2016; Zhao *et al.*, 2015). They provide more interpretable results with controlled Type I error rate, and yield good performance comparable to the best choice of distance metric. However, previous methods could only analyze and test bivariate relations.

Recent studies revealed that there is a complex interplay among the environmental (or genetic) factors, the human microbiome and health. For example, the facts that long-term dietary intake influences the composition of microorganisms residing in the human gut (David *et al.*, 2014) and the composition of the gut microbiota determines the efficacy of nutrients harvest from food (Cotillard *et al.*, 2013) strongly suggest that microbiome may mediate the effect of long-term diet on human health (Sonnenburg and Bäckhed, 2016). Co-localization of genetic variants associated with both the microbiome and disease indicates a potential mediation role of the microbiome in conferring the genetic susceptibility to disease (Snijders *et al.*, 2016). Nonetheless, few statistical methods have been developed to efficiently analyze such trivariate relationship in the microbiome data, and to test whether the effect of some independent variable on an outcome variable is mediated by the microbiome.

Traditional mediation analysis, which tests if a single variable mediates the relationship between a known exposure and an outcome, has been widely applied in biomedical, behavioral, and psychosocial studies (Baron and Kenny, 1986; MacKinnon, 2008; Zhang *et al.*, 2016). Recently, Boca *et al.* (2014) proposed a permutation-based approach to test multiple putative mediators between a known risk factor and a disease, which controlled the family-wise error rate (FWER). Zhang *et al.* (2016) extended the multiple mediator model to the high-dimensional setting and studied how the methylation markers mediate the relationship between smoking and lung function. However, none of the above-mentioned methods could be directly applied to test the mediation effect of the microbiome. Microbiome data are highly skewed and zero-inflated, which violates the basic assumptions (e.g. normality or linearity) of existing methods. Due to the complex relationship among the microbiome, environment and disease, linear models are not appropriate. In addition, the large number of rare and low-abundance OTUs makes the individual OTU-based testing less powerful. The power is exacerbated by multiple testing correction. To increase the power of the test, one commonly used strategy is to group the OTU data based on either within-data correlation structure or prior structure inferred from an external source. For microbiome data, OTUs are evolutionarily related to each other by a phylogenetic tree, and environmental factors usually affect bacterial clade (a cluster of OTUs) due to the sharing of a similar biological function. Thus using the tree to guide the OTU grouping could potentially improve the statistical power.

In this article, we propose a powerful and robust statistical tool for testing the mediation effect of the human microbiome. Instead of working with the original OTU data, the method uses the sample-wise distance matrices. By using distance metrics, we achieve effective dimension reduction by pooling individually weak signals. Moreover, the distance approach provides the flexibility of incorporating the prior structure information so that the information pooling is prior-guided. Since each distance metric implicitly represents some (non-linear) transformation of the OTU abundances, the use of distance metrics could model a wide range of mediation effects. In the framework, we consider multiple distance metrics, including both the phylogenetic tree-based and non-tree-based distances, to capture diverse types of mediation effect. We adopt permutation to evaluate the significance. Permutation allows adjusting to an unknown distribution of the test statistic, and thus properly controls the Type I error. Extensive simulation studies showed that our method could precisely control the Type I error and is robust and powerful under different mediation models. Finally, we analyze a real dataset to demonstrate that our method is powerful in detecting the mediation effect.

2 Materials and methods

2.1 Mediation model

Let $\boldsymbol{M}$ be an $n \times m$ count matrix, which measures the abundances of m OTUs on n microbiota samples. Let X be an $n \times 1$ vector for the independent variable and Y be an $n \times 1$ vector for the outcome variable. We assume that the microbiome mediates the effect of X (e.g. environmental exposure) on Y (e.g. disease phenotype) through some unknown microbiome feature vector $f_M^{(l)}(l = 1, \ldots, L)$. A microbiome feature could be the abundance or prevalence of a taxonomic group, the weighted average of several functionally related OTUs or even the richness of the entire microbial community. More generally, a microbiome feature could be defined as a scalar function of the original OTU abundance vector

$$f_M^{(l)} : \mathbb{R}^m \to \mathbb{R}.$$

Due to the multivariate nature of the microbiome data, it is possible to have multiple microbiome features that mediate the effect $(l = 1, \ldots, L)$. Following previous literature (Boca *et al.*, 2014; Zhang *et al.*, 2016), we assume the following mediation model

$$\begin{aligned} Y &= X\gamma^* + \boldsymbol{\varepsilon} \\ f_M^{(l)} &= X\alpha_l + \boldsymbol{\varepsilon}_l' \quad (l = 1 \cdots L) \\ Y &= \textstyle\sum_l^L f_M^{(l)}\beta_l + X\gamma + \boldsymbol{\varepsilon}'' \end{aligned} \tag{1}$$

where γ^* and γ represent the total effect and the direct effect of the independent variable X on the outcome Y, respectively; $\boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon}'$ and $\boldsymbol{\varepsilon}''$ are random errors. The mediation effect is denoted by the path $X \to \boldsymbol{M} \to Y$ (Fig. 1). Note that, by using f_M, the model could explore potential nonlinear relationship between the OTU abundances and X or Y. Though we assume a linear relationship between X and Y, it could be extended to generalized linear model using a link function. Also note that potential confounders Z could be easily adjusted in the model and, for simplicity of notations, we omit the potential confounders in the model.

We consider the classical three-step method for mediational analysis (Baron and Kenny, 1986; Judd and Kenny, 1981; MacKinnon *et al.*, 2007). We assume that a significant relation between the independent variable X and the dependent variable Y, which is required by the first equation in (1), has already been satisfied, as this is the basis for mediation analysis. To establish mediation, we need to test whether there is a significant relation of the independent variable X to some mediating variable $f_M^{(l)}$, and whether the mediating variable $f_M^{(l)}$ is significantly related to the dependent variable Y, when adjusted by the independent variable.

