Biometrika (XXXX), **XXX**, XX, *pp*. 1–21
Printed in Great Britain Advance Access publication on XX XX XXXX

Phylogenetic Association Analysis with Conditional Rank Correlation

BY SHULEI WANG, BO YUAN

Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820 USA

shuleiw@illinois.edu, boyuan5@illinois.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

Phylogenetic association analysis is an essential and powerful tool for studying the association between microbial compositions and the outcome of interest in microbiome studies. However, existing methods for testing such associations are more sensitive to a linear association in a high-dimensional setting and the assumptions of confounding effects. Methods that are capable of characterizing complex association, including non-monotonic association, are therefore needed. This paper proposes a new phylogenetic association analysis framework to address these challenges. The new framework introduces conditional rank correlation as a measure of association to detect a wide range of dependencies, which is robust to the outlier, accounts for confounders in a fully nonparametric way. The new framework aggregates conditional rank correlations for subtrees as the weighted sum and maximum to capture dense and sparse signals. To determine the significance level, we calibrate the test statistics by a nearest neighbor bootstrapping method, which is easy to use and can incorporate extra data sets when available. The practical merits of the new framework are demonstrated by numerical experiments using both simulated and real microbiome data sets.

Some key words: Compositional Data, Association Analysis, Phylogenetic Tree, Covariate Adjustment.

1. INTRODUCTION

The microbial communities inhabiting the human body play an essential role in human health and are associated with many human diseases, such as obesity, inflammatory bowel disease, and type II diabetes (Li, 2015). Understanding the association between human diseases and microbiotas can help discover diagnostic biomarkers and develop efficient treatments for diseases (Pflughoeft & Versalovic, 2012). The recent advances in high throughput sequencing technolo-

gies make it possible to characterize the microbial communities in a high resolution and thus provide an opportunity for a more comprehensive understanding of the roles of microbial communities in disease onset and progression.

One key step towards understanding the microbiome's role in human health is to detect an association between the human microbiome and the outcome of interest after adjusting potential measured confounders. We can naturally formulate the association analysis as a conditional independence hypothesis testing problem (or independence hypothesis testing problem when no confounder is present) and apply some existing generic conditional independence tests (or independence tests) to the microbiome data directly (Székely et al., 2007; Gretton et al., 2008; Lyons, 2013; Wang et al., 2015; Azadkia & Chatterjee, 2019; Huang et al., 2020). However, this might not be the most efficient strategy as microbiome data have several unique characteristics, including prevalent zero counts (Wang, 2022), compositional data (Li, 2015), and a phylogenetic tree structure among the microbes (Washburne et al., 2018; Wang, 2021). Many association analysis methods are designed for the microbial compositional data set to account for these characteristics and incorporate a phylogenetic tree (McArdle & Anderson, 2001; Pan, 2011; Zhao et al., 2015; Wu et al., 2016; Tang et al., 2016, 2017; Koh et al., 2017; Song et al., 2020). These methods have been very successful in many human microbiome studies and helped decipher the role of microbiota in different human diseases.

Most state-of-the-art microbiome association analysis methods are distance-based (kernel-based) methods or their generalization. The advantage of distance-based dependence metrics is that they can detect any type of association between random vectors when the dimension of the random vector is not very large (Székely et al., 2007). However, recent results suggest that distance or kernel-based dependence metric mainly captures linear association under a high dimensional setting, which microbiome data usually fall into (El Karoui, 2010; Zhu et al., 2020). In addition, these association analysis methods usually adjust confounding variables by a parametric regression model, such as a linear model. It is not immediately clear if existing methods can detect an association robustly when the parametric model is misspecified. This paper proposes a new phylogenetic association analysis framework for detecting complex associations in microbiome studies that is capable to detect a wide range of dependencies and accounts for confounders in a fully nonparametric way.

We first introduce a new phylogenetic independence test by incorporating the idea of rank correlation (Weihs et al., 2018; Shi et al., 2020) and phylogenetic tree structure. Instead of evaluating pairwise distance, the new phylogenetic independence test directly measures the association between the outcome of interest and the total microbial abundance in each lineage (subtree) by a rank correlation. Because of rank correlation, the new test is robust to the outlier, invariant to monotonic transformation, and captures a wide range of associations. To detect global association, the new phylogenetic independence test aggregates the rank correlations for different lineages in two ways: the weighted sum and the maximum. The power analysis shows that the weighted sum aggregation is more sensitive to the dense dependence structure, while the maximum aggregation is more powerful for the sparse dependency.

We then generalize the phylogenetic independence test to a phylogenetic conditional independence test when confounder adjustment is needed. To adjust the confounders nonparametrically, we introduce conditional rank correlation and its corresponding estimator based on the idea of the nearest neighbor method. As a generalization of rank correlation, conditional rank correlation inherits many desired properties, including robustness to outliers and the ability of detecting nonlinear association. Similar to the independence test, the new phylogenetic conditional independence test also takes the weighted sum or maximum of the conditional rank correlation for different lineages. We investigate the power of the new conditional independence

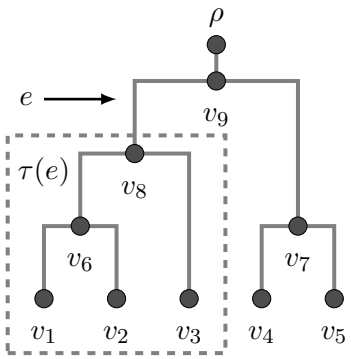
test under the conditional randomization test (CRT) framework (Candes et al., 2018), where the conditional distribution of outcome given covariates is assumed to be known. We show that its performance is determined by the smoothness of the conditional distribution and the phylogenetic tree structure. In addition, we discuss several different test calibration methods, including a nearest neighbor bootstrapping method for the new phylogenetic conditional independence test. The method takes advantage of extra data sets when available and is easy to implement. All newly introduced methods in this paper are implemented in the DAFOT package, available at `https://github.com/lakerwsl/DAFOT`.

2. Model and Phylogenetic Association Analysis

2.1. *Model and Notation*

Two of the most popular sequencing methods to quantify the abundance of different bacteria in microbial communities are 16S rRNA sequencing and shotgun metagenomics sequencing. The sequencing reads can be placed on a reference phylogenetic tree by phylogenetic placement methods, such as pplacer (Matsen et al., 2010) or SEPP (Mirarab et al., 2012; Janssen et al., 2018). More specifically, let $T = (V, E)$ be the phylogenetic tree of microbe species, where V is the collection of microbe species and their ancestors, and E represents the collection of edges of the phylogenetic tree T. We assume the tree T is rooted at the node ρ, the common ancestor of all microbe species. For each edge e, L_e denotes the corresponding branch length. A toy example of the phylogenetic tree is given in Figure 1. The relative abundance of a microbial community can be represented by a discrete distribution on the nodes of tree T. More specifically, let p_v be the relative abundance of microbial taxon v and write all possible discrete distributions on T as

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



$$P_e = p_{v_1} + p_{v_2} + p_{v_3} + p_{v_6} + p_{v_8}$$

Fig. 1: A toy example of phylogenetic tree and relative abundance on subtree $\tau(e)$.

In microbiome studies, we are interested in testing an association between microbial community composition and outcome of interest $Y \in \mathbb{R}$, which can be naturally formulated as the following independence hypothesis testing problem

$$H_0 : \boldsymbol{P} \perp Y \qquad \text{vs} \qquad H_1 : \boldsymbol{P} \not\perp Y. \tag{1}$$

Besides the true relative abundance of different microbial taxa $\boldsymbol{P}$ and outcome of interest Y, we usually also observe several covariates $\boldsymbol{X} \in \mathbb{R}^d$ for each sample, which might be linked to both $\boldsymbol{P}$ and Y. In the presence of the available covariates, we need to adjust potential confounding effects from these observed covariates. The association analysis can be formulated as the following conditional independence hypothesis in such a case

$$H_0 : \boldsymbol{P} \perp Y | \boldsymbol{X},\ \forall \boldsymbol{X} \qquad \text{vs} \qquad H_1 : \boldsymbol{P} \not\perp Y | \boldsymbol{X},\ \exists \boldsymbol{X}. \tag{2}$$

If there are no observed covariates (i.e., $d = 0$), hypothesis (2) is reduced to hypothesis (1).

In order to test the hypothesis in (1) or (2), n individuals/samples are independently drawn from a given population. $(\boldsymbol{P}_i, Y_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$ represent their true relative abundance of microbe taxa, the outcome of interest, and covariates. While the outcome of interest Y and covariates $\boldsymbol{X}$ are usually known exactly, the true relative abundance of different bacteria taxa $\boldsymbol{P}$ cannot be observed directly. Sequencing techniques are usually applied to assess each sample's relative abundance of microbes. After phylogenetic placement, $N_{i,v}$ reads are placed to node v in the ith sample. The number of sequencing reads $N_{i,v}$ is assumed to follow a Poisson distribution

$$N_{i,v} \sim \text{Pois}(N_i p_{i,v}), \qquad v \in V \quad \text{and} \quad 1 \le i \le n,$$

where N_i is the total number of reads in the ith sample, and $p_{i,v}$ is the relative abundance of microbe taxon v in the ith sample. We usually normalize the count data $N_{i,v}$ as a relative abundance vector, that is, $\hat{\boldsymbol{P}}_i = \{\hat{p}_{i,v}\}_{v \in V}$, where $\hat{p}_{i,v} = N_{i,v}/N_i$. For simplicity, we always assume $N_1 = \ldots = N_n = N$ throughout the paper. The observed relative abundance $\hat{\boldsymbol{P}}_i$ can be seen as an empirical version of true relative abundance $\boldsymbol{P}_i$. The goal of association analysis is to test the hypothesis in (1) or (2) based on the observed data $(\hat{\boldsymbol{P}}_i, Y_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$.

