The Annals of Applied Statistics
2018, Vol. 12, No. 1, 1–26
https://doi.org/10.1214/17-AOAS1086

A PHYLOGENETIC SCAN TEST ON A DIRICHLET-TREE MULTINOMIAL MODEL FOR MICROBIOME DATA

BY YUNFAN TANG*, LI MA†,[1] AND DAN L. NICOLAE*,[2]

*University of Chicago** *and Duke University*†

In this paper, we introduce the phylogenetic scan test (PhyloScan) for investigating cross-group differences in microbiome compositions using the Dirichlet-tree multinomial (DTM) model. DTM models the microbiome data through a cascade of independent local DMs on the internal nodes of the phylogenetic tree. Each of the local DMs captures the count distributions of a certain number of operational taxonomic units at a given resolution. Since distributional differences tend to occur in clusters along evolutionary lineages, we design a scan statistic over the phylogenetic tree to allow nodes to borrow signal strength from their parents and children. We also derive a formula to bound the tail probability of the scan statistic, and verify its accuracy through simulations. The PhyloScan procedure is applied to the American Gut dataset to identify taxa associated with diet habits. Empirical studies performed on this dataset show that PhyloScan achieves higher testing power in most cases.

1. Introduction. Microbiome refers to the full collection of genes of all microbes in a community; for example, all bacteria in a sample from the gut of a healthy individual. The advent of next generation sequencing technologies, such as Illumina Solexa, has allowed researchers to investigate the microbiome communities at an unprecedented level of quantification. The focus of this paper is on targeted amplicon sequencing and not on metagenome, but the ideas introduced here can be easily extended to metagenome data. A typical analysis pipeline involves sequencing one or a few of the variable regions of 16s ribosomal RNA, clustering the sequences into operational taxonomic units (OTU), assigning taxonomy to OTUs according to a reference database, and constructing a phylogenetic tree [e.g., Caporaso et al. (2010)]. There have been burgeoning efforts devoted to the study of human microbiome in the past decade, many of which aim at establishing evidence between microbiome and treatment effects or environmental covariates. Examples of such include associating gut microbiome with diet [David et al. (2014)], autism spectrum disorder [McDonald et al. (2015b)] and hormones [Neuman et al. (2015)]. Several of these studies are very large-scale initiatives such as the Human Microbiome Project [Human Microbiome Project Consortium

Received October 2016; revised March 2017.
[1]Supported in part by NSF Grant DMS-1612889 and a Google Faculty Research Award.
[2]Supported in part by NIH Grants R01-MH101820 and R01-HL129735.
Key words and phrases. Dirichlet-multinomial, microbiome, Dirichlet-tree multinomial, phylogenetic tree, PhyloScan, scan statistics, union probability.

(2012)] and American Gut [McDonald, Birmingham and Knight (2015a)], providing a broader understanding of the microbial variability. These studies jointly point to the fact that a microbiome plays an integral part to our health, and much still remains to be explored in this area.

The vast improvement in experimental tools contrasts with the slower development of statistical methods to analyze microbiome data. Typically, the majority of taxa can be observed in only a very small subset of samples, which causes the data table to be highly sparse. In addition, the within-group heterogeneity among samples leads to pronounced overdispersion in taxa proportions. Since standard multinomial distributions fail at capturing these features, Dirichlet-multinomial (DM) has been used as a natural extension. DM was originally proposed by Mosimann (1962) and introduced into the microbiome context by La Rosa et al. (2012) and Holmes, Harris and Quince (2012). Applying DM to test a cross-group variation suffers from a number of drawbacks such as inability to localize any signal to a subgroup of taxa and reduced test power when a large number of taxa is present. Recent efforts to tackle these issues focus on incorporating a phylogenetic tree into the model [Tang et al. (2016), Silverman et al. (2017) and Wang and Zhao (2017)]. In particular, Wang and Zhao (2017) applied an extension of DM, namely the Dirichlet-tree multinomial (DTM), first proposed by Dennis (1991) under the name hyper-Dirichlet type 1 distribution. DTM is based on a decomposition of the sample space through a cascade of nested partitions similar to a Polya tree process [Lavine (1992)]. Instead of placing a single global DM on all taxa, DTM consists of a collection of independent local DMs, each corresponding to a particular internal node on the phylogenetic tree. Since descendants of each internal node on the phylogenetic tree share a certain degree of evolutionary affinity, such decomposition strategy allows one to assign meaningful interpretation to each of the local DM distributions. An additional benefit is that the local DMs target only particular groups of taxa, and consequently enjoy much lower degrees of freedom. This breaking down of the global distribution on all taxa counts allows testing each branch of the phylogeny individually, hence locating the signals to a certain taxonomic rank. The global cross-group test is therefore represented by a number of independent and biologically relevant constituents. For a more general application of the Polya tree decomposition to hypothesis testing, see Ma and Wong (2011), Chen and Hanson (2014), Holmes et al. (2015) and Soriano and Ma (2017).

Although standard multiple testing procedures could be applied to results from testing all nodes, it is usually not the best practice to treat each hypothesis as a segregated entity. Soriano and Ma (2017) pointed out that cross-group distributional variations tend to cluster, which causes hypotheses defined on nearby and/or nested windows more likely to be jointly true or false. This observation also holds in the microbiome data; cross-group differences in a certain ancestor node are more frequently accompanied with similar differences in its descendants. To take advantage of this structure and optimize test power, we adopt ideas from scan tests through constructing a collection of triplet statistics, each incorporating evidence

from an internal node on the phylogenetic tree along with its parent and one of its children. The maximum of all these triplet statistics is used to test the global null hypothesis. Since the exact distribution of the maximum statistic is intractable, we derive an upper and lower bound on its tail probability based on existing results on union probability [e.g., Hunter (1976), Efron (1997) and Taylor, Worsley and Gosselin (2007)]. Our improved strategy first finds a subset consisting of independent components from the union, followed by bounding the probability of remaining components conditioned on the complement of that subset. A decay rate of the relative error of our approximation is also provided.

Section 2 briefly reviews the DM model. Section 3 formulates the DTM model and establishes its relation to the DM. Section 4 develops p-value approximation on the scan statistic for the DTM and verifies the result through simulation. Section 5 applies the DTM model on the American Gut dataset to test the association of gut microbiome with a number of dietary habits. It also empirically demonstrates improvement of DTM over DM through likelihood ratio tests and comparing simulated test power. Section 6 concludes with further discussions on potential DTM extensions.

2. Dirichlet-multinomial for microbiome data. In this section, we briefly recap the cross-group testing procedures on microbiome data using the Dirichlet-multinomial model, as presented in La Rosa et al. (2012).

Consider a microbiome dataset with n samples and let Ω be the collection of a total of $K = |\Omega|$ OTUs. Without loss of generality, we assume $\Omega = \{1, 2, \ldots, K\}$. Each sample is a K-dimensional count vector representing the number of sequences in each of the K OTUs. Let $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iK})$ be the taxa count vector of the ith sample for $i = 1, 2, \ldots, n$. In addition, define $N_{i\cdot} = \sum_{j=1}^{K} x_{ij}$ to be the total number of sequences in the ith sample, $N_{\cdot j} = \sum_{i=1}^{n} x_{ij}$ to be the total number of sequences in the jth OTU, and $N_{\cdot\cdot} = \sum_{i=1}^{n} N_{i\cdot} = \sum_{j=1}^{K} N_{\cdot j}$. The Dirichlet-multinomial (DM) model assumes

$$\boldsymbol{q}_i \overset{\text{i.i.d.}}{\sim} \text{Dir}(\nu\boldsymbol{\pi}), \qquad \boldsymbol{x}_i | \boldsymbol{q}_i \sim \text{Multinomial}(N_{i\cdot}, \boldsymbol{q}_i),$$

where $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K)$ satisfies $\sum_{j=1}^{K} \pi_j = 1$, $\pi_j > 0$ denotes the mean taxa proportions and $\nu > 0$ is a dispersion parameter that controls the level of variation across samples. Alternatively, one may use $\theta = \frac{1}{1+\nu}$ to parametrize the dispersion so that $0 \le \theta < 1$. Integrating out the $\boldsymbol{q}_i$ gives

$$f(\boldsymbol{x}_i) = \binom{N_{i\cdot}}{\boldsymbol{x}_i} \frac{\Gamma(\nu)}{\Gamma(N_{i\cdot} + \nu)} \prod_{j=1}^{K} \frac{\Gamma(x_{ij} + \nu\pi_j)}{\Gamma(\nu\pi_j)}. \tag{1}$$

Throughout this paper, we use $f(\cdot)$ exclusively to denote the DM probability mass function. When $\nu = \infty$ ($\theta = 0$), the DM degenerates to the standard multinomial distribution. Smaller values of ν indicate larger degrees of overdispersion.

Assuming $\boldsymbol{x}_i$'s are independent, the likelihood function is simply the product of probabilities over all samples:

$$\mathcal{L}(\boldsymbol{\pi}, \nu) = \prod_{i=1}^{n}\left[\binom{N_{i\cdot}}{\boldsymbol{x}_i}\frac{\Gamma(\nu)}{\Gamma(N_{i\cdot}+\nu)}\prod_{j=1}^{K}\frac{\Gamma(x_{ij}+\nu\pi_j)}{\Gamma(\nu\pi_j)}\right]. \tag{2}$$

As is shown in Weir and Hill (2002), the method of moments (MoM) estimates of the mean proportion $\boldsymbol{\pi}$ and dispersion θ are respectively

$$\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_K) \qquad \text{with } \hat{\pi}_j = N_{\cdot j}/N_{\cdot\cdot},$$

$$\hat{\theta} = \frac{\sum_{j=1}^{K}(S_j - G_j)}{\sum_{j=1}^{K}(S_j + (N_c - 1)G_j)},$$

where we have $N_c = \frac{N_{\cdot\cdot} - (N_{\cdot\cdot})^{-1}\sum_{i=1}^{n} N_{i\cdot}^2}{n-1}$, $S_j = \frac{\sum_{i=1}^{n} N_{i\cdot}(\hat{\pi}_{ij} - \hat{\pi}_j)^2}{n-1}$ and $G_j = \frac{\sum_{i=1}^{n} N_{i\cdot}\hat{\pi}_{ij}(1-\hat{\pi}_{ij})}{\sum_{i=1}^{n}(N_{i\cdot}-1)}$ with $\hat{\pi}_{ij} = x_{ij}/N_{i\cdot}$.