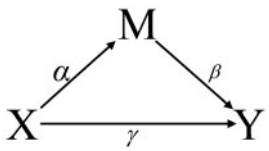

Fig. 1. Mediation model

The null hypothesis H_0 can be expressed as

$$H_0 : \alpha_l \beta_l = 0 \text{ for } \forall f_M^{(l)},$$

and the alternative hypothesis H_1 is that there exists some $f_M^{(l)}$ such that $\alpha_l \beta_l \neq 0$ (Zhang *et al.*, 2016).

2.2 A distance-based test for mediation effect

Apparently, if the mediating microbiome features ($f_M^{(l)}$) are known a priori, we could apply traditional mediation tests. In practice, we have little knowledge about the specific microbiome features that mediate the effect. The power of the mediation test thus relies on a good choice of microbiome features that capture the mediation relationship as precisely as possible. One simple strategy is to treat the abundance of each OTU as the microbiome feature, perform tests on all the OTUs and apply Bonferroni correction to control the FWER. However, due to the extreme sparsity in the OTU data, individual tests are usually underpowered. To enrich signals and reduce multiple testing burden, community-level analysis, which considers all OTUs jointly, has been proposed to improve the power (Zhao *et al.*, 2015). One possible approach is to perform principal component analysis (PCA) and the principal components (PC) are used as microbiome features for mediation test. PCA defines the microbiome features based on the within-data correlations. However, for microbiome data, we are more interested in defining the microbiome features based on the phylogenetic tree of OTUs. Environmental exposure or disease usually affects a cluster of phylogenetically related OTUs, which share a similar biological function. To accommodate the tree structure, we propose to form PCs respecting the tree structure so that the PCs could capture the variation of evolutionarily related OTUs. One way to achieve this is through multidimensional scaling (also known as 'principal coordinate analysis') on a distance matrix, where the distance incorporates the tree structure information. Given the availability of many ecological distances, this approach is particularly appealing.

We thus apply a distance-based non-parametric method to test the mediation effects. The test consists of two parts: a distance-based test statistic and a permutation scheme to approximate the distribution under the null. Let $\boldsymbol{D} = (d_{ij}) \in \mathbb{R}^{n \times n}$ be the distance matrix that measures the dissimilarity between the samples based on their microbiota profiles. The microbiome features could be formed by performing eigen-decompostion on the double centered matrix of squared distances, which is defined as

$$\boldsymbol{G} = \left(\boldsymbol{I} - \frac{11'}{n}\right)\boldsymbol{A}\left(\boldsymbol{I} - \frac{11'}{n}\right),$$

where $\boldsymbol{I}$ is the identity matrix and 1 is a vector of 1's and $\boldsymbol{A} = (a_{ij}) = \left(-\frac{1}{2}d_{ij}^2\right)$. Let $(u_1, u_2, \ldots, u_L)$ be the L eigenvectors associated with positive eigenvalues $(\lambda_1, \lambda_2, \ldots, \lambda_L)$. We take these eigenvectors $(u_1, u_2, \ldots, u_L)$ as the microbiome features.

Suppose that X_Z is the residual vector after the effects of confounder Z has been regressed out, and $Y_{X,Z}$ the residual vector after the effects of X, Z have been regressed out. The test statistic for mediation is formulated as

$$\begin{aligned} T &= \sum_{l=1}^{L} \lambda_l |\widehat{\alpha}_l \widehat{\beta}_l| \\ &= \sum_{l=1}^{L} \lambda_l | < X_Z, u_l > < Y_{X,Z}, u_l > |, \end{aligned} \tag{2}$$

where $|.|$ is the absolute value function and $< ., . >$ is the inner product. Note that λ_j is proportional to the percentage of explained variance, and we assign more weights to the microbiome features that account for larger variance.

We apply permutation test to calculate the P-value based on the proposed test statistic. For jth permutation, $j \in \{1, 2, \ldots, B\}$, we permute the residual vectors X_Z and $Y_{X,Z}$ separately and denote the permuted residual vectors as $X_Z^{(j)}$ and $Y_{X,Z}^{(j)}$. We calculate the following statistics

$$T_X^{(j)} = \sum_{l=1}^{L} \lambda_l | < X_Z^{(j)}, u_l > < Y_{X,Z}, u_l > |,$$

$$T_Y^{(j)} = \sum_{l=1}^{L} \lambda_l | < X_Z, u_l > < Y_{X,Z}^{(j)}, u_l > |,$$

$$T_{X,Y}^{(j)} = \sum_{l=1}^{L} \lambda_l | < X_Z^{(j)}, u_l > < Y_{X,Z}^{(j)}, u_l > |.$$

The final test statistic under the jth permutation is calculated as

$$T^{(j)} = \max\{T_X^{(j)}, \quad T_Y^{(j)}, \quad T_{X,Y}^{(j)}\}.$$

The P-value is obtained as the proportion of $\{T, T^{(1)}, T^{(2)}, \ldots, T^{(B)}\}$ that is equal to or larger than the observed statistic T. Note that the permutation strategy reflects the three different types of null hypotheses ($X \rightarrow \boldsymbol{M} \nrightarrow Y$, $X \nrightarrow \boldsymbol{M} \rightarrow Y$, and $X \nrightarrow \boldsymbol{M} \nrightarrow Y$).