2.2. *Phylogenetic Association Analysis*

Most existing distance or kernel-based phylogenetic association analysis methods adopt a phylogenetic distance between two microbial communities. The phylogenetic distance is defined on the phylogenetic tree structure and thus can reflect the evolution information among microbe species. Commonly used phylogenetic distances include unweighted and weighted UniFrac distance (Lozupone & Knight, 2005; Lozupone et al., 2007) and their Zolotarev-type generalizations (Evans & Matsen, 2012; Wang et al., 2021b). The weighted UniFrac distance is defined as

$$D(\boldsymbol{P}, \mathbf{Q}) = \sum_e L_e |P_e - Q_e|,$$

where L_e is the branch length of the edge e. Here, $P_e = \sum_{v \in \tau(e)} p_v$ is the total relative abundance on the subtree $\tau(e)$, and $\tau(e)$ is the subtree below the edge e. To illustrate the concepts, Figure 1 shows a typical example of $\tau(e)$ and P_e. As illustrated by Wang et al. (2021a), the phylogenetic distance-based methods essentially consider each P_e, $e \in E$ as the analysis unit instead of each p_v, $v \in V$ and thus try to detect the dependency between the outcome of interest Y and each P_e, $e \in E$. Following the same idea, we will use P_e as our analysis unit for the rest of the paper.

The phylogenetic distance-based methods are designed to capture possible nonlinear dependency between $\hat{\boldsymbol{P}}$ and Y when the dimension of $\hat{\boldsymbol{P}}$ is not very large. However, a recent result in Zhu et al. (2020) shows that the distance-based dependence metric reflects linear association in the high dimensional setting. This indicates that the phylogenetic distance-based methods might not be the most powerful for detecting nonlinear associations when the number of microbes is large. To adjust the covariate $\boldsymbol{X}$, most existing phylogenetic association analysis methods rely on a parametric model, such as a linear model, and thus can only remove a particular type of confounding effect. It is not immediately clear if the conditional independence hypothesis in (2)

can still be tested consistently when such a parametric model is misspecified. We introduce a new phylogenetic association analysis framework for the microbial compositional data. We first discuss the independence testing problem in (1) and then extend the methods and results to the conditional independence testing problem in (2).

3. Phylogenetic Independence Test with Rank Correlation

3.1. *Rank Correlation*

Testing independence between two real-valued random variables is a classical statistical problem, and many methods have been developed (Rényi, 1959; Breiman & Friedman, 1985; Hoeffding, 1948; Blum et al., 1961; Yanagimoto, 1970; Romano, 1988; Gretton et al., 2005; Székely et al., 2007; Gretton et al., 2008; Székely & Rizzo, 2009; Reshef et al., 2011; Bergsma & Dassios, 2014; Weihs et al., 2018; Chatterjee, 2020). In order to capture a wide range of dependencies, we focus on the following three consistent rank correlation measures: Hoeffding's D (Hoeffding, 1948), Blum-Kiefer-Rosenblatt's R (Blum et al., 1961), and Bergsma-Dassios-Yanagimoto's τ^* (Bergsma & Dassios, 2014; Yanagimoto, 1970). These three rank correlation measures have the following advantages: 1) they are zero if and only if the two random variables are independent; 2) they can be estimated by a rank statistic, so they are invariant to any monotone transformation and robust to potential outliers; 3) their estimator can be computed in $O(n \log n)$ time when random variables are continuous (Even-Zohar, 2020). See more discussions in Weihs et al. (2018); Shi et al. (2020). To present formal definitions of these rank correlation measures, we focus on the association between outcome and abundance on a given subtree $\tau(e)$ and let $(\hat{P}_{1,e}, Y_1), \ldots, (\hat{P}_{n,e}, Y_n)$ be independent copies of $(\hat{P}_e, Y)$.

Definition 1 (Hoeffding's D). *The Hoeffding's correlation coefficient D_n between $\hat{P}_e$ and Y is defined as*

$$D_n(\hat{P}_e, Y) = \binom{n}{5}^{-1} \sum_{1 \le i_1 < \ldots < i_5 \le n} \frac{1}{16} g_D(\hat{P}_{i_1,e}, \ldots, \hat{P}_{i_5,e}) g_D(Y_{i_1}, \ldots, Y_{i_5}),$$

where $g_D(z_1, \ldots, z_5) = (\mathbf{I}_{(z_1 \le z_5)} - \mathbf{I}_{(z_2 \le z_5)})(\mathbf{I}_{(z_3 \le z_5)} - \mathbf{I}_{(z_4 \le z_5)})$.

Definition 2 (Blum-Kiefer-Rosenblatt's R). *Blum-Kiefer-Rosenblatt's R_n between $\hat{P}_e$ and Y is defined as*

$$R_n(\hat{P}_e, Y) = \binom{n}{6}^{-1} \sum_{1 \le i_1 < \ldots < i_6 \le n} \frac{1}{32} g_R(\hat{P}_{i_1,e}, \ldots, \hat{P}_{i_4,e}, \hat{P}_{i_5,e}) g_R(Y_{i_1}, \ldots, Y_{i_4}, Y_{i_6}),$$

where $g_R(z_1, \ldots, z_5) = (\mathbf{I}_{(z_1 \le z_5)} - \mathbf{I}_{(z_2 \le z_5)})(\mathbf{I}_{(z_3 \le z_5)} - \mathbf{I}_{(z_4 \le z_5)})$.

Definition 3 (Bergsma-Dassios-Yanagimoto's τ^*). *Bergsma-Dassios-Yanagimoto's τ_n^* between $\hat{P}_e$ and Y is defined as*

$$\tau_n^*(\hat{P}_e, Y) = \binom{n}{4}^{-1} \sum_{1 \le i_1 < \ldots < i_4 \le n} \frac{1}{16} g_{\tau^*}(\hat{P}_{i_1,e}, \ldots, \hat{P}_{i_4,e}) g_{\tau^*}(Y_{i_1}, \ldots, Y_{i_4}),$$

where $g_{\tau^*}(z_1, \ldots, z_4) = \mathbf{I}_{(z_1,z_3<z_2,z_4)} + \mathbf{I}_{(z_2,z_4<z_1,z_3)} - \mathbf{I}_{(z_1,z_4<z_2,z_3)} - \mathbf{I}_{(z_2,z_3<z_1,z_4)}$. *Here,* $\mathbf{I}_{(y_1,y_2<y_3,y_4)} = \mathbf{I}_{(y_1<y_3)}\mathbf{I}_{(y_1<y_4)}\mathbf{I}_{(y_2<y_3)}\mathbf{I}_{(y_2<y_4)}$. *As long as* $n \ge 6$, $\tau_n^*(\hat{P}_e, Y) = 12 D_n(\hat{P}_e, Y) + 24 R_n(\hat{P}_e, Y)$.

All these statistics above can be written as U-statistic

$$\psi_n(\hat{P}_e, Y) = \binom{n}{l}^{-1} \sum_{1 \le i_1 < \ldots < i_l \le n} h_\psi(Z_{i_1,e}, \ldots, Z_{i_l,e}),$$

where h_ψ is a symmetric bounded kernel depending on $\psi \in \{D, R, \tau^*\}$ and $Z_e = (\hat{P}_e, Y)$. These statistics are called rank correlation coefficients because $\psi_n(\hat{P}_e, Y) = \psi_n(r(\hat{P}_e), r(Y))$, where $r(\hat{P}_e)$ and $r(Y)$ are the ranks of $\hat{P}_e$ and Y. Based on $\psi_n(\hat{P}_e, Y)$, the corresponding rank correlation measure is then defined as $\psi(\hat{P}_e, Y) = \mathbb{E}(\psi_n(\hat{P}_e, Y)) = \mathbb{E}(h_\psi(\hat{P}_e, Y))$, where $h_\psi(\hat{P}_e, Y) = h_\psi(Z_{1,e}, \ldots, Z_{l,e})$. When $\psi \in \{D, R, \tau^*\}$, $\psi(\hat{P}_e, Y)$ is always larger than or equal to 0, and $\psi(\hat{P}_e, Y) = 0$ if and only if $\hat{P}_e$ and Y are independent. The conditional expectations of kernel h_ψ are defined as

$$h_\psi^{(k)}(z_1, \ldots, z_k) = \mathbb{E}\left(h_\psi(z_1, \ldots, z_k, Z_{k+1}, \ldots, Z_l)\right), \qquad 1 \le k \le l.$$

When $\hat{P}_e$ and Y are independent, h_ψ is mean zero and degenerate kernel, i.e., $h_\psi^{(1)}(z_1) = 0$ almost surely with respect to z_1. For simplicity, we also write $h_\psi(\hat{P}_e, Y) = h_\psi(Z_{1,e}, \ldots, Z_{l,e})$ and $h_\psi^{(k)}(\hat{P}_e, Y) = h_\psi(Z_{1,e}, \ldots, Z_{k,e})$.

3.2. *Phylogenetic Independence Test in the Absence of Covariates*

In this section, we assume no observed covariate and focus on testing if the relative abundance of different microbial species is independent of the outcome of interest. As noted in Section 2, existing phylogenetic independence tests are mainly designed to detect a linear association between Y and P_e for each edge e in the high dimensional setting. To capture their possible nonlinear associations, the association between Y and P_e can be measured by a rank correlation $\psi_n(\hat{P}_e, Y)$, where $\psi_n \in \{D_n, R_n, \tau_n^*\}$. To test the hypothesis in (1), we aggregate these rank correlation $\psi_n(\hat{P}_e, Y)$ into a single test statistic. Depending on the forms of the alternative hypothesis, we consider two different ways to integrate rank correlation over edges. If the outcome of interest Y is associated with most microbial taxa, we consider the following weighted sum type test

$$\Psi_1 = n \sum_{e \in E} w_e \psi_n(\hat{P}_e, Y),$$

since the rank correlation measures are non-negative. There are multiple ways to choose the weights $w_e > 0$ in Ψ_1, e.g., we can choose $w_e = L_e$. In practice, it is also possible that only a small number of microbe species are associated with the outcome (Wu et al., 2016; Wang et al., 2021a). In such a sparse case, we consider the following maximum type test

$$\Psi_\infty = n \max_{e \in E} \psi_n(\hat{P}_e, Y).$$

Another advantage of such a maximum type test is that it provides a way of identifying the microbial lineage that is associated with the outcome when the null hypothesis is rejected.

