Genome analysis

A general framework for association analysis of microbial communities on a taxonomic tree

Zheng-Zheng Tang^{1,*}, Guanhua Chen¹, Alexander V. Alekseyenko^{2,3} and Hongzhe Li⁴

¹Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, TN 37203, USA, ²Department of Public Health Sciences and ³Department of Oral Health Sciences, Medical University of South Carolina, Charleston, SC 29403, USA and ⁴Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on August 15, 2016; revised on October 31, 2016; editorial decision on December 9, 2016; accepted on December 14, 2016

Abstract

Motivation: Association analysis of microbiome composition with disease-related outcomes provides invaluable knowledge towards understanding the roles of microbes in the underlying disease mechanisms. Proper analysis of sparse compositional microbiome data is challenging. Existing methods rely on strong assumptions on the data structure and fail to pinpoint the associated microbial communities.

Results: We develop a general framework to: (i) perform robust association tests for the microbial community that exhibits arbitrary inter-taxa dependencies; (ii) localize lineages on the taxonomic tree that are associated with covariates (e.g. disease status); and (iii) assess the overall association of the whole microbial community with the covariates. Unlike existing methods for microbiome association analysis, our framework does not make any distributional assumptions on the microbiome data; it allows for the adjustment of confounding variables and accommodates excessive zero observations; and it incorporates taxonomic information. We perform extensive simulation studies under a wide-range of scenarios to evaluate the new methods and demonstrate substantial power gain over existing methods. The advantages of the proposed framework are further demonstrated with real datasets from two microbiome studies. The relevant R package *miLineage* is publicly available.

Availability and Implementation: *miLineage* package, manual and tutorial are available at <https://medschool.vanderbilt.edu/tang-lab/software/miLineage>.

Contact: z.tang@vanderbilt.edu

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

A promising area of genomic research focuses on the sequencing and analysis of the genomes of microorganisms (Gilbert *et al.*, 2016; Human Microbiome Project Consortium, 2012). Many human microbiome sequencing studies have been conducted to survey the bacterial composition in different body sites and investigate associations with various covariates (e.g. disease status). Research in this area has begun to uncover relationships between human microbiota

and diseases such as obesity and psoriasis (Alekseyenko *et al.*, 2013; Sanderson *et al.*, 2006). These studies suggest that microbes play an important role in human health and disease, and that understanding the microbiome composition provides insight into the functions of microbes and their effects on the human host.

Microbiome data consist of quantification of abundances of microorganisms with specific taxonomic identity. Typically, the data comes from high-throughput sequencing of a marker gene,

such as 16S ribosomal ribonucleic acid (rRNA) gene, and assignment of the sequencing reads to taxonomy through pre-processing pipelines such as QIIME (Caporaso *et al.*, 2010). A taxonomic tree is a directed acyclic graph representing the hierarchical structure through which organisms classify into taxonomic groups (taxa). The main levels in the taxonomic hierarchy listed from lowest to highest are species, genus, family, order, class, phylum and kingdom. A node in the taxonomic tree represents a taxon at a certain level. A lineage contains a cluster of taxa that form a line of descent. It is desirable to incorporate the taxonomic information into the analyses of the microbiome composition because the ecological roles played by the microbial communities are different at various taxonomic levels, and changes in the structure of the microbiota that are associated with covariates can occur at any taxonomic levels and along any relevant lineages of the tree (Gilbert *et al.*, 2016; Liu *et al.*, 2008).

Association analysis of the microbiome composition proves challenging for several reasons: the number of taxa is large relative to the sample size, the count data are sparse for many taxa with excessive zero observations, and total counts across taxa are constrained by the sequencing depth (Li, 2015). Two types of methods exist to test the association of the microbiome composition at the community level. The first type models microbiome non-parametrically and are often used to assess the overall association of the microbial community with covariates. PERMANOVA (Anderson, 2001), PERMANOVA-S (Tang *et al.*, 2016) and MiRKAT (Zhao *et al.*, 2015) summarize the taxa counts into a distance matrix and utilize that information to evaluate associations. MiSPU (Wu *et al.*, 2016) generates the generalized taxon proportions that combine taxon abundance and tree information, and adaptively weights those proportions in the association testing. These methods are not designed to pinpoint the relevant taxa that drive the association. The second type directly models the compositional taxa counts using Dirichlet-multinomial (DM) distribution (Chen and Li, 2013; La Rosa *et al.*, 2012). DM distribution was shown to fit the overdispersed count data better than the multinomial distribution by assuming that the proportion parameters follow a Dirichlet distribution. The DM test is often applied to taxa at high taxonomic levels (e.g. phylum) because it is underpowered when testing the large numbers of taxa present at low levels (e.g. genus). The DM assumption may not be appropriate for microbiome data because such models intrinsically impose a negative correlation among taxa counts, while the actual microbiome data display both positive and negative correlations (Mandal *et al.*, 2015). In addition, current implementation of the DM association test cannot accommodate continuous covariates or potential confounders (Rosa *et al.*, 2016).

