Biometrika (XXXX), **XXX**, XX, *pp.* 1–20
Printed in Great Britain Advance Access publication on XX XX XXXX

Hypothesis Testing for Phylogenetic Composition: A Minimum-cost Flow Perspective

BY SHULEI WANG

Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, U.S.A.

Shulei.Wang@pennmedicine.upenn.edu

T. TONY CAI

Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104, U.S.A.

tcai@wharton.upenn.edu

AND HONGZHE LI

Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, U.S.A.

hongzhe@upenn.edu

SUMMARY

Quantitative comparison of microbial composition from different populations is a fundamental task in various microbiome studies. We consider two-sample testing for microbial compositional data by leveraging the phylogenetic tree information. Motivated by existing phylogenetic distances, we take a minimum-cost flow perspective to study such testing problems. Our investigation shows that multivariate analysis of variance with permutation using phylogenetic distances, one of the most commonly used methods in practice, is essentially a sum-of-squares type test and has better power for dense alternatives. However, empirical evidence from real data sets suggests that the phylogenetic microbial composition difference between two populations is usually sparse. Motivated by this observation, we propose a new maximum type test, Detector of Active Flow on a Tree, and investigate its properties. It is shown that the proposed method is particularly powerful against sparse phylogenetic composition difference and enjoys certain optimality. The practical merit of the proposed method is demonstrated by simulation studies and an application to a human intestinal biopsy microbiome data set for patients with ulcerative colitis.

Some key words: Microbiome and Metagenomics; Phylogenetic tree; Sparse alternative.; Wasserstein distance.

1. INTRODUCTION

High throughput sequencing technologies make it possible to survey the microbiome communities from multiple samples, resulting in a need for statistical methods to quantitatively compare samples from different populations/experiments. Testing whether two groups of samples have the same microbiome composition is a key step to decipher the quantitative difference between populations and to identify the dysbiotic components. In this paper, we consider two-sample testing for the means of relative abundance from two populations. Although the problem is mainly mo-

© 2018 Biometrika Trust

tivated by microbiome and metagenomic data analysis, it, as a general problem, also arises in other high-throughput sequencing data, e.g. single cell RNA sequencing data.

The microbial community from one sample is usually represented by discrete distributions with the relative abundance of microbe species organized in taxonomy, or operational taxonomic units in some applications. To assess the quantitative difference between groups of samples, various methods have been proposed for the taxonomic compositional data, including global two-sample tests (Zhao et al., 2015; Cao et al., 2017) and differential abundance tests (Robinson et al., 2010; Wagner et al., 2011; Love et al., 2014; Mandal et al., 2015). These methods, however, neglect the degree of similarity between microbe species, due to the fact that the analysis units in these methods, microbe species, are implicitly assumed to be equally distinct (Fukuyama, 2017). Furthermore, the classification of microbes by contemporary microbial taxonomy is coarse, which results in loss of power to detect subtle difference in a higher resolution (Washburne et al., 2018). To alleviate these issues, the phylogeny of the bacterial species is usually incorporated into the analyses of microbiome data (Fukuyama, 2017; Washburne et al., 2018).

In order to capture the phylogenetic microbial compositional difference between populations, one of the most widely used two-sample testing methods is multivariate analysis of variance with permutation (PERMANOVA) equipped with phylogenetic distance (McArdle & Anderson, 2001; Anderson, 2014; Xia & Sun, 2017). In microbiome data analysis, the popular choices of phylogenetic distances include the unweighted or weighted UniFrac distances (Lozupone & Knight, 2005; Lozupone et al., 2007) and their L^α Zolotarev-type generalized variants (Evans & Matsen, 2012). Through studying these phylogenetic composition distances, we show that they are closely related to a minimum-cost flow problem on the underlying phylogenetic tree and the phylogenetic composition difference between samples can be fully characterized by the optimal flow at each edge. Motivated by this observation, we consider the optimal flow at each edge as the analysis unit, instead of each microbe species. The main goal of the present paper is to study the problem of two-sample testing on a phylogenetic tree from this minimum-cost flow perspective.

We first investigate PERMANOVA equipped with L^2 Zolotarev-type phylogenetic distance. Due to its flexibility and ease of computation, PERMANOVA using phylogenetic distance has been applied in a wide range of microbiome studies (Smith et al., 2015; Wu et al., 2016; Chen et al., 2016), but it still lacks theoretical justification. Following the minimum-cost flow perspective, we show that PERMANOVA is essentially a sum-of-squares type test, which has been widely used to test the difference between the means of two populations in high-dimensional problems (Bai & Saranadasa, 1996; Srivastava & Du, 2008; Chen & Qin, 2010). We establish its asymptotic normality under the null hypothesis and show that its power is indeed determined by the phylogenetic distance between the group means. It is known that sum-of-squares type tests are effective in detecting the dense alternatives, but not powerful against sparse alternatives (Cai et al., 2014; Chen et al., 2019). However, in most microbiome studies, only a small fraction of taxa may have different mean abundances (Cao et al., 2017), resulting in optimal flows on a small number of edges that are active, i.e. non-zero. Moreover, PERMANOVA, as a global method, is not able to identify the specific location of the significant differences even when the null hypothesis is rejected. Therefore, there is a need for a more powerful and interpretable test to detect sparse phylogenetic composition difference between two populations.

To fill this need, we introduce a new test, Detector of Active Flow on the Tree (DAFOT), to detect the sparse phylogenetic composition difference between two populations. To detect sparse signals, the maximum type statistics are usually adopted in various settings because of their simplicity, effectiveness and optimality (see, e.g. Dumbgen & Spokoiny, 2001; Arias-Castro et al., 2005; Jeng et al., 2010; Arias-Castro et al., 2011; Cai et al., 2014, and references therein). Motivated by this, we construct DAFOT as the maximum of the standardized statistics for optimal flow at each edge. When the null hypothesis is rejected by DAFOT, it is also able to identify the edges that the active optimal flows lie on. Thus, different from PERMANOVA, DAFOT can not only detect the difference, it is also able to identify the branches of the phylogenetic tree that show difference in relative abundance between the populations. It is shown that DAFOT is the minimax optimal test against sparse alternatives and the optimal detection boundary of phylogenetic composition difference relies on both the structure of phylogenetic tree and heteroskedastic variance of microbe species. The practical merits of DAFOT are further demonstrated through a real data example. The method is implemented in the R package `DAFOT` available from `CRAN`.

Transformation of compositional data is often employed in order to account for the compositional nature of the data. For example, the centered log-ratio transformation is one of the commonly used transformation methods for the analysis of compositional data (Aitchison, 1982). To account for possible data transformation, we introduce f-generalized optimal flow for any given strictly increasing transformation function f defined on $[0, 1]$. The original form of optimal flow corresponds to the special case with $f(x) = x$. Another special case of f-generalized optimal flow is the difference of balance between populations (Egozcue & Pawlowsky-Glahn, 2016; Rivera-Pinto et al., 2018), when $f(x) = \log(x)$, equivalent to adopting centered log-ratio transformation. After introducing this new concept, we show that all methodology and theory discussed previously can be generalized accordingly.

2. A Hierarchical Model for Microbiome Count Data and Phylogenetic Distance

Human microbiome can be quantified using 16S rRNA sequencing or shotgun metagenomic sequencing. Such 16S rRNA gene sequences of the bacterial genomes or the sequencing of evolutionarily conserved universal marker genes can be used to construct the phylogenetic tree of the bacterial species. The microbe species and their ancestors are usually organized in such a phylogenetic tree based on their evolutionary relationships. Let $T = (V, E)$ be the phylogenetic tree of microbe species. Here, V is the collection of microbe species and their ancestors and E represents the collection of edges of the phylogenetic tree T. For any $e \in E$, L_e is the corresponding branch length. We assume the phylogenetic tree is rooted at ρ, which can be seen as the common ancestor of all microbe species. For any pair of nodes $v_1, v_2 \in V$, the unique shortest path between them is denoted by $[v_1, v_2]$ and the corresponding distance between them is defined as

$$d(v_1, v_2) := \sum_{e \in [v_1, v_2]} L_e. \tag{1}$$

The dissimilarity between two microbe species v_1 and v_2 can thus be quantified by $d(v_1, v_2)$. The height of tree is then defined as the maximum of distances between the root and other nodes of the tree $d(T) = \max_{v \in V} d(\rho, v)$.

