Efficient Exact Tests in Linear Mixed Models for Longitudinal Microbiome Studies

by

Jing Zhai

A Thesis Submitted to the Faculty of the

MEL AND ENID ZUCKERMAN

COLLEGE OF PUBLIC HEALTH

in Partial Fulfillment of the Requirements

For the Degree of

MASTER OF SCIENCE

with a Major in Biostatistics

in the Graduate College

THE UNIVERSITY OF ARIZONA

2016

STATEMENT BY AUTHOR

The thesis titled *Efficient Exact Tests in Linear Mixed Models for Longitudinal Microbiome Studies* prepared by *Jing Zhai* has been submitted in partial fulfillment of requirements for a master's degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this thesis are allowable without special permission, provided that an accurate acknowledgement of the source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: Jing Zhai

APPROVAL BY THESIS DIRECTOR

This thesis has been approved on the date shown below:

Jin Zhou	April 28th, 2016
Assistant Professor of Biostatistics	Date

Acknowledgements

First and foremost, I would like thank my thesis committee chair Dr. Jin Zhou. Without her assistance and dedicated involvement in every step throughout the process, this paper would have never been accomplished. I appreciate her support and understanding over the past year. I would also like to show gratitude to my committee, including Dr. Denise Roe and Dr. Chengcheng Hu. Thanks for being on my thesis committee and giving me help and advice during my thesis work.

I am using this opportunity to express my gratitude to everyone who helped me throughout my study in the Biostatistics Master program. I am grateful for their aspiring guidance, invaluably constructive criticism and friendly advice during my two years' study. I am sincerely thankful to them for sharing their truthful and illuminating views on a number of issues related to my life and study here.

Thanks,

Jing Zhai

Contents

ABSTRACT

Microbiome plays an important role in human health. The analysis of association between microbiome and clinical outcome has become an active direction in biostatistics research. Testing the microbiome effect on clinical phenotypes directly using operational taxonomic unit abundance data is a challenging problem due to the high dimensionality, non-normality and phylogenetic structure of the data. Most of the studies only focus on describing the change of microbe population that occur in patients who have the specific clinical condition. Instead, a statistical strategy utilizing distance-based or similarity-based non-parametric testing, in which a distance or similarity measure is defined between any two microbiome samples, is developed to assess association between microbiome composition and outcomes of interest. Despite the improvements, this test is still not easily interpretable and not able to adjust for potential covariates. A novel approach, kernel-based semi-parametric regression framework, is applied in evaluating the association while controlling the covariates. The framework utilizes a kernel function which is a measure of similarity between samples' microbiome compositions and characterizes the relationship between the microbiome and the outcome of interest. This kernel-based regression model, however, cannot be applied in longitudinal studies since it could not model the correlation between the repeated measurements.

We proposed microbiome association exact tests (MAETs) in linear mixed model can deal with longitudinal microbiome data. MAETs can test not only the effect of overall microbiome but also the effect from specific cluster of the OTUs while controlling for others by introducing more random effects in the model. The current methods for multiple variance component testing are based on either asymptotic distribution or parametric bootstrap which require large sample size or high computational cost. The exact (R)LRT tests, an computational efficient and powerful testing methodology, was derived by $Crainiceanu$. Since the exact (R)LRT can only be used in

testing one variance component, we proposed an approach that combines the recent development of exact (R)LRT and a strategy for simplifying linear mixed model with multiple variance components to a single case.

The Monte Carlo simulation studies present correctly controlled type I error and provided superior power in testing association between microbiome and outcomes in longitudinal studies. Finally, the MAETs were applied to longitudinal pulmonary microbiome datasets to demonstrate that microbiome composition is associated with lung function and immunological outcomes. We also successfully found two interesting genera $Prevotella$ and $Veillonella$ which are associated with forced vital capacity.

Key Words: Longitudinal study; Microbiome composition; Kernel-based regression; Multiple variance components; Exact tests.

1

Background and Introduction

1.1 Microbiome and Human Health

The human body contains trillions of microbial cells which is over 10 times more than human cells. The collection of genomes of these microbial symbionts is defined as the microbiome[1]. Microbiome is important for maintaining human health, and disorder of microbiome might contribute to disease. Since late 1990s, studies showed that microbiota in the gut might play an important part in the human immune system and the microbiome of the gut has been characterised as a "forgotten organ"[2, 3, 4]. Currently, most studies are focusing on describing the variant microbe populations that occur in patients who have a specific disease, or the temporal microbial changes are observed over the course of a disease. Larger populations of Fusobacterium $spp.$ was discovered in colorectal carcinoma patients[5]. The ratio of $Bacteroidetes$ to $Firmicutes$ in gut microbiome of obsity patients was reduced based on a number of studies[4, 6]. Absent gastric $H.pylori$ may cause childhood-onset asthma[7, 8].

1.2 Microbiome Profiling

To address any question about the human microbiome, the microbiota first needs to be sampled. The traditional culture approach works well for microbes that grow easily in the lab, but most microbes cannot be grown in culture[9]. The culture-independent techniques, such as the next generation sequencing, can help us to know more about the microbial make up of the human body and how microbes influence human disease[10]. These techniques enable high-throughput profiling of microbial communities via direct DNA sequencing[11]. For the 16S rRNA gene targeted sequencing approach, the 16S tags from the same species are highly similar[12]. The tags

can be clustered based on their sequence similarity to form the so-called operational taxonomic units (OTUs)[13, 14]. At 97% similarity level, OTUs are considered as the biological species.