In this article, we provide a general framework to test associations between covariates and microbial communities on a taxonomic tree. Our methods model multivariate taxa counts but do not rely on any distributional assumptions, which provides flexibility to various correlation structures among taxa. Our methods allow for confounder adjustment, which reduces false positive findings particularly for the current microbiome studies with more complex designs and heterogenous samples. Our methods handle excessive zero counts, which potentially increases power for the detection of changes in presence/absence frequencies of multiple taxa. Harnessing the taxonomic tree and the proposed methods, our framework powerfully and robustly identifies covariate-associated microbial lineages and evaluates the overall association of the whole microbial community. The advantages of these new methods over existing ones are demonstrated through extensive simulation studies. Application to two real microbiome datasets shows the usefulness of

our framework in the detection of associations between the human microbiome and two disorders, psoriasis and obesity.

2 Methods

2.1 Quasi-conditional association test (QCAT)

Suppose that we have n subjects measured on $m + 1$ taxa. We let Y_{ij} denote the count of taxon j for subject i , and $\mathbf{X}_i$ denote the design matrix that includes covariates of interest plus the potential confounding variables and a unit component for the intercept. We assume the mean of the count for subject i takes the form

$$E(Y_{ij}) = c_i \mu_{ij}, \quad j = 1, \dots, m, m + 1, \quad (1)$$

where c_i is the subject-level component uniformly applied to all taxa, and μ_{ij} is the taxon-specific component. In general, we are not interested in c_i because it is driven by the experimental artifact (e.g. variation of the sequencing depth among samples). For the count data, it is natural to model the taxon-specific component as $\mu_{ij} = \exp(\beta_j^T \mathbf{X}_i)$, where $\beta_j = (\beta_{j1}, \dots, \beta_{jd})^T$ is the vector of coefficients associated with the d components in $\mathbf{X}_i$ for taxon j . Here, we assume $\mathbf{X}_i$'s are the same for all taxa but allow the regression coefficients β_j to be taxon-specific. To make the parameters identifiable, we constrain $\beta_{(m+1)} = 0$, which essentially picks the last taxon as the reference. We show in [Supplementary Data Method B](#) that QCAT is invariant to the selection of the reference taxon.

We let $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{im})^T$ denote the vector of counts for the first m taxa. In order to make robust statistical inference on the parameters $\beta = (\beta_1^T, \dots, \beta_m^T)^T$, we consider the following estimating equations

$$\sum_{i=1}^n S_i(\beta) = \sum_{i=1}^n (\mathbf{Y}_i - N_i \mathbf{p}_i) \otimes \mathbf{X}_i = 0, \quad (2)$$

where $\otimes$ denotes the Kronecker product, $N_i = \sum_{j=1}^{m+1} Y_{ij}$, $\mathbf{p}_i = (p_{i1}, \dots, p_{im})^T$, and

$$p_{ij} = \frac{E(Y_{ij})}{\sum_{k=1}^{m+1} E(Y_{ik})} = \frac{\exp(\beta_j^T \mathbf{X}_i)}{\sum_{k=1}^m \exp(\beta_k^T \mathbf{X}_i) + 1}, \quad j = 1, \dots, m.$$

These estimating equations are essentially the score functions of the multinomial logistic regression with size N_i and proportion parameters $\mathbf{p}_i$ for each subject i . Provided that the mean assumption (1) holds, consistent estimation of β can be obtained by solving these estimating equations, even if the actual distribution of $\mathbf{Y}_i$ is not multinomial (Wooldridge, 1999). This quasi-conditional approach takes into account the compositional nature of the microbiome data by conditioning on the total counts N_i 's, so that the unknown subject-level components c_i 's are removed from the estimating equations (2).

Let β_1 denote the parameters of interest among β . We use the generalized score statistics (Boos, 1992) to test the null hypothesis $H_0 : \beta_1 = 0$. The derivation of the generalized score test is shown in the [Supplementary Data Method A](#). As an example, we show here a simple case of testing differential abundances in $(m + 1)$ taxa between two groups. Suppose we have n_1 and n_2 subjects in group 1 and 2, respectively. Without loss of generality, we order the subjects such that the first n_1 subjects belong to group 1 and the remaining subjects belong to group 2. We construct the design matrix as $\mathbf{X}_i = [X_{i1} \ 1]$, where X_{i1} takes a value of 0 or 1 to indicate membership in group 1 or 2. The null hypothesis can be written as $H_0 : \beta_1 = (\beta_{11}, \dots, \beta_{m1})^T = 0$, where β_{j1} is the coefficient

associated with X_{j1} for taxon j . In this setting, the generalized score test statistic can be written as

$$Q_1 = \left(\sum_{i=1}^{n_1} \tilde{S}_i \right)^T \left\{ \left(\frac{N_{(1)}}{N_{(1)} + N_{(2)}} \right)^2 \sum_{i=1}^n \tilde{S}_i \tilde{S}_i^T - \frac{N_{(1)} - N_{(2)}}{N_{(1)} + N_{(2)}} \sum_{i=1}^{n_1} \tilde{S}_i \tilde{S}_i^T \right\}^{-1} \left(\sum_{i=1}^{n_1} \tilde{S}_i \right), \quad (3)$$

where $N_{(1)} = \sum_{i=1}^{n_1} N_i$, $N_{(2)} = \sum_{i=n_1+1}^n N_i$, and $\tilde{S}_i = \mathbf{Y}_i - N_i \sum_{i=1}^n \mathbf{Y}_i / (N_{(1)} + N_{(2)})$. The reference distribution for the test statistic is chi-square with m degrees of freedom. The test is referred to as the one-part test to parallel with the two-part test described in section 2.2.