The relative abundance of a microbial community can be represented by a discrete distribution on the nodes of tree T. More specifically, write all possible discrete distributions on T as

$$\mathcal{P} = \left\{ \mathrm{P} = \{p_v\}_{v \in V} : \sum_{v \in V} p_v = 1 \quad \text{and} \quad p_v \geq 0 \right\}.$$

Here, p_v is the relative abundance of microbial species v and $\mathcal{P}$ is a simplex of $|V|$ dimension, where $|\cdot|$ is the number of elements. Suppose there are two populations of interest on $\mathcal{P}$, e.g., treated and control groups. These two populations can be represented by two probability distributions on $\mathcal{P}$, $\pi_1(\mathrm{P})$ and $\pi_2(\mathrm{P})$, respectively. We are interested in comparing the mean of relative abundance between these two populations

$$H_0 : \mathrm{P}_{1,\mu} = \mathrm{P}_{2,\mu} \qquad \text{v.s.} \qquad H_1 : \mathrm{P}_{1,\mu} \neq \mathrm{P}_{2,\mu} \tag{2}$$

where $\mathrm{P}_{k,\mu}$ is the mean of $\pi_k(\mathrm{P})$

$$\mathrm{P}_{k,\mu} := \int \mathrm{P} d\pi_k(\mathrm{P}), \qquad k = 1, 2.$$

The covariance matrix of $\pi_k(\mathrm{P})$ is defined similarly

$$\Sigma_{k,\mu} := \int (\mathrm{P} - \mathrm{P}_{k,\mu}) (\mathrm{P} - \mathrm{P}_{k,\mu})^T d\pi_k(\mathrm{P}), \qquad k = 1, 2.$$

To test the mean equality hypothesis, m_1 and m_2 samples are drawn from each of two populations

$$\mathrm{P}_{1,1}, \mathrm{P}_{1,2}, \ldots, \mathrm{P}_{1,m_1} \sim \pi_1(\mathrm{P}) \qquad \text{and} \qquad \mathrm{P}_{2,1}, \mathrm{P}_{2,2}, \ldots, \mathrm{P}_{2,m_2} \sim \pi_2(\mathrm{P}).$$

However, the true relative abundance of each sample, $\mathrm{P}_{k,j}$, $1 \leq j \leq m_k, k = 1, 2$, is unknown in practice. Sequencing microbial DNAs are then applied to each sample to assess the relative abundance of microbe in the sample. In microbiome studies, the sequencing read data can be modeled by a Poisson distribution. To be specific, the number of sequencing reads $N_{k,j,v}$ that can be assigned to species v from jth sample of kth group is assumed to follow a Poisson distribution

$$N_{k,j,v} \sim \mathrm{Pois}(n_{k,j} p_{k,j,v}), \qquad v \in V, \ 1 \leq j \leq m_k \text{ and } k = 1, 2,$$

where, $n_{k,j}$ is the total number of reads in jth sample of kth group and $p_{k,j,v}$ is the relative abundance of microbe species v in sample $\mathrm{P}_{k,j}$. Thus, the reads count is assumed to be drawn from the following hierarchical model

$$\mathrm{P}_{k,j} \sim \pi_k(\mathrm{P}) \qquad \text{and} \qquad N_{k,j,v} \sim \mathrm{Pois}(n_{k,j} p_{k,j,v})$$

for any $v \in V$, $j = 1, \ldots, m_k$ and $k = 1, 2$. The goal of this paper is to test the hypothesis in (2) based on the count data $\mathrm{N}_{k,j} = \{N_{k,j,v}\}_{v \in V}$.

Following this hierarchical model, the empirical distribution of each $\mathrm{P}_{k,j}$ is written as $\hat{\mathrm{P}}_{k,j} = \{\hat{p}_{k,j,v}\}_{v \in V} := \{N_{k,j,v}/n_{k,j}\}_{v \in V}$. Due to the hierarchical structure of model, the covariance matrix of empirical distribution $\hat{\mathrm{P}}_{k,j}$ is

$$\begin{aligned} \Sigma_{k,j} : &= \mathbb{E}\left\{ \left(\hat{\mathrm{P}}_{k,j} - \mathrm{P}_{k,\mu}\right) \left(\hat{\mathrm{P}}_{k,j} - \mathrm{P}_{k,\mu}\right)^T \right\} \\ &= \Sigma_{k,\mu} + \mathrm{diag}(\mathrm{P}_{k,\mu}/n_{k,j}). \end{aligned}$$

Here, diag represents the diagonal matrix of a vector. It is clear that the covariance matrix of the empirical distribution depends on both mean and covariance matrix of $\pi_k(\mathrm{P})$ and the variance of empirical distribution is inflated because of sequencing steps. The difference between $\Sigma_{k,j}$ and $\Sigma_{k,\mu}$ vanishes when $n_{k,j}$ goes to infinity. For simplicity of analysis, in the rest of paper, we always assume the total number of reads in each sample is equal, i.e. $n_{k,j} = n$ for any $1 \leq j \leq m_k$ and $k = 1, 2$. For brevity, we write $\Sigma_k = \Sigma_{k,j}$ when $n_{k,j} = n$ and $m = m_1 + m_2$. In the rest of paper, we assume that there exits $t \in (0, 1)$ such that

$$m_1/(m_1 + m_2) \to t. \tag{3}$$

This assumption implies that proportion of samples from either population does not vanish. We also write $\bar{\Sigma}(t) = (1-t)\Sigma_1 + t\Sigma_2$.

In microbiome studies, a phylogenetic distance that reflects the evolution relationships among microbe species is often used in defining the distance between two microbial communities. Examples include unweighted and weighted UniFrac distance (Lozupone & Knight, 2005; Lozupone et al., 2007). As shown in Evans & Matsen (2012), the weighted UniFrac distance is a plugin estimator of Wasserstein distance of probability masses on the tree, which can be generalized to L^α Zolotarev-type variants (for brevity, we call them L^α Zolotarev-type phylogenetic distance). In the present paper, we focus on L^2 Zolotarev-type phylogenetic distance

$$D(\mathrm{P}_1, \mathrm{P}_2) = \left\{ \sum_{e \in E} L_e (P_{1,e} - P_{2,e})^2 \right\}^{1/2}, \tag{4}$$

where $P_{k,e}$ is the total probability of all descendants of edge e

$$P_{k,e} = \sum_{v \in \tau(e)} p_{k,v},$$

where $k = 1, 2$ and $\tau(e)$ is a subtree below edge e: $\tau(e) = \{v \in V : e \in [\rho, v]\}$. We also use the following notation in this paper

$$\hat{P}_{k,j,e} = \sum_{v \in \tau(e)} \hat{p}_{k,j,v}.$$

3. A Minimum-cost Flow Perspective for Two Sample Testing

The phylogenetic distance between two discrete distributions is closely related to optimal transport theory (Evans & Matsen, 2012). To be specific, the weighted UniFrac/Wasserstein distance between $\mathrm{P}_{1,\mu} = \{p_{1,\mu,v}\}_{v \in V}$ and $\mathrm{P}_{2,\mu} = \{p_{2,\mu,v}\}_{v \in V}$ is equal to the solution of the following optimal transport problem

$$\begin{aligned} &\min_{\{r_{v_1,v_2}\}_{v_1,v_2 \in V}} \sum_{v_1, v_2 \in V} d(v_1, v_2) r_{v_1,v_2} \\ &\text{s.t.} \sum_{v_2} r_{v_1,v_2} = p_{1,\mu,v_1}, \quad \sum_{v_1} r_{v_1,v_2} = p_{2,\mu,v_2} \quad \text{and} \quad r_{v_1,v_2} \geq 0. \end{aligned} \tag{5}$$

In this optimal transport problem, the objective function is the total cost of transport $\{r_{v_1,v_2}\}_{v_1,v_2 \in V}$ and $d(v_1, v_2)$ is the cost per unit transported from v_1 to v_2.

Different from the general optimal transport problem, the distance $d(v_1, v_2)$ in (1) is defined as the geodesic distance along the path $[v_1, v_2]$ on tree T. Therefore, this optimal transport problem can be naturally cast as a minimum-cost flow problem on a network, as illustrated in Figure 1. The tree T can be seen as a special network and there are source with capacity $p_{1,\mu,v}$ and sink with capacity $p_{2,\mu,v}$ at each node $v \in V$. Then, the optimization problem in (5) aims to find a way with the minimum cost of sending an amount of flow from the sources to sinks, through the network T. In particular, for any given transport $\{r_{v_1,v_2}\}_{v_1,v_2 \in V}$, define the flows on each edge as

$$\Delta_e^+ = \sum_{v_1 \in \tau(e), v_2 \notin \tau(e)} r_{v_1,v_2} \qquad \text{and} \qquad \Delta_e^- = \sum_{v_1 \notin \tau(e), v_2 \in \tau(e)} r_{v_1,v_2}.$$

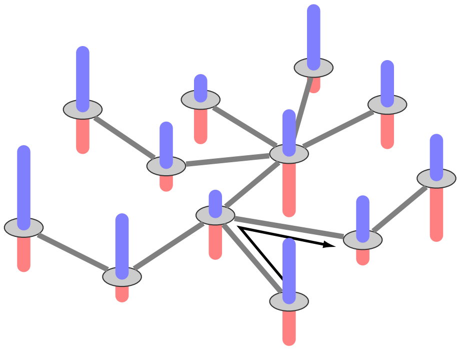


Fig. 1: An illustration of the minimum-cost flow problem on the tree: blue bars are source and red bars are sink.