To make the above test statistic as decision rules, we still need to choose appropriate critical values or transform them into P-values. Due to the complex dependency structure among $\hat{P}_e$, it is difficult to derive an asymptotic distribution for the above test. So we opt to adopt a resampling method. More specifically, let A_n be the set of permutations on $\{1, \ldots, n\}$, i.e., $A_n = \{\phi : \{1, \ldots, n\} \to \{1, \ldots, n\} | \phi(i) \ne \phi(j) \text{ if } i \ne j\}$. We write $\Psi_1(\hat{\boldsymbol{P}}, \phi Y)$ and $\Psi_\infty(\hat{\boldsymbol{P}}, \phi Y)$ as test statistics calculated on $(\hat{\boldsymbol{P}}_1, Y_{\phi(1)}), \ldots, (\hat{\boldsymbol{P}}_n, Y_{\phi(n)})$ given a map $\phi \in A_n$. When ϕ is an identity map, we also write $\Psi_1(\hat{\boldsymbol{P}}, \phi Y)$ and $\Psi_\infty(\hat{\boldsymbol{P}}, \phi Y)$ as $\Psi_1(\hat{\boldsymbol{P}}, Y)$ and $\Psi_\infty(\hat{\boldsymbol{P}}, Y)$. Let $\phi_1, \ldots, \phi_B$

be B maps drawn from A_n randomly. Then, the P-value can be calculated by

$$\hat{P}(\Psi) = \frac{1 + \sum_{b=1}^{B} \mathbf{I}_{(\Psi(\hat{\boldsymbol{P}}, \phi_b Y) \geq \Psi(\hat{\boldsymbol{P}}, Y))}}{1 + B},$$

where Ψ is Ψ_1 or Ψ_∞ test. Here, if $\Psi(\hat{\boldsymbol{P}}, Y)$ is a tie with some $\Psi(\hat{\boldsymbol{P}}, \phi_b Y)$, the tie is broken randomly. We then make the decision based on P-value, i.e., the null hypothesis is rejected when $\hat{P}(\Psi) \leq \alpha$. Since $\Psi(\hat{\boldsymbol{P}}, \phi_b Y)$ and $\Psi(\hat{\boldsymbol{P}}, Y)$ have the same distribution under the null hypothesis, the type I error can be controlled at the desired level, that is,

$$\mathbb{P}\left(\hat{P}(\Psi) \leq \alpha \middle| H_0\right) \leq \alpha.$$

3.3. *Power Analysis for Phylogenetic Independence Test*

We now investigate the power of the proposed phylogenetic independence tests. The test $\hat{P}(\Psi)$ is called consistent if

$$\mathbb{P}\left(\hat{P}(\Psi) > \alpha\right) \to 0.$$

The main difficulty in studying the behavior of Ψ_1 or Ψ_∞ is the strong and complex dependence structure among $\psi_n(\hat{P}_e, Y)$ for different $e \in E$. Owing to the tree structure, there are only two possibilities for a pair of different subtrees $\tau(e)$ and $\tau(e')$: $\tau(e) \subset \tau(e')$ ($\tau(e') \subset \tau(e)$) or $\tau(e) \cap \tau(e') = \emptyset$. Here, we assume the dependence structure among $\psi_n(\hat{P}_e, Y)$ is mainly due to the overlapping between subtrees $\tau(e)$. Specifically, we assume that if $\tau(e) \subset \tau(e')$, there are constants c_0 and α such that

$$\mathbb{E}\left(h_\psi(\hat{P}_e, Y) - h_\psi(\hat{P}_{e'}, Y)\right)^2 \leq c_0 \left(\frac{\lambda(\tau(e')) - \lambda(\tau(e))}{\lambda(\tau(e'))}\right)^\alpha, \tag{3}$$

where λ is a measure on the tree T to quantify the level of overlapping between two subtrees. For example, $\lambda(\tau(e))$ could be the total number of nodes or the total relative abundance in the subtree $\tau(e)$. The assumption in (3) suggests that $h_\psi(\hat{P}_e, Y)$ and $h_\psi(\hat{P}_{e'}, Y)$ are highly correlated if there is large overlapping between corresponding subtrees. On the other hand, we assume that if there is no overlapping between two subtrees, i.e., $\tau(e) \cap \tau(e') = \emptyset$, $h_\psi(\hat{P}_e, Y)$ and $h_\psi(\hat{P}_{e'}, Y)$ are almost uncorrelated, that is, there exists a constant Δ such that

$$\mathbb{E}(h_\psi(\hat{P}_e, \hat{P}_{e'})) \leq \Delta \qquad \text{and} \qquad \left|\mathrm{Cov}\left(h_\psi(\hat{P}_e, Y), h_\psi(\hat{P}_{e'}, Y)\right)\right| \leq \Delta. \tag{4}$$

Here, Δ reflects the level of dependence between $h_\psi(\hat{P}_e, Y)$ and $h_\psi(\hat{P}_{e'}, Y)$. In particular, we can choose $\Delta = 0$ if $\hat{P}_e$, $\hat{P}_{e'}$, and Y are mutually independent. Besides the assumptions in (3) and (4), we also need to make another assumption on $h_\psi^{(1)}(\hat{P}_e, Y)$ when the distribution is under the alternative hypothesis. Specifically, we assume there is a constant γ such that

$$\mathrm{Var}\left(h_\psi^{(1)}(\hat{P}_e, Y)\right) \leq \gamma \mathbb{E}\left(h_\psi(\hat{P}_e, Y)\right) \quad \text{and} \quad \left|\mathrm{Corr}\left(h_\psi^{(1)}(\hat{P}_e, Y), h_\psi^{(1)}(\hat{P}_{e'}, Y)\right)\right| \leq \Delta. \tag{5}$$

A similar condition also appears in Drton et al. (2018), showing that Gaussian distribution satisfies such a condition. The following theorem characterizes the power of our proposed tests.

THEOREM 1. *Suppose the weighted sum type test Ψ_1 and the maximum type test Ψ_∞ are defined based on a kernel h_ψ and the number of permutations $B \to \infty$. If we assume* (4)*,* (5)

hold, $1/M \leq w_e \leq M$ *for some fixed constant* M *and*

$$\sum_{e \in E} w_e \psi(\hat{P}_e, Y) \gg \frac{\sqrt{D_T|E| + \Delta|E|^2}}{n}, \tag{6}$$

then the weighted sum type test Ψ_1 *is consistent. In addition, if* (3), (5) *hold,* $\log^4 |E| \ll n$ *and*

$$\max_{e \in E} \psi(\hat{P}_e, Y) \gg \frac{\log S_T}{n},$$

then the maximum type test Ψ_∞ *is a consistent test.* S_T *is a quantity depending on the measure* λ *and tree structure* T*, which is defined in Appendix.*

Theorem 1 suggests Ψ_1 is more sensitive to weak and dense dependence between outcome and subtree, and Ψ_∞ is designed to capture strong but sparse dependence. In practice, we usually have no access to the sparsity level of dependence. In order to capture both dense and sparse dependence for robustness, the natural idea is to combine the weighted sum type test and maximum type test, e.g., taking the smallest P-value

$$\hat{P} = \min\left(\hat{P}(\Psi_1), \hat{P}(\Psi_\infty)\right).$$

In addition, Theorem 1 also helps characterize the effect of a strong dependency structure among subtrees for both tests. Specifically, the power of the weighted sum type test is reduced because of dependence among $\hat{P}_e$, while the power of the maximum type test increases. The required signal strength in Theorem 1 relies on assumptions (3) and (4). Without such assumptions, sufficient signal strength conditions for consistent tests are

$$\sum_{e \in E} w_e \psi(\hat{P}_e, Y) \gg \frac{|E|}{n} \qquad \text{and} \qquad \max_{e \in E} \psi(\hat{P}_e, Y) \gg \frac{\log |E|}{n}.$$

4. Phylogenetic Conditional Independence Test with Conditional Rank Correlation

4.1. *Conditional Rank Correlation*

The previous section shows that rank correlation can detect a nonlinear association between the outcome of interest and microbial composition. This section mainly focuses on adjusting potential confounding effects from the observed covariates. Recall that the rank correlation measure between $\hat{P}_e$ and Y is defined as

$$\psi(\hat{P}_e, Y) = \mathbb{E}(h_\psi(\hat{P}_e, Y)),$$

where the expectation is with respect to the joint distribution of $\hat{P}_e$ and Y. Following this notation, we can naturally define the local conditional rank correlation measure

$$\psi(\hat{P}_e, Y|\boldsymbol{X}) = \mathbb{E}(h_\psi(\hat{P}_e, Y)|\boldsymbol{X}),$$

where the expectation is with respect to the conditional joint distribution of $\hat{P}_e$ and Y given $\boldsymbol{X}$. We can show that $\psi(\hat{P}_e, Y|\boldsymbol{X}) = 0$ if and only if $\hat{P}_e \perp Y|\boldsymbol{X}$. In order to test $\hat{P}_e \perp Y|\boldsymbol{X}$ for all possible $\boldsymbol{X}$, we define the global conditional rank correlation measure as

$$\psi_\Lambda(\hat{P}_e, Y) = \int \psi(\hat{P}_e, Y|\boldsymbol{X})\Lambda(\boldsymbol{X})d\boldsymbol{X}$$

for some measure $\Lambda(\boldsymbol{X})$. A natural choice of $\Lambda(\boldsymbol{X})$ is the probability density function of $\boldsymbol{X}$. The following proposition suggests that $\psi_\Lambda(\hat{P}_e, Y)$ can measure conditional independence.