The term in the middle of the test statistic (3) is the empirical covariance estimator of the score statistics. This estimator is robust to arbitrary inter-taxa relationships. The vector of the score statistics among taxa might generally be expected to exhibit negative correlations because the total taxa count is bounded by the sequencing depth. This negative correlation, which is implied by the information matrix of the multinomial regression, does not necessarily hold under the mean assumption (1). Therefore, the empirical covariance estimator that reflects the true dependence among taxa in the community is advantageous. In addition, as in any multivariate tests, the covariance matrix is not always invertible when the sample size is smaller than the number of taxa. This problem can be greatly alleviated by performing tests on each lineage of a taxonomic tree such that the number of taxa involved in each test will be significantly reduced (details in section 2.3).

2.2 Two-part distribution-free association test

With microbiome data, we often observe excessive zero counts. Two-part models are commonly used to handle such data by assuming that the data have a probability mass at zero and a response of positive values. We use δ_{ij} to indicate if Y_{ij} is positive or zero by the values 1 versus 0, respectively. In the two-part model, the probability distribution function of Y_{ij} can be expressed as

$$f(Y_{ij}) = \underbrace{\{Pr(Y_{ij} = 0)\}^{1-\delta_{ij}} \{1 - Pr(Y_{ij} = 0)\}^{\delta_{ij}}}_{\text{zero part}} \underbrace{b(Y_{ij}|\delta_{ij})}_{\text{positive part}},$$

where $b(\cdot)$ is the probability distribution function of Y_{ij} given δ_{ij} . The zero part models presence/absence of a taxon in the samples and the positive part models non-zero microbial abundance (counts at the presence). We will perform hypothesis tests on the zero and the positive parts, separately, and then combine the tests.

For the positive part, we apply the QCAT approach to the conditional mean of the count

$$E(Y_{ij}|\delta_{ij}) = c_i \delta_{ij} \mu_{ij}, \quad j = 1, \dots, m, m+1,$$

where $\mu_{ij} = \exp(\gamma_j^T \mathbf{X}_i)$. We adopt the estimating Equation (2) and replace p_{ij} by the formula

$$p_{ij} = \frac{E(Y_{ij}|\delta_{ij})}{\sum_{k=1}^{(m+1)} E(Y_{ik}|\delta_{ik})} = \frac{\delta_{ij} \exp(\gamma_j^T \mathbf{X}_i)}{\sum_{k=1}^m \delta_{ik} \exp(\gamma_k^T \mathbf{X}_i) + \delta_{i(m+1)}}.$$

The rest of the inference is the same as that in the previous section. We refer to the resulting test as the positive-part test and its statistic as Q_{POS} .

For the zero part, we employ the generalized estimating equation (GEE) method (Zeger and Liang, 1986) for the vector of the binary indicators of absence $\mathbf{Y}_i^* = (1 - \delta_{i1}, \dots, 1 - \delta_{im}, 1 - \delta_{i(m+1)})^T$. The covariates for the zero part are not necessarily the same as those for the positive part. But for the ease of presentation, we keep them the same in this paper. The detailed derivation of the GEE-based

generalized score test is shown in [Supplementary Data Method C](#). We refer to the test as the zero-part test and its statistic as Q_{ZERO} .

The Q_{POS} and Q_{ZERO} are independent under the two-part model. Thus, we can combine them by direct summation $Q_2 = Q_{\text{POS}} + Q_{\text{ZERO}}$ and obtain the P -value based on the chi-square distribution with the total degrees of freedom from both parts. We refer to the resulting combined test as the two-part test.

The asymptotic approximation of the test statistics Q_1 and Q_{POS} may not be accurate with very rare taxa when most of the observations are zeros. On the other hand, the asymptotic approximation of Q_{ZERO} may not be accurate with very common taxa when almost all observations are positive. In these scenarios, we resort to resampling techniques to obtain P -values. Our methods adopt score-type statistics, which are computationally efficient because the null model does not involve covariates of interest and needs to be fit only once in the resampling procedure. To further improve the computational speed, we employ a multi-stage procedure that uses a small number of resampling steps in the initial stage and sequentially increases the number if the estimated P -value is small.

2.3 Localization of covariate-associated lineages on the taxonomic tree

Suppose we have a rooted taxonomic tree with nodes $v_1, \dots, v_M$ representing the taxonomic units of the tree. Without loss of generality, we assume that the first K nodes $v_1, \dots, v_K$ are internal nodes (nodes with at least one child). We let $Y(v_k)$ denote the count of reads assigned to the node v_k and $\mathbf{Y}(S) = (Y(v_{k_1}), \dots, Y(v_{k_j}))^T$ denote a vector of counts of reads assigned to the set of nodes $S = \{v_{k_1}, \dots, v_{k_j}\}$. For each split from a parent node to a child node, the reads on the parent node are either assigned to the child node or remain unassigned. We let $\tau(v_k)$ denote the set of all direct child nodes of the internal node v_k . Then, $\mathbf{Y}(\tau(v_k))$ is the vector of counts of reads assigned to its direct child nodes, and $Y(v_k) - \sum_{v_{k_i} \in \tau(v_k)} Y(v_{k_i})$ is the count of unassigned reads. For each lineage represented by its root node v_{k_s} , the observed counts for subject i consist of $\mathbf{Y}_i^{(k)} = (\mathbf{Y}_i^T(\tau(v_k)), Y_i(v_k) - \sum_{v_{k_i} \in \tau(v_k)} Y_i(v_{k_i}))^T$. See [Supplementary Figure S1](#) for an example on a simple mock tree.