Here, Δ_e^+ is the flow through edge $e \in E$ towards the root ρ and Δ_e^- is the flow through edge e in opposite direction. Then, the optimization problem in (5) can be reformulated as

$$\begin{aligned} &\min_{\{\Delta_e^+,\Delta_e^-\}_{e\in E}} \sum_{e\in E} L_e\left(\Delta_e^+ + \Delta_e^-\right) \\ &\text{s.t. } \Delta_e^+ + \Delta_e^- \leq 1, \quad \Delta_e^+ - \Delta_e^- = P_{1,\mu,e} - P_{2,\mu,e} \quad \text{and} \quad \Delta_e^+, \Delta_e^- \geq 0. \end{aligned} \tag{6}$$

Although the optimal transport r_{v_1,v_2} in (5) might not be unique, the optimal flow in (6) is unique and has a closed form solution

$$\Delta_e^+ = \max\left(P_{1,\mu,e} - P_{2,\mu,e}, 0\right) \qquad \text{and} \qquad \Delta_e^- = \max\left(P_{2,\mu,e} - P_{1,\mu,e}, 0\right).$$

The optimal net flows on each edge is then defined as $\Delta_e^* = \Delta_e^+ - \Delta_e^- = P_{1,\mu,e} - P_{2,\mu,e}$. The weighted UniFrac distance and corresponding L^α Zolotarev-type variants can thus be seen as the weighted L^α norm of optimal flows. It is clear from the above discussion that the phylogenetic composition difference between P_1 and P_2 can be fully characterized by the optimal flow Δ_e^*. If we write $\Delta^* = \{\Delta_e^*\}_{e\in E}$, the hypothesis in (2) can be rewritten in the following equivalent form

$$H_0 : \Delta^* = 0 \qquad \text{and} \qquad H_1 : \Delta^* \neq 0. \tag{7}$$

This suggests that the basic unit to quantify phylogenetic composition difference shall be the optimal flow at each edge Δ_e^*. L^2 Zolotarev-type phylogenetic distance $D(\mathrm{P}_1, \mathrm{P}_2)$ can be seen as L_2 norm of these optimal flows.

4. Permutational Multivariate Analysis of Variance

4.1. *Introduction of* PERMANOVA

To incorporate phylogenetic tree information in comparing two populations, one of the most commonly used two-sample tests is PERMANOVA equipped with some phylogenetic distance (McArdle & Anderson, 2001; Anderson, 2014). Specifically, let $D(\mathrm{P}_1, \mathrm{P}_2)$ be the phylogenetic distance defined in (4). The average empirical distance within and between group is defined as

$$\bar{D}_{k,w} = \frac{2}{m_k(m_k-1)} \sum_{1\leq j_1<j_2\leq m_k} D^2\left(\hat{\mathrm{P}}_{k,j_1}, \hat{\mathrm{P}}_{k,j_2}\right), \qquad k = 1, 2$$

and

$$\bar{D}_b = \frac{1}{m_1 m_2} \sum_{1 \le j_1 \le m_1, 1 \le j_2 \le m_2} D^2\left(\hat{\mathrm{P}}_{1,j_1}, \hat{\mathrm{P}}_{2,j_2}\right).$$

Similar to analysis of variance (ANOVA), PERMANOVA defines the total sum-of-squares as

$$SS_T = \sum_{k=1}^{2} \frac{m_k(m_k-1)}{2(m_1+m_2)} \bar{D}_{k,w} + \frac{m_1 m_2}{m_1+m_2} \bar{D}_b,$$

within-group sum-of-squares and between-group sum-of-squares as

$$SS_W = \sum_{k=1}^{2} \frac{m_k - 1}{2} \bar{D}_{k,w} \qquad \text{and} \qquad SS_A = SS_T - SS_W.$$

The pseudo F-statistic for two-sample testing is then defined as the normalized ratio of SS_A to SS_W

$$F = \frac{SS_A}{SS_W/(m-2)}.$$

To evaluate the significance of the F-statistic, the P-value is calculated by permutations. To be more specific, the m samples are permuted randomly B times and F-statistic is calculated on these permuted data, denoted by $F^{(1)}, \ldots, F^{(B)}$. Then, the estimated P-value is

$$P = \frac{1 + |\{i : F^{(i)} > F^{(B)}\}|}{1+B}.$$

One implicitly prerequisite assumption of a valid permutation test is the exchangeability of samples under the null hypothesis. Thus, the hypothesis required by valid permutation test is

$$H_0 : \pi_1(\mathrm{P}) = \pi_2(\mathrm{P}) \qquad \text{v.s.} \qquad H_1 : \pi_1(\mathrm{P}) \neq \pi_2(\mathrm{P}).$$

Compared with mean equality hypothesis in (2), this is a more restrictive hypothesis. In the next section, we present another way to estimate P-value of PERMANOVA based on the asymptotic results.

4.2. *Properties of* PERMANOVA

We investigate the properties of pseudo F-statistic under L^2 Zolotarev-type phylogenetic distance $D(\mathrm{P}_1, \mathrm{P}_2)$. A simple calculation decomposes SS_A into two parts

$$SS_A = -\frac{m_1 m_2}{2(m_1+m_2)} \left(\bar{D}_{1,w} + \bar{D}_{2,w} - 2\bar{D}_b\right) + \frac{m_2 \bar{D}_{1,w} + m_1 \bar{D}_{2,w}}{2(m_1+m_2)}.$$

In particular, the second term is asymptotically equal to $SS_W/(m_1+m_2-2)$

$$\left(\frac{m_2 \bar{D}_{1,w} + m_1 \bar{D}_{2,w}}{2(m_1+m_2)}\right) \Big/ \left(\frac{SS_W}{m_1+m_2-2}\right) \to 1, \qquad m_1, m_2 \to \infty.$$

This shows that F is a scaled version of difference between average within group distance and across group distance. In other words, in terms of two–sample testing, F is asymptotically equivalent to the following energy distance statistic (Székely & Rizzo, 2005; Sejdinovic et al., 2013)

$$L := \frac{1}{2}\left(2\bar{D}_b - \bar{D}_{1,w} - \bar{D}_{2,w}\right).$$

Due to the fact that the distance in (4) is a negative type, the energy distance statistic can also be written as a kernel-based test statistic (Sejdinovic et al., 2013)

$$L=\sum_{k=1}^{2}\frac{1}{m_k(m_k-1)}\sum_{j_1\neq j_2=1}^{m_k}K\left(\hat{\mathrm{P}}_{k,j_1},\hat{\mathrm{P}}_{k,j_2}\right)-\frac{2}{m_1m_2}\sum_{j_1,j_2=1}^{m_1,m_2}K\left(\hat{\mathrm{P}}_{1,j_1},\hat{\mathrm{P}}_{2,j_2}\right),$$

where the kernel of P_1 and P_2 is defined as $K\left(\mathrm{P}_1,\mathrm{P}_2\right):=\sum_{e\in E}L_eP_{1,e}P_{2,e}$. The kernel form of PERMANOVA suggests

$$\mathbb{E}(L)=D^2(\mathrm{P}_{1,\mu},\mathrm{P}_{2,\mu}).$$

Therefore, F statistic in PERMANOVA is a reasonable statistic for testing the phylogenetic composition difference in hypothesis (2), as the mean of L only depends on $\mathrm{P}_{1,\mu}$ and $\mathrm{P}_{2,\mu}$.

Clearly, the behavior of L depends on both the covariance matrix of $\hat{\mathrm{P}}_{k,j}$ and tree structure T. The structure of tree T can be expressed as a transformation matrix $H\in\mathbb{R}^{|E|\times|V|}$, where $H_{ev}=\sqrt{L_e}$ if $v\in\tau(e)$ and $H_{ev}=0$ if $v\notin\tau(e)$. We assume

$$\mathrm{tr}\{H\Sigma_{i_1}H^TH\Sigma_{i_2}H^TH\Sigma_{i_3}H^TH\Sigma_{i_4}H^T\}=o\left(\mathrm{tr}^2\{(H\bar{\Sigma}(t)H^T)^2\}\right), \tag{8}$$

where $i_1,i_2,i_3,i_4=1$ or 2. Such a moment assumption is a common condition in high dimension statistics, e.g., condition (3.6) in Chen & Qin (2010). Condition (8) is true when eigenvalues of both $H\Sigma_1H^T$ and $H\Sigma_2H^T$ are bounded. Besides the assumption on the moment, another assumption we make is

$$\mathbb{P}\left(K(\hat{\mathrm{P}}_1,\hat{\mathrm{P}}_2)\geq\sqrt{r_m}\right)\leq r_m\,\mathrm{tr}\{(H\bar{\Sigma}(t)H^T)^2\}/d(T)^4 \tag{9}$$

where $\hat{\mathrm{P}}_1$ and $\hat{\mathrm{P}}_2$ are empirical distribution drawn from the 1st or 2nd population and r_m is a sequence of number such that $r_m=o(m\,\mathrm{tr}\{(H\bar{\Sigma}(t)H^T)^2\})$. This is a fairly weak condition. For example, this is a trivial assumption when the tree is not too high, i.e. $d(T)=o(m\,\mathrm{tr}\{(H\bar{\Sigma}(t)H^T)^2\})^{1/2}$, because $K(\hat{\mathrm{P}}_1,\hat{\mathrm{P}}_2)=O(d(T))$.