For hypothesis testing, suppose we collect G groups and the gth group data is given by $\boldsymbol{x}_1^{(g)}, \boldsymbol{x}_2^{(g)}, \ldots, \boldsymbol{x}_{n_g}^{(g)}$ with $N_{i\cdot}^{(g)} = \sum_{j=1}^{K} x_{ij}^{(g)}$ and $N_{\cdot\cdot}^{(g)} = \sum_{i=1}^{n} N_{i\cdot}^{(g)}$. Similarly, we define the gth group parameters as $\boldsymbol{\pi}^{(g)}$, $\nu^{(g)}$ with $\theta^{(g)} = \frac{1}{1+\nu^{(g)}}$. We wish to test the equality of mean proportion across all groups:

$$H_0 : \boldsymbol{\pi}^{(1)} = \boldsymbol{\pi}^{(2)} = \cdots = \boldsymbol{\pi}^{(G)} \text{ vs. } H_a : \text{otherwise}.$$

Let $\hat{\boldsymbol{\pi}}^{(g)}$ and $\hat{\theta}^{(g)}$ be the MoM estimates of $\boldsymbol{\pi}^{(g)}$ and $\theta^{(g)}$, respectively. The cross-group pooled estimate of $\boldsymbol{\pi}$ is $\hat{\boldsymbol{\pi}}^{(\text{Pool})} = \sum_{g=1}^{G} \bar{s}_g \hat{\boldsymbol{\pi}}^{(g)}$ with

$$\bar{s}_g = \frac{(N_{\cdot\cdot}^{(g)})^2 C(\hat{\theta}^{(g)}, N_{\cdot\cdot}^{(g)})^{-1}}{\sum_{r=1}^{G}(N_{\cdot\cdot}^{(r)})^2 C(\hat{\theta}^{(r)}, N_{\cdot\cdot}^{(r)})^{-1}},$$

where

$$C\big(\hat{\theta}^{(g)}, N_{\cdot\cdot}^{(g)}\big) = \hat{\theta}^{(g)}\left(\sum_{i=1}^{n_g}\big(N_{i\cdot}^{(g)}\big)^2 - N_{\cdot\cdot}^{(g)}\right) + N_{\cdot\cdot}^{(g)}.$$

Finally, the test statistic is defined as

$$T = \sum_{g=1}^{G}\big(\hat{\boldsymbol{\pi}}^{(g)} - \hat{\boldsymbol{\pi}}^{(\text{Pool})}\big)^T (\bar{S}_g)^{-1}\big(\hat{\boldsymbol{\pi}}^{(g)} - \hat{\boldsymbol{\pi}}^{(\text{Pool})}\big), \tag{3}$$

where $\bar{S}_g$ is a diagonal matrix given by

$$\bar{S}_g = \big(\big(N_{\cdot\cdot}^{(g)}\big)^2 C\big(\hat{\theta}_g, N_{\cdot\cdot}^{(g)}\big)^{-1}\big)^{-1} D\big(\hat{\boldsymbol{\pi}}^{(\text{Pool})}\big),$$

and $D(\hat{\boldsymbol{\pi}}^{(\text{Pool})})$ is also diagonal with diagonal elements given by $\hat{\boldsymbol{\pi}}^{(\text{Pool})}$. The asymptotic distribution of T under H_0 is $\chi^2_{(K-1)(G-1)}$ as $n_g \to \infty$ for all g.

3. Dirichlet-tree multinomial and hypothesis testing. To incorporate the phylogenetic tree into the model, Wang and Zhao (2017) considered an extension named Dirichlet-tree multinomial (DTM). DTM allows us to separately test cross-group differences in each internal node, locating the source of overall difference within particular subgroups of OTUs. Each of the local test, by design, has the benefit of reduced degrees of freedom.

3.1. *Model formulation.* Let $\mathcal{T} = (\Omega, \mathcal{I})$ be a rooted phylogenetic tree where the set of OTUs Ω are placed on the leaves and $\mathcal{I}$ is the set of all internal nodes. We represent the elements in $\mathcal{I}$ to be subsets of Ω since each internal node is uniquely identified by the subset of all OTUs underneath it, and vice versa. Each subset of OTU that corresponds to an internal node shares a hypothetical ancestor along the lineage. Additionally, each leaf node is uniquely identified by a singleton set consisting of that particular OTU.

Figure 1 shows an example of a simple phylogenetic tree over 5 OTUs and 4 internal nodes. This tree has $\Omega = \{1, 2, 3, 4, 5\}$ and $\mathcal{I} = \{\{1, 2, 3, 4, 5\}, \{1, 2, 3\}, \{4, 5\}, \{2, 3\}\}$.

Now for $\forall A \in \mathcal{I}$, let $\mathcal{C}(A)$ be the collection of A's child nodes in $\mathcal{T}$. The elements of $\mathcal{C}(A)$ are also subsets of Ω. Also, $\forall A \in \mathcal{I} \cup \{\{\omega\} | \omega \in \Omega\}$, $A \neq \Omega$, let $R(A)$ denote the parent node of A. In Figure 1, for example, $\mathcal{C}(\{1, 2, 3\}) = \{\{1\}, \{2, 3\}\}$ and $R(\{1, 2, 3\}) = \{1, 2, 3, 4, 5\} = \Omega$. Notice that certain $\mathcal{C}(A)$'s contain singletons of Ω since some children are leaves. Let $k(A) = |\mathcal{C}(A)|$ be the number of children under A and write $\mathcal{C}(A) = \{\mathcal{C}(A)_1, \mathcal{C}(A)_2, \ldots, \mathcal{C}(A)_{k(A)}\}$. For each $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, k(A)$, let

$$x_{ij}(A) = \sum_{\omega \in \mathcal{C}(A)_j} x_{i\omega}$$

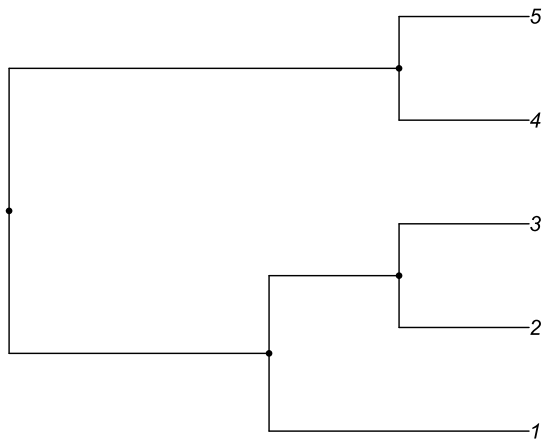

FIG. 1. *An example of a phylogenetic tree with five OTUs.*

be the count of the jth child of A in the ith sample. The count vector associated with A is therefore

$$\boldsymbol{x}_i(A) = \big(x_{i1}(A), x_{i2}(A), \ldots, x_{ik(A)}(A)\big)$$

with the sum $N_i(A) = \sum_{j=1}^{k(A)} x_{ij}(A) = \sum_{\omega \in A} x_{i\omega}$. It is straightforward to see that $N_i(\mathcal{C}(A)_j) = x_{ij}(A)$ for $j = 1, 2, \ldots, k(A)$. In addition, we always have $N_i(\Omega) = N_{i\cdot}$.

The DTM distribution separately models the count vector $\boldsymbol{x}_i(A)$ conditional on $N_i(A)$ for each A. Specifically, for $\forall A \in \mathcal{I}$,

$$\begin{aligned} &\boldsymbol{q}_{A,i} \overset{\text{i.i.d.}}{\sim} \text{Dir}(\nu_A \boldsymbol{\pi}_{\boldsymbol{A}}), \qquad \boldsymbol{x}_i(A)|N_i(A), \\ &\boldsymbol{q}_{A,i} \sim \text{Multinomial}\big(N_i(A), \boldsymbol{q}_{A,i}\big), \end{aligned} \tag{4}$$

where $\nu_A > 0$ is the overdispersion parameter of the counts of A's children and $\boldsymbol{\pi}_A = (\pi_{A,1}, \pi_{A,2}, \ldots, \pi_{A,k(A)})$ satisfying $\sum_{i=1}^{k(A)} \pi_{A,i} = 1$ denotes their mean proportion. The Dirichlet prior distribution of all A's are mutually independent. Integrating out $q_{A,i}$ gives

$$f\big(\boldsymbol{x}_i(A)|N_i(A)\big) = \binom{N_i(A)}{\boldsymbol{x}_i(A)} \frac{\Gamma(\nu_A)}{\Gamma(N_i(A) + \nu_A)} \prod_{j=1}^{k(A)} \frac{\Gamma(x_{ij}(A) + \nu_A \pi_{A,j})}{\Gamma(\nu_A \pi_{A,j})}, \tag{5}$$

which ultimately yields

$$f_T(\boldsymbol{x}_i) = \prod_{A \in \mathcal{I}} f\big(\boldsymbol{x}_i(A)|N_i(A)\big) \tag{6}$$

and likelihood function

$$\mathcal{L}_T\big(\{(\nu_A, \boldsymbol{\pi}_A) : A \in \mathcal{I}\}\big) = \prod_{i=1}^{n} \prod_{A \in \mathcal{I}} f\big(\boldsymbol{x}_i(A)|N_i(A)\big) \tag{7}$$

with $f_T(\cdot)$ and $\mathcal{L}_T(\cdot)$ denoting the DTM probability mass function and likelihood function, respectively. The representations in (5) and (6) naturally lead to a top-down generative scheme of the count data on the nodes, as each layer of DM models a subset of OTU counts at increased level of resolution conditioned on their sum.