In order to identify the covariate-associated lineages, we apply the proposed one-part and two-part tests to $\mathbf{Y}_i^{(k)}$ and consider testing the set of hypotheses

$$\text{one-part test } H_0^{(k)} : \beta_{\cdot 1}^{(k)} = 0, \quad (k = 1, \dots, K),$$

$$\text{two-part test } H_0^{(k)} : \gamma_{\cdot 1}^{(k)} = 0, \alpha_{\cdot 1}^{(k)} = 0 \quad (k = 1, \dots, K),$$

where $\beta_{\cdot 1}^{(k)}$ are the parameters of interest under the one-part model on lineage v_k , and $\gamma_{\cdot 1}^{(k)}$ and $\alpha_{\cdot 1}^{(k)}$ are the parameters of interest under the two-part model corresponding to the positive and zero parts, respectively, on lineage v_k . Similarly, the DM test can be applied to $\mathbf{Y}_i^{(k)}$ defined for each lineage. We use the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001) to control for the false discovery rate (FDR) of multiple comparisons.

2.4 Assessment of the overall association of the whole microbial community

It is also of interest to assess the association of the whole microbial community with the covariates of interest. In order to do that, we consider testing the composite hypothesis

$$H_0 = \cap_{k=1}^K H_0^{(k)} \quad \text{versus} \quad H_a = \cup_{k=1}^K (H_0^{(k)})^c.$$

We first compute P -values for testing the individual lineages $H_0^{(k)}$. The one-part test is conditional on the total read counts of the lineage (the count on its root node), so the Q_1 P -values over the lineages on the tree are mutually independent under H_0 . We can obtain the P -value of the global test using standard P -value combination methods. Simes' method (Simes, 1986) is powerful when the association signals concentrate in a few lineages. In contrast, Fisher's method (Fisher et al., 1950) is more powerful when the signals spread over multiple lineages. In our simulation studies and real data analyses, we combine the Q_1 P -values on the tree using Fisher's method and Simes' method, and we refer to the resulting global tests as Fisher- Q_1 and Simes- Q_1 , respectively. Similar to the one-part test, we can combine the DM P -values across lineages to obtain the corresponding global tests, which we refer to as Fisher-DM and Simes-DM. The Q_2 P -values over the lineages are not independent because the zero-part test is not a conditional test. Fisher's method does not apply for dependent P -values, but Simes' method is largely valid (Sarkar and Chang, 1997). We refer to the global two-part test as Simes- Q_2 .

3 Results

We conducted extensive numerical studies (simulation studies and real data analyses) to investigate the performance of the proposed and existing methods. The DM test was performed using the R package HMP (v1.4.3) (Rosa et al., 2016). The global tests PERMANOVA, MiRKAT, MiSPU tests were performed using the R packages vegan (v2.2-1) (Oksanen et al., 2015), MiRAKT (v0.02) and MiSPU (v1.0), respectively. The PERMANOVA and MiRKAT used L^1 Kantorovich-Rubinstein (K-R) distance with unit branch length (Evans and Matsen, 2012). For each pair of samples, this distance is calculated as the sum of the absolute difference between two compositional vectors over each branch of the tree. For MiSPU, we used a range of values $(2, 3, \dots, 8, \infty)$ for the parameter γ and reported the one that yields the highest power. All the global tests were performed based on the same taxonomic tree.

3.1 Simulation studies

We simulated taxa counts for two groups with an equal number of samples in each group and tested differential abundances in multiple taxa between the two groups. We considered total sample sizes of 50 and 100 in all simulation studies.

Our first series of simulation studies was designed to evaluate the performance of the asymptotic tests on a certain set of taxa. We simulated taxa counts from parametric distributions to control the relative abundance and the presence/absence frequency of each taxon. This allowed us to clearly see how the different methods perform under various scenarios. In particular, we first generated the vector of proportions for 7 taxa from Dirichlet, log-normal, or zero-inflated log-normal (ZILN) distributions, and then simulated counts from multinomial distribution based on those proportions.

Supplementary Table S1 summarizes the simulation strategies under different parametric models. In the view of the fact that the sum of the proportions over taxa is bounded by 1, we induced an abundance difference between two groups by applying a swapping procedure to change abundances of two taxa at a time (increase the abundance in one taxon and decrease in the other) without disturbing other taxa. This procedure is illustrated by the diagram in Supplementary Table S1 and is explained below for each parametric models.