The following theorem shows the asymptomatic behavior of PERMANOVA statistic L.

THEOREM 1. *Under the null hypothesis, i.e.* $\mathrm{P}_{1,\mu}=\mathrm{P}_{2,\mu}$*, and assumptions* (3)*,* (8) *and* (9)*,*

$$L/\sigma_L\to N(0,1),\qquad m\to\infty,$$

where σ_L can be written as

$$\sigma_L^2=\frac{\mathrm{tr}\{(H\bar{\Sigma}(t)H^T)^2\}}{2m^2t^2(1-t)^2}.$$

Furthermore, if $d(T)=o(m)$ and assumptions (3)*,* (8) *and* (9) *hold, the test is consistent when*

$$D^2(\mathrm{P}_{1,\mu},\mathrm{P}_{2,\mu})\gg\frac{\sqrt{\mathrm{tr}\{(H\bar{\Sigma}(t)H^T)^2\}}}{m}. \tag{10}$$

This theorem suggests that the PERMANOVA is a consistent test if the phylogenetic distance between the means of two populations is large enough. As $\bar{\Sigma}(t)$ can be written explicitly as

$$\bar{\Sigma}(t)=(1-t)\Sigma_{1,\mu}+t\Sigma_{2,\mu}+\frac{1}{n}\mathrm{diag}\left((1-t)\mathrm{P}_{1,\mu}+t\mathrm{P}_{2,\mu}\right),$$

the power of PERMANOVA depends on both the number of samples m and the number of reads per sample n. The power becomes larger if we increase either m or n. However, (10) also suggests that a larger number of samples is a more efficient way to increase power than a larger number of reads per sample. This theorem also suggests that P-value can be calculated based on asymptomatic distribution instead of conducting permutations. For instance, σ_L^2 can be estimated based on a similar U-statistic $\hat{\sigma}_L^2$ in Chen & Qin (2010). Then, the P-value can be then calculated by $1 - \Phi(L/\hat{\sigma}_L)$. It is worth noting that this way to calculate P-value does not require $\pi_1(\mathrm{P}) = \pi_2(\mathrm{P})$ under the null hypothesis. In practice, we recommend permutation test when m is not large and asymptotic critical value if m is large.

4.3. PERMANOVA *under sparse setting*

As we see in previous sections, PERMANOVA is a sum-of-squares type statistic. However, the interesting setting in practice, e.g., in microbiome studies, is a sparse case where only a small number of microbe species may have different relative mean abundance (Cao et al., 2017). This suggests that only a small fraction of optimal flow Δ_e^* is active, i.e. $\Delta_e^* \neq 0$. To investigate the performance of PERMANOVA under this sparse setting, we consider a simple case where there is active optimal flow on one edge, denoted by $e_s \in E$. As suggested by Theorem 1, the condition for a consistent PERMANOVA test is

$$L_{e_s}|\Delta_{e_s}^*|^2 \gg \frac{\sqrt{\mathrm{tr}\{(H\bar{\Sigma}(t)H^T)^2\}}}{m}.$$

On the other hand, we consider an oracle test that has knowledge of active flow location e_s. Since the location of e_s is known, we consider a two sample t-test for $\Delta_{e_s}^*$

$$M_{e_s} = \frac{\bar{P}_{1,e_s} - \bar{P}_{2,e_s}}{\sqrt{\hat{\sigma}_{1,e_s}^2/m_1 + \hat{\sigma}_{2,e_s}^2/m_2}}, \tag{11}$$

where

$$\bar{P}_{k,e_s} = \frac{1}{m_k}\sum_{j=1}^{m_k}\hat{P}_{k,j,e_s} \qquad \text{and} \qquad \hat{\sigma}_{k,e_s}^2 = \frac{1}{m_k - 1}\sum_{j=1}^{m_k}(\hat{P}_{k,j,e_s} - \bar{P}_{k,e_s})^2.$$

With central limit theorem, we know that M_{e_s} is a consistent test if

$$|\Delta_{e_s}^*|^2 \gg \frac{(H\bar{\Sigma}(t)H^T)_{e_s,e_s}}{m}.$$

A comparison of two detection boundaries indicates that the oracle test is able to detect a much smaller difference between two group of samples than PERMANOVA. This naturally leads to the question of whether it is possible to develop a more powerful test under sparse flow setting.

5. ACTIVE OPTIMAL FLOW DETECTION

5.1. *Detector of Active Flow on the Tree*

As shown in the last section, the two sample t-test could improve the power to detect the difference between two populations when location of active optimal flow is known. In practice, the location information is usually unknown. To address this issue, we consider the maximum of two sample t-test at each edge

$$M^* = \max_{e \in E}|M_e|,$$

where M_e is defined in the same way as (11). The use of this maximum type statistic for detecting sparse signals is very common in a wide range of applications and it leads to construction of rate-optimal test in many problems (Dumbgen & Spokoiny, 2001; Arias-Castro et al., 2005; Jeng et al., 2010; Arias-Castro et al., 2011; Cai et al., 2014; Cao et al., 2017; Wang et al., 2019).

To evaluate the statistical significance of M^*, one still needs to choose an appropriate critical value for M^*. However, it is difficult to derive the asymptotic distribution of M^* due to the complex dependency structure among M_e. To overcome this problem, one adopts resampling method to assign appropriate critical value for M^*. In particular, a common resampling method to choose critical value is permutation test as in PERMANOVA (Good, 2013; Anderson, 2014). Although the permutation test requires $\pi_1(\mathrm{P}) = \pi_2(\mathrm{P})$ under the null hypothesis as we discussed before, its performance is robust when sample size is small.

We propose another resampling method, bootstrap, to choose critical value for M^* in order to avoid the condition $\pi_1(\mathrm{P}) = \pi_2(\mathrm{P})$. Let $\tilde{P}_{k,j,e} = \hat{P}_{k,j,e} - \bar{P}_{k,e} + \bar{P}_e$ for $1 \leq j \leq m_k$ and $k = 1, 2$, where $\bar{P}_{k,e}$ is mean of $\hat{P}_{k,j,e}$ within each group and $\bar{P}_e$ is mean of the combined samples. We randomly draw m_k samples with replacement from each group and then calculate M^* with shifted data $\tilde{P}_{k,j,e}$. This procedure is repeated B times and the corresponding statistics are denoted by $M^*_{(1)}, \ldots, M^*_{(B)}$. Finally, the approximated q_α is chosen as $(1-\alpha)$ quantile of empirical distribution of $M^*_{(1)}, \ldots, M^*_{(B)}$ or the P-value is calculated as

$$P = \frac{1 + |\{i : M^*_{(i)} > M^*\}|}{1 + B}.$$

We then make decision under the null hypothesis based critical value or P-value.

When the null hypothesis is rejected, M^* also provides a natural way to identify the set of edges that optimal flow on is not zero $\{e : \Delta^*_e \neq 0\}$. To be specific, we consider the following set of edges $\{e : |M_e| > q_\alpha\}$ as the active edges identified. Due to the construction of M^*, the family-wise error rate (FWER) of active edges identification is naturally controlled at α level. The identified edge indicates that all microbe species below this edge, as a whole object, are differentially abundant between the two populations. In this way, the active edges can be identified as a microbial signature associated with the difference of the two populations.

5.2. *Asymptotic Behavior and Optimality of M^**

In this section, we investigate the behavior of M^* under the null and alternative hypothesis. The main difficulty to study the behavior of M^* is the strong and complex dependence among different M_e. This complexity of dependency structure mainly comes from two sources: the high overlapping structure of subtree $\tau(e)$ and the unknown heteroskedastic variance/covariance structure among different microbe species. Our investigation shows that the complexity of dependence structure among M_e can be characterized by a single quantity that depends on both the tree structure and the heteroskedastic variance.

Since each M_e is defined on a subtree below edge e, the asymptotic behavior of M^* clearly depends on "effective" number of subtrees. We still assume each group of samples has no vanishing proportion, i.e.(3). Let $\sigma_v^2(t)$ be the element of vth row and vth column of $\bar{\Sigma}(t)$. To decouple the complex dependence structure among M_e, we appeal to the following proposition.

PROPOSITION 1. *For any given $a > 0$, let $E(a)$ be a subset of edges of the tree such that $E(a) := \{e : a < \sum_{v \in \tau(e)} \sigma_v^2(t) \leq 2a\}$. Then, $E(a)$ can be decomposed into a collection of disjoint paths*

$$E(a) = \bigcup_{r=1}^{R} \left[v_r, v'_r\right],$$

where v_r and v_r' are nodes of the tree and R is the number of disjoint paths. In addition, any two edges from different paths in the above decomposition do not share any descendants.