2.3 Distances

Numerous distance measures have been proposed to quantify the difference between microbial composition profiles (Kuczynski *et al.*, 2010; Swenson, 2011). On the one hand, they could be generally categorized into abundance-based distances (or quantitative measures) and presence-absence-based distances (or qualitative measures) (Tang *et al.*, 2016). The presence-absence distances only consider the presence and absence information of the species, while abundance distances utilize species abundance data (e.g. counts or relative proportions) to compare microbial communities. Thus they have different efficiency in detecting changes in community composition or community structure. On the other hand, based on whether a phylogenetic tree is involved in computing the dissimilarity matrix, they could be divided into tree-based and non-tree-based distances. Distance measures, which incorporate phylogenetic information, account for the degree of divergence between different sequences (Lozupone and Knight, 2005; Lozupone *et al.*, 2007) and hence is most powerful to detect change of clustered signals (i.e. bacterial clades). In contrast, non-tree-based distance is more powerful to detect randomly distributed signals. We thus use standard ecological distances from each category to have a more comprehensive view of the microbiome. We assume the data are rarefied to the same depth before calculating the distances. We consider the following distances.

2.3.1 Jaccard and Bray–Curtis distances

Among the non-tree-based distances, Jaccard distance is a qualitative measure that utilizes presence–absence data of the species. Let n_{jk}^{10}, n_{jk}^{01} and n_{jk}^{11} denote the count of the species that present in sample j only, sample k only, and both samples, respectively. Jaccard distance between sample j and k is defined as

$$d_J = \frac{n_{jk}^{10} + n_{jk}^{01}}{n_{jk}^{10} + n_{jk}^{01} + n_{jk}^{11}}.$$

In contrast, Bray–Curtis distance is a quantitative measure based on the abundance of species (Bray and Curtis, 1957). Let p_{ji} and p_{ki}, for $i = 1, \ldots, m$, be the relative abundance of OTU i in samples j and k, respectively. As defined by Bray and Curtis, the index of dissimilarity is

$$d_{\mathrm{BC}} = \frac{\sum_{i=1}^{m} |p_{ji} - p_{ki}|}{\sum_{i=1}^{m} (p_{ji} + p_{ki})}.$$

Note that the presence-absence version of Bray–Curtis distance is actually equivalent to the Jaccard distance, as the difference is usually ignorable (Tang *et al.*, 2016).

2.3.2 Unweighted, weighted and generalized UniFrac distances

The unique fraction metric, or UniFrac distance, which takes into account the phylogenetic relationship between OTUs, is frequently used to summarize the overall microbiota variability. The original UniFrac distance comes in two versions: the unweighted UniFrac uses only the presence/absence data (Lozupone and Knight, 2005), while weighted UniFrac is based on the relative abundance of each taxon (Lozupone *et al.*, 2007). Let b_i, for $i = 1, \ldots, m_b$, denote the length of ith branch of the phylogenetic tree, and p_i^A and p_i^B denote the cumulative proportions of all OTUs descending from the ith branch for community A and B, respectively. The unweighted UniFrac distance is mathematically defined as

$$d_{\mathrm{U}} = \frac{\sum_{i=1}^{m_b} b_i |I(p_i^{\mathrm{A}} > 0) - I(p_i^{\mathrm{B}} > 0)|}{\sum_{i=1}^{m_b} b_i},$$

where $I(\cdot)$ is the binary indicator function. The unweighted UniFrac is the most efficient to capture the variability in community membership or the abundance of rare lineages, since the probability of being sequenced for these rare taxa is directly related to the presence/absence of species (Chen *et al.*, 2012). On the contrary, weighted UniFrac distance, which is defined as

$$d_{\mathrm{W}} = \frac{\sum_{i=1}^{m_b} b_i |p_i^{\mathrm{A}} - p_i^{\mathrm{B}}|}{\sum_{i=1}^{m_b} b_i (p_i^{\mathrm{A}} + p_i^{\mathrm{B}})},$$

is the most efficient to capture the variability in the abundant lineages, because these abundant lineages contribute the most weights (Chen *et al.*, 2012). However, both unweighted and weighted UniFrac distances have limited ability to capture the variability of taxa in the middle of the abundance spectrum, where a significant portion of the taxa lie. Chen *et al.* (2012) proposed a generalized version of UniFrac distance to address the limitations of the traditional UniFrac distances. The generalized UniFrac distance is defined as

$$d_G^{(\alpha)} = \frac{\sum_{i=1}^{m_b} b_i (p_i^{\mathrm{A}} + p_i^{\mathrm{B}})^{\alpha} \left| \frac{p_i^{\mathrm{A}} - p_i^{\mathrm{B}}}{p_i^{\mathrm{A}} + p_i^{\mathrm{B}}} \right|}{\sum_{i=1}^{m_b} b_i (p_i^{\mathrm{A}} + p_i^{\mathrm{B}})^{\alpha}}.$$

The distance, $d_{\mathrm{G}}^{(0.5)}$, where $\alpha = 0.5$, has been shown to be robust and is very efficient to capture the microbiota variability in these moderately abundant lineages. Besides unweighted and weighted UniFrac, we also use $d_{\mathrm{G}}^{(0.5)}$ (we will drop the superscript for simplicity) in the mediation model to summarize the microbiota variability. By using different UniFrac distances to summarize the overall microbiota variability, more insights can be gained about the source of microbiota variability.

2.4 An omnibus test based on multiple distance measures

Each distance represents a distinct view of the microbiota and is expected to be the most powerful to detect a specific mediation pattern. However, in real applications, we may have little knowledge about the underlying mediation mechanism. Sticking to a single distance could miss important mediation effect. Therefore, considering different distance measures is key to the robustness and power of the test. Here we consider phylogenetic tree-based distances, i.e. unweighted, weighted and generalized UniFrac distances, together with non-tree-based Jaccard and Bray–Curtis distances. One possible solution for ensembling different distances is that we compute the testing statistic for each distance, apply permutation tests to calculate P-values, and report the final P-value as the minimum P-value across multiple distances adjusted by Bonferroni correction for multiple comparisons. However, this method may lose some power due to the correlations between the distance measures. Thus we propose an omnibus test to ensemble multiple distances so that it is more powerful in detecting mediation effects.