Interestingly, Dennis (1991) showed that the global DM distribution on OTU counts is nested in the DTM family. A simple explanation of this relation is that both the global Dirichlet prior and the multinomial probabilities can be factorized over $\mathcal{I}$, that is,

$$\begin{aligned} &\boldsymbol{q}_i \sim \text{Dir}(\nu \boldsymbol{\pi}) \\ &\quad \Leftrightarrow \quad \forall A \in \mathcal{I}, \qquad \frac{\boldsymbol{q}_i(A)}{\sum_{\omega \in A} q_{i\omega}} \sim \text{Dir}\Big(\nu \sum_{\omega \in A} \pi_\omega \cdot \frac{\boldsymbol{\pi}(A)}{\sum_{\omega \in A} \pi_\omega}\Big) \text{ independently,} \end{aligned}$$

$$\begin{aligned}\boldsymbol{x}_i|\boldsymbol{q}_i &\sim \text{Multinomial}(N_{i,}, \boldsymbol{q}_i)\\ &\Leftrightarrow \quad \forall A \in \mathcal{I}, \boldsymbol{x}_i(A)|\boldsymbol{q}_i,\\ &\qquad N_i(A) \sim \text{Multinomial}\Big(N_i(A), \frac{\boldsymbol{q}_i(A)}{\sum_{\omega\in A} q_{i\omega}}\Big),\end{aligned}$$

where we similarly defined

$$q_{ij}(A) = \sum_{\omega\in\mathcal{C}(A)_j} q_{i\omega}, \boldsymbol{q}_i(A) = \big(q_{i1}(A), q_{i2}(A), \ldots, q_{ik(A)}(A)\big),$$

$$\pi_j(A) = \sum_{\omega\in\mathcal{C}(A)_j} \pi_\omega, \boldsymbol{\pi}(A) = \big(\pi_1(A), \pi_2(A), \ldots, \pi_{k(A)}(A)\big).$$

In the DTM representation of global DM, the overdispersion and mean proportion of the counts of A's children are respectively $\nu_A = \nu\sum_{\omega\in A}\pi_\omega$ and $\boldsymbol{\pi}_A = \boldsymbol{\pi}(A)/\sum_{\omega\in A}\pi_\omega$. It is not hard to notice that there is a bijective correspondence between $\boldsymbol{\pi}$ and $\{\boldsymbol{\pi}_A = \boldsymbol{\pi}(A)/\sum_{\omega\in A}\pi_\omega : A \in \mathcal{I}\}$. In addition, all of these local dispersions are governed by a single global ν, which is highly restrictive as it does not allow any node-specific characterization of within-group variation. Section 5.2 provides likelihood ratio test results supporting this claim.

3.2. *Hypothesis testing.* The DTM model in (6) and (7) motivates a node-by-node testing strategy for cross-group comparison. To compare the proportion across G groups of observations, we carry out an MoM test using (3) individually for each A, that is,

$$H_{0,A} : \boldsymbol{\pi}_A^{(1)} = \boldsymbol{\pi}_A^{(2)} = \cdots = \boldsymbol{\pi}_A^{(G)} \text{ vs. } H_{a,A} : \text{otherwise}$$

Each of the MoM test statistics is calculated conditional on $\{N_i^{(g)}(A)|1 \le g \le G, 1 \le i \le n_g\}$, where $N_i^{(g)}(A)$ is the sum of OTU counts under A in the ith sample of gth group. The test statistic for $H_{0,A}$ has degrees of freedom $(G-1)(k(A)-1)$, much smaller than the degrees of freedom for DM test as $(G-1)(K-1)$. The local DM test is therefore more powerful than the global DM test, provided that the extent of cross-group difference on the internal nodes is not diluted too much as we group multiple OTUs together. Obviously, the extent of dilution is largely determined by the tree structure. The ideal scenario is that OTUs placed under the same internal node A demonstrate increasing or decreasing abundance simultaneously for all samples in a certain group, so $H_{0,R(A)}$ will be most effective. This also motivates using the phylogenetic tree to carry out the decomposition, as functionally similar OTUs tend to exhibit similar abundance changes within the same group.

The mean proportion of all OTUs across G groups are equal if and only if $H_{0,A}$ is true for all $A \in \mathcal{I}$. Therefore, we define the global null as $H_0 = \bigcap_{A\in\mathcal{I}} H_{0,A}$.

Controlling the Type-I error on the global null is simply equivalent to controlling the family-wise error rate (FWER) across $H_{0,A}$'s.

The following theorem makes controlling FWER straightforward.

THEOREM 1. *Let p_A be the MoM p-value for testing $H_{0,A}$. Under the global null $H_0 = \bigcap_{A\in\mathcal{I}} H_{0,A}$, p_A's are asymptotically mutually independent as the number of subjects in each group goes to infinity.*

The online supplementary material [Tang, Ma and Nicolae (2018)] has a proof of this theorem.

The independence of p-value under the null grants one of the following procedures to control the exact FWER at level α: (i) Sidak's procedure, in which one assigns equal Type I error $\alpha(A) = 1 - (1-\alpha)^{1/\mathcal{I}}$ to all A's (ii) allocate α_A according to the tree structure while constraining $1 - \prod_{A\in\mathcal{I}}(1-\alpha_A) = \alpha$. After choosing the individual Type I error thresholds, one can report the collection of nodes $\{A : p_A < \alpha_A, A \in \mathcal{I}\}$ as being significant.

4. PhyloScan: Scan statistic over the tree tuples. Cross-group difference in distributions of taxa counts often occurs in clusters or chains on the phylogenetic tree. If one internal node exhibits significant difference in relative proportion across several groups, then this is often associated with signals from at least one of its children or parent. Figure 4 shows four examples of signal clusters on American Gut data using the top 100 OTUs with the highest counts. In each graph, subjects are divided into two groups according to different ingestion frequencies in one of the following diets: milk and cheese, seafood, sugary sweets and vegetable. (details in Section 5.1). The size of the circle on internal node A is proportional to -log(p_A) from the cross-group comparison (the circle colors are irrelevant here). It is apparent that large circles tend to form in chaining patterns, which motivates scanning for signals in chains or clusters instead of on each node separately. Moreover, the partitioning nature of the phylogenetic tree always leads to a much smaller sample size on the bottom nodes (farthest from the root placed on top). Sharing information across nodes would alleviate the limitation to detect distributional differences on the bottom level.

Without prior knowledge of the length and shape of signal clusters, we focus on only triplets formulated by a certain internal node, its parent and one of its children. Each triplet has its own statistic defined as the sum of all the node statistics within, pooling signal strength from its members. The maximum of these statistics on all the triplets is then used to test the global null hypothesis. Our method belongs to the class of scan statistics [Glaz, Naus and Wallenstein (2001)], in which one searches for signals over varying sizes of windows. In our case, each window denotes a particular branch of the phylogenetic tree. The shape of our designed triplet reflects our knowledge of correlated signals on the tree, while the size of the triplet achieves a compromise between signal pooling around neighboring nodes

and the ability to detect alternatives in short chains. Since the exact distribution of the maximum statistic is unknown, we design a novel method to calculate the upper and lower bound of its tail probability using low dimensional integrals that can be efficiently evaluated through standard numerical integration techniques. Since this entire hypothesis testing procedure is established on the phylogenetic tree decomposition, we call it PhyloScan.

4.1. *Overview.* For each $A \in \mathcal{I}$ such that $R(A) \in \mathcal{I}$ and $\mathcal{C}(A) \cap \mathcal{I} \neq \varnothing$, we define a triplet to be the set of three consecutive internal nodes $\{A, R(A), \mathcal{C}(A)_i\}$ where $i \in \{1, 2, \ldots, k(A)\}$ satisfies $\mathcal{C}(A)_i \in \mathcal{I}$. Let $\mathcal{B}$ be the set of all such triplets, and without loss of generality we write $\mathcal{B} = \{\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_b\}$ where each $\mathcal{B}_i$ is a triplet and $b = |\mathcal{B}|$ depends on both K and the structure of the tree. We assume the ordering of elements in $\mathcal{B}$ obeys the following rule: $\{A, R(A), \mathcal{C}(A)_i\}$ always has a smaller index than (or appear in front of) $\{\tilde{A}, R(\tilde{A}), \mathcal{C}(\tilde{A})_j\}$ if $\tilde{A} \subset A$. Now we proceed to define the test statistic for $\mathcal{B}_i$ as follows. First, each of the p-values on the internal nodes can be inverted to a chi-square random variable with 1 degree of freedom, namely

$$Z_A = F_1^{-1}(p_A) \qquad \text{for all } A \in \mathcal{I},$$

where F_j denotes the cumulative distribution function (CDF) of χ_j^2 distribution. Theorem 1 states that under the global null H_0, Z_A's are asymptotically mutually independent. In order to test the following hypothesis on each triplet,

$$H_{0,\mathcal{B}_i} = \bigcap_{A \in \mathcal{B}_i} H_{0,A} \text{ vs. } H_{a,\mathcal{B}_i} : \text{otherwise},$$

we define the statistic to be the sum of Z_A's within:

$$W_i = \sum_{A \in \mathcal{B}_i} Z_A \qquad \text{for } i = 1, 2, \ldots, b. \tag{8}$$

It is apparent that each $W_i \sim \chi_3^2$ under $H_{0,\mathcal{B}_i}$. For the global null hypothesis $H_0 = \bigcap_{A \in \mathcal{I}} H_{0,A} = \bigcap_{i=1}^b H_{0,\mathcal{B}_i}$, we use the maximum of W_i's as the test statistic:

$$W = \max_{1 \le i \le b} W_i. \tag{9}$$

Since $\mathcal{B}_i$ overlaps with each other, W_i's are heavily correlated and the exact distribution of W is hard to derive. For testing purposes, it suffices to calculate the tail probability of W. Suppose our observed value of the maximum statistic is w, and let $B_i(w) = \{W_i > w\}$ be the event of ith triplet statistic exceeding w. Without incurring any confusion, we may drop w and simply write B_i. We are mainly interested in the global p-value $P(\bigcup_{i=1}^b B_i)$, which boils down to the problem of bounding the union probability.

The simplest upper bound of the union probability is the Bonferroni inequality:

$$P\left(\bigcup_{i=1}^{b} B_i\right) \leq \sum_{i=1}^{b} P(B_i).$$

Several authors have provided sharper bounds over the Bonferroni inequality in the past few decades. The results in Hunter (1976), Worsley (1982) and Efron (1997) suggest the following improvement:

$$P\left(\bigcup_{i=1}^{b} B_i\right) = P(B_1) + P\big(B_2 \cap B_1^c\big) + P\big(B_3 \cap B_1^c \cap B_2^c\big) + \cdots$$

$$\leq P(B_1) + \sum_{i=2}^{b} \min_{j<i} P\big(B_i \cap B_j^c\big). \tag{10}$$

In particular, $\min_{j<i} P(B_i \cap B_j^c)$ is achieved at $j = i-1$ when the neighboring variables (W_{i-1}, W_i) have the highest pairwise correlation. Each of the terms inside summation can be easily evaluated by numerical integration. It can be easily generalized to the union of more than two sets to improve approximation. More generally, the above inequality belongs to the class of approximations with the following representation:

$$P\left(\bigcup_{i=1}^{b} B_i\right) \leq \sum_{J\in S} (-1)^{|J|-1} f(J) P\left(\bigcap_{j\in J} B_j\right), \tag{11}$$

where $f(J)$ is some nonnegative function on subset of $S = \{1, 2, \ldots, b\}$. Naiman and Wynn (1992) and Naiman and Wynn (1997) gave results regarding when (11) achieves equality. Following their work, Dohmen (2000) and Dohmen (2002) gave further improvement on the Bonferroni inequalities. There is also research on Bonferroni inequalities for particular applications, such as Dohmen and Tittmann (2004) on partition lattice and Taylor, Worsley and Gosselin (2007) on maxima over Gaussian random fields.