Intuitively, for any e_1 and e_2 from the same path defined above, the subtrees $\tau(e_1)$ and $\tau(e_2)$ are highly overlapped and thus the M_e defined on them are expected to behave similarly. On the other hand, for the edges from different paths, the corresponding subtrees are "distinct" in that they do not share any descendants. In general, the number of disjoint paths in above proposition characterizes "effective" number of subtrees. Motivated by this observation, we define S_j as the number of different path in $E(2^{-j})$ for $j = 0, 1, \ldots, \infty$ and $S(T, \pi_1, \pi_2)$ as the sum of S_j

$$S(T, \pi_1, \pi_2) := \sum_{j=0}^{\infty} S_j.$$

Clearly, $S(T, \pi_1, \pi_2)$ is an integer between 1 and $|E|$ and depends on the structure of tree T and the distribution π_1 and π_2. For example, when the tree T is a fully balanced binary tree and π_1 and π_2 are Dirichlet distribution with equal parameters for the leaves, $S(T, \pi_1, \pi_2) \asymp |E|$. If π_1 and π_2 are Dirichlet distribution with equal parameters for all nodes of a chain tree T, i.e. one dimensional lattice, then $S(T, \pi_1, \pi_2) \asymp \log |E|$. More specific examples can be found in Section S3. We later show that the asymptotic behavior of M^* can be fully characterized by this quantity $S(T, \pi_1, \pi_2)$.

When two subtrees highly overlap, we expect the correlation between M_e on them to be very strong. We shall assume there exit constants $c > 0$ and $\alpha > 0$ such that

$$\mathrm{Corr}\left((1-t)\hat{P}_{1,j_1,e} - t\hat{P}_{2,j_2,e}, (1-t)\hat{P}_{1,j_1,e'} - t\hat{P}_{2,j_2,e'}\right) \geq 1 - c\left(1 - \sqrt{\frac{\sum_{v\in\tau(e)} \sigma_v^2(t)}{\sum_{v\in\tau(e')} \sigma_v^2(t)}}\right)^{\alpha}, \tag{12}$$

where $1 \leq j_1 \leq m_1$, $1 \leq j_2 \leq m_2$ and e, e' are a pair of edges such that $\tau(e) \subset \tau(e')$. Here, Corr represents the correlation between two random variables. This is not a strong condition. For instance, (12) is satisfied when the relative abundance of different microbe species p_v in $\tau(e')$ are mutually independent or drawn from a Dirichlet distribution. Furthermore, we also assume $\hat{P}_{k,j,e}$ is sub-Gaussian distribution, i.e there exist constants η and K such that

$$\mathbb{E}\left(\exp(\eta(\hat{P}_{k,j,e} - P_{k,\mu,e})^2/\sigma_{k,e}^2)\right) \leq K, \tag{13}$$

where $\sigma_{k,e}^2$ is variance of $\hat{P}_{k,j,e}$, the elements of eth row and eth column of $H\Sigma_k H^T$. We now show the asymptotic behavior of M^* under the null hypothesis.

THEOREM 2. *Suppose* $\log^5 |E| = o(m)$. *Under the null hypothesis and assumptions* (3), (12) *and* (13), *we have*

$$M^* \leq \sqrt{2\log S(T, \pi_1, \pi_2)} + O_p(\sqrt{\log\log S(T, \pi_1, \pi_2)}), \qquad \text{as } m \to \infty. \tag{14}$$

Theorem 2 suggests that the amplitude of M^* is $\sqrt{2\log S(T, \pi_1, \pi_2)}$ when null hypothesis is true, where $S(T, \pi_1, \pi_2)$ here plays the same role as the number of variables when each component are almost independent (Cai et al., 2014; Cao et al., 2017). In other words, although M^* is constructed from $|E|$ different subtrees, it is equivalent to take maximum of roughly $S(T, \pi_1, \pi_2)$ independent variables because of high dependency among the statistics M_e. For the one active flow example discussed in previous section, Theorem 2 suggests that M^* is a consistent test when

$$|\Delta^*_{e_s}|^2 \gg \log S(T, \pi_1, \pi_2) \frac{(H\bar{\Sigma}(t)H^T)_{e_s,e_s}}{m}.$$

A comparison between above and oracle test's detection boundary suggests that the price we pay for unknown location of e_s is $\log S(T, \pi_1, \pi_2)$. With the same argument as for PERMANOVA, we know that either a larger number of samples m or a larger number of reads per sample n can increase the power of DAFOT. More detailed discussions are given in Section 6.

We next turn to the power analysis of the test based on M^* from a minimax point view (Ingster, 1993; Ingster & Suslina, 2012). The parameter space of the null hypothesis is

$$\mathcal{H}_{0,S} = \{(T, \pi_1, \pi_2) : \Delta^*_e = 0,\ \forall e \in E, S(T, \pi_1, \pi_2) \leq S\}$$

and the parameter space of the alternative hypothesis is

$$\mathcal{H}_{1,S}(h) = \left\{ (T, \pi_1, \pi_2) : \sup_e \frac{|\Delta^*_e|}{\sqrt{(1-t)\sigma^2_{1,e} + t\sigma^2_{2,e}}} \geq h, S(T, \pi_1, \pi_2) \leq S \right\}.$$

The worst-case risk of any given test Λ is then defined as

$$R_S(\Lambda; h) := \sup_{(T,\pi_1,\pi_2)\in\mathcal{H}_{0,S}} \mathbb{P}(\Lambda = 1|T, \pi_1, \pi_2) + \sup_{(T,\pi_1,\pi_2)\in\mathcal{H}_{1,S}(h)} \mathbb{P}(\Lambda = 0|T, \pi_1, \pi_2)$$

We say a test Λ is consistent for separating $\mathcal{H}_{0,S}$ and $\mathcal{H}_{1,S}(h)$ if $R_S(\Lambda; h) \to 0$ and $\mathcal{H}_{0,S}$ and $\mathcal{H}_{1,S}(h)$ are separable if there exists a consistent test Λ for them. On the other hand, a test Λ is powerless for separating $\mathcal{H}_{0,S}$ and $\mathcal{H}_{1,S}(h)$ if $R(\Lambda; h) \to 1$ and $\mathcal{H}_{0,S}$ and $\mathcal{H}_{1,S}(h)$ are inseparable if $\inf_\Lambda R_S(\Lambda; h) \to 1$. Here h is a parameter to control the distance between $\mathcal{H}_{0,S}$ and $\mathcal{H}_{1,S}(h)$. Clearly, if h is smaller, $\mathcal{H}_{0,S}$ and $\mathcal{H}_{1,S}(h)$ are more difficult to separate. Λ_M is defined as test that rejects the null hypothesis if and only if $M^* > q_\alpha$ for some $\alpha \to 0$. The following theorem characterizes the power of Λ_M and separability of $\mathcal{H}_{0,S}$ and $\mathcal{H}_{1,S}(h)$.

THEOREM 3. *Consider testing $\mathcal{H}_{0,S}$ and $\mathcal{H}_{1,S}(h)$ by Λ_M. Suppose* $\log^5 |E| = o(m)$*,* (3)*,* (12) *and* (13) *hold. Then there exist constants C and c such that*

$$R_S\left(\Lambda_M; C\sqrt{\frac{\log S}{m}}\right) \to 0 \qquad \text{and} \qquad \inf_\Lambda R_S\left(\Lambda; c\sqrt{\frac{\log S}{m}}\right) \to 1.$$

This theorem shows that the optimal rate of detection boundary between $\mathcal{H}_{0,S}$ and $\mathcal{H}_{1,S}(h)$ is

$$\sqrt{\frac{\log S}{m}}.$$

This optimal rate suggests that the difficulty of this problem is mainly determined by the single quantity S, which relies on both the tree structure and heteroskedastic variance structure. This theorem also suggests that M^* is minimax rate optimal.

6. NUMERICAL EXPERIMENTS

6.1. *Simulation Studies*

We first investigate the properties of PERMANOVA and DAFOT on simulated data sets. We choose a phylogenetic tree of bacteria species within the class *Gammaproteobacteria* as the underlying tree T, which is extracted from reference tree of Greengenes 16S rRNA database version 13.8 clustered at 85% similarity by the R package *metagenomeFeatures* (DeSantis et al., 2006). This tree T has a total of 247 leaves, denoted by V_L, and 246 internal nodes, denoted by V_I. Figure 2 shows the structure of the tree with each leaf labeled with a number.