Suppose that there are K distances, which are denoted as $d_1, d_2, \ldots, d_K$. Algorithm 1 shows the detailed steps of the proposed testing procedure. To integrate different distances simultaneously, we compute the minimum P-value across multiple distances as the test statistic and simulate its distribution under null hypothesis through permutation procedures. As mentioned by Tang *et al.* (2016), a larger B is usually desired for achieving accurate P-values.

3 Results

3.1 Simulation strategy

We conduct extensive simulation studies to investigate the performance of the proposed method. We demonstrate that our method,

Algorithm 1 A Distance-based omnibus test for mediation effect

1. Compute the test statistic T_k for each distance d_k.
2. For each distance, generate B permuted statistics $T_k^{(j)}$, for $j = 1, 2, \ldots, B$, by residual permutations.
3. Calculate the P-value P_k for each distance d_k as the proportion of statistics $T_k, T_k^{(1)}, \ldots, T_k^{(B)}$ that is equal to or exceeds T_k.
4. Compute the minimum P-value across K distances $P_{\min} = \min(P_1, P_2, \ldots, P_K)$
5. For each distance, compute B permuted 'P-values' $P_k^{(j)}$ $1 - (\mathrm{rank}(T_k^{(j)}) - 1)/B$, where $\mathrm{rank}(T_k^{(j)})$ is the rank of the statistic $T_k^{(j)}$ among B permuted statistics $T_k^{(1)}, T_k^{(2)}, \ldots, T_k^{(B)}$.
6. Compute the permuted minimum P-values $P_{\min}^{(j)} = \min(P_1^{(j)}, P_2^{(j)}, \ldots, P_K^{(j)})$, for $j = 1, 2, \ldots, B$.
7. Report the final P-value as the proportion of permuted minimum P-values $P_{\min}^{(j)}$ exceeding minimum P-value $P_{\min}$.

which considers multiple distances simultaneously, could precisely control the Type I error rate and yields a competitive power compared with the best distance measure.

Following Chen *et al.* (2012), we simulate the data by mimicking a real throat microbiome data (Charlson *et al.*, 2010). Dirichlet-multinomial (DM) model (Mosimann, 1962) is applied to model the overdispersion of the real data and generate the simulated OTU counts. Note that the model parameters are estimated from the real throat data (Chen *et al.*, 2012; Tang *et al.*, 2016; Zhao *et al.*, 2015).

We consider two representative mediation models, where the effect is mediated through OTU abundance and presence/absence, respectively. In the first scenario (A), the independent variable X affects the relative abundance of some OTU set and the abundance data of the OTUs affect the final outcome Y. In the second scenario (B), we vary the procedure by allowing X affecting the presence/absence of OTUs, and the presence/absence of OTUs affecting Y.

Suppose that the microbiome counts of m taxa OTUs are generated by a DM distribution with the proportion parameter $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_m)$ and dispersion parameter θ. Let $\mathcal{A} \subseteq \{1, 2, \ldots, m\}$ denote the indices of the mediating OTUs.

The simulation procedure for scenario A:

1. Generate $X = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$ from a standard normal distribution.
2. Compute the proportion parameters $\boldsymbol{\pi}^{(i)} = (\pi_1^{(i)}, \pi_2^{(i)}, \ldots, \pi_m^{(i)})$, for $i = 1, 2, \ldots, n$, as

$$\pi_j^{(i)} = \begin{cases} e^{(ax_i + \varepsilon_i')}\pi_j & \text{if } j \in \mathcal{A}, \\ \pi_j & \text{if } j \notin \mathcal{A}, \end{cases}$$

where coefficient a measures the relation between X and M, and $\varepsilon_i' \sim N(0, 1)$ is the random error.

3. Renormalize $\boldsymbol{\pi}^{(i)}$ for each sample i such that the total proportion is equal to 1.
4. Generate the count matrix $\boldsymbol{M}$ through DM distributions with the proportion parameter $\boldsymbol{\pi}^{(i)}$, for $i = 1, 2, \ldots, n$, and dispersion θ.
5. Calculate the outcome Y as

$$Y = b \cdot f(\Pi_{\mathcal{A}}) + cX + \varepsilon_i''$$

where

$$\Pi_{\mathcal{A}} = \left(\sum_{j \in \mathcal{A}} \pi_j^{(1)}, \sum_{j \in \mathcal{A}} \pi_j^{(2)}, \ldots, \sum_{j \in \mathcal{A}} \pi_j^{(n)} \right)^{\mathrm{T}},$$

f is a scaling function that standardizes the OTU abundance to have mean 0 and SD 1, and $\varepsilon_i'' \sim N(0, 1)$.

The observed data are X, M and Y, and a and b jointly control the mediating effects. We vary the above procedure to simulate presence/absence data for scenario B.

The simulation procedure for scenario B:

1. Generate $X = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$ from a standard normal distribution.
2. Generate the count matrix $\boldsymbol{M}$ through DM distributions with the proportion parameter $\boldsymbol{\pi}$ and dispersion θ.
3. Update M by randomly changing presence to absence for each mediating OTU j with a probability proportional to the independent variable such that

$$M_{ij} = \begin{cases} M_{ij} & \text{if } s_i = 1, \\ 0 & \text{if } s_i = 0, \end{cases}$$

where $j \in \mathcal{A}$, $s_i \sim \text{Bernoulli}(p_i)$ and $p_i = \frac{e^{ax_i}}{1 + e^{ax_i}}$.

4. Calculate the outcome Y as

$$Y = b \cdot f(\Pi_{\mathcal{A}}') + cX + \varepsilon_i''$$

where

$$\Pi_{\mathcal{A}}' = \left(\sum_{j \in \mathcal{A}} I(M_{1j} \neq 0), \ldots, \sum_{j \in \mathcal{A}} I(M_{nj} \neq 0) \right)^{\mathrm{T}},$$

$I(\cdot)$ is the binary indicator function, and f is a scaling function.