4.2. *Bounding the union probability.* Our upper bound of the union probability involves a decomposition of $\bigcup_{i=1}^{b} B_i$ into (i) a union of independent events and (ii) their complement in $\bigcup_{i=1}^{b} B_i$. The probability of the union of independent events can be exactly evaluated, while a similar strategy to (10) is applied to estimate its complement.

Specifically, let $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_m\}$ be a class of disjoint nonempty subsets of $\mathcal{I}$ satisfying $\forall i \leq m, \exists j \leq b$ s.t. $\mathcal{M}_i \subset \mathcal{B}_j$. For each i, define $M_i = \{\sum_{A\in\mathcal{M}_i} Z_A > w\}$ to be the exceeding event on $\mathcal{M}_i$. It follows that $\forall i \leq m, \exists j \leq b$ s.t. $M_i \subset B_j$. Write $M = \bigcup_{i=1}^{m} M_i$. This leads to

$$P\left(\bigcup_{i=1}^{b} B_i\right) = P(M) + P\big(M^c\big) \cdot P\left(\bigcup_{i=1}^{b} B_i | M^c\right) \tag{12}$$

since $M \subset \bigcup_{i=1}^{b} B_i$. The independence of M_i's leads to a straightforward calculation of $P(M)$ as $P(M) = 1 - F_1(w)^{t_1} F_2(w)^{t_2} F_3(w)^{t_3}$ where $t_l = |\{i \le m : |\mathcal{M}_i| = l\}|$ and $F_i(\cdot)$ is the CDF of χ^2_i distribution. Next, we approximate $P(\bigcup_{i=1}^{b} B_i | M^c)$ using a similar strategy to (10). It is apparent that enlarging M will always decrease $P(\bigcup_{i=1}^{b} B_i \cap M^c)$ and most likely the error of its upper bound, which makes our strategy superior to directly applying (10) to the B_i's.

The next question is how to choose an M as large as possible. An obvious optimality condition is that $\bigcup_{i=1}^{m} \mathcal{M}_i = \mathcal{I}$, because otherwise we can always enlarge M by $\{Z_A > w\}$ for a certain $A \in \mathcal{I} \setminus \bigcup_{i=1}^{m} \mathcal{M}_i$. Moreover, the elements in $\mathcal{M}$ should not be able to combine together and still belong to a certain element in $\mathcal{B}$, that is, $\forall i_1, i_2 \le m, \nexists j \le b$ s.t. $\mathcal{M}_{i_1} \cup \mathcal{M}_{i_2} \subset \mathcal{B}_j$. This is because merging $\mathcal{M}_{i_1}$ and $\mathcal{M}_{i_2}$ enlarges M, that is, $M_{i_1} \cup M_{i_2} \subset \{\sum_{A \in \mathcal{M}_{i_1} \cup \mathcal{M}_{i_2}} Z_A > w\}$. Since exhaustively searching through overall combinations is computationally infeasible for large trees, we propose the following greedy algorithm that satisfies these optimality conditions:

(a) Order the elements in $\mathcal{I}$ as $A_1, A_2, \ldots, A_{|\mathcal{I}|}$ such that each internal node always appears in front of its children.
(b) Set $\mathcal{M} = \varnothing$ and $i = 1$.
(c) For each $i = 1, 2, \ldots, |\mathcal{I}|$, sequentially go through the following steps:

(i) If $\exists j \le m$ s.t. $A_i \in \mathcal{M}_j$, set $i \leftarrow i + 1$ and go back to the beginning of step (c).

(ii) If $\exists j_1, j_2$ s.t. $\mathcal{C}(A_i)_{j_1} \in \mathcal{I}$ and $\mathcal{C}(\mathcal{C}(A_i)_{j_1})_{j_2} \in \mathcal{I}$, set $\mathcal{M} \leftarrow \mathcal{M} \cup \{A_i, \mathcal{C}(A_i)_{j_1}, \mathcal{C}(\mathcal{C}(A_i)_{j_1})_{j_2}\}$ and $i \leftarrow i + 1$. Go back to the beginning of step (c).

(iii) If $\exists j_1$ s.t. $\mathcal{C}(A_i)_{j_1} \in \mathcal{I}$, set $\mathcal{M} \leftarrow \mathcal{M} \cup \{A_i, \mathcal{C}(A_i)_{j_1}\}$ and $i \leftarrow i + 1$. Go back to the beginning of step (c).

(iv) Otherwise, set $\mathcal{M} \leftarrow \mathcal{M} \cup \{A_i\}$ and $i \leftarrow i + 1$. Go back to the beginning of step (c).

The above greedy algorithm seeks to incorporate the longest chain (with maximum of 3 nodes) starting from A_i and use its descendants as subsequent nodes, if A_i has not been included in $\mathcal{M}$ so far. Since the parent node is always considered ahead of its children, the resulting $\mathcal{M}$ will always satisfy the two aforementioned optimality conditions. As the algorithm prioritizes longer chains at each step, it effectively produces a large M that yields relatively accurate estimates of the union probability for our applications (numerical results to be shown later).

Figure 2 shows an example of $\mathcal{M}$ on a simple phylogenetic tree with $K = 13$ OTUs. Each internal node in $\mathcal{M}_i$ is assigned the same number i for $i = 1, 2, \ldots, 7$.

The remaining task is to put an upper bound on $P(\bigcup_{i=1}^{b} B_i | M^c)$. For each $i \le b$, let $\mathcal{N}_i = \{j : |\mathcal{B}_j \cap \mathcal{B}_i| = 2 \text{ and } j < i\}$. Apparently, $|\mathcal{N}_i| \le 2$ for all i because of the

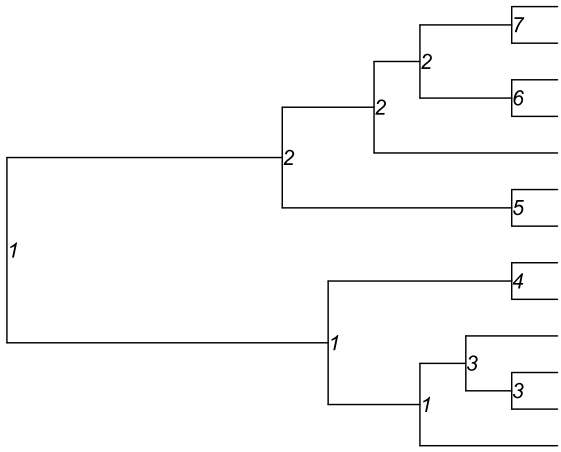

FIG. 2. *Example configuration of $\mathcal{M}$ using the greedy algorithm.*

ordering of $\mathcal{B}_i$'s. Write $B_{\mathcal{N}_i} = \bigcup_{j\in\mathcal{N}_i} B_j$ for short. Now we proceed as follows:

$$P\left(\bigcup_{i=1}^{b} B_i | M^c\right) \leq \sum_{i=1}^{b} P\left(B_i \cap B_{\mathcal{N}_i}^c | M^c\right). \tag{13}$$

The equation in (13) is very similar to (10) in that for each B_i, it includes only the highest correlated events $B_{\mathcal{N}_i}$, which will minimize the right-hand side of the equation. It is worth noting that $P(B_i|M^c) = 0$ if $\mathcal{B}_i \in \mathcal{M}$, hence the strategy of prioritizing triplets while constructing $\mathcal{M}$. To efficiently evaluate each of the terms in the right-hand side of (13), notice that the distributions of Z_A conditioned on M^c are the same as the product of a truncated chi-square and an independent Dirichlet random variable, so their density function can be expressed using chi-square CDFs. Let $f_i(\cdot)$ and $F_i(\cdot)$ denote the density and CDF of χ_i^2 distribution, respectively, then the marginal density of Z_A conditional on M^c becomes

$$f_A\left(z|M^c\right) = \begin{cases} \dfrac{f_1(z)}{F_1(w)} & \text{if } \left|\mathcal{M}(A)\right| = 1 \wedge w \leq z, \\ \dfrac{F_1(w-z)\,f_1(z)}{F_2(w)} & \text{if } \left|\mathcal{M}(A)\right| = 2 \wedge w \leq z, \\ \dfrac{F_2(w-z)\,f_1(z)}{F_3(w)} & \text{if } \left|\mathcal{M}(A)\right| = 3 \wedge w \leq z, \\ 0 & \text{otherwise,} \end{cases} \tag{14}$$

where we define $\mathcal{M}(A) = \mathcal{M}_i$ if $A \in \mathcal{M}_i$. The existence and uniqueness of $\mathcal{M}(A)$ is guaranteed by the fact that $\{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_m\}$ are disjoint with $\bigcup_{i=1}^{m} \mathcal{M}_i = \mathcal{I}$.

The joint density of Z_{A_1} and Z_{A_2} for $\forall A_1, A_2 \in \mathcal{I}$ is

$$f_{A_1,A_2}(z_1,z_2|M^c) = \begin{cases} f_{A_1}(z_1|M^c) f_{A_2}(z_2|M^c) \\ \qquad \text{if } \mathcal{M}(A_1) \cap \mathcal{M}(A_2) = \varnothing, \\ \dfrac{\prod_{i=1}^2 f_1(z_i)}{F_2(w)} \\ \qquad \text{if } \mathcal{M}(A_1) = \mathcal{M}(A_2) \\ \qquad\quad \wedge \big|\mathcal{M}(A_1)\big| = 2 \wedge \sum_{i=1}^2 z_i \le w, \\ \dfrac{F_1(w - \sum_{i=1}^2 z_i) \prod_{i=1}^2 f_1(z_i)}{F_3(w)} \\ \qquad \text{if } \mathcal{M}(A_1) = \mathcal{M}(A_2) \\ \qquad\quad \wedge \big|\mathcal{M}(A_1)\big| = 3 \wedge \sum_{i=1}^2 z_i \le w, \\ 0 \qquad \text{otherwise.} \end{cases} \tag{15}$$

Given w, we pre-calculate the density functions in (14) and (15), and the CDFs of Z_A using (14) and of $Z_{A_1} + Z_{A_2}$ using (15) up to a certain precision and store them into the memory. This turns each term in the right-hand side of (13) into at most two-dimensional integrals. We evaluate these integrals using the functions `cuhre` and `suave` in R package `R2Cuba` [Hahn (2005)].