Under the Dirichlet model, the proportions of taxa were directly generated from Dirichlet distribution with mean parameters (0.58,

0.12, 0.10, 0.08, 0.06, 0.04, 0.02) and dispersion parameter of 3. Then, we swapped the mean parameter of the last taxon (i.e. 0.02) with that of another taxon in one group. As the values of the mean parameters are in order, the magnitude of the swapped parameter dictates the degree of perturbation. Under the log-normal model, we set the means of the 7 normal distributions to $\mu = (\mu_1, \dots, \mu_7) = (3, 1.4, 1.2, 1, 0.8, 0.5, 0)$ and variances to 1. The proportions of the 7 taxa are calculated from the relative abundance among the 7 log-normal samples. Similar to the DM model, we swapped the mean parameter of the last taxon with that of another taxon in one group to induce an abundance difference. Under the ZILN model, we set the mean parameters of the normal distributions the same as those in the log-normal model and considered two sets of zero proportion parameters: (a) $\pi = (0.6, 0.55, 0.5, 0.45, 0.4, 0.35, 0.2)$; and (b) $\pi = (0, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4)$. We referred to these two settings as ZILN(a) and ZILN(b), respectively. In addition, we considered three patterns of differentiation under the ZILN model. In the 'Zero' pattern, π was swapped, while μ was fixed. In the 'Positive' pattern, μ was swapped, while π was fixed. In the 'Zero + Positive' pattern, both π and μ were swapped simultaneously.

For the type I error evaluation, we considered a situation with the presence of a confounder. Specifically, we generated another group consisting of half of the members of group 1. For those members, we perturbed the abundances of the taxa at level 5. In particular, we swapped the Dirichlet parameter 0.02 with 0.12 in the Dirichlet model; we swapped the normal mean parameter 0 with 1.4 in the log-normal and ZILN models. Clearly, the additional group is correlated with the group of interest and the abundance of the taxa.

For all scenarios in the first set of simulation studies, the number of total counts was simulated from a Poisson distribution with mean 1000. We used 5000 simulated datasets to evaluate type I error and power of the asymptotic test at $\alpha = 0.05$.

The empirical type I errors are shown in Table 1. The results for Q_2 , Q_{ZERO} and Q_{POS} tests are not shown under the log-normal model because the two-part test degenerates to the one-part test

Table 1. Type I error of asymptotic tests

n	Test	Dirichlet	Log-Normal	ZILN(a)	ZILN(b)
Without confounders					
50	Q_1	0.032	0.038	0.040	0.038
	Q_2	0.044	–	0.044	0.038
	Q_{ZERO}	0.045	–	0.044	0.039
	Q_{POS}	0.038	–	0.044	0.036
	DM	0.030	0.021	0.037	0.021
100	Q_1	0.040	0.043	0.041	0.048
	Q_2	0.044	–	0.044	0.046
	Q_{ZERO}	0.046	–	0.044	0.048
	Q_{POS}	0.038	–	0.043	0.040
	DM	0.042	0.017	0.035	0.021
With a confounder					
50	Q_1	0.020	0.014	0.018	0.010
	Q_2	0.021	–	0.015	0.024
	Q_{ZERO}	0.032	–	0.020	0.028
	Q_{POS}	0.018	–	0.015	0.014
	DM	0.17	0.17	0.077	0.12
100	Q_1	0.039	0.035	0.053	0.028
	Q_2	0.039	–	0.039	0.032
	Q_{ZERO}	0.044	–	0.047	0.041
	Q_{POS}	0.041	–	0.036	0.030
	DM	0.38	0.59	0.17	0.35

Fig. 1. Power as a function of the degree of perturbation under Dirichlet and log-normal models

when all the counts are positive. When no confounders exist, all tests preserve the type I error. When a confounder exists, our proposed tests that adjust for the confounder control the type I error. The DM test yields inflated type I error due to the inability to accommodate confounders.

The power results under the Dirichlet and the log-normal models when $n = 50$ are displayed in Figure 1 (see Supplementary Fig. S2 for results when $n = 100$). Each plot demonstrates the power curve as a function of the degree of perturbation. Under the Dirichlet model, the one-part test is slightly less powerful than the DM test. The two-part test is dramatically more powerful than the one-part and DM tests, as the differentiation is dominated by the change of presence/absence frequencies. Under the log-normal model, the one-part test is much more powerful than the DM test.

The power results under the ZILN model when $n = 50$ are displayed in Figure 2 (see Supplementary Fig. S3 for results when $n = 100$). In the ‘Zero’ pattern, the two-part test is substantially more powerful than the one-part test and less powerful than the zero-part test. In the ‘Positive’ pattern, the two-part test is less powerful than the positive-part test. The two-part test is more powerful than the one-part test in ZILN(a) but less powerful in ZILN(b) because the zero proportions are generally higher in the former setting. In the ‘Zero + Positive’ pattern, the two-part test is more powerful than both the zero-part and positive-part tests. With the ZILN(a), the perturbation increases/decreases the zero proportion and the mean of the positive count simultaneously, in which case the overall means do not change much and the tests of the overall means (i.e. the one-part test and the DM test) have low power. With ZILN(b), the perturbation makes the zero proportion and the mean of the positive count change in different directions, in which case both the one-part and the two-part tests have good power. In all scenarios under the ZILN, the DM test is always less powerful than the proposed tests.

Our second series of studies was devoted to comparing the performance of the new methods with existing methods for lineage identification and overall association assessment. In particular, we applied the one-part, two-part and DM tests to each lineage of the tree and combined the lineage-specific P -values to produce the global P -values; we also applied PERMANOVA, MiRKAT and MiSPU to assess the overall association. We used permutation to obtain P -values for all tests.