For each scenario, we partition the m OTUs into 20 clusters via the partitioning around medoids algorithm (Maechler *et al.*, 2017) and select clusters with different abundance levels as mediating OTUs. We also consider a setting that randomly selects 40 OTUs to form a group, which ignores the phylogenetic relationship (Chen *et al.*, 2012). We use a sample size $n = 150$, and the sequencing depth is 1000 reads on average per sample. The simulated count matrix M, which is generated with parameters estimated from the real dataset, contains 90% zeros, similar to the percentage of zeros in real data (93% zeros). We vary a and b from 0 to 1 to evaluate the Type I error rate and the statistical power of the proposed method. We compare the power of the omnibus test with tests based on individual distance measures. In addition, a simpler method, which selects the minimum P-value of multiple single distance-based tests and adjusts it with Bonferroni correction, is also implemented. We conduct 1000 simulations for each parameter setting.

3.2 Simulation results

Figures 2 and 3 show two representative results of different tests for detecting the mediation effects. Phylogenetic tree-based clusters 5 and 6 are selected to mediate the effects of independent variable X on the outcome Y, respectively. The dashed lines represent the results from single distance-based tests. Specifically, BC, JAC, UniFrac, WUniFrac and GUniFrac represent the test results from Bray–Curtis distance, Jaccard distance, unweighted UniFrac distance, weighted UniFrac distance and generalized UniFrac distance, respectively. The solid lines, Omnibus and Bonferroni, denote the test results of the proposed omnibus test and the method with Bonferroni correction, respectively. We vary coefficients a and b from 0 to 1, and consider the two scenarios for each choice of a and b. All the results are reported at the significance level 0.05. Note that the power of the setting that $a = 0$ or $b = 0$ is the Type I error rate.

We can see that the proposed omnibus test could control the Type I error rate at the significance level 0.05, when the null hypothesis is true ($a = 0$ or $b = 0$). In contrast, when the null hypothesis is not true, the power increases with the increase of mediating effects. This observation is consistent for all methods and settings. The omnibus test dominates the Bonferroni-corrected test. Single distance-based tests demonstrate completely different power in Scenarios A and B. For the abundance data (Scenario A), weighted UniFrac distance and generalized UniFrac distance deliver much better results than the other distances. In the Scenario B, where only the presence/absence data affect the mediation effects, the UniFrac distance stands out. However, none of the above-mentioned distance metrics could maintain their good performance in both scenarios. For example, weighted distance and generalized distance have little power to detect the mediation effects in Scenario B, and UniFrac distance has a much worse performance in Scenario A. When compared with these single distance-based tests, the proposed omnibus test, which incorporates all distance metrics, could maintain a good performance in both scenarios. It yields considerably better results than

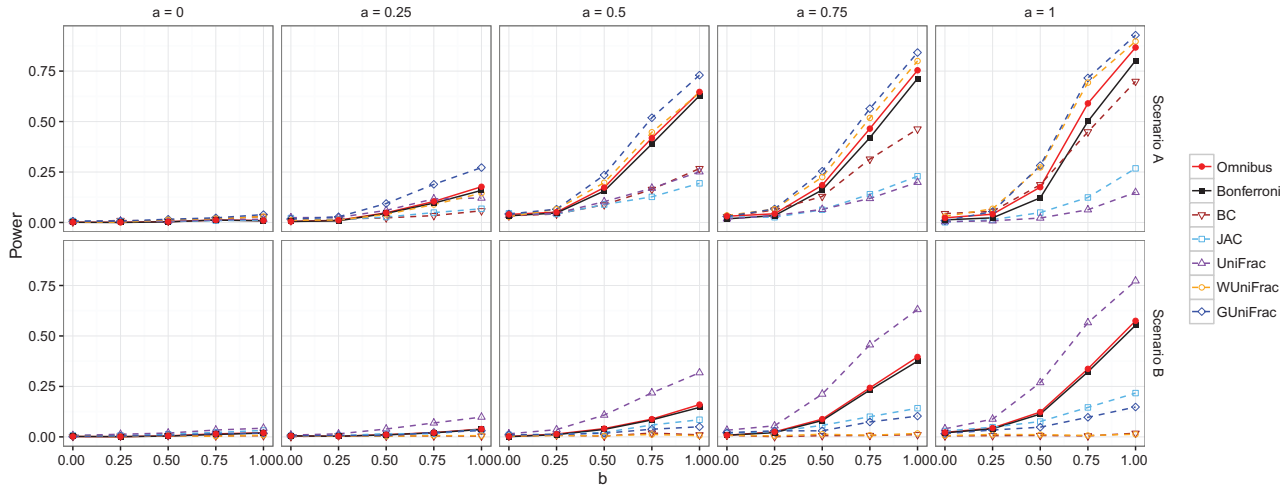

Fig. 2. Type I error and power comparison of different distance metrics for detecting the mediation effects (Cluster 5)