Substituting (13) into (12) gives

$$\begin{aligned} P_0 &= P\left(\bigcup_{i=1}^b B_i\right) \le P_U \\ &= P(M) + P(M^c) \cdot \sum_{i=1}^b P\big(B_i \cap B_{\mathcal{N}_i}^c | M^c\big), \end{aligned} \tag{16}$$

where P_0 is the actual p-value and P_U is its upper bound. Let $\varepsilon_U = P_U - P_0$ be the error of our approximation. Using a similar strategy to Theorem A1 in Taylor, Worsley and Gosselin (2007), it follows that

$$\begin{aligned} \varepsilon_U &= P\big(M^c\big) \sum_{i=1}^b \big(P\big(B_i \cap B_{\mathcal{N}_i}^c | M^c\big) \\ &\quad - P\big(B_i \cap B_{i-1}^c \cap B_{i-2}^c \cap \cdots \cap B_1^c | M^c\big)\big) \\ &= P\big(M^c\big) \sum_{i=1}^b P\left(B_i \cap B_{\mathcal{N}_i}^c \cap \left(\bigcup_{j<i,\, j \notin \mathcal{N}_i} B_j\right) \middle| M^c\right) \end{aligned} \tag{17}$$

$$\leq P(M^c)\sum_{i=1}^{b} P\Big(\bigcup_{j<i,\, j\notin\mathcal{N}_i}(B_i\cap B_j)|M^c\Big)$$

$$\leq P(M^c)\sum_{i=1}^{b}\sum_{j<i,\, j\notin\mathcal{N}_i} P(B_i\cap B_j|M^c).$$

Each term in (17) can be evaluated by a numerical integral with at most three dimensions using the pre-calculated densities and CDFs. This also establishes

$$P_0 \in \left(P_U - P(M^c)\sum_{i=1}^{b}\sum_{j<i,\, j\notin\mathcal{N}_i} P(B_i\cap B_j|M^c),\, P_U\right). \tag{18}$$

In the next section, we give the numerical results of P_U and upper bound of ε_U using the phylogenetic tree from the American Gut dataset. In addition, we have the following theorem on the convergence rate of the relative error with regards to the observed statistic w.

THEOREM 2. *Given the set of all triplets $\mathcal{B}$ and the partition $\mathcal{M}$ on the internal nodes $\mathcal{I}$, define the following quantities*:

- $\xi_1 = |\{(i,j) : \mathcal{B}_i\cap\mathcal{B}_j = \varnothing, \mathcal{B}_i\notin\mathcal{M}, \mathcal{B}_j\notin\mathcal{M}$ *and* $1\leq j<i\leq b|$.
- $\xi_2 = |\{(i,j) : |\mathcal{B}_i\cap\mathcal{B}_j| = 1, \mathcal{B}_i\notin\mathcal{M}, \mathcal{B}_j\notin\mathcal{M}$ *and* $1\leq j<i\leq b|$.
- $\xi_3 = |\{i : |\mathcal{M}_i| = 3$ *and* $1\leq i\leq m\}|$.

Then under the condition that $(\xi_3-1)(1-F_3(w_T)) < 0.1$ *and* $w_T\geq 12$, *we have*

$$\frac{\varepsilon_U}{P_U} < \frac{\xi_1}{0.95\xi_3+\xi_T}\big(1-F_3(w)\big) + \frac{0.9\xi_2}{0.95\xi_3+\xi_T}\sqrt{\frac{\pi}{2}}\cdot\frac{1}{w}$$
$$= \mathcal{O}\big(e^{-\frac{w}{6}}\big) + \mathcal{O}\big(w^{-1}\big)$$

for all $w\geq w_T$, *where*

$$\xi_T = \frac{\sum_{i=1}^{b} P(B_i\cap B^c_{\mathcal{N}_i}\cap M^c)}{1-F_3(w)} \qquad \textit{is evaluated at } w = w_T.$$

See the supplementary material [Tang, Ma and Nicolae (2018)] for the proof.

4.3. *Comparison with Monte-Carlo simulation.* We compared the lower and upper bound in (18) with Monte-Carlo simulated p-values. Each round of simulation produces 5×10^4 simulated maximum triplet statistics, and we use their proportion of exceeding w as the estimated p-value. The maximum triplet statistic is simulated through generating the χ^2_1 distributed Z_A's for all $A\in\mathcal{I}$ and then applying (8) and (9). We draw the comparison for a variety of scenarios with different

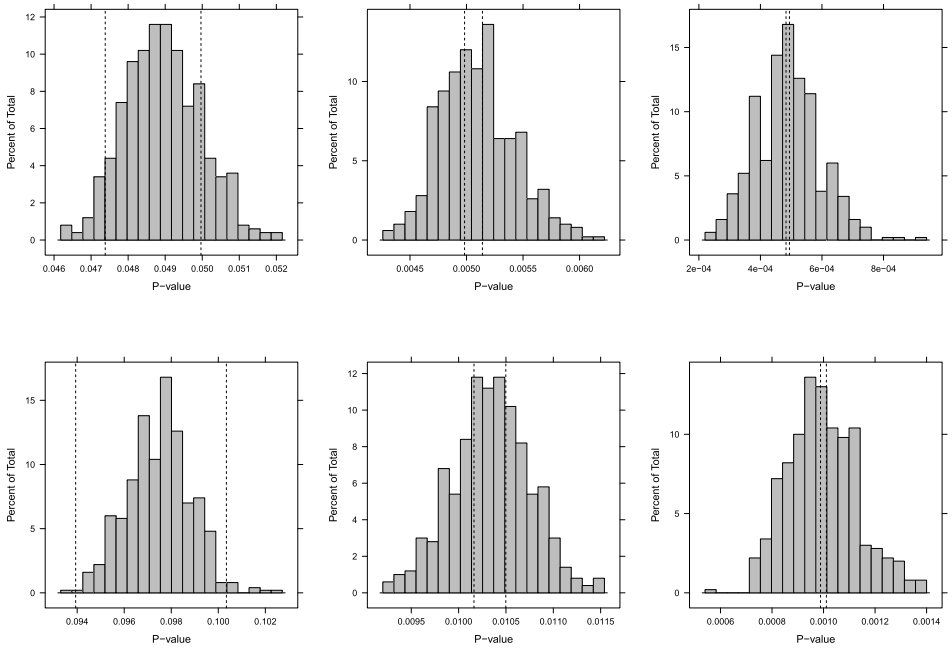

FIG. 3. *Comparison between the interval bound and simulated p-values. Each simulated p-value is the proportion exceeding w over 5×10^4 runs. Dashed lines indicate the upper and lower bound as in (18). Top row and bottom row indicate $K = 50$ and $K = 100$, respectively. Left column: $w = 15$, middle column: $w = 20$, and right column: $w = 25$.*

numbers of OTU $K \in \{50, 100\}$ and different observed statistic $w \in \{15, 20, 25\}$. Given K, the tree structure is obtained from keeping the top K OTUs with the highest count in all feces samples from the American Gut dataset (introduced later). In each scenario, we provide the histogram of simulated p-values in over 5000 rounds.

Figure 3 shows that our bounds consistently contain the center of the simulated p-values. Since the Monte Carlo p-values are merely binomial proportions, the ratio of their spread (measured by standard deviation) to P_0 goes to infinity as $P_0 \to 0$. In contrast, our method gives a ratio that tends to zero by Theorem 2. This makes our approach particularly useful for scenarios where a large number of tests leads to very small p-value threshold after multiple testing correction. In order to keep a fixed relative error, the computation time of the Monte Carlo method needs to scale up much faster with w than our method.

5. Application to American Gut Project. The American Gut Project [McDonald, Birmingham and Knight (2015a)] is an open-access and crowdsourced initiative that involves the public into the research of human microbiome and aims at providing a much more comprehensive reference set than the previous Human Microbiome Project [Human Microbiome Project Consortium (2012)].

After contributing to the project fund, participants complete a questionnaire and ship their microbiome sample to the sequencing lab currently located at University of California, San Diego. The questionnaire covers a wide range of topics regarding demographic information, diet, lifestyle, etc. Sampling sites include skin, tongue and feces, although the vast majority of participants provided the feces sample. The samples are sequenced on 16s rRNA and further processed by QIIME [Caporaso et al. (2010)] pipeline to produce the OTUs and the phylogenetic tree. The 2016 May 16 cohort of public dataset includes more than eight thousands of subjects, with median of sequences per individual as 14680 and standard deviation as 32,455.

5.1. *Cross-group comparison*. Our focus is comparison of the feces microbiome across different diet habits. We pick the top 100 OTUs with the highest count summing over all feces samples. The phylogenetic tree on these OTUs is fully binary. We also select a total of seven categories of diet from the questionnaire. Each diet divides the samples into two groups; group 1 consists of individuals with ingestion rate less than three times per week, and group 2 corresponds to more than or equal to three times per week. Since the questions are not compulsory, a large number of subjects do not leave any response. The diet names and their sample sizes in both groups are as follows: fermented plant (880 vs. 3024), fruit (2336 vs. 1660), milk and cheese (1743 vs. 2261), poultry (1421 vs. 2611), seafood (556 vs. 3452), sugary sweet (1542 vs. 2493) and vegetable (3422 vs. 577).

For each diet type, we test the equality of mean proportions between two groups using three methods: DTM with PhyloScan, DTM with Sidak correction and global DM. Table 1 presents their p-values using the 100 OTUs. The DTM(PhyloScan) column contains P_U and the upper bound of its error ε_U in the parenthesis, both derived in Section 4.2. DTM(Sidak) column is calculated as the Sidak multiple testing correction $1-(1-\min_{A\in\mathcal{I}} p_A)^{|\mathcal{I}|} \approx |\mathcal{I}| \min_{A\in\mathcal{I}} p_A$. We also provide DM p-values after grouping the 100 OTUs into family and class levels, respectively. The grouping operation based on taxonomy is a common practice in recent papers including La Rosa et al. (2012) and Chen and Li (2013). At each taxonomic level, all OTUs with missing taxa information are placed into the same group. This leads to a total of 22 categories on family level and 9 categories on class level. The DM p-values are calculated using the R package `HMP`.