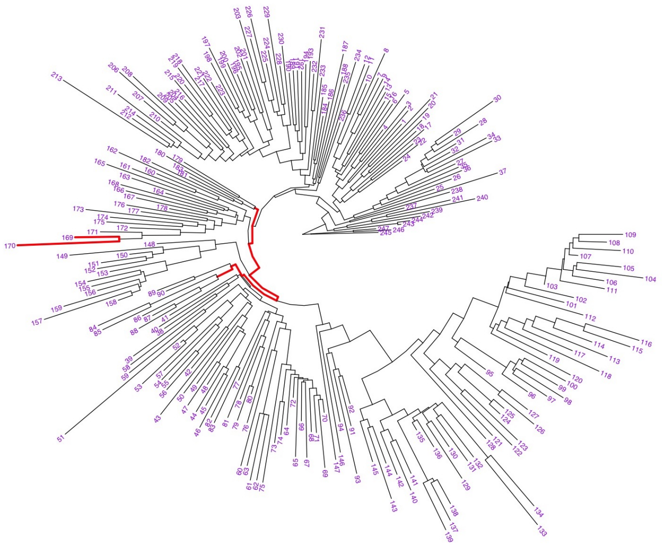


Fig. 2: Phylogenetic tree of bacteria within the class *Gamma proteobacteria* used in the simulation studies with each leaf node labeled with a number. The edges with active optimal flow in the numerical examples are colored in red.

To simulate the numbers of the reads on this tree T, we adopt the Dirichlet-multinomial distribution. More specifically, the true relative abundance $\mathrm{P}_{k,j}$, where $k = 1, 2$ and $1 \leq j \leq m_k$, is drawn from some Dirichlet distribution, i.e. $\pi_k(\mathrm{P})$ follows a Dirichlet distribution indexed by $\boldsymbol{\alpha}_k = \{\alpha_{k,v}\}_{v \in V}$

$$\pi_k(\mathrm{P}) = \frac{\Gamma(\sum_{v \in V} \alpha_{k,v})}{\prod_{v \in V} \Gamma(\alpha_{k,v})} \prod_{v \in V} p_v^{\alpha_{k,v}-1}.$$

For each sample, the reads are then drawn from a multinomial distribution with respect to the true relative abundance. Under this model, the mean of relative abundance for kth group can be written as

$$p_{k,\mu,v} = \frac{\alpha_{k,v}}{\sum_{v \in V} \alpha_{k,v}}.$$

Thus, under the null hypothesis, we assume $\alpha_{1,v} = \alpha_{2,v} = 1$ if $v \in V_L$ and $\alpha_{1,v} = \alpha_{2,v} = 0$ if $v \in V_I$. Under alternative hypothesis, we perturb $\boldsymbol{\alpha}_2$ and consider two different scenarios: A) the difference is at node 169 and 170, i.e. $\alpha_{2,169} = 1 + \delta$ and $\alpha_{2,170} = 1 - \delta$; B) the difference is at clades $V_{c1} = (84, 85, 86, 87, 88)$ and $V_{c2} = (179, 180, 181, 182, 183)$, i.e. $\alpha_{2,v} = 1 + 0.4 * \delta$ if $v \in V_{c1}$ and $\alpha_{2,v} = 1 - 0.4 * \delta$ if $v \in V_{c2}$. The parameter δ is specified later. In particular, the edges with active optimal flow under the scenarios A) and B) are colored in red in Figure 2.

The first set of simulation experiments is designed to compare the performance of DAFOT and PERMANOVA under scenario A). To be specific, we compare 4 different methods: DAFOT , DAFOT after center log-ratio transformation(DAFOT-log), PERMANOVA equipped with weighted

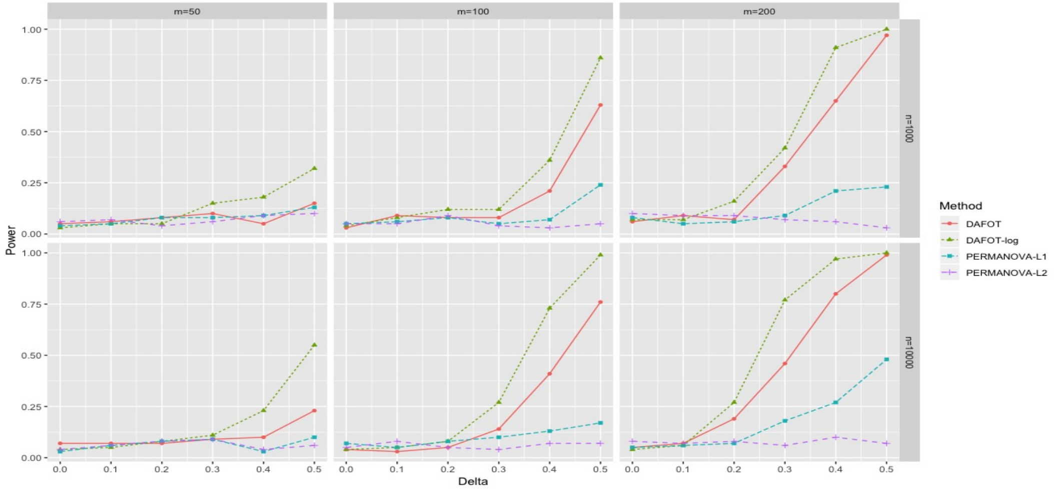


Fig. 3: Power comparisons between DAFOT and PERMANOVA under scenario (A), where the difference is at nodes 169 and 170. DAFOT: proposed method based on proportions; DAFOT-LOG: proposed method based on log-proportions; PERMANOVA-L1: PERMANOVA with L_1 Zolotarev-type phylogenetic distance; PERMANOVA-L2: PERMANOVA with L_2 Zolotarev-type phylogenetic distance.

UniFrac distance (PERMANOVA-L1) and PERMANOVA equipped with L_2 Zolotarev-type phylogenetic distance (PERMANOVA-L2). To make the comparisons fair, the critical values for all tests are chosen by permutations at significance level 5%. To investigate the effect of the sample size $m_1 = m_2 = m$, the sequence depth n and the signal strength δ, we chose $m = 50, 100$ and 200, $n = 1000$ and 10000 and δ=0,0·1,0·2,0·3,0·4 and 0·5 in the simulation experiments. The experiment is repeated 100 times for each combination of the m, n and δ. The performance of the tests is evaluated by the power of test, i.e. the probability of rejecting the null hypothesis, which can be estimated by the proportion of the null hypothesis rejections among the 100 simulation experiments.

The results are summarized in Figure 3. These results show that the type I error is under control when the null hypothesis is true ($\delta = 0$) and the power of DAFOT is larger than PERMANOVA when alternative hypothesis is true ($\delta \neq 0$). Figure 3 implies that the observed effects of m, n and δ on the power of the tests are consistent with the theoretical results. We observe similar improved power DAFOT over PERMANOVA for scenario B) when the active flows connect two clades (See Figure S1 in Supplementary Material for details).

The sequence count data in real microbiome studies are usually very sparse, i.e. there are a lot of zero values. The next set of simulation experiments is designed to assess the performance of DAFOT and PERMANOVA when there are a lot of zero values. More concretely, we set $\alpha_{1,v} = \alpha_{2,v} = 0$ for $1 \leq v \leq 160$ under scenarios A) and for $1 \leq v \leq 80$ under scenarios B). In other words, probability on nearly $2/3$ nodes are zeros in scenarios A) and probability on nearly $1/3$ nodes are zeros in scenarios B). In order to avoid undefined value of $\log 0$, zero counts are replaced by 0.5 in DAFOT-log (Aitchison, 2003; Lin et al., 2014). The sequence depth of each sample is drawn uniformly between 1000 and 10000 instead of being fixed as in previous experiments. The sample size m and the difference between populations δ are varied in the same way as in previous simulation experiments. Figure S2 in Supplementary Material summarizes the results based on 100 runs for each combination of m and δ. A comparison between Figure 3, S1 and Figure S2 suggests that DAFOT, PERMANOVA-L1 and PERMANOVA-L2 are relatively robust against a lot of zero values, however, the power of DAFOT-log is affected by these zero values.

We further compare the performance of DAFOT and PERMANOVA under a wide range of sparseness. Specifically, we adopt a similar setting as in scenario A) and choose $m = 100$, $n = 10000$ and $\delta = 0.3$. To assess the effect of sparsity, we randomly choose $2k$ leaves at each simulation experiment and set $\alpha_{2,v} = 1 + \delta$ for the first k leaves and $\alpha_{2,v} = 1 - \delta$ for the last k leaves. k is chosen equal to $1, 5, 10, 15, 20, 25$ and 30. The results based on 100 times simulation experiments are summarized in Figure 4. These results show that the DAFOT outperforms PERMANOVA when fewer leaves are perturbed. When the signal becomes denser, the PERMANOVA-L1 can gain more power than DAFOT.