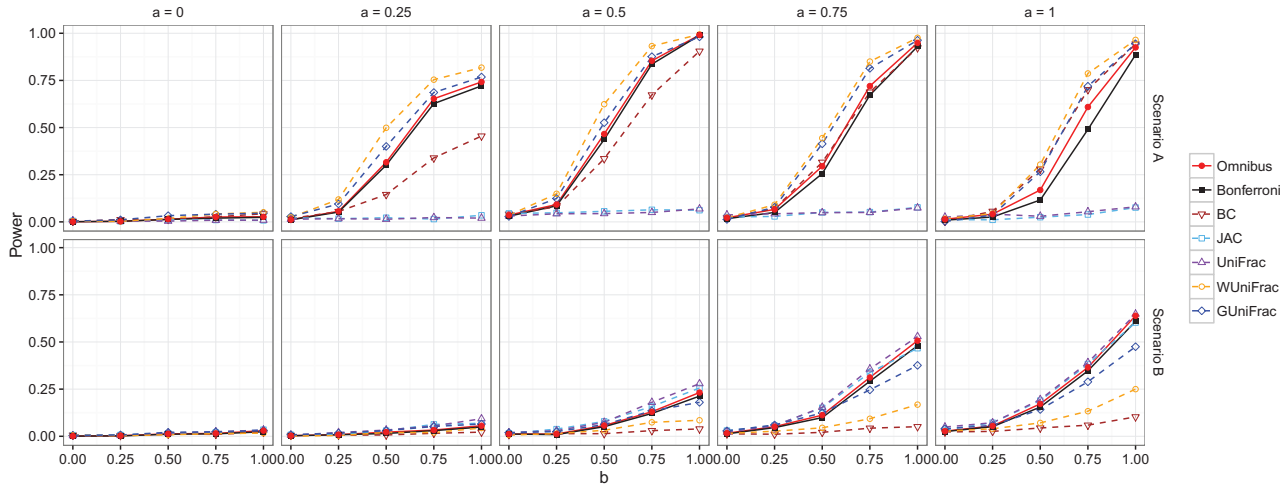

Fig. 3. Type I error and power comparison of different distance metrics for detecting the mediation effects (Cluster 6)

most single distance-based tests, and only loses moderate power when compared with the one using optimal distance, which is usually unknown in advance.

The mediating effects may also be built upon a set of OTUs (or species) which do not have any phylogenetic relationship. We further conduct experiments with randomly selected 40 OTUs as the mediating predictors, and show the detailed results in Figure 4. As expected, the Bray–Curtis distance and Jaccard distance, which do not utilize phylogenetic information, present the best performance in Scenarios A and B, respectively. It should be also noted that our omnibus test is consistently better than the Bonferroni-adjusted method and all other single distance-based methods.

Full simulation results for all clusters ($n = 150$) are provided in Supplementary Data (Supplementary Figs S1–S20). In addition, we further reduce the sample size to $n = 50$ to evaluate the performance of the proposed method under small sample sizes. As shown in the Supplementary Data (Supplementary Figs S21–S40), the results remain similar and the proposed omnibus test demonstrates a more pronounced improvement over the method based on Bonferroni correction.

3.3 Real data application

Diet strongly affects human health, partly by modulating gut microbiome composition (Wu *et al.*, 2011). Wu *et al.* (2011) studied the association of long-term dietary and environmental variables with the gut microbiota. Ninety eight healthy volunteers were enrolled in the cross-sectional study. The volunteers' long-term diet information was collected through food frequency questionnaire and converted to intake amounts of 214 nutrient categories. At the same time, their stool samples were collected, and the DNA samples were analyzed by 454/Roche pyrosequencing of 16S rDNA gene segments (Wu *et al.*, 2011). The pyrosequences were denoised (Quince *et al.*, 2009) prior to taxonomic assignment and then analyzed by the QIIME pipeline (Caporaso *et al.*, 2010) with the default parameter settings. The data also has measurements of body mass index (BMI).

Previously, we observed significant association between dietary fiber intake (as assessed by percent calories from dietary fiber) and BMI, the association between the gut microbiota and BMI, and the association between fiber intake and the gut microbiota (Wu *et al.*, 2011). We want to know whether the association between the fiber intake and BMI is mediated by the gut microbiota. This problem is

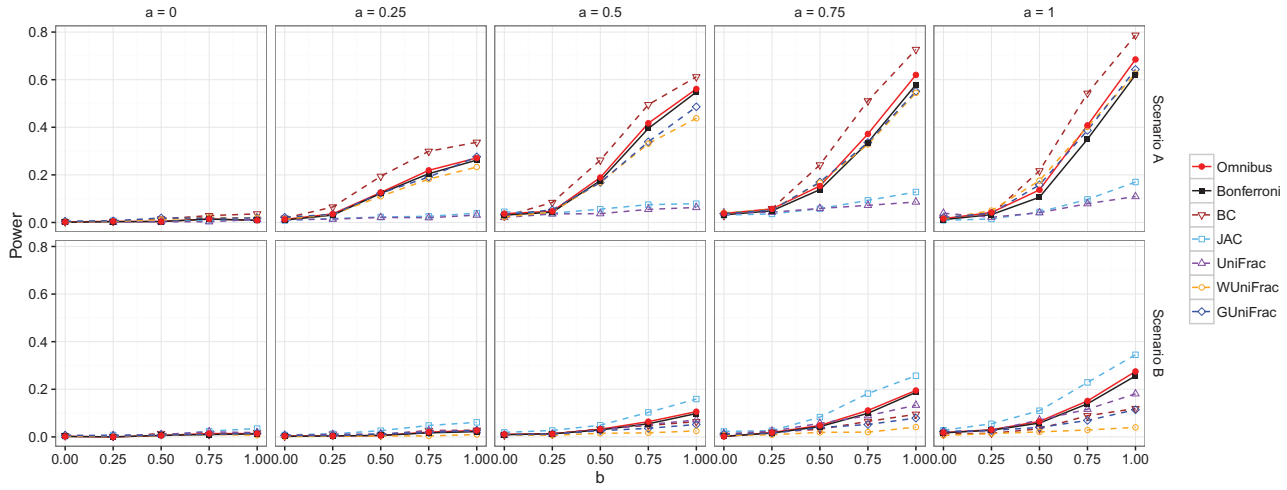

Fig. 4. Type I error and power comparison of different distance metrics for detecting the mediation effects (Random 40 OTUs)