To closely resemble reality, we generated taxa counts on a taxonomic tree based on a real dataset (Flores et al., 2014). Those investigators collected 638 gut microbial samples from 85 college-age adults over a three-month period and then characterized the microbiome composition using 16S rRNA sequencing. Counts of reads were summarized on a taxonomic tree that had 1050 nodes from kingdom to species (see Supplementary Fig. S4b). Because there is

Fig. 2. Power as a function of the degree of perturbation under zero-inflated log-normal models. The left panels pertain to ZILN (a) and the right panels pertain to ZILN (b). The pattern of perturbation is indicated above each graph

no interference over the three-month study period, we considered these samples as coming from the same null population.

We constructed a matrix $P_0 = \{P_{ij}\}_{i=1,j=1}^{638, 1050}$ with each row corresponding to the relative abundances (proportions) of the 1050 nodes for each sample. We then simulated count data from multinomial distribution with the vector of proportion parameters sampled from the rows of P_0 and size parameter N_i sampled from the sequencing depths of the 638 microbial samples. To introduce perturbation, we added count E_i to the taxa under genus *Streptococcus* in one group, where E_i was simulated from $\text{Binomial}(N_i, p_e)$, and p_e varied from 0 to 0.002%. For each scenario, we repeated the simulation 100 times. For lineage detection on the tree, we controlled the FDR at $\alpha = 0.05$. For the global test, we rejected the null composite hypothesis at $\alpha = 0.05$.

Supplementary Figure S4a displays the heatmap of the counts for the 40 taxa under order *Lactobacillales* when $p_e = 0.002\%$. Clearly, the counts of the taxa under genus *Streptococcus* are different between the two groups. As a perturbation is made to the counts under genus *Streptococcus*, the counts on all the parent nodes of *Streptococcus* also change. Therefore, the differential lineages denoted by their root nodes are kingdom *Bacteria*, phylum *Firmicutes*, class *Bacilli*, order *Lactobacillales*, family *Streptococcaceae* and genus *Streptococcus* (see Supplementary Fig. S4b). The discovery rate at genus *Streptococcus* is displayed in the panels of Figure 3a. The two-part test becomes more powerful than the one-part test with higher degrees of perturbation because of the remarkable difference in the presence/absence frequencies (see Supplementary Fig. S4a). Discovery rates for lineages above genus are very low (results not shown) because differences in counts on these lineages are diluted by the counts from the non-differential

Fig. 3. Simulation results for association tests on the taxonomic tree. (a) Discovery rate at genus *Streptococcus*. (b) Power of the global test. The total sample size is indicated above each graph

child nodes. The FDR is under controlled for all tests (see [Supplementary Table S2](#)).

The power of the global test is displayed in the panels in [Figure 3b](#). The Simes-Q₁ and Simes-Q₂ tests are always more powerful than the Fisher-Q₁ test because the perturbation occurs on a small number of lineages. The PERMANOVA, MiRKAT and MiSPU tests have low power because the differentiation is too subtle to be detected by jointly considering all the taxa on the tree. All global tests preserve the type I error. The DM test is underpowered in the detection of differential lineages, as well as the global association assessment.

3.2 Cutaneous microbiome in psoriasis

Psoriasis is a common chronic inflammatory disease of the skin. Alekseyenko *et al.* investigated the community differentiation of cutaneous microbiota in psoriasis, where a total of 51 patients with psoriasis were studied by swabbing the lesion and a contralateral sample of normal tissue from each subject (Alekseyenko *et al.*, 2013). The sample DNA was analyzed by sequencing the V1-V3 region of the 16S rRNA gene. The microbial operational taxonomic units (OTUs) were constructed through the QIIME pipeline. After removing samples with sequencing depth less than 1000 and discarding taxa with only one read, we ended up with 100 samples for the two groups (lesion and normal tissue) and 13429 OTUs. These OTUs were taxonomically classified up to the genus level. Taxa that appear in less than two samples have been dropped resulting in 477 genera. These genera were then mapped to a taxonomic tree with 304 lineages from family to kingdom.

We compared the microbial communities among the two groups over all lineages on the taxonomic tree at the family, order, class, phylum and kingdom levels. We performed the association analysis using the proposed tests and the DM test. The DM test does not discover any differential lineages after controlling FDR at 0.05 level. Our proposed tests identify seven lineages on the tree, including *Streptococcaceae*, *Thermaceae*, *Actinomycetaceae*, *Paenibacillaceae* at family level, *Bacillales* at order level, *Deinococci* at class level and

Fig. 4. Taxonomic tree for the cutaneous microbiome. The root nodes of the discovered lineages are annotated by the names of the corresponding taxa. Four phyla that dominate the skin microbial community are also annotated

Acidobacteria at phylum level. Their permutation *P*-values are listed in [Supplementary Table S3](#), and their locations on the tree are marked in [Figure 4](#). The differentiation of the identified lineage is driven mainly by the differences in the presence/absence frequencies of taxa in those lineages. For instance, [Supplementary Figure S5](#) displays the proportion of the presence and the count at those presence in the families of order *Bacillales*. The lesion and normal groups differ in the percentages of the presence (Q_{ZERO} *P*-value = 0.000075) but do not differ in the counts at those presence (Q_{POS} *P*-value = 0.34). The family *Paenibacillaceae* that shows a prominent difference between groups is also detected by the two-part test at the family level.