PROPOSITION 1. *If we assume $\Lambda(\boldsymbol{X})$ has the same support as probability density function of $\boldsymbol{X}$, then $\psi_\Lambda(\hat{P}_e, Y) = 0$ if and only if $\hat{P}_e \perp Y|\boldsymbol{X}$ almost surely. In addition, $\psi_\Lambda(\hat{P}_e, Y) = \psi_\Lambda(f(\hat{P}_e), g(Y))$ for any strictly increasing functions f and g.*

In order to estimate $\psi_\Lambda(\hat{P}_e, Y)$, we adopt an idea from the nearest neighbor method (Biau & Devroye, 2015). Specifically, let $(\boldsymbol{P}_{(1),i}, Y_{(1),i}, \boldsymbol{X}_{(1),i}), \ldots, (\boldsymbol{P}_{(n),i}, Y_{(n),i}, \boldsymbol{X}_{(n),i})$ be a permutation of $(\boldsymbol{P}_1, Y_1, \boldsymbol{X}_1), \ldots, (\boldsymbol{P}_n, Y_n, \boldsymbol{X}_n)$ such that $\|\boldsymbol{X}_{(1),i} - \boldsymbol{X}_i\| \leq \ldots \leq \|\boldsymbol{X}_{(n),i} - \boldsymbol{X}_i\|$. The level of local conditional rank correlation at $\boldsymbol{X}_i$ can be estimated by its k nearest neighbors

$$\psi_k(\hat{P}_e, Y|\boldsymbol{X}_i) = \binom{k}{l}^{-1} \sum_{1 \leq i_1 < \ldots < i_l \leq k} h_\psi(Z_{(i_1),i,e}, \ldots, Z_{(i_l),i,e}).$$

A natural estimator for the global conditional rank correlation coefficient is then given by

$$\psi_{\Lambda,n}(\hat{P}_e, Y) = \frac{1}{n} \sum_{i=1}^{n} \psi_k(\hat{P}_e, Y|\boldsymbol{X}_i).$$

We call this estimator the nearest neighbor conditional rank correlation coefficient. Another potential choice of the estimator for $\psi_\Lambda(\hat{P}_e, Y)$ is a kernel-based estimator. For example, a kernel-based conditional distance correlation is proposed in Wang et al. (2015). The nearest neighbor-based estimator is more computationally efficient compared with the kernel-based estimator. Specifically, the computational complexity of $\psi_{\Lambda,n}(\hat{P}_e, Y)$ is almost linear, i.e., $O(n \log n)$. We now investigate the statistical properties of this estimator.

THEOREM 2. *Suppose $\mathbb{P}(\|\boldsymbol{X}\| \geq t) \leq C_1 e^{-c_1 t}$ for some constants c_1 and C_1 and there exists β, γ and C such that*

$$|F_{P,Y}(t, s|\boldsymbol{X}_1) - F_{P,Y}(t, s|\boldsymbol{X}_2)| \leq C \left(1 + (\|\boldsymbol{X}_1\| + \|\boldsymbol{X}_2\|)^\gamma\right) \|\boldsymbol{X}_1 - \boldsymbol{X}_2\|^\beta, \tag{7}$$

where $F_{P,Y}(t, s|\boldsymbol{X})$ is the conditional cumulative distribution function of $\hat{P}_e, Y$ given $\boldsymbol{X}$. Then,

$$\left|\mathbb{E}(\psi_{\Lambda,n}(\hat{P}_e, Y)) - \psi_\Lambda(\hat{P}_e, Y)\right| = O\left(\sqrt{\psi_\Lambda(\hat{P}_e, Y)} \frac{\log^{\beta(d+2)} n}{n^{\beta/d}} + \frac{\log^{2\beta(d+2)} n}{n^{2\beta/d}}\right),$$

and

$$\left|\psi_{\Lambda,n}(\hat{P}_e, Y) - \mathbb{E}(\psi_{\Lambda,n}(\hat{P}_e, Y))\right| = O_p\left(n^{-1/2}\right).$$

Theorem 2 suggests that the error of the nearest neighbor conditional rank correlation coefficient is mainly dominated by bias when the dimension of covariate d is high. Here, we also briefly compare $\psi_{\Lambda,n}(\hat{P}_e, Y)$ with the other nearest neighbor-based conditional independence coefficient proposed in Azadkia & Chatterjee (2019); Huang et al. (2020). Compared with Azadkia & Chatterjee (2019), $\psi_{\Lambda,n}(\hat{P}_e, Y)$ cannot capture the function relationship between variables but is designed to detect more subtle conditional dependency. In particular, the bias of $\psi_{\Lambda,n}(\hat{P}_e, Y)$ only depends on the dimension of $\boldsymbol{X}$, but not the dimension of Y or $\hat{P}_e$, while the bias of the coefficient of Azadkia & Chatterjee (2019) depends on the sum of the dimensions of $\boldsymbol{X}$ and Y.

4.2. *Phylogenetic Conditional Independence Test in the Presence of Covariates*

Following a similar idea in the phylogenetic independence test, we can test hypothesis (2) by aggregating conditional rank correlation over different edges of the phylogenetic tree. Specifically, we still consider the weighted sum type test

$$\Psi_{\Lambda,1} = \sum_{e\in E} w_e \psi_{\Lambda,n}(\hat{P}_e, Y),$$

and maximum type test

$$\Psi_{\Lambda,\infty} = \max_{e\in E} \psi_{\Lambda,n}(\hat{P}_e, Y).$$

The weighted sum type test is more powerful for dense alternatives, while the maximum type test is more suitable for sparse alternatives.

To make the decision based on the above test statistics $\Psi_{\Lambda,1}$ and $\Psi_{\Lambda,\infty}$, we adopt the conditional randomization test (CRT) framework proposed in Candes et al. (2018). The conditional distribution $\mathbb{P}(Y|\boldsymbol{X})$ is assumed to be known in the CRT framework. Given the conditional distribution $\mathbb{P}(Y|\boldsymbol{X})$, independent samples $Y_i^{(1)}$ are drawn from $\mathbb{P}(\cdot|\boldsymbol{X}_i)$ for $i = 1, \ldots, n$. The intuition behind the CRT framework is that $(\hat{\boldsymbol{P}}_i, Y_i^{(1)}, \boldsymbol{X}_i)$ follows the same distribution as $(\hat{\boldsymbol{P}}_i, Y_i, \boldsymbol{X}_i)$ under the null hypothesis H_0 in (2), and the difference between $(\hat{\boldsymbol{P}}_i, Y_i^{(1)}, \boldsymbol{X}_i)$ and $(\hat{\boldsymbol{P}}_i, Y_i, \boldsymbol{X}_i)$ can be seen as evidence against the null hypothesis. To implement this idea, we draw independent samples from $\mathbb{P}(\cdot|\boldsymbol{X}_i)$ B times, that is,

$$Y_i^{(b)} \sim \mathbb{P}(\cdot|\boldsymbol{X}_i), \qquad b = 1, \ldots, B \text{ and } i = 1, \ldots, n.$$

Given $(\hat{\boldsymbol{P}}_i, Y_i^{(b)}, \boldsymbol{X}_i)$, $b = 1, \ldots, B$, we evaluate test statistics and write them as $\Psi_{\Lambda,1}(\hat{\boldsymbol{P}}, Y^{(b)})$ and $\Psi_{\Lambda,\infty}(\hat{\boldsymbol{P}}, Y^{(b)})$. Similarly, the test statistics evaluated on $(\hat{\boldsymbol{P}}_i, Y_i, \boldsymbol{X}_i)$ are written as $\Psi_{\Lambda,1}(\hat{\boldsymbol{P}}, Y)$ and $\Psi_{\Lambda,\infty}(\hat{\boldsymbol{P}}, Y)$. Then, the P-value can be calculated by

$$\hat{P}(\Psi_\Lambda) = \frac{1 + \sum_{b=1}^{B} \mathbf{I}_{(\Psi_\Lambda(\hat{\boldsymbol{P}}, Y^{(b)}) \geq \Psi_\Lambda(\hat{\boldsymbol{P}}, Y))}}{1 + B},$$

where Ψ_Λ is $\Psi_{\Lambda,1}$ or $\Psi_{\Lambda,\infty}$, and the tie is broken randomly. We reject the null when $\hat{P}(\Psi_\Lambda) \leq \alpha$.

4.3. *Power Analysis for Phylogenetic Conditional Independence Test*

This section focuses on the power analysis for the phylogenetic conditional independence test. Similar to the power analysis for the phylogenetic independence test, we assume that the overlapping between subtrees $\tau(e)$ is the main source of the dependence structure among $\psi_{\Lambda,n}(\hat{P}_e, Y)$. More concretely, if $\tau(e) \subset \tau(e')$, there exists a measure λ defined on tree T such that

$$\mathbb{E}\left(\tilde{\psi}_k(\hat{P}_e, Y^{(1)}|\boldsymbol{X}_i) - \tilde{\psi}_k(\hat{P}_{e'}, Y^{(1)}|\boldsymbol{X}_i)\Big|\boldsymbol{X}_1 \ldots, \boldsymbol{X}_n\right)^2 \leq c_0 \left(\frac{\lambda(\tau(e')) - \lambda(\tau(e))}{\lambda(\tau(e'))}\right)^\alpha \quad (8)$$

for some constant c_0 and α. Here, we define $\tilde{\psi}_k(\hat{P}_e, Y|\boldsymbol{X}_i) = \psi_k(\hat{P}_e, Y|\boldsymbol{X}_i) - \mathbb{E}(\psi_k(\hat{P}_e, Y|\boldsymbol{X}_i))$. Moreover, if $\tau(e) \cap \tau(e') = \emptyset$, then there is a constant Δ_Λ such that

$$\left|\mathrm{Cov}\left(\psi_k(\hat{P}_{e_1}, \tilde{Y}|\boldsymbol{X}_j), \psi_k(\hat{P}_{e_2}, \tilde{Y}|\boldsymbol{X}_j)\Big|\boldsymbol{X}_1 \ldots, \boldsymbol{X}_n\right)\right| \leq \Delta_\Lambda, \quad (9)$$

where $\tilde{Y}$ is Y or $Y^{(1)}$. Besides these two assumptions, we also assume that the conditional distribution of $(\hat{P}_e, Y)$ given $\boldsymbol{X}$ is smooth in the sense of (7). We characterize the power of the proposed test in the following theorem.