of clinical significance. If the gut microbiota plays a mediation role, we could potentially modulate the gut microbiota to treat obesity. We use the proposed distance-based approach to test for mediation. Distance matrices are calculated based on the rarefied OTU table (rarefied to 2387 counts per sample) to reduce potential sequence depth-dependent bias. As shown in Supplementary Data (Supplementary Fig. S41), rarefaction could improve the power of mediation test with unweighted distance measures. For weighted measures, the rarefaction does not affect the power much (Weiss *et al.*, 2017). Overall, using rarefaction slightly improves the performance of the proposed omnibus test. Table 1 summarizes the *P*-values for the association tests between the fiber intake (X), gut microbiota (M) and BMI (Y). The *P*-value of 'Y–X' is calculated using *F*-test, while 'M–X' and 'Y–M' are computed via the Microbiome Regression-Based Kernel Association Test (MiRKAT) (Zhao *et al.*, 2015) with kernels built upon above-mentioned tree-based and non-tree-based distance metrics. As shown in Table 1, the fiber intake demonstrates significant (negative) associations with BMI, and the gut microbiota is significantly associated with both fiber intake and BMI. Thus we further perform the mediation test to check whether the gut microbiota mediates the effects of the fiber intake on the BMI. Table 2 summarizes the results for the single distance-based mediation tests, the Bonferroni-adjusted test and the omnibus test. For the single distance-based mediation tests, only the one based on the Jaccard distance could detect the mediation effects (P-value < 0.05). The UniFrac distance also achieves a P-value of 0.09. These two distance measures are efficient in capturing the patterns in presence/absence data. In contrast, none of the abundance-based distances, Bray–Curtis, weighted UniFrac and generalized UniFrac distances, report significant mediation effects. Mediation analysis on individual distances indicates that, besides phylogenetically related OTUs, probably there are more OTUs with distant relationships mediate the effect. As we do not know the underlying mediation model in advance, using a single distance alone could miss important mediating effects. The proposed omnibus test, which simultaneously considers multiple distance metrics, achieves a single *P*-value of 0.0309. The proposed omnibus test is more powerful than Bonferroni correction, whose *P*-value is 0.0410.

We are also interested to know which taxa play the mediator roles. We conduct mediation tests on individual taxa at different taxonomic ranks. We extracted the OTUs belonging to the taxonomic groups and performed the same omnibus test as using the whole data. We excluded taxa with few counts from testing. Table 3 summarizes the number of taxa tested at each level. The phylogeny of the tested taxa is visualized in Supplementary Data (Supplementary Fig. S42). As shown in Supplementary Figure S42 and Table 4, Lachnospiraceae Family, Ruminococcaceae Family and *Lachnospira* genus (under family Lachnospiraceae) are nominally significant with *P*-values 0.0129, 0.0468 and 0.0430, respectively. However, none of the taxa are significant if multiple testing correction such as false discovery rate control is applied, indicating the increased power by jointly analyzing the OTUs. Though not truly

Table 1. The *P*-values for association tests between fiber intake (X), gut microbiome (M) and BMI (Y)

	Y∼X	M∼X	Y∼M
P-value	0.0219	0.0385	0.0300

Table 2. The *P*-values for single distance-based and multiple distance-based mediation tests

	BC	JAC	UniFrac	WUniFrac	GUniFrac	Bonferroni	Omnibus
P-value	0.5568	0.0082	0.0901	0.7859	0.5768	0.0410	0.0309

Table 3. The number of taxa tested at each taxonomic level

	Phylum	Class	Order	Family	Genus
Tested taxa	6	13	16	31	57

Table 4. Taxonomies of potential mediating taxa (Unadjusted $P < 0.05$)

Phylum	Class	Order	Family	Genus	*P*-value
Firmicutes	Clostridia	Clostridiales	Lachnospiraceae		0.0129
Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	*Lachnospira*	0.0430
Firmicutes	Clostridia	Clostridiales	Ruminococcaceae		0.0468

statistically significant, the *Lachnospira* genus has important literature support. Clarke *et al.* (2012) observed that the gut microbiota of lean individuals contained elevated proportions of sequences corresponding to the *Lachnospira* compared with that of obese and gastric bypass individuals. In addition, the *Lachnospira* are known pectin degraders (Rode *et al.*, 1981) and play important roles in the colonic fermentation of dietary fibers (Zhang *et al.*, 2009). Thus, the *Lachnospira* genus could help explain the association between fiber intake and individual's BMI.

4 Discussion

In this article, we propose a novel distance-based omnibus test of mediation effect, and apply it to microbiome data as a special case. We show that our method is robust and powerful in detecting the structured mediators. Our method is very general and can be applied to any genomics data with different structures (e.g. LD structure for genetic data).

We simulate two scenarios that the effects of predictors X on response Y are mediated by the abundance of OTUs and the presence/absence of OTUs, respectively. Both phylogenetic tree-based and non-tree-based settings are investigated to thoroughly evaluate the performance of the proposed method. Our method is naturally capable of accommodating the confounding variables, though we do not explicitly simulate such settings. In addition, possible alternatives (e.g. testing the mediation effect on individual taxon, followed by FWER control) are not compared in the experiments, as they have no power.

The proposed method focuses on detecting an overall mediation effect by using an ensemble of distance measures. The next step is to identify specific taxa or OTUs accounting for the mediation effect. The identified taxa can provide deep insights into the underlying biological mechanisms. Our framework can be easily adapted to perform a hierarchical taxa mapping, where we start from the phylum down to the OTU level. At each taxonomic level, we extract the OTUs belonging to the same taxonomic group (e.g. Bacteroides genus), construct the distances based on the subset of OTUs and apply the proposed method, coupled with multiple testing correction. The proposed method could also extend to multivariate X, where we sum up the test statistics for individual orthogonal components of X after singular value decomposition.