The four lineages identified at the family level were completely missed by the investigators' initial analysis (Alekseyenko *et al.*, 2013). The association of family *Streptococcaceae* with psoriasis was recently confirmed by another study (Drago *et al.*, 2016), and the presence of species *Streptococcus* was previously linked to the pathogenesis of psoriasis (Weisenseel *et al.*, 2002). This suggests the advantage of our methods, especially in the detection of associations at lower taxonomic levels.

For the global association analysis, our Fisher-Q₁ test produced significant results with *P*-values of 0.0016. The Simes' tests are less significant than the Fisher's test (Simes-Q₁ *P*-value = 0.073, Simes-Q₂ *P*-value = 0.016) because the differential signals spread over many lineages in the taxonomic tree. We also applied global tests PERMANOVA, MiRKAT and MiSPU to the same data. These tests produced significant results with *P*-values of 0.00081, 0.00068 and 0.0088, respectively. The *P*-values of the DM global tests are not significant (Fisher-DM *P*-value = 0.06, Simes-DM *P*-value = 0.22).

3.3 Gut microbiome and body mass index

Gut microbiota play an important role in obesity by contributing to nutrient digestion and absorption in humans. Wu *et al.* examined the relationship between micronutrients and gut microbiome composition, by collecting fecal samples from 98 healthy volunteers, along with their demographic data and diet information, and then sequencing the V1-V2 region of the 16S rRNA genes in the fecal

samples (Wu et al., 2011). A total of 3068 OTUs were obtained from the QIIME pipeline. These OTUs were taxonomically classified up to the genus level. Taxa that appear in less than two samples have been dropped resulting in 80 genera. These genera were then mapped to a taxonomic tree with 74 lineages from family to kingdom.

Because dysbiosis of the gut microbiome has been shown to be associated with obesity (Sanderson et al., 2006), we were interested in identifying the bacterial lineages that are associated with body mass index (BMI) after adjusting for total fat and caloric intakes. We performed the proposed tests over all lineages on the taxonomic tree at the level of family, order, class, phylum and kingdom. The DM test could not be performed because it is unable to handle continuous covariates or adjust for confounders. Our tests identify two BMI-associated lineages: family *Ruminococcaceae* (Q_1 P -value = 0.00011, Q_2 P -value = 0.025, Q_{ZERO} P -value = 0.38, Q_{POS} P -value = 0.0070) and family *Veillonellaceae* (Q_1 P -value = 0.0036, Q_2 P -value = 0.0018, Q_{ZERO} P -value = 0.0011, Q_{POS} P -value = 0.20). As indicated by the results of the zero-part and the positive-part tests, the association of *Ruminococcaceae* is driven mainly by the changes in the positive abundances and the association of *Veillonellaceae* is driven mainly by the changes of the presence/absence frequencies. For family *Ruminococcaceae*, the signal concentrates in its direct child genus *Oscillibacter* (Q_1 P -value becomes 0.13 after leaving *Oscillibacter* out). Of relevance to our findings, *Oscillibacter* was previously identified to be negatively associated with BMI and potentially regulates the maintenance of gut barrier integrity (Lam et al., 2012; Wu et al., 2011). In addition, family *Veillonellaceae* was previously identified to be positively associated with adiposity and the deterioration of metabolic factors (Krajmalnik-Brown et al., 2012; Wu et al., 2011).

For the global association analysis, all of our global tests produce significant results with Fisher- Q_1 P -value of 0.010, Simes- Q_1 P -value of 0.0028, and Simes- Q_2 P -value of 0.042. We also applied global tests PERMANOVA, MiRKAT and MiSPU to the same data. The results are not significant because the association signals are subtle and sparse on the tree (PERMANOVA P -value = 0.39, MiRKAT P -value = 0.57, MiSPU P -value = 0.24).

4 Discussion

In this paper, we develop a general framework for association analysis between microbial communities on a taxonomic tree and covariates (e.g. clinical outcomes, disease status). This framework exploits the hierarchical structure of the microbial taxonomy and enables us to identify the covariate-associated lineages, as well as to assess the overall association of the microbial community with covariates. Our methods accommodate the compositional and sparse nature of microbiome data. The proposed tests are distribution-free to allow for complex dependency among taxa. Moreover, the ability to adjust for confounding variables facilitates robust discoveries and reduces false positive findings. We demonstrate that our proposed methods outperform existing methods under a wide range of scenarios. We apply our framework to real microbiome datasets to discover associations of the microbial communities with psoriasis and obesity and to localize the relevant microbial lineages on the taxonomic tree.

The performance of different proposed tests depends on the underlying distributions of the microbiome data and the association patterns. If the taxa counts have many zeros, then the two-part test tends to be more desirable than the one-part test. If most of the taxa

counts are positive, then the one-part test tends to be more attractive than the two-part test. In terms of the global tests, if the association signals are very sparse on the tree, Simes' method would be more desirable than Fisher's method to combine P -values across lineages. If the association signals spread over many lineages on the tree, Fisher's method would be more desirable.