All diet habit comparisons exhibit significant DTM(PhyloScan) p-values at 0.05 level except fermented plant. This is consistent with the findings in Turnbaugh et al. (2014) and David et al. (2014), both of which established that the human gut microbiome is highly sensitive to the dietary nutrient composition. DTM(Sidak) also produces similar significance results. Although in five out of seven comparisons its p-values are larger than PhyloScan. The largest relative difference occurs at seafood comparison (Sidak p-value about 100 times greater than PhyloScan). The rest of the two comparisons (fermented plant and vegetable) are likely to have

TABLE 1

DM and DTM p-values for testing microbiome compositions across different diet habits. DTM(PhyloScan) contains P_U and the upper bound on ε_U shown in parenthesis. DTM(Sidak) contains the Sidak-corrected p-values $1-(1-\min_{A\in\mathcal{I}} p_A)^{|\mathcal{I}|}$. For DM, we provide p-values directly on the 100 OTUs as well as after grouping the 100 OTUs into family and class levels, respectively

	DTM		DM		
Diet	**PhyloScan**	**Sidak**	**OTU**	**Family**	**Class**
Fermented plant	0.308 (0.036)	0.239	0.377	0.147	0.038
Fruit	8.75×10^{-5} (1.52×10^{-6})	1.64×10^{-4}	2.81×10^{-3}	0.012	0.218
Milk and cheese	1.07×10^{-4} (1.86×10^{-6})	6.48×10^{-3}	0.029	0.262	0.285
Poultry	0.023 (8.71×10^{-4})	0.111	0.158	0.287	0.691
Seafood	6.85×10^{-5} (1.17×10^{-6})	6.40×10^{-3}	1.75×10^{-4}	0.194	0.772
Sugary sweet	5.13×10^{-3} (1.43×10^{-4})	0.015	0.719	0.558	0.815
Vegetable	7.39×10^{-5} (1.27×10^{-6})	4.79×10^{-5}	3.77×10^{-3}	1.88×10^{-3}	0.014

either a single dominating signal or weak clustering pattern, both of which hurt testing power after signal pooling. Still, PhyloScan has only mildly larger p-values under these circumstances. This data analysis concludes that PhyloScan is superior to Sidak correction in most cases. Note that p-values of DM on OTUs fail to reach significance for fermented plant, poultry and sugary sweet. This happens in even more comparisons on family and class levels.

We further visualize the significant internal nodes in Figure 4 for four of the diet comparisons: milk and cheese, seafood, sugary sweets and vegetable. Using a simple binary search, we find that $w = 16.579$ yields $P_U = 0.05$ with $\varepsilon_U \le 2.43\times10^{-3}$. All triplets with test statistics greater than the above threshold, that is, $W_i > 16.579$, are plotted in dark gray. In some cases, the triplets overlap with each other, leading to a much longer chain than the original setup. We also provide the taxonomy for all internal nodes that belong to a certain significant triplet in Table 2. The internal node taxon is defined according to the following algorithm: starting from kingdom, repeatedly decrease the rank by one level until the descendant OTUs of that particular internal node no longer share the same taxa on the next lower rank (missing taxa on OTUs are excluded). The algorithm then

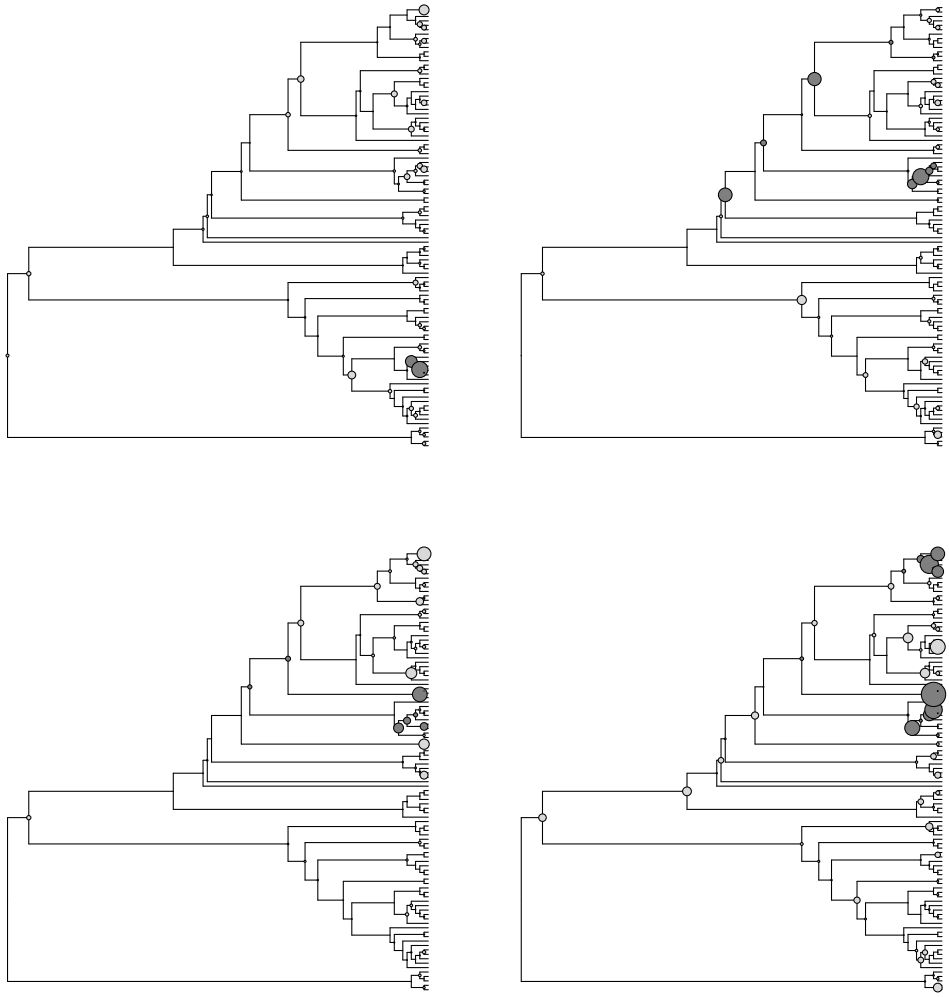

FIG. 4. *Significant triplets from DTM testing. Top left*: *Milk and cheese*, *top right*: *seafood*, *bottom left*: *sugary sweets and bottom right*: *vegetable. The size of the circle on internal node A is proportional to* $-\log(p_A)$. *Triplets with* $W_i > 16.579$ *are plotted in dark gray.*

picks the common taxon of the descendant OTUs on the rank at which the algorithm stops. In other words, the internal node taxon reflects the finest classification upon which all of its descendant OTUs agree.

5.2. *DM vs. DTM test.* We can also test the model fit of DTM against DM directly on the OTUs. Since DM is nested in the DTM family, we can use the likelihood ratio test (LRT) for

$$H_0 : \exists \nu > 0 \text{ s.t. } \forall A \in \mathcal{I},\ \nu_A = \nu \sum_{\omega \in A} \pi_\omega \text{ (DM) vs. } H_a : \text{otherwise (DTM)}$$

TABLE 2

Taxa on significant triplets from PhyloDM hypothesis testing for each diet comparison. Each internal node that belongs to a certain significant triplet is assigned a taxon based on its descendant OTUs (details described in Section 5.1). Only the lowest level taxon is reported for each internal node. We omit the class rank since there are no significant internal nodes on such level in any cross-group comparisons

Diet	Phylum	Order	Family	Genus
Fermented plant	–	–	–	–
Fruit	Firmicutes	Clostridiales	Clostridiaceae	Clostridium
	–	–	Ruminococcaceae	Faecalibacterium
Milk and cheese	–	–	–	Bacteroides
Poultry	–	Clostridiales	–	Coprococcus
Seafood	Firmicutes	Clostridiales	Lachnospiraceae	Coprococcus
	–	–	Ruminococcaceae	Ruminococcus
Sugary sweet	Firmicutes	Clostridiales	Lachnospiraceae	Coprococcus
	–	–	Ruminococcaceae	–
Vegetable	Firmicutes	Clostridiales	Lachnospiraceae	Blautia
	–	–	Ruminococcaceae	Coprococcus
	–	–	–	Lachnospira

with the test statistic defined as

$$\Lambda(\boldsymbol{x}) = -2\log\frac{\mathcal{L}(\hat{\nu},\hat{\boldsymbol{\pi}})}{\mathcal{L}_T(\{(\hat{\nu}_A,\hat{\boldsymbol{\pi}}_A) : A\in\mathcal{I}\})} \sim \chi^2_{|\mathcal{I}|-1} \qquad \text{under } H_0, \tag{19}$$

where $(\hat{\nu},\hat{\boldsymbol{\pi}})$ in the numerator of (19) are MLEs of the DM model, and each $(\hat{\nu}_A,\hat{\boldsymbol{\pi}}_A)$ in the denominator are obtained through maximizing the DTM conditional likelihood (5). We use the low-storage BFGS optimization implemented in package `nloptr` to calculate the MLE estimates. The degrees of freedom in (19) is $|\mathcal{I}|-1$ for a binary phylogenetic tree since (i) $\dim(\boldsymbol{\pi}) = \dim(\{\boldsymbol{\pi}_{\boldsymbol{A}} : A\in\mathcal{I}\}) = K-1$, and (ii) $\dim(\{\nu_A : A\in\mathcal{I}\}) = |\mathcal{I}|$.

Table 3 shows the LRT result. The test is separately applied to male and female Caucasians living in a variety of geographic regions. Each region consists of certain states in the U.S. defined according to the Bureau of Economic Analysis. The degree of freedom for all tests is 98 since our phylogenetic tree is binary and $|\mathcal{I}| = |K|-1 = 99$. All scenarios yield LRT p-values less than 10^{-10}, which indicates significantly improved fit on the data using DTM. We also note that $\Lambda(x)$ in general increases with the sample size, as evidence towards heterogeneity in OTU dispersion strengthens with more available data.

5.3. *Simulation*. We use two simulation strategies to evaluate the power of PhyloScan test under various conditions. From the American Gut dataset, we extracted a total of 662 individuals who identified themselves as male Caucasians

TABLE 3
Likelihood ratio test for DM versus DTM. The test is separately applied to male and female Caucasians in a variety of geographic regions. Each test is accompanied by the LRT statistic $\Lambda(x)$ and the sample size

	Male		Female	
Region	$\Lambda(x)$	**Sample size**	$\Lambda(x)$	**Sample size**
Far West	14,179.76	663	15,768.10	775
Great Lakes	4025.36	180	5541.45	276
Mideast	7497.80	328	8112.80	396
New England	5630.21	239	5793.38	269
Rocky Mountain	6084.90	244	6676.25	300
Southeast	7030.43	324	7383.72	366
Southwest	3442.69	153	3675.52	189

living in the far west (Alaska, California, Hawaii, Nevada, Oregon and Washington). In each round of simulation, these selected samples are randomly divided into two equal-sized groups to generate data under the global null. For data under the alternative, the first simulation strategy randomly selects an OTU and increases its count by a fixed percentage for all samples in the second group, whereas the second simulation strategy randomly selects an internal node and increases the count of all of its descendant OTUs equally by a fixed percentage for all samples in the second group. We use the same 100 OTUs as before and produce 5000 rounds of simulation.