One limitation of our method is the inability to quantify the relative contribution of direct and indirect effects since we do not directly test mediation effects for individual OTUs. In addition, as all statistical mediation models, the proposed model depends on many assumptions such as no unmeasured confounders and it remains subject to the same rules that association does not prove causality (MacKinnon and Fairchild, 2009; VanderWeele and Vansteelandt, 2009). Nevertheless, the proposed mediation test provides important statistical evidence, which justifies the efforts for deeper mechanistic study or experimental validation such as using randomized-control trials.

Funding

The work was supported by Mayo Clinic Gerstner Family Career Award and Mayo Clinic Center of Individualized Medicine.

Conflict of Interest: none declared.

References

Arumugam,M. *et al.* (2011) Enterotypes of the human gut microbiome. *Nature*, **473**, 174–180.

Baron,R.M. and Kenny,D.A. (1986) The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J. Person. Soc. Psychol.*, **51**, 1173.

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

Bray,J.R. and Curtis,J.T. (1957) An ordination of the upland forest communities of southern wisconsin. *Ecol. Monogr.*, **27**, 325–349.

Caporaso,J.G. *et al.* (2010) Qiime allows analysis of high-throughput community sequencing data. *Nat. Methods*, **7**, 335–336.

Charlson,E.S. *et al.* (2010) Disordered microbial communities in the upper respiratory tract of cigarette smokers. *PloS One*, **5**, e15216.

Chen,J. *et al.* (2012) Associating microbiome composition with environmental covariates using generalized unifrac distances. *Bioinformatics*, **28**, 2106–2113.

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

Cotillard,A. *et al.* (2013) Dietary intervention impact on gut microbial gene richness. *Nature*, **500**, 585–588.

David,L.A. *et al.* (2014) Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, **505**, 559–563.

Faith,J.J. *et al.* (2013) The long-term stability of the human gut microbiota. *Science*, **341**, 1237439.

Grice,E.A. *et al.* (2009) Topographical and temporal diversity of the human skin microbiome. *Science*, **324**, 1190–1192.

Judd,C.M. and Kenny,D.A. (1981) Process analysis estimating mediation in treatment evaluations. *Eval. Rev.*, **5**, 602–619.

Kuczynski,J. *et al.* (2010) Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nat. Methods*, **7**, 813–819.

Le Chatelier,E. *et al.* (2013) Richness of human gut microbiome correlates with metabolic markers. *Nature*, **500**, 541–546.

Lozupone,C. and Knight,R. (2005) Unifrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*, **71**, 8228–8235.

Lozupone,C.A. *et al.* (2007) Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Appl. Environ. Microbiol.*, **73**, 1576–1585.

Lozupone,C.A. *et al.* (2012) Diversity, stability and resilience of the human gut microbiota. *Nature*, **489**, 220–230.

MacKinnon,D.P. (2008) *Introduction to Statistical Mediation Analysis*. Routledge, Abingdon, UK.

MacKinnon,D.P. and Fairchild,A.J. (2009) Current directions in mediation analysis. *Curr. Direct. Psychol. Sci.*, **18**, 16–20.

MacKinnon,D.P. *et al.* (2007) Mediation analysis. *Annu. Rev. Psychol.*, **58**, 593.

Maechler,M. *et al.* (2017) *cluster: Cluster Analysis Basics and Extensions*. R package version 2.0.6—For new features, see the 'Changelog' file (in the package source).

Mosimann,J.E. (1962) On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika*, **49**, 65–82.

Nunes-Alves,C. (2016) Microbiome: microbiota-based nutrition plans. *Nat. Rev. Microbiol.*, **14**, 1–1.

Qin,J. *et al.* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, **464**, 59–65.

Quince,C. *et al.* (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat. Methods*, **6**, 639.

Rode,L. *et al.* (1981) Syntrophic association by cocultures of the methanol-and co2-h2-utilizing species Eubacterium limosum and pectin-fermenting Lachnospira multiparus during growth in a pectin medium. *Appl. Environ. Microbiol.*, **42**, 20–22.

Snijders,A.M. *et al.* (2016) Influence of early life exposure, host genetics and diet on the mouse gut microbiome and metabolome. *Nat. Microbiol.*, **2**, 16221.

Sonnenburg,J.L. and Bäckhed,F. (2016) Diet-microbiota interactions as moderators of human metabolism. *Nature*, **535**, 56–64.

Swenson,N.G. and Hector,A. (2011) Phylogenetic beta diversity metrics, trait evolution and inferring the functional beta diversity of communities. *PloS One*, **6**, e21264.

Tang,Z.-Z. *et al*. (2016) Permanova-s: association test for microbial community composition that accommodates confounders and multiple distances. *Bioinformatics*, **32**, 2618–2625.

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

Weiss,S. *et al*. (2017) Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, **5**, 27.

Wu,G.D. *et al*. (2011) Linking long-term dietary patterns with gut microbial enterotypes. *Science*, **334**, 105–108.

Yatsunenko,T. *et al*. (2012) Human gut microbiome viewed across age and geography. *Nature*, **486**, 222–227.

Zhang,H. *et al*. (2009) Human gut microbiota in obesity and after gastric bypass. *Proc. Natl. Acad. Sci. USA*, **106**, 2365–2370.

Zhang,H. *et al*. (2016) Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics*, btw351.

Zhang,J. and Wei,Z. (2016) An empirical bayes change-point model for identifying 3 and 5 alternative splicing by next-generation rna sequencing. *Bioinformatics*, **32**, 1823–1831.

Zhang,J. *et al.* (2017) A feature sampling strategy for analysis of high dimensional genomic data. *IEEE/ACM Transa. Comput. Biol. Bioinformatics*, http://doi.ieeecomputersociety.org/.

Zhao,N. *et al.* (2015) Testing in microbiome-profiling studies with mirkat, the microbiome regression-based kernel association test. *Am. J. Hum. Genet.*, **96**, 797–807.