THEOREM 3. *Suppose the number of permutations $B \to \infty$ and the number of neighbors k is upper bounded. If* (7)*and* (9) *hold, $1/M \le w_e \le M$ for some fixed constant M and*

$$\sum_{e\in E} w_e \psi_\Lambda(\hat{P}_e, Y) \gg \sqrt{\frac{D_T|E| + |E|^2\Delta_\Lambda}{n}} + |E|\frac{\log^{\beta(d+2)} n}{n^{\beta/d}}, \tag{10}$$

then the weighted sum type test $\Psi_{\Lambda,1}$ is consistent. In addition, if (7)-(8) *hold, $\log^4 |E| \ll n$ and*

$$\max_{e\in E} \psi_\Lambda(\hat{P}_e, Y) \gg \sqrt{\frac{\log S_T}{n}} + \frac{\log^{\beta(d+2)} n}{n^{\beta/d}},$$

then the maximum type test $\Psi_{\Lambda,\infty}$ is a consistent test. S_T is a quantity depending on the measure λ, which is defined in Appendix.

Similar to the phylogenetic independence test, Theorem 3 suggests that $\Psi_{\Lambda,1}$ is more powerful when Y is associated with most $\hat{P}_e$; $\Psi_{\Lambda,\infty}$ is more suitable when Y is associated with a few $\hat{P}_e$. We can still combine these two types of tests if we would like to capture both dense and sparse signals. The results in Theorem 3 rely on the assumptions on dependency structure, that is, (8) and (9). When such assumptions are not satisfied in practice, the sufficient signal strength is

$$\sum_{e\in E} w_e \psi_\Lambda(\hat{P}_e, Y) \gg |E|\left(\frac{1}{\sqrt{n}} + \frac{\log^{\beta(d+2)} n}{n^{\beta/d}}\right) \quad \text{and} \quad \max_{e\in E} \psi_\Lambda(\hat{P}_e, Y) \gg \sqrt{\frac{\log|E|}{n}} + \frac{\log^{\beta(d+2)} n}{n^{\beta/d}}.$$

Unlike the phylogenetic independence test, the required signal in phylogenetic association analysis is mainly dominated by the bias in the test statistics when the dimension is high.

4.4. *Practical Consideration for Phylogenetic Conditional Independence Test*

The phylogenetic conditional independence tests are previous discussed in the conditional randomization test framework where a key assumption is that the conditional distribution $\mathbb{P}(Y|\boldsymbol{X})$ is known or can be estimated accurately. In some microbiome studies, the conditional distribution $\mathbb{P}(Y|\boldsymbol{X})$ can be estimated from an extra-large data set of $(Y, \boldsymbol{X})$. For example, the covariates $\boldsymbol{X}$ usually include demographic variables, such as age, gender, and smoking status, and the variable of interest Y is also observed easily, such as body mass index (BMI). In such a case, a much larger data set of $(Y, \boldsymbol{X})$ is usually available, e.g., in census data, although microbial compositional data is scarce. When such a large data set of $(Y, \boldsymbol{X})$ is available, the conditional distribution $\mathbb{P}(Y|\boldsymbol{X})$ can be estimated, and $Y_i^{(b)}$ is drawn from the estimated conditional distribution.

In some microbiome studies, it is not always easy to estimate the conditional distribution accurately due to the lack of a much larger extra data set or an efficient conditional distribution estimator. In such cases, we also introduce two other ways to make the decision based on test statistics $\Psi_{\Lambda,1}$ and $\Psi_{\Lambda,\infty}$: the asymptotic method and the nearest neighbor bootstrapping method. The proof of Theorem 3 also suggests theoretical critical values for $\Psi_{\Lambda,1}$ and $\Psi_{\Lambda,\infty}$ when (7) holds for $\beta = 1$. Specifically, we can reject the null hypothesis when

$$\Psi_{\Lambda,1} \ge s|E|\left(\frac{1}{\sqrt{n}} + \frac{\log^{(d+2)} n}{n^{1/d}}\right) \quad \text{and} \quad \Psi_{\Lambda,\infty} \ge s\left(\sqrt{\frac{\log|E|}{n}} + \frac{\log^{(d+2)} n}{n^{1/d}}\right), \tag{11}$$

where s is any sequence going to infinity, e.g., $s = \log\log n$. These choices of critical values rely on the assumption in (7) and can also be conservative in practice. However, these critical values are easy to calculate practically, and they provide an alternative and asymptotically valid way to test the conditional independence hypothesis in (2) when estimating the conditional distribution

$\mathbb{P}(Y|\boldsymbol{X})$ is not easy. Such an asymptotic method is also used in other conditional independence tests (Su & White, 2008; Wang et al., 2015; Zhou et al., 2020; Shah & Peters, 2020).

Besides the asymptotic method, we introduce a nearest neighbor bootstrapping (NNB) method for $\Psi_{\Lambda,1}$ and $\Psi_{\Lambda,\infty}$. The key idea of the CRT framework is to draw random samples from the conditional distribution $\mathbb{P}(Y|\boldsymbol{X})$. When $\mathbb{P}(Y|\boldsymbol{X})$ is unavailable, we can randomly draw Y from the neighbors of $\boldsymbol{X}$ by combining ideas in the nearest neighbor method and bootstrapping. To be specific, let $(\tilde{Y}_i, \tilde{\boldsymbol{X}}_i)$, $i = 1, \ldots, \tilde{n}$ be a data set of $(Y, \boldsymbol{X})$, which can come from $(\boldsymbol{P}_i, Y_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$ or an extra data set. We permute the data as $(\tilde{Y}_{(1),i}, \tilde{\boldsymbol{X}}_{(1),i}), \ldots, (\tilde{Y}_{(\tilde{n}),i}, \tilde{\boldsymbol{X}}_{(\tilde{n}),i})$ such that $\|\tilde{\boldsymbol{X}}_{(1),i} - \boldsymbol{X}_i\| \leq \ldots \leq \|\tilde{\boldsymbol{X}}_{(\tilde{n}),i} - \boldsymbol{X}_i\|$. Then, we can draw $Y_i^{(b)}$ randomly from the $\tilde{k}$-nearest neighbors $\{\tilde{Y}_{(1),i}, \ldots, \tilde{Y}_{(\tilde{k}),i}\}$. After we have $Y_i^{(b)}$, $b = 1, \ldots, B$ and $i = 1, \ldots, n$, we can follow the same procedure in the CRT framework. The main difference between the NNB method and CRT framework is that we use the empirical distribution of $\{\tilde{Y}_{(1),i}, \ldots, \tilde{Y}_{(\tilde{k}),i}\}$ to approximate unknown $\mathbb{P}(Y|\boldsymbol{X}_i)$ and do not need to estimate the conditional distribution. The NNB method shares many similarities with the local permutation method (Doran et al., 2014; Sen et al., 2017; Kim et al., 2021), where Y is permuted locally within the small bins divided based on $\boldsymbol{X}$. Compared with the local permutation method, the NNB method has two advantages: finding the $\tilde{k}$-nearest neighbor is usually much simpler than bins construction; the NNB method can take advantage of the extra data set of $(Y, \boldsymbol{X})$ when available.

In summary, this section discusses three ways to calibrate test statistics $\Psi_{\Lambda,1}$ and $\Psi_{\Lambda,\infty}$: CRT framework, asymptotic method, and NNB method. Among these three, we recommend the NNB method as it is easy to use and can incorporate data sets from other studies.

5. Numerical Experiments

5.1. *Simulation Studies*

We now investigate the properties of the proposed independence and conditional independence tests through simulation studies. We consider a phylogenetic tree of microbes within the class *Gammaproteobacteria* as the underlying tree T. This phylogenetic tree is a subtree from a commonly used reference tree of Greengenes 16S rRNA database version 13.8 clustered at 85% similarity. There are 247 leaves, 246 internal nodes, and 492 edges in tree T, shown in Figure 2.

To simulate compositional/relative abundance data, we consider the Dirichlet-multinomial model in this section. Specifically, given a non-negative vector $\boldsymbol{\alpha} = \{\alpha_v\}_{v \in V}$, the true relative abundance $\boldsymbol{P}$ is drawn from a Dirichlet distribution indexed by $\boldsymbol{\alpha}$. The count data of each type of microbes $\boldsymbol{N} = \{N_v\}_{v \in V}$ are drawn from a multinomial distribution with respect to the true relative abundance $\boldsymbol{P}$. Then, we normalize the count data $\boldsymbol{N}$ as empirical relative abundance data $\hat{\boldsymbol{P}}$. Given the above model, we only need to specify $\boldsymbol{\alpha}$, Y, and $\boldsymbol{X}$ in simulation studies of independence and conditional independence test.