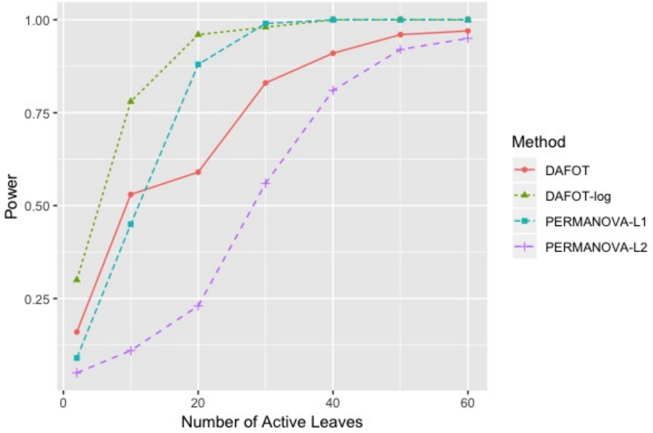


Fig. 4: The power comparisons between DAFOT and PERMANOVA for different sparsity levels.

Table 1: The performance of edge identification by DAFOT and DAFOT-log. AFP is the average number of false positive edges, FWER is the probability of making at least one type I error and ATP is number of true positive edges.

		$m = 50$		$m = 100$		$m = 200$	
		$n = 10^3$	$n = 10^4$	$n = 10^3$	$n = 10^4$	$n = 10^3$	$n = 10^4$
	AFP	0·09	0·06	0·08	0·13	0·10	0·08
DAFOT	FWER	0·04	0·06	0·06	0·09	0·10	0·06
	ATP	0·11	0·15	0·56	0·78	1·49	1·66
	AFP	0·08	0·05	0·08	0·13	0·13	0·36
DAFOT-log	FWER	0·04	0·05	0·07	0·11	0·05	0·21
	ATP	0·35	0·50	0·90	1·35	1·81	1·94

The final set of simulation experiments aims to evaluate the performance of edge identification by DAFOT and DAFOT-log. In particular, we consider scenario A) with $\delta = 0 \cdot 5$ and vary m and n as in previous simulation experiments. For each combination of m and n, we repeat the experiment 100 times and the active edges detected by two methods are recorded. The results of average number of false positive edges, the probability of making at least one type I error and true positive edges are summarized in Table 1, showing that the FWER is under control regardless of signal strength and the two active edges can be identified successfully when signal is strong enough. In Figure 3 and Table 1, DAFOT-log performs better than DAFOT, as the log transformation is suitable for non-zero data.

6.2. *Analysis of Ulcerative Colitis Disease Microbiome Data*

To further demonstrate the performance of DAFOT, we apply the method to a 16S rRNA data set of 47 human intestinal biopsy samples collected at the University of Pennsylvania. These samples are divided into 3 groups:

A) 18 control samples (control),
B) 14 samples with ulcerative colitis who did not receive treatment (unexposed),
C) 15 samples with ulcerative colitis who received treatment (exposed).

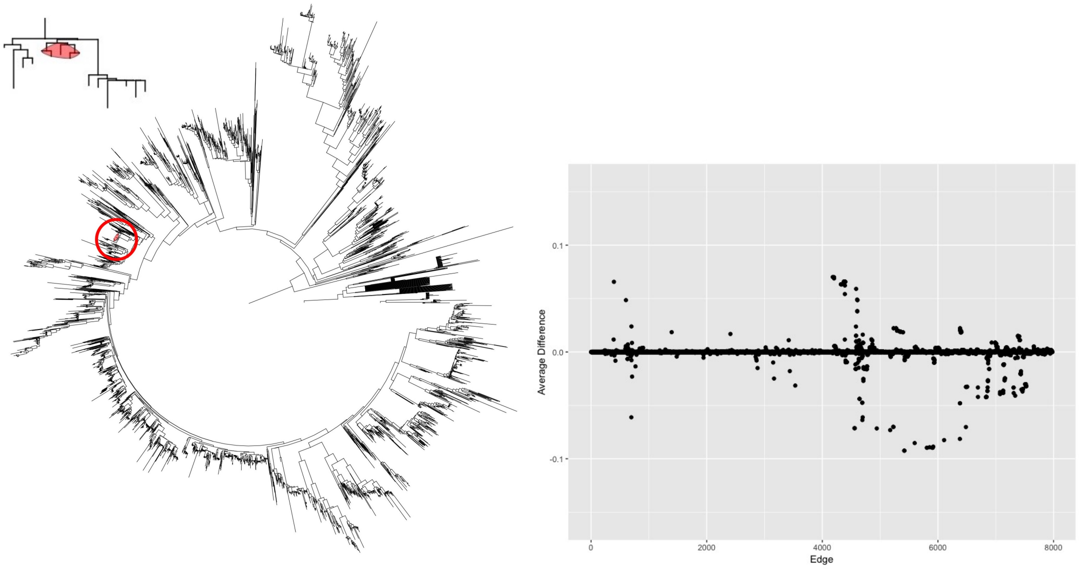
To compare microbiome communities of these groups, the raw sequence reads of each sample are placed into a reference phylogenic tree from Greengenes 16S rRNA database version 13.8 with a 99% similarity by using SEPP (Mirarab et al., 2012; Janssen et al., 2018). The reference phylogenetic tree is then trimmed by keeping all nodes related to the operational taxonomic units observed in the samples. The trimmed phylogenetic tree is shown in Figure 5a, including 7980 edges. Figure 5b shows the empirical optimal flow between group A and B on each edge, i.e. difference of probability on subtree indexed by edges $\bar{P}_{1,e} - \bar{P}_{2,e}$. The order of edge is ranked automatically by *R* class *'phylo'*. It is clear from Figure 5b that most of empirical active flows are small, indicating sparse flow on the tree is a reasonable assumption. In addition, we estimate the "effective" number of subtrees $S(T, \pi_1, \pi_2)$ for pair-wise comparison: $S(T, \pi_1, \pi_2) = 2687$ for group A and B; $S(T, \pi_1, \pi_2) = 3172$ for group A and C; $S(T, \pi_1, \pi_2) = 2676$ for group B and C. The estimation of $S(T, \pi_1, \pi_2)$ is based on the reference phylogenetic tree structure and estimated variance $\hat{\sigma}_v^2(t)$ at each node $v \in V$.

Table 2: The P-values of comparing different groups using DAFOT and PERMANOVA based on 1000 permutations.

	DAFOT	DAFOT-log	PERMANOVA-L1	PERMANOVA-L2
Group A vs B	0·007	0·256	0·378	0·460
Group A vs C	0·147	0·495	0·305	0·270
Group B vs C	0·648	0·832	0·639	0·677

To test the phylogenetic composition difference between the groups, we apply the same four two-sample testing methods as in the first simulation experiment. The resulting P-values estimated by permutation test with 1000 permutations are summarized in Table 2. No methods identify any significant phylogenetic composition difference between group C and A or group C and B (P-value>0·05). However, only DAFOT indicates an overall difference in intestinal biopsy microbiome composition between group A and B with a P-value=0·007, while other methods do not detect such a difference.

Besides the overall difference in microbiome compositions between groups A and B, DAFOT also identifies that the overall difference is due to the active flow on one edge. The subtree indexed by this edge is shown in Figure 5a, colored by red in the original phylogenetic tree and zoomed in a side figure. There are a total of 31 operational taxonomic units placed on this subtree, 18 of which are annotated as *Ruminococcaceae* family and *Oscillospira* genus and 13 of which are annotated as *Ruminococcaceae* family and unknown genus. Figure 6a shows the box



(a) Reference phylogenetic tree.

(b) The empirical optimal flow.

Fig. 5: Left: reference phylogenetic tree used in analysis of intestinal biopsy samples. The branches in red are those identified as active edge and zoomed in subtree below active edge. Right: the empirical optimal flow between group A and group B on each edge.

plot of the combined relative abundance on these 31 operational taxonomic units, indicating that the relative abundance on this subtree decreased in ulcerative colitis patients, but is increased partially after receiving treatments. This is consistent with previous finding that the proportion of *Oscillospira* genus and *Ruminococcaceae* family in gut microbiota deceases in inflammatory bowel disease patients (Konikoff & Gophna, 2016; Santoru et al., 2017; Morgan et al., 2012). For the purpose of comparison, the box plot of the combined relative abundance of all operational taxonomic units assigned to *Oscillospira* genus is shown in Figure 6b, showing that the pattern of relative abundance found in Figure 6a is not that clear any more. This suggests that the finer species classification by microbial phylogeny can provide more power to detect the subtle difference between populations than standard taxonomic classification (Washburne et al., 2018).

ACKNOWLEDGEMENT

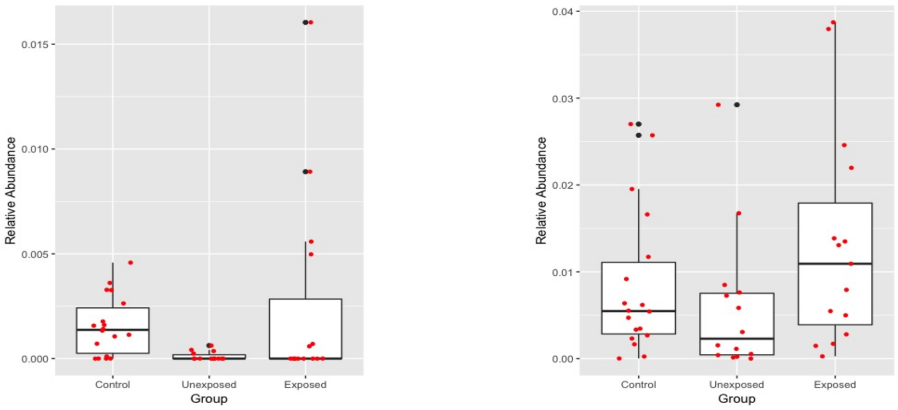
This project is supported by grants from the National Institutes of Health.