Figure 5 demonstrates the distribution of DM and DTM p-values under the global null. We fit three separate DM on 100 OTUs, family level and class level. The histogram of DTM p-values is produced by using all the p-values on the internal nodes. Surprisingly, the distribution of p-values for DM on the OTUs is far from being uniform on $(0, 1)$, which leads to conservative inference and loss of power. The discrepancy alleviates as we group more OTUs into family or class level, although its empirical distribution is still noticeably skewed. This phenomenon reflects the fact that DM is severely under-parametrized for microbiome data even in low dimensions, as it fits a single dispersion parameter that simultaneously controls all categories. In contrast, DTM solves this issue through fitting a family of dispersion parameters $\{\nu_A : A \in \mathcal{I}\}$ that leads to better calibrated p-values.

Figures 6 and 7 show the ROC and power curves when we use the first simulation strategy to increase the count of a random OTU. We provide the result for (i) DM on the OTUs (ii) DTM using the maximum of the single node statistic, or $\max_{A\in\mathcal{I}} Z_A$ and (iii) DTM using the maximum of the triplet statistic, or $\max_{i\leq b} W_i$. The last strategy is the one employed in the PhyloScan procedure.

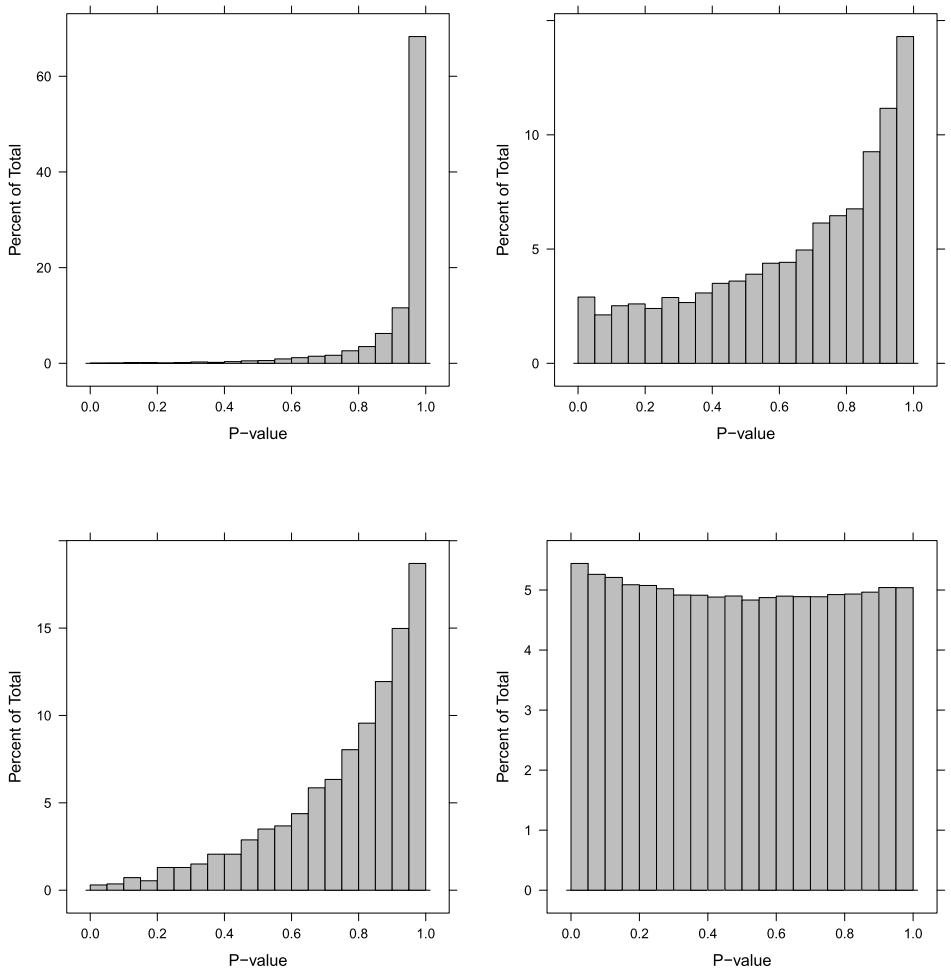

FIG. 5. *P-value histograms under the global null. We randomly place the* 662 *samples from Caucasian males living in the far west into two equal-sized groups and produce their p-values for* 5000 *rounds. Top left is DM on the OTUs, top row right is DM on the family level, bottom left is DM on the class level, and bottom right is DTM.*

Both DTM methods give improved performance compared to DM due to a highly localized signal.

Figures 8 and 9 show the ROC and power curves when we use the second simulation strategy to increase the count of all OTUs under a random internal node. The minimum number of OTUs under the randomly selected internal node controls the degree of localization in the signal. This simulation setup reflects the more biologically meaningful scenario in which a number of taxa exhibit differences in the between-group comparison. In all cases, DTM consistently provides higher power than DM. The DTM 3-node method also provides higher power than the DTM

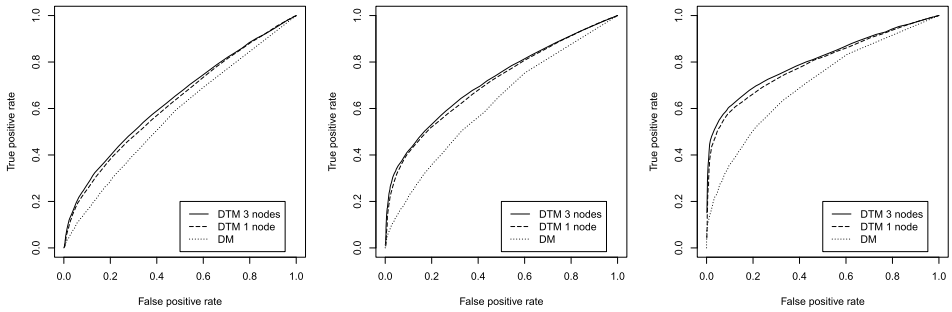

FIG. 6. *ROC curves from increasing the count of a random OTU. For left to right, the percentage increment is set as* 100%, 150% *and* 250%.

1-node at moderate increment levels. When the increment level is high, there will be a certain Z_A whose value dominates all other node statistics, so the extra gain from pooling signal strength within triplets diminishes.

6. Discussion. DTM models the microbiome data through a cascade of local DMs with varying degrees of resolutions on the phylogenetic tree. We take advantage of the correlated signals on the tree through a scan statistic approach and provide an upper and lower bound on its tail probability for testing cross-group differences. Both empirical results on American Gut data and simulations demonstrated the efficiency and accuracy of our method. We also developed the PhyloScan R package, which can be found at https://github.com/yunfantang/PhyloScan.

DTM is a generalization of DM with $|\mathcal{I}| - 1$ more dispersion parameters. An interesting question is whether one could stepwise tune the model (hence the number of parameters) from DM to DTM. To start, the DTM representation in (4) shrinks to the degenerate DM if $\exists \nu > 0$ s.t. $\nu_A = \nu \sum_{w \in A} \pi_w$ for all $A \in \mathcal{I}$. This

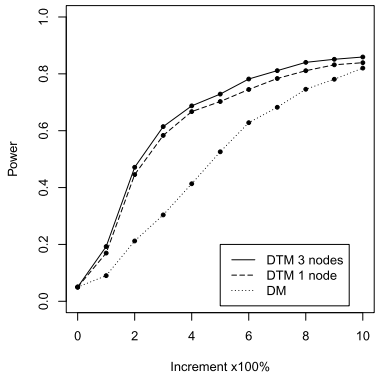

FIG. 7. *Power of DM and DTM with regard to different increments in a random OTU at a false positive rate* = 0.05.

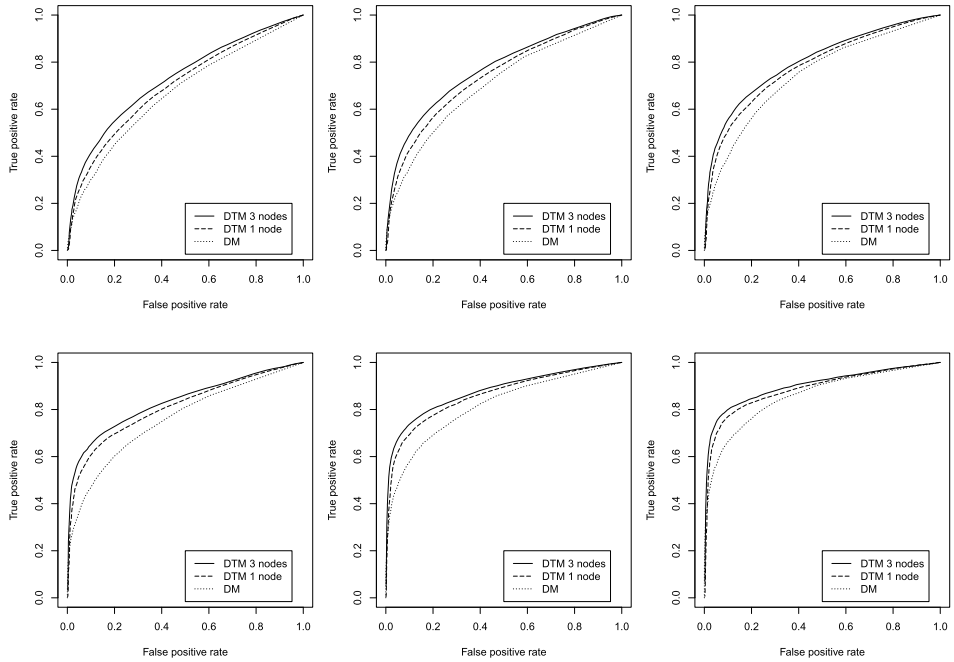

FIG. 8. *ROC curves from increasing the count of all OTUs under a random internal node. The top row and the bottom row have the percentage increment set as* 50% *and* 75%, *respectively. From left to right column, the minimum number of OTUs under the chosen internal node is* 2, 3 *and* 5.

condition is equivalent to $\nu_A = \nu_{R(A)}\pi_{R(A),i}$ for $\forall A \in \mathcal{I}\backslash\{\Omega\}$ with $A = \mathcal{C}(R(A))_i$. Stepwise tuning can be achieved through requiring only $\nu_A = \nu_{R(A)}\pi_{R(A),i}$ to hold over $A \in \tilde{\mathcal{I}}$ where $\tilde{\mathcal{I}} \subset \mathcal{I}\backslash\{\Omega\}$ controls the effective degrees of freedom. Apparently, $\tilde{\mathcal{I}} = \varnothing$ leads to DTM and $\tilde{\mathcal{I}} = \mathcal{I}\backslash\{\Omega\}$ leads to DM, so any choice of $\tilde{\mathcal{I}}$ in the middle yields a model between the two extremes. Standard model selection