Independence Test In the simulation experiments of the independence test, we first draw Y from a uniform distribution between 0 and 1. Given Y and two clades S_1 and S_2 of tree T, we choose $\boldsymbol{\alpha}$ in the following way

$$\alpha_v = \begin{cases} 1 + f(Y), & v \in S_1 \\ 1 - f(Y), & v \in S_2 \\ 1, & v \in V_L \setminus (S_1 \cup S_2) \\ 0, & v \in V_I \end{cases}.$$

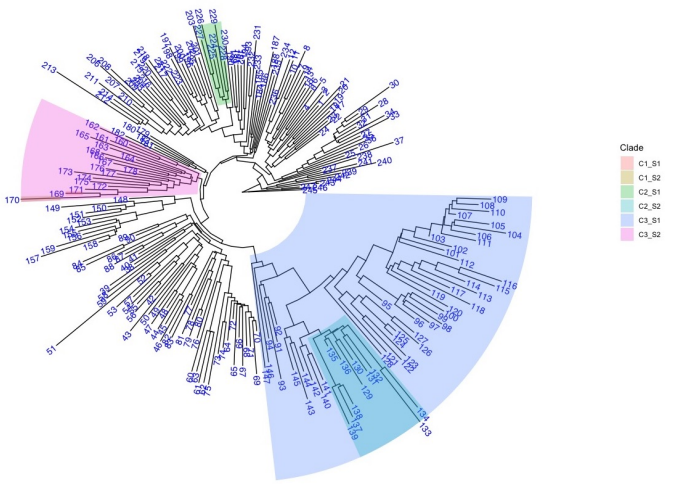


Fig. 2: The phylogenetic tree of microbes within the class *Gammaproteobacteria*. The leaf nodes are labeled by numbers and clade used in simulation studies are highlighted by different colors.

Here, $f : \mathbb{R} \to \mathbb{R}$ is a link function, V_I is the collection of internal nodes of tree T, and V_L is the collection of leaves of tree T. We can choose different combinations of f, S_1, and S_2 to study different aspects of our proposed independence tests. We compare eight tests in the simulation experiments of the independence test. In particular, we consider six rank-based phylogenetic independence tests proposed in Section 3: $\Psi_1(D_n)$, $\Psi_1(\tau_n^*)$, $\Psi_1(R_n)$, $\Psi_\infty(D_n)$, $\Psi_\infty(\tau_n^*)$, and $\Psi_\infty(R_n)$. We also consider one of the most popular distance-based association analysis methods, PERMANOVA (implemented in R package `vegan`), equipped with weighted Unifrac distance, its L_2 Zolotarev-type generalization, generalized UniFrac distances ($\alpha = 0, 0.5$ in Chen et al. (2012)) and Bray-Curtis distance, denoted by PL_1, PL_2, $PL_{G,0}$, $PL_{G,0.5}$ and PL_{BC}, respectively. We adopt the permutation test in all eight tests and reject the null when the P-value is smaller than 0.05.

The first set of simulation experiments is designed to study the performance of different independence tests when the dependence structure between $\hat{P}$ and Y is linear or nonlinear. More concretely, we consider the clade $S_1 = \{224, 228, 229, 225, 226, 227\}$, $S_2 = \{139, 135, 136, 137, 138, 129, 130, 131, 132, 133, 134\}$, and three different link functions:

(a) M1: $f(Y) = 3/4 - \delta Y$;
(b) M2: $f(Y) = \delta \sin(\pi Y)$;
(c) M3: $f(Y) = \delta \sin(2\pi Y)$.

Here, δ is a number between 0 and 1 to characterize the strength of the dependency. There is a stronger dependency between $\hat{P}$ and Y when δ is larger. The dependence structure between $\hat{P}$ and Y is linear under M1 and nonlinear under M2 and M3. To investigate the effect of sample size n and signal strength δ, we choose $n = 50, 100$, and $\delta = 0.2, 0.4, 0.6, 0.8, 1$ in the simulation experiments. We repeat the simulation experiments 500 times for each combination of f, n, and δ. The performance of different tests is evaluated by the power of the test, that is, the probability of rejecting the null hypothesis. The experiment results are summarized in Table S1. Table S1 shows that rank-based phylogenetic independence tests achieve similar power with PERMANOVA when the association between $\hat{P}$ and Y is linear (M1), especially when the sample size is small. However, for non-linear associations (M1 and M2), Table S1 shows that the

proposed rank-based phylogenetic independence tests are in general more or equally powerful than the distance-based method PERMANOVA.

In the second set of simulation studies, we mainly investigate the effect on the signal's sparsity level. In particular, we consider the link function $f(Y) = \delta \sin(3\pi Y)$, where δ is signal strength and three choices of the clade

(a) C1: $S_1 = \{169\}, S_2 = \{170\}$;
(b) C2: $S_1 = \{224, 228, 229, 225, 226, 227\}, S_2 = \{139, 135, 136, 137, 138, 129, 130, 131, 132, 133, 134\}$;
(c) C3: $S_1 = \{147, 146, 94, 93, 128, 91, 92, 145, 121, 142, 143, 144, 95, 139, 140, 141, 96, 112, 120, 122, 123, 135, 136, 137, 138, 101, 119, 124, 125, 126, 127, 129, 130, 97, 98, 99, 100, 102, 118, 131, 132, 133, 134, 103, 117, 111, 113, 114, 106, 107, 115, 116, 104, 105, 110, 108, 109\}, S_2 = \{162, 160, 161, 178, 165, 168, 163, 164, 166, 167, 176, 177, 172, 173, 171, 174, 175, 169, 170\}$.

These three choices of the clade are highlighted in Figure 2. In these three choices of the clade, C1 represents the sparsest setting, and C3 represents the densest setting. We set $n = 100, 200$ and $\delta = 0.2, 0.4, 0.6, 0.8, 1$ in the simulation experiments and still repeat the simulation experiments 500 times for each combination of S_1, S_2, n, and δ. We summarize the power of the eight tests in Table S2. Table S2 suggests that the maximum type test Ψ_∞ can capture the dependency more efficiently than the weighted sum type test Ψ_1 when the signal is sparse, while Ψ_1 is more powerful than Ψ_∞ when the signal is dense. However, under the non-linear link function, PERMANOVA has almost no power in detecting the dependency structure.

Conditional Independence Test In the conditional independence test simulation experiments, we choose $d = 5$ and draw each entry of $\boldsymbol{X} = (X_1, \ldots, X_d)$ from a uniform distribution between 0 and 1. Given $\boldsymbol{X}$, we draw $Y \sim \mathcal{N}(\mu(\boldsymbol{X}), 1)$. Given $\boldsymbol{X}$, Y, and two clades S_1 and S_2 of tree T, we choose $\boldsymbol{\alpha}$ in the following way

$$\alpha_v = \begin{cases} 1 + g(\boldsymbol{X}, Y), & v \in S_1 \\ 1 - g(\boldsymbol{X}, Y), & v \in S_2 \\ 1, & v \in V_L \setminus (S_1 \cup S_2) \\ 0, & v \in V_I \end{cases},$$

where $g(\boldsymbol{X}, Y)$ is a link function. In the following simulation experiments, we can choose different combinations of $\mu(\cdot)$, $g(\cdot, \cdot)$, S_1, and S_2.

We first investigate if different conditional independence tests can control the type I error under the null hypothesis. We choose $S_1 = \{224, 228, 229, 225, 226, 227\}$, $S_2 = \{139, 135, 136, 137, 138, 129, 130, 131, 132, 133, 134\}$, $\mu(\boldsymbol{X}) = dg(\boldsymbol{X}, Y) = h(\boldsymbol{X})$ for function h. In particular, we consider four different choices of h:

(a) M1: $h(\boldsymbol{X}) = \sum_{i=1}^d X_i$;
(b) M2: $h(\boldsymbol{X}) = \sum_{i=1}^d \sin(\pi X_i)$;
(c) M3: $h(\boldsymbol{X}) = \sum_{i=1}^d \sin(2\pi X_i)$;
(d) M4: $h(\boldsymbol{X}) = \sum_{i=1}^d \sin(3\pi X_i)$.

In these four choices of $h(\boldsymbol{X})$, the degree of nonlinear increases from M1 to M4. Under these settings, we have $\boldsymbol{P} \perp Y | \boldsymbol{X}$ for any $\boldsymbol{X}$. Based on the simulation independence test results, we consider the phylogenetic association analysis methods defined by Blum-Kiefer-Rosenblatt's R: $\Psi_{\Lambda,1}(R_n)$ and $\Psi_{\Lambda,\infty}(R_n)$. We consider two ways of making decisions: the CRT framework and the NNB method. We compare the newly proposed method with two state-of-the-art association

analysis methods: PERMANOVA and MiRKAT (both can adjust the confounding effect in $\boldsymbol{X}$). Here we construct a distance matrix by weighted Unifrac distance and its L_2 Zolotarev-type generalization in PERMANOVA and MiRKAT. To compare different association analysis methods, we set $n = 100$, the significance level as 0.05, and estimate type I error from 300 times simulation experiments. The simulation results are summarized in Table 1. In Table 1, only CRT can control type I error in all settings as we know the exact conditional distribution $\mathbb{P}(Y|\boldsymbol{X})$. When the confounding effect is linear, all methods can adjust the confounding effect very well. In the absence of knowledge on $\mathbb{P}(Y|\boldsymbol{X})$, NNB can help better control false discoveries than PERMANOVA and MiRKAT when the confounding effect is highly nonlinear.

Table 1: Comparison of type I error control by different conditional independence test methods.