SUPPLEMENTARY MATERIAL

Supplementary material available at Biometrika online includes proofs of all theorems, additional simulation results and section on generalized optimal flow. The software package DAFOT is available at https://cran.r-project.org/web/packages/DAFOT/index.html

REFERENCES

AITCHISON, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)* **44**, 139–160.



(a) Detected subtree.

(b) *Oscillospira* genus

Fig. 6: Box plots of relative abundance on operational taxonomic units placed on detected subtree in Figure 5a and on all operational taxonomic units assigned to *Oscillospira* genus. Red dots are raw relative abundance.

AITCHISON, J. (2003). *The Statistical Analysis of Compositional Data.* Blackburn Press.

ANDERSON, M. J. (2014). Permutational multivariate analysis of variance (PERMANOVA). *Wiley Statsref: Statistics Reference Online* , 1–15.

ARIAS-CASTRO, E., CANDES, E. J. & DURAND, A. (2011). Detection of an anomalous cluster in a network. *The Annals of Statistics* **39**, 278–304.

ARIAS-CASTRO, E., DONOHO, D. L. & HUO, X. (2005). Near-optimal detection of geometric objects by fast multiscale methods. *IEEE Transactions on Information Theory* **51**, 2402–2425.

BAI, Z. & SARANADASA, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica* , 311–329.

CAI, T. T., LIU, W. & XIA, Y. (2014). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 349–372.

CAO, Y., LIN, W. & LI, H. (2017). Two-sample tests of high-dimensional means for compositional data. *Biometrika* **105**, 115–132.

CHEN, J., RYU, E., HATHCOCK, M., BALLMAN, K., CHIA, N., OLSON, J. E. & NELSON, H. (2016). Impact of demographics on human gut microbial diversity in a us midwest population. *PeerJ* **4**, e1514.

CHEN, S., LI, J. & ZHONG, P. (2019). Two-sample and ANOVA tests for high dimensional means. *The Annals of Statistics* **47**, 1443–1474.

CHEN, S. & QIN, Y. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics* **38**, 808–835.

DESANTIS, T. Z., HUGENHOLTZ, P., LARSEN, N., ROJAS, M., BRODIE, E. L., KELLER, K., HUBER, T., DALEVI, D., HU, P. & ANDERSEN, G. L. (2006). Greengenes, a chimera-checked 16s rRNA gene database and workbench compatible with ARB. *Applied and environmental microbiology* **72**, 5069–5072.

DUMBGEN, L. & SPOKOINY, V. G. (2001). Multiscale testing of qualitative hypotheses. *Annals of Statistics* , 124–152.

EGOZCUE, J. J. & PAWLOWSKY-GLAHN, V. (2016). Changing the reference measure in the simplex and its weighting effects. *Austrian Journal of Statistics* **45**, 25–44.

EVANS, S. & MATSEN, F. (2012). The phylogenetic kantorovich–rubinstein metric for environmental sequence samples. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74**, 569–592.

FUKUYAMA, J. (2017). Adaptive gPCA: A method for structured dimensionality reduction. *arXiv preprint arXiv:1702.00501* .

GOOD, P. (2013). *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer Science & Business Media.

INGSTER, Y. I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives. i, ii, iii. *Mathematical Methods in Statistics* **2**, 85–114, 171–189, 249–268.

INGSTER, Y. I. & SUSLINA, I. A. (2012). *Nonparametric goodness-of-fit testing under Gaussian models*, vol. 169. Springer Science & Business Media.

Janssen, S., McDonald, D., Gonzalez, A., Navas-Molina, J. A., Jiang, L., Xu, Z., Winker, K., Kado, D. M., Orwoll, E., Manary, M., Mirarab, S. & Knight, R. (2018). Phylogenetic placement of exact amplicon sequences improves associations with clinical information. *MSystems* **3**, e00021–18.

Jeng, X. J., Cai, T. T. & Li, H. (2010). Optimal sparse segment identification with application in copy number variation analysis. *Journal of the American Statistical Association* **105**, 1156–1166.

Konikoff, T. & Gophna, U. (2016). Oscillospira: a central, enigmatic component of the human gut microbiota. *Trends in Microbiology* **24**, 523–524.

Lin, W., Shi, P., Feng, R. & Li, H. (2014). Variable selection in regression with compositional covariates. *Biometrika* **101**, 785–797.

Love, M. I., Huber, W. & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550.

Lozupone, C., Hamady, M., Kelley, S. & Knight, R. (2007). Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Applied and Environmental Microbiology* **73**, 1576–1585.

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

Mandal, S., Van Treuren, W., White, R. A., Eggesbø, M., Knight, R. & Peddada, S. D. (2015). Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial Ecology in Health and Disease* **26**, 27663.

McArdle, B. H. & Anderson, M. J. (2001). Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology* **82**, 290–297.

Mirarab, S., Nguyen, N. & Warnow, T. (2012). SEPP: SATé-enabled phylogenetic placement. In *Biocomputing 2012*. pp. 247–258.

Morgan, X. C., Tickle, T. L., Sokol, H., Gevers, D., Devaney, K. L., Ward, D. V., Reyes, J. A., Shah, S. A., LeLeiko, N., Snapper, S. B., Bousvaros, A., Korzenik, J., Sands, B. E., Xavier, R. J. & Huttenhower, C. (2012). Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment (article) author. *Genome biology* **13**, R79–R79.

Rivera-Pinto, J., Egozcue, J., Pawlowsky-Glahn, V., Paredes, R., Noguera-Julian, M. & Calle, M. (2018). Balances: a new perspective for microbiome analysis. *MSystems* **3**, e00053–18.

Robinson, M. D., McCarthy, D. J. & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140.

Santoru, M. L., Piras, C., Murgia, A., Palmas, V., Camboni, T., Liggi, S., Ibba, I., Lai, M. A., Orrù, S., Blois, S., Loizedda, A., Griffin, J. L., Usai, P., Caboni, P., Atzori, L. & Manzin, A. (2017). Cross sectional evaluation of the gut-microbiome metabolome axis in an italian cohort of IBD patients. *Scientific reports* **7**, 9523.

Sejdinovic, D., Sriperumbudur, B., Gretton, A. & Fukumizu, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics* **41**, 2263–2291.

Smith, C. C., Snowberg, L. K., Caporaso, J. G., Knight, R. & Bolnick, D. I. (2015). Dietary input of microbes and host genetic variation shape among-population differences in stickleback gut microbiota. *The ISME journal* **9**, 2515.

Srivastava, M. S. & Du, M. (2008). A test for the mean vector with fewer observations than the dimension. *Journal of Multivariate Analysis* **99**, 386–402.

Székely, G. J. & Rizzo, M. L. (2005). A new test for multivariate normality. *Journal of Multivariate Analysis* **93**, 58–80.

Wagner, B. D., Robertson, C. E. & Harris, J. K. (2011). Application of two-part statistics for comparison of sequence variant counts. *PloS One* **6**, e20296.

Wang, S., Fan, J., Pocock, G., Arena, E. T., Eliceiri, K. W. & Yuan, M. (2019). Structured correlation detection with application to colocalization analysis in dual-channel fluorescence microscopic imaging. *Statistica Sinica* .

Washburne, A. D., Morton, J. T., Sanders, J., McDonald, D., Zhu, Q., Oliverio, A. M. & Knight, R. (2018). Methods for phylogenetic analysis of microbiome data. *Nature Microbiology* **3**, 652.

Wu, G. D., Compher, C., Chen, E. Z., Smith, S. A., Shah, R. D., Bittinger, K., Chehoud, C., Albenberg, L. G., Nessel, L., Gilroy, E., Star, J., Weijie, A. M., Flint, H. J., Metz, D., Bennett, M., Li, H., Bushman, F. & Lewis, J. (2016). Comparative metabolomics in vegans and omnivores reveal constraints on diet-dependent gut microbiota metabolite production. *Gut* **65**, 63–72.

Xia, Y. & Sun, J. (2017). Hypothesis testing and statistical analysis of microbiome. *Genes & diseases* **4**, 138–148.

Zhao, N., Chen, J., Carroll, I. M., Ringel-Kulka, T., Epstein, M. P., Zhou, H., Zhou, J. J., Ringel, Y. & Li, H.and Wu, M. C. (2015). Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. *The American Journal of Human Genetics* **96**, 797–807.