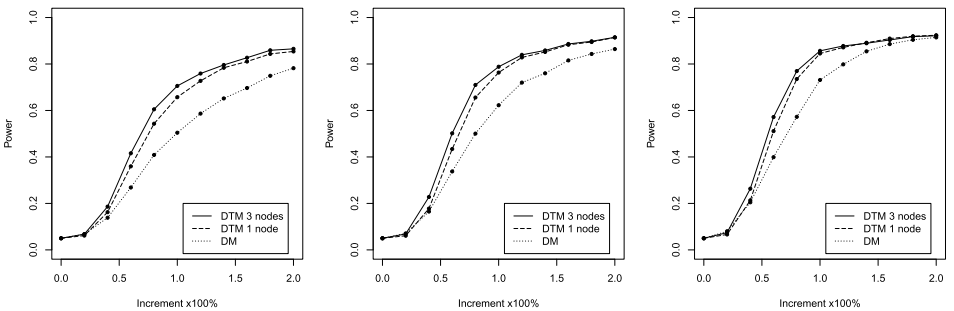

FIG. 9. *Power of DM and DTM with regard to different increments in all OTUs under a random internal node at a false positive rate* = 0.05. *From left to right, the minimum number of OTUs under the chosen internal node is* 2, 3 *and* 5.

techniques such as information criterion or cross validation can then be applied. Although the existence of such spectrum grants substantial flexibility, we note that it can be computationally infeasible to examine the model fit of all $2^{|\mathcal{I}|-1}$ possible configurations. A potential work-around is to enlarge $\mathcal{T}$ stepwise by a greedy algorithm or use dynamic programming, but it is not clear under which conditions we are guaranteed to recover the global optimum.

Since our PhyloScan procedure requires only the p-value as input, it can be easily applied to any extensions or other distributions. For example, the DTM framework can be adapted to incorporate continuous variable of interest and adjust for the effects of confounders. When the tree is fully binary, we let $\lambda_A = \pi_{A,1}$ to fully represent $\boldsymbol{\pi}_A = (\pi_{A,1}, 1-\pi_{A,1})$ in (4). Then we can build separate logistic regression models for each A:

$$\log \frac{\lambda_A}{1-\lambda_A} = \beta_{A,0} + \beta_{A,1} u + \sum_{i=1}^{s} \beta_{A,s+1} c_s, \tag{20}$$

where $\beta_{A,i}$ is the ith regression coefficient, u denotes the continuous variable of interest and $c_1, c_2, \ldots, c_s$ are the confounders. After obtaining maximum likelihood estimates of the coefficients as well as ν_A, we test the significance of u's coefficient to produce p-values and use them as input to PhyloScan in order to borrow strength from neighboring nodes. Another possible extension is related to the issue of zero-adjustment. In Figure 5, DM p-values exhibit apparent right skew under the global null hypothesis. A follow-up inspection shows that MoM estimation tends to produce a higher expected zero count than observed. Both right-skewness and zero-deflation of the global DM are likely caused by underestimation of ν, which makes the Dirichlet prior more dispersed. In DTM, we still observe a mild level of zero-deflation, although the extent is much less severe than global DM. It is also possible to have zero-inflation when one switches to a different sequencing technology or the OTU construction algorithm. Incorporating zero-adjustment into the existing model can lead to significantly better fit while easily handled by PhyloScan.

Acknowledgments. We thank the journal editors and an anonymous referee whose comments have substantially improved the quality of this manuscript. Part of the research was completed when L. Ma was a Visiting Scholar in the Department of Statistics at University of Chicago in 2016.

SUPPLEMENTARY MATERIAL

Theorem proofs (DOI: 10.1214/17-AOAS1086SUPP; .pdf). This supplementary file contains proofs of independence of DTM p-value under the global null (Theorem 1) and error bound of PhyloScan statistic approximation (Theorem 2).

REFERENCES

CAPORASO, J. G., KUCZYNSKI, J., STOMBAUGH, J., BITTINGER, K., BUSHMAN, F. D., COSTELLO, E. K., FIERER, N., PENA, A. G., GOODRICH, J. K., GORDON, J. I. et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7** 335–336.

CHEN, Y. and HANSON, T. E. (2014). Bayesian nonparametric k-sample tests for censored and uncensored data. *Comput. Statist. Data Anal.* **71** 335–346. MR3131974

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. MR3086425

DAVID, L. A., MAURICE, C. F., CARMODY, R. C., GOOTENBERG, D. B., BUTTON, J. E., WOLFE, B. E., LING, A. V., DEVLIN, A. S., VARMA, Y., FISCHBACH, M. A. et al. (2014). Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505** 559–563.

DENNIS, S. Y. III (1991). On the hyper-Dirichlet type 1 and hyper-Liouville distributions. *Comm. Statist. Theory Methods* **20** 4069–4081. MR1158563

DOHMEN, K. (2000). Improved Bonferroni inequalities via union-closed set systems. *J. Combin. Theory Ser. A* **92** 61–67. MR1783939

DOHMEN, K. (2002). Improved inclusion-exclusion identities and Bonferroni inequalities with reliability applications. *SIAM J. Discrete Math.* **16** 156–171. MR1972081

DOHMEN, K. and TITTMANN, P. (2004). Bonferroni–Galambos inequalities for partition lattices. *Electron. J. Combin.* **11** Article ID 85. MR2114189

EFRON, B. (1997). The length heuristic for simultaneous hypothesis tests. *Biometrika* **84** 143–157. MR1450198

GLAZ, J., NAUS, J. and WALLENSTEIN, S. (2001). *Scan Statistics*. Springer, New York. MR1869112

HAHN, T. (2005). Cuba—A library for multidimensional numerical integration. *Comput. Phys. Commun.* **168** 78–95. MR2136794

HOLMES, I., HARRIS, K. and QUINCE, C. (2012). Dirichlet multinomial mixtures: Generative models for microbial metagenomics. *PLoS ONE* **7** Article ID e30126.

HOLMES, C. C., CARON, F., GRIFFIN, J. E. and STEPHENS, D. A. (2015). Two-sample Bayesian nonparametric hypothesis testing. *Bayesian Anal.* **10** 297–320. MR3420884

HUMAN MICROBIOME PROJECT CONSORTIUM (2012). A framework for human microbiome research. *Nature* **486** 215–221.

HUNTER, D. (1976). An upper bound for the probability of a union. *J. Appl. Probab.* **13** 597–603. MR0415722

LAVINE, M. (1992). Some aspects of Pólya tree distributions for statistical modelling. *Ann. Statist.* **20** 1222–1235. MR1186248

LA ROSA, P. S., BROOKS, J. P., DEYCH, E., BOONE, E. L., EDWARDS, D. J., WANG, Q., SODERGREN, E., WEINSTOCK, G. and SHANNON, W. D. (2012). Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PLoS ONE* **7** Article ID e52078. DOI:10.1371/journal.pone.0052078.

MA, L. and WONG, W. H. (2011). Coupling optional Pólya trees and the two sample problem. *J. Amer. Statist. Assoc.* **106** 1553–1565. MR2896856

MCDONALD, D., BIRMINGHAM, A. and KNIGHT, R. (2015a). Context and the human microbiome. *Microbiome* **3** 1–8.

MCDONALD, D., HORNIG, M., LOZUPONE, C., DEBELIUS, J., GILBERT, J. and KNIGHT, R. (2015b). Towards large-cohort comparative studies to define the factors influencing the gut microbial community structure of ASD patients. *Microb. Ecol. Health Dis.* **26** 26555.

MOSIMANN, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika* **49** 65–82. MR0143299

NAIMAN, D. Q. and WYNN, H. P. (1992). Inclusion-exclusion-Bonferroni identities and inequalities for discrete tube-like problems via Euler characteristics. *Ann. Statist.* **20** 43–76. MR1150334

NAIMAN, D. Q. and WYNN, H. P. (1997). Abstract tubes, improved inclusion-exclusion identities and inequalities and importance sampling. *Ann. Statist.* **25** 1954–1983. MR1474076

NEUMAN, H., DEBELIUS, J. W., KNIGHT, R. and KOREN, O. (2015). Microbial endocrinology: The interplay between the microbiota and the endocrine system. *FEMS Microbiol. Rev.* **39** 509–521.

SILVERMAN, J. D., WASHBURNE, A., MUKHERJEE, S. and DAVID, L. A. (2017). A phylogenetic transform enhances analysis of compositional microbiota data. *ELife* **6** Article ID e21887.

SORIANO, J. and MA, L. (2017). Probabilistic multi-resolution scanning for two-sample differences. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 547–572. MR3611759

TANG, Y., MA, L. and NICOLAE, D. L. (2018). Supplement to "A phylogenetic scan test on a Dirichlet-tree multinomial model for microbiome data." DOI:10.1214/17-AOAS1086SUPP.

TANG, Z., CHEN, G., ALEKSEYENKO, A. V. and LI, H. (2017). A general framework for association analysis of microbial communities on a taxonomic tree. *Bioinformatics* **33** 1278–1285. DOI:10.1093/bioinformatics/btw804.

TAYLOR, J. E., WORSLEY, K. J. and GOSSELIN, F. (2007). Maxima of discretely sampled random fields, with an application to 'bubbles'. *Biometrika* **94** 1–18. MR2307898

TURNBAUGH, P. J., RIDAURA, V. K., FAITH, J. J., REY, F. E., KNIGHT, R. and GORDON, J. I. (2014). The effect of diet on the human gut microbiome: A metagenomic analysis in humanized gnotobiotic mice. *Sci. Transl. Med.* **1** Article ID 6ra14. DOI:10.1126/scitranslmed.3000322.

WANG, T. and ZHAO, H. (2017). A Dirichlet-tree multinomial regression model for associating dietary nutrients with gut microorganisms. *Biometrics* **73** 792–801.

WEIR, B. S. and HILL, W. G. (2002). Estimating F-statistics. *Annu. Rev. Genet.* **36** 721–750.

WORSLEY, K. J. (1982). An improved Bonferroni inequality and applications. *Biometrika* **69** 297–302. MR0671966

Y. TANG
D. L. NICOLAE
DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
CHICAGO, ILLINOIS 60637
USA
E-MAIL: yunfantang@uchicago.edu
nicolae@galton.uchicago.edu

L. MA
DEPARTMENT OF STATISTICAL SCIENCE
DUKE UNIVERSITY
DURHAM, NORTH CAROLINA 27708
USA
E-MAIL: li.ma@duke.edu