		M1	M2	M3	M4
NNB	$\Psi_{\Lambda,\infty}(R_n)$	0.047	0.053	0.050	0.063
	$\Psi_{\Lambda,1}(R_n)$	0.070	0.066	0.113	0.246
	$\Psi_{\Lambda,\infty}(R_n)$	0.043	0.076	0.223	0.485
PERMANOVA	L_1	0.080	0.405	0.748	1.000
	L_2	0.070	0.306	0.731	1.000
MiRKAT	L_1	0.080	0.405	0.767	1.000
	L_2	0.073	0.319	0.757	1.000

The next simulation experiment studies the effect of sample size $\tilde{n}$ and the number of neighborhood $\tilde{k}$ in the NNB method. We adopt the same setting as in the last simulation experiment and only focus on two choices of h: M2 and M3. We choose $\tilde{n} = 100, 1000, 10000$ and $\tilde{k} = 10, 30, 100$ in NNB method. The type I errors estimated from 300 repeated simulation experiments is summarized in Table 2. Table 2 suggests a large extra data set is helpful for type I error control, and we shall not choose the number of the neighborhood too large.

Table 2: Comparison of type I error control by different choices of parameters in NNB.

		$\Psi_{\Lambda,1}(R_n)$			$\Psi_{\Lambda,\infty}(R_n)$		
	$\tilde{n}$	$\tilde{k} = 10$	$\tilde{k} = 30$	$\tilde{k} = 100$	$\tilde{k} = 10$	$\tilde{k} = 30$	$\tilde{k} = 100$
M2	100	0.070	0.076	0.066	0.070	0.047	0.033
	1000	0.073	0.073	0.076	0.100	0.093	0.063
	10000	0.063	0.076	0.063	0.080	0.090	0.083
M3	100	0.316	0.375	0.385	0.708	0.684	0.691
	1000	0.193	0.269	0.312	0.515	0.618	0.674
	10000	0.106	0.126	0.163	0.233	0.322	0.458

We now investigate the performance of different methods when there is a highly nonlinear conditional association. We still choose S_1 and S_2 as the previous two simulation experiments. To construct a highly nonlinear case, we choose $\mu(\boldsymbol{X}) = \sum_{i=1}^{d} \sin(2\pi X_i)$ and

$$g(\boldsymbol{X}, Y) = (1-\delta)\frac{\sum_{i=1}^{d}\cos(2\pi X_i)}{d} + \delta\cos(\pi Y/3)$$

for some $0 < \delta < 1$. These choices suggest $\boldsymbol{P} \not\perp Y|\boldsymbol{X}$ for some $\boldsymbol{X}$, and the conditional dependency is stronger when δ is larger in this setting. We still compare the same eight phylogenetic conditional independence tests as we did in the previous simulation experiments and chose

$n = 100$, the significance level as 0.05. Here, we compare the power of different methods estimated by 300 times simulation experiments. Table 3 summarizes different methods' power when $\delta = 0.2, 0.4, 0.6, 0.8, 1$. From Table 3, we can conclude that the newly proposed tests are more sensitive to the nonlinear conditional association than existing methods.

Table 3: Power comparison when the conditional association is highly nonlinear.

		$\delta = 1$	$\delta = 0.8$	$\delta = 0.6$	$\delta = 0.4$	$\delta = 0.2$
CRT	$\Psi_{\Lambda,1}(R_n)$	1.000	0.977	0.854	0.336	0.113
	$\Psi_{\Lambda,\infty}(R_n)$	1.000	1.000	0.977	0.558	0.076
NNB	$\Psi_{\Lambda,1}(R_n)$	1.000	0.990	0.894	0.355	0.103
	$\Psi_{\Lambda,\infty}(R_n)$	1.000	1.000	0.980	0.575	0.076
PERMANOVA	L_1	0.033	0.053	0.063	0.037	0.086
	L_2	0.037	0.063	0.060	0.040	0.076
MiRKAT	L_1	0.033	0.056	0.060	0.043	0.090
	L_2	0.043	0.060	0.066	0.037	0.073

5.2. *Analysis of Real Data*

To demonstrate its practical merit, we apply the proposed method to a gut microbiome data set of 7009 samples collected in the Guangdong Gut Microbiome Project (He et al., 2018). The raw sequence reads are denoised and placed on a reference phylogenic tree from the Greengenes database. The resulting data sets, including 37532 ASVs and a corresponding phylogenetic tree, can be downloaded from Qiita (https://qiita.ucsd.edu/) under study ID 11757. In this section, we focus on the investigation of the association between the human gut microbiome and the host's diet, including red wine and salt consumption. In the association study, we also adjust the potential confounding effect of age, as middle-aged people have higher salt diets and drink more red wine than other age levels (Figure 3).

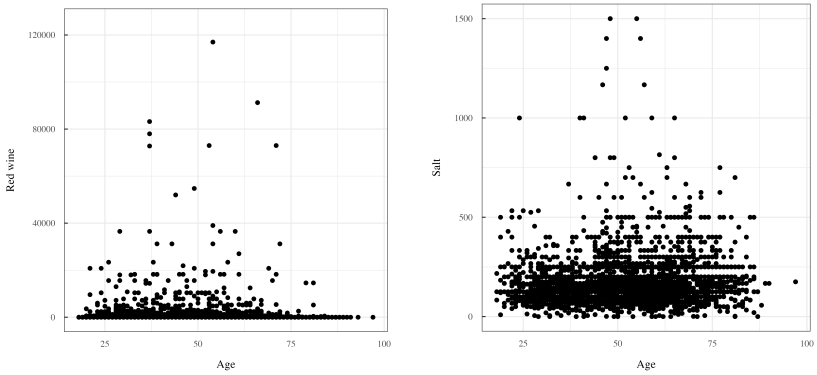


Fig. 3: Scatter plot of red wine and salt consumption against age. Left figure is age vs red wine and right figure is age vs salt.

To test the association between the human gut microbiome and the host's diet, we consider the six rank-based phylogenetic conditional independence tests: $\Psi_{\Lambda,1}(D_n)$, $\Psi_{\Lambda,1}(\tau_n^*)$, $\Psi_{\Lambda,1}(R_n)$, $\Psi_{\Lambda,\infty}(D_n)$, $\Psi_{\Lambda,\infty}(\tau_n^*)$, and $\Psi_{\Lambda,\infty}(R_n)$, and MiRKAT equipped with weighted Unifrac distance and its L_2 Zolotarev-type generalization. The P-values in rank-based phylogenetic conditional

independence tests are calculated using 200 permutations. To compare the performance in different sample size, we consider three different data set: 1) 2000 randomly selected samples, 2) 4000 randomly selected samples, and 3) all samples. Table 4 summarizes the resulting P-values.

Table 4: The P-values for association analysis using different methods.

Method	Red Wine			Salt		
	$n = 2000$	$n = 4000$	All	$n = 2000$	$n = 4000$	All
$\Psi_{\Lambda,1}(D_n)$	0.1940	0.0149	0.0050	0.0647	0.0050	0.0050
$\Psi_{\Lambda,\infty}(D_n)$	0.6269	0.1393	0.1244	0.8856	0.6269	0.6070
$\Psi_{\Lambda,1}(R_n)$	0.1294	0.0100	0.0050	0.0050	0.0050	0.0050
$\Psi_{\Lambda,\infty}(R_n)$	0.6418	0.0597	0.1244	0.7662	0.5373	0.1940
$\Psi_{\Lambda,1}(T_n)$	0.0896	0.0100	0.0050	0.0299	0.0050	0.0050
$\Psi_{\Lambda,\infty}(T_n)$	0.5871	0.1194	0.1343	0.8507	0.5572	0.3134
MiRKAT (L_1)	0.4049	0.3146	0.4690	0.5934	0.0580	0.0002
MiRKAT (L_2)	0.3208	0.1970	0.4148	0.6277	0.0691	0.0003

When all samples are used in the analysis, all sum type rank-based phylogenetic conditional independence tests and MiRKAT indicate an association between salt diet and the gut microbiome, which is consistent with previous findings that gut microbial composition can be reshaped by a high-salt diet and is a key linkage in the relationship between high blood pressure and sodium intake (Smiljanec & Lennon, 2019; Jama & Marques, 2020; Yan et al., 2020; Chen et al., 2020). However, if the sample size is reduced to 2000 or 4000, only rank-based tests can identify the association between sodium intake and gut microbiome, while MiRKAT does not detect such a nonlinear association (Figure 4). Besides the salt dietary's effect, the rank-based tests also identify an association between red wine consumption and gut microbiome when the sample size is larger than 4000. The red wine's effect on the gut microbiome has been well studied, and it is believed that red wine polyphenols are related to several beneficial and pathogenic bacteria (Cardona et al., 2013; Nash et al., 2018; Le Roy et al., 2020). Nevertheless, the MiRKAT does not report such an association as the relationship between red wine consumption and the gut microbiome is highly nonlinear (Figure 4). The results in Table 4 suggest that the proposed method is more powerful in detecting nonlinear association and can potentially lead to more scientific discoveries in microbiome studies.

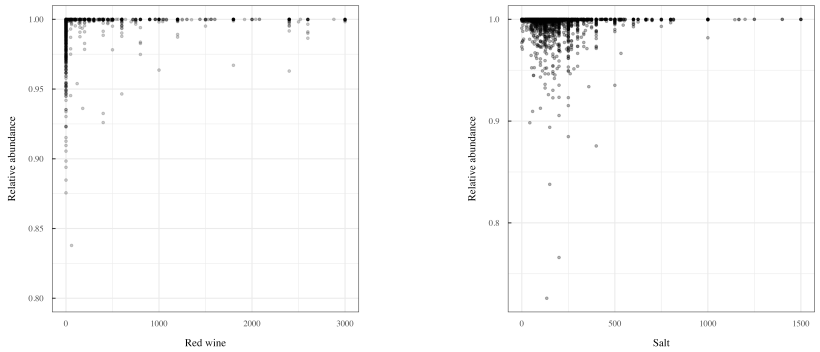


Fig. 4: Scatter plot of red wine and salt consumption against the relative abundance of the most significant subtree. Left figure is red wine and right figure is salt.

ACKNOWLEDGMENT

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

SUPPLEMENTARY MATERIAL

Supplementary Material available at Biometrika online includes proofs of all theorems, technical assumptions, and numerical experiments results.