Many distance-based methods have been developed to test the overall association of the whole microbial community with covariates, but they are unable to pinpoint taxa that drive the associations. On the other hand, association testing with a single taxon (Love et al., 2014; Paulson et al., 2013; Robinson et al., 2010) can be underpowered because the signal in one taxon is often too weak to be powerfully detected, and there are a large number of tests to be adjusted for. Our framework leverages taxonomic structure to define the unit of the tests in a biologically meaningful way. Our multivariate tests concerning multiple taxa in a lineage can enrich association signals and reduce the penalty of multiple testing. For example, the 80 genera were reduced to 34 families in the analysis of the gut microbiome data. The tree-guided discoveries from our framework are highly interpretable and are expected to facilitate the precise delineation of molecular mechanisms and functions of the relevant bacteria that interact with the human host.

Besides the 16S rRNA sequencing, the shotgun metagenomic sequencing is a popular high-throughput techniques commonly used in microbiome studies. The shotgun sequencing platform surveys all microbial genomes instead of one marker gene. Because of the ambiguity of sequence alignment among multiple bacterial genomes, the metagenomic data are often summarized into proportions rather than counts. Even though we presented our methods using count data, the framework can be modified easily to handle proportion data from metagenomic sequencing.

Funding

This work was supported by the National Institutes of Health, Building Interdisciplinary Research Careers in Women's Health (BIRCWH) award 5K12HD043483-14.

Conflict of Interest: none declared.

References

- Alekseyenko, A.V. et al. (2013) Community differentiation of the cutaneous microbiota in psoriasis. *Microbiome*, **1**, 31.
- Anderson, M.J. (2001) A new method for non-parametric multivariate analysis of variance. *Austral. Ecol.*, **26**, 32–46.
- Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **57**, 289–300.
- Benjamini, Y. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, **29**, 1165–1188.
- Boos, D.D. (1992) On generalized score tests. *Am. Stat.*, **46**, 327–333.
- Caporaso, J.G. et al. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, **7**, 335–336.
- Chen, J. and Li, H. (2013) Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *Ann. Appl. Stat.*, **7**, 418–442.
- Drago, L. et al. (2016) Skin microbiota of first cousins affected by psoriasis and atopic dermatitis. *Clin. Mol. Allergy*, **14**, 1.
- Evans, S.N. and Matsen, F.A. (2012) The phylogenetic Kantorovich–Rubinstein metric for environmental sequence samples. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **74**, 569–592.

- Fisher, R.A. *et al.* (1950). *Statistical methods for research workers. Biological monographs and manuals*. No. V. 11th ed. Oliver and Boyd, Edinburgh and London.
- Flores, G.E. *et al.* (2014) Temporal variability is a personalized feature of the human microbiome. *Genome Biol.*, **15**, 531.
- Gilbert, J.A. *et al.* (2016) Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature*, **535**, 94–103.
- Human Microbiome Project Consortium. (2012) A framework for human microbiome research. *Nature*, **486**, 215–221.
- Krajmalnik-Brown, R. *et al.* (2012) Effects of gut microbes on nutrient absorption and energy regulation. *Nutr. Clin. Pract.*, **27**, 201–214.
- La Rosa, P.S. *et al.* (2012) Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PLoS ONE*, **7**, e52078.
- Lam, Y.Y. *et al.* (2012) Increased gut permeability and microbiota change associate with mesenteric fat inflammation and metabolic dysfunction in diet-induced obese mice. *PLoS ONE*, **7**, e34233.
- Li, H. (2015) Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annu. Rev. Stat. Appl.*, **2**, 73–94.
- Liu, Z. *et al.* (2008) Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res.*, **36**, e120–e120.
- Love, M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 1.
- Mandal, S. *et al.* (2015) Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb. Ecol. Health D*, **26**.
- Oksanen, J. *et al.* (2015) *vegan: community ecology package*. R package version 2.2-1.
- Paulson, J.N. *et al.* (2013) Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods*, **10**, 1200–1202.
- Robinson, M.D. *et al.* (2010) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
- Rosa, P.S.L. *et al.* (2016). *HMP: hypothesis testing and power calculations for comparing metagenomic samples from HMP*. R package version 1.4.3.
- Sanderson, S. *et al.* (2006) Human gut microbes associated with obesity. *Nature*, **444**, 1022–1023.
- Sarkar, S.K. and Chang, C.K. (1997) The Simes method for multiple hypothesis testing with positively dependent test statistics. *JASA*, **92**, 1601–1608.
- Simes, R.J. (1986) An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, **73**, 751–754.
- Tang, Z.Z. *et al.* (2016) PERMANOVA-S: association test for microbial community composition that accommodates confounders and multiple distances. *Bioinformatics*, **31**, 311.
- Weissensteil, P. *et al.* (2002) Streptococcal infection distinguishes different types of psoriasis. *J. Med. Genet.*, **39**, 767–768.
- Wooldridge, J.M. (1999) Distribution-free estimation of some nonlinear panel data models. *J. Econom.*, **90**, 77–97.
- Wu, C. *et al.* (2016) An adaptive association test for microbiome data. *Genome Med.*, **8**, 56.
- Wu, G.D. *et al.* (2011) Linking long-term dietary patterns with gut microbial enterotypes. *Science*, **334**, 105–108.
- Zeger, S.L. and Liang, K.Y. (1986) Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42**, 121–130.
- Zhao, N. *et al.* (2015) Testing in microbiome profiling studies with the Microbiome Regression-based Kernel Association Test (MiRKAT). *Am. J. Hum. Genet.*, **96**, 797–807.