REFERENCES

AZADKIA, M. & CHATTERJEE, S. (2019). A simple measure of conditional dependence. *arXiv preprint arXiv:1910.12327* .

BERGSMA, W. & DASSIOS, A. (2014). A consistent test of independence based on a sign covariance related to kendall's tau. *Bernoulli* **20**, 1006–1028.

BIAU, G. & DEVROYE, L. (2015). *Lectures on the nearest neighbor method*, vol. 246. Springer.

BLUM, J. R., KIEFER, J. & ROSENBLATT, M. (1961). Distribution free tests of independence based on the sample distribution function. *The Annals of Mathematical Statistics* , 485–498.

BREIMAN, L. & FRIEDMAN, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association* **80**, 580–598.

CANDES, E., FAN, Y., JANSON, L. & LV, J. (2018). Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80**, 551–577.

CARDONA, F., ANDRÉS-LACUEVA, C., TULIPANI, S., TINAHONES, F. J. & QUEIPO-ORTUÑO, M. (2013). Benefits of polyphenols on gut microbiota and implications in human health. *The Journal of Nutritional Biochemistry* **24**, 1415–1422.

CHATTERJEE, S. (2020). A new coefficient of correlation. *Journal of the American Statistical Association* , 1–21.

CHEN, J., BITTINGER, K., CHARLSON, E. S., HOFFMANN, C., LEWIS, J., WU, G. D., COLLMAN, R. G., BUSHMAN, F. D. & LI, H. (2012). Associating microbiome composition with environmental covariates using generalized unifrac distances. *Bioinformatics* **28**, 2106–2113.

CHEN, L., HE, F. J., DONG, Y., HUANG, Y., WANG, C., HARSHFIELD, G. A. & ZHU, H. (2020). Modest sodium reduction increases circulating short-chain fatty acids in untreated hypertensives: a randomized, double-blind, placebo-controlled trial. *Hypertension* **76**, 73–79.

DORAN, G., MUANDET, K., ZHANG, K. & SCHÖLKOPF, B. (2014). A permutation-based kernel conditional independence test. In *UAI*.

DRTON, M., HAN, F. & SHI, H. (2018). High-dimensional consistent independence testing with maxima of rank correlations. *arXiv* , arXiv–1812.

EL KAROUI, N. (2010). The spectrum of kernel random matrices. *The Annals of Statistics* **38**, 1–50.

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.

EVEN-ZOHAR, C. (2020). independence: Fast rank tests. *arXiv preprint arXiv:2010.09712* .

GRETTON, A., BOUSQUET, O., SMOLA, A. & SCHÖLKOPF, B. (2005). Measuring statistical dependence with hilbert-schmidt norms. In *International Conference on Algorithmic Learning Theory*. Springer.

GRETTON, A., FUKUMIZU, K., TEO, C. H., SONG, L., SCHÖLKOPF, B. & SMOLA, A. J. (2008). A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*.

HE, Y., WU, W., ZHENG, H., LI, P., MCDONALD, D., SHENG, H., CHEN, M., CHEN, Z., JI, G., ZHENG, Z., MUJAGOND, P., CHEN, X., RONG, Z., CHEN, P., LYU, L., WANG, X., WU, C., YU, N., XU, Y., YIN, J., RAES, J., KNIGHT, R., MA, W. & ZHOU, H. (2018). Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nature Medicine* **24**, 1532–1535.

HOEFFDING, W. (1948). A non-parametric test of independence. *The Annals of Mathematical Statistics* , 546–557.

HUANG, Z., DEB, N. & SEN, B. (2020). Kernel partial correlation coefficient–a measure of conditional dependence. *arXiv preprint arXiv:2012.14804* .

JAMA, H. A. & MARQUES, F. Z. (2020). Don't take it with a pinch of salt: how sodium increases blood pressure via the gut microbiota. *Circulation Research* **126**, 854–856.

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.

KIM, I., NEYKOV, M., BALAKRISHNAN, S. & WASSERMAN, L. (2021). Local permutation tests for conditional independence. *arXiv preprint arXiv:2112.11666* .

KOH, H., BLASER, M. J. & LI, H. (2017). A powerful microbiome-based association test and a microbial taxa discovery framework for comprehensive association mapping. *Microbiome* **5**, 1–15.

LE ROY, C. I., WELLS, P. M., SI, J., RAES, J., BELL, J. T. & SPECTOR, T. D. (2020). Red wine consumption associated with increased gut microbiota α-diversity in 3 independent cohorts. *Gastroenterology* **158**, 270–272.

LI, H. (2015). Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application* **2**, 73–94.

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.

LYONS, R. (2013). Distance covariance in metric spaces. *The Annals of Probability* **41**, 3284–3305.

MATSEN, F. A., KODNER, R. B. & ARMBRUST, E. V. (2010). pplacer: linear time maximum-likelihood and bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC bioinformatics* **11**, 1–16.

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*. World Scientific, pp. 247–258.

NASH, V., RANADHEERA, C. S., GEORGOUSOPOULOU, E. N., MELLOR, D. D., PANAGIOTAKOS, D. B., MCKUNE, A. J., KELLETT, J. & NAUMOVSKI, N. (2018). The effects of grape and red wine polyphenols on gut microbiota–a systematic review. *Food Research International* **113**, 277–287.

PAN, W. (2011). Relationship between genomic distance-based regression and kernel machine regression for multimarker association testing. *Genetic epidemiology* **35**, 211–216.

PFLUGHOEFT, K. J. & VERSALOVIC, J. (2012). Human microbiome in health and disease. *Annual Review of Pathology: Mechanisms of Disease* **7**, 99–122.

RÉNYI, A. (1959). On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica* **10**, 441–451.

RESHEF, D. N., RESHEF, Y. A., FINUCANE, H. K., GROSSMAN, S. R., MCVEAN, G., TURNBAUGH, P. J., LANDER, E. S., MITZENMACHER, M. & SABETI, P. C. (2011). Detecting novel associations in large data sets. *Science* **334**, 1518–1524.

ROMANO, J. P. (1988). A bootstrap revival of some nonparametric distance tests. *Journal of the American Statistical Association* **83**, 698–708.

SEN, R., SURESH, A. T., SHANMUGAM, K., DIMAKIS, A. G. & SHAKKOTTAI, S. (2017). Model-powered conditional independence test. *Advances in Neural Information Processing Systems* **30**.

SHAH, R. D. & PETERS, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics* **48**, 1514–1538.

SHI, H., DRTON, M. & HAN, F. (2020). On the power of chatterjee rank correlation. *arXiv preprint arXiv:2008.11619* .

SMILJANEC, K. & LENNON, S. L. (2019). Sodium, hypertension, and the gut: does the gut microbiota go salty? *American Journal of Physiology-Heart and Circulatory Physiology* **317**, H1173–H1182.

SONG, Y., ZHAO, H. & WANG, T. (2020). An adaptive independence test for microbiome community data. *Biometrics* **76**, 414–426.

SU, L. & WHITE, H. (2008). A nonparametric hellinger metric test for conditional independence. *Econometric Theory* **24**, 829–864.

SZÉKELY, G. J. & RIZZO, M. L. (2009). Brownian distance covariance. *The Annals of Applied Statistics* , 1236–1265.

SZÉKELY, G. J., RIZZO, M. L. & BAKIROV, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics* **35**, 2769–2794.

TANG, Z., CHEN, G. & ALEKSEYENKO, A. V. (2016). Permanova-s: association test for microbial community composition that accommodates confounders and multiple distances. *Bioinformatics* **32**, 2618–2625.

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

WANG, S. (2021). Multi-scale adaptive differential abundance analysis in microbial compositional data. *bioRxiv* .

WANG, S. (2022). Robust differential abundance test in compositional data. *Biometrika* **accepted**.

WANG, S., CAI, T. T. & LI, H. (2021a). Hypothesis testing for phylogenetic composition: a minimum-cost flow perspective. *Biometrika* **108**, 17–36.

WANG, S., CAI, T. T. & LI, H. (2021b). Optimal estimation of wasserstein distance on a tree with an application to microbiome studies. *Journal of the American Statistical Association* **116**, 1237–1253.

WANG, X., PAN, W., HU, W., TIAN, Y. & ZHANG, H. (2015). Conditional distance correlation. *Journal of the American Statistical Association* **110**, 1726–1734.

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**, 652–661.

WEIHS, L., DRTON, M. & MEINSHAUSEN, N. (2018). Symmetric rank covariances: a generalized framework for nonparametric measures of dependence. *Biometrika* **105**, 547–562.

WU, C., CHEN, J., KIM, J. & PAN, W. (2016). An adaptive association test for microbiome data. *Genome Medicine* **8**, 56.

YAN, X., JIN, J., SU, X., YIN, X., GAO, J., WANG, X., ZHANG, S., BU, P., WANG, M., ZHANG, Y., WANG, Z. & ZHANG, Q. (2020). Intestinal flora modulates blood pressure by regulating the synthesis of intestinal-derived corticosterone in high salt-induced hypertension. *Circulation research* **126**, 839–853.

YANAGIMOTO, T. (1970). On measures of association and a related problem. *Annals of the Institute of Statistical Mathematics* **22**, 57–63.

ZHAO, N., CHEN, J., CARROLL, I. M., RINGEL-KULKA, T., EPSTEIN, M. P., ZHOU, H., ZHOU, J. J., RINGEL, Y., LI, H. & 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.

ZHOU, Y., LIU, J. & ZHU, L. (2020). Test for conditional independence with application to conditional screening. *Journal of Multivariate Analysis* **175**, 104557.

ZHU, C., ZHANG, X., YAO, S. & SHAO, X. (2020). Distance-based and rkhs-based dependence metrics in high dimension. *The Annals of Statistics* **48**, 3366–3394.