1.3 UniFrac Distance

Knowledge of how microbial communities differ across individuals can provide key insights on the role of the microbiome in relation to variation in biological and clinical variables[15]. The OTUs are related by their phylogenetic tree which provides their lineage information. It would be inappropriate to study the difference of microbiome composition between samples simply using the OTUs abundance data. Thus, pairwise distance matrix are built to measure this dissimilarity between each pair of samples by taking the phylogenetic relationship into account. The most popular distance matrix is the UniFrac distance. There are several types of UniFrac distance widely applied to microbiome studies recently, such as unweighted UniFrac [16], weighted UniFrac distance[17], variance adjusted weighted UniFrac distance[18] and generalized UniFrac distance[19]. Unweighted UniFrac distance proposed by $Lozupone$ is most efficient in detecting abundance change in rare lineages[16], while the weighted UniFrac distance is most sensitive to detect difference[17]. However, for detecting change in moderate abundant lineages, the unweighted/weighted UniFrac distances may not be as powerful as variance adjusted weighted UniFrac distance (VAW-UniFrac). The VAW-UniFrac distance moderates the change of branch proportion by its variance[18]. The new generalized UniFrac distances introduced by $J\ Chen\ et\ al$ contain a series of distances ranging from unweighted to weighted UniFrac via adjusting the weight on the phylogenetic tree branches. The generalized UniFrac distances are a powerful tool for detecting a wider range of changes in microbiome composition by tuning the branch control parameter[19].

1.4 Statistical Method Based on UniFrac Distance

Due to non-normality[20], high dimensionality[21] and phylogenetic structure of the OTU data, it is difficult to test the association of microbiome composition with potential biomedical conditions directally using OTU abundances. A prevalent strategy, called distance-based non-parametric test, is developed to identify potential biological and exposure factors that shape

the microbiome composition[22, 23, 24]. The permutational multivariate analysis of variance procedure is applied in the test utilizing the distance matrix introduced above in which the dissimilarity between samples is defined[25]. The procedure includes partitioning the distance matrix among sources of variation, fitting linear models to distance matrices and performing a permutation test with pseudo-F ratios to obtain the p-values. Despite the improvement of the distance-based method, the permutation procedure is computationally expensive and not able to adjust for potential covariates. Besides, the power of the distance-based non-parametric test rests on a proper choice of the distance type. An inappropriate distance matrix has poor performance on evaluating the true associations.

1.5 Statistical Regression Method Based on Microbiome Profile Kernel

$J\ Chen\ et\ al.$ then extended such approaches to the kernel-based semi-parametric regression method in order to accommodate more-sophisticated outcomes and adjust for the potential covariates[26]. It evaluates the association of overall microbiome and outcomes of interest by directly regressing the microbiome profiles represented by a kernel matrix transformed from the pairwise distance matrix[19]. In this linear mixed effect model, the coefficients for covariates are fixed effects and the non-parametrically kernel function captured microbiome effect are modeled as subject-specific random effects with the variance component. Therefore, testing no microbiome effect on the phenotype becomes testing the zero variance component. A score test and the score statistic is used in the variance component test[26]. This kind of overall test can outperform than the multiple testing of individual associations between microbial taxon and outcomes especially when the association are very weak. However, it does not overcome the problem of power reduction caused by improper distance measure. Choosing a particular type of distance matrix which does not meet the true state of nature may reduce power of association testing.

$N\ Zhao\ et\ al$ provided a more flexible and powerful regression approach, the optimal microbiome regression based kernel association test (optimal MiRKAT), for testing the association between microbiome diversity and biomedical conditions[27]. The idea behind the optimal

MiRKAT is that it will consider multiple kernel candidates simultaneously if the researchers have no prior knowledge about which to chose. It firstly tests the association with each candidate kernel to obtain the p-value for each. Then the minimum p-value is selected and a multiple-comparison technique is used to adjust for selecting the minimum p- value. The optimal MiRKAT, however, cannot be applied in longitudinal studies since it could not model the correlation between the repeated measurements. Besides, it may give false positive results when evaluating associations between part of the microbiome composition and outcomes since it doesn't allow researchers to control for the effect from the other part of microbiome composition.

1.6 Longitudinal microbiome data and LMM

A longitudinal study is a correlational research study that involves repeated measurements of the same variables over long periods of time, often weeks or years. The key advantage to longitudinal studies is the ability to show the patterns of a variable over time. Depending on the scope of the study, longitudinal data can also help to discover "sleeper effects" or connections between different events over a long period of time, events that might otherwise not be linked. Longitudinal microbiome studies track the changes of phenotype and microbiome composition in each participant. It is one powerful way in which we come to learn more about cause-and-effect relationships.

A linear mixed model (LMM) is a statistical model containing both fixed effects and random effects[28]. It is particularly useful in settings where repeated measurements are made on the same statistical units (longitudinal study), or where measurements are made on clusters of related statistical units. Because of their advantage in dealing with missing values, mixed effects models are more flexible in terms of repeated measures. It does not need have same number of repeated measurements per subject.

1.7 Testing for Zero Variance Components

Testing variance components is one of the most challenging problems in the context of linear mixed effects models. (Restricted) likelihood ratio test (LRT, RLRT) have been introduced in variance component tests since such tests have desirable asymptotic properties[29, 30]. Par-

ticularly, the standard likelihood ratio test statistics is assumed asymptotically to follow a χ^2 distribution with degrees of freedom equal to the number of parameters[31]. The usual asymptotic theory does not hold in many cases, such as having boundary constraints. This non-standard problem is that the true value of variance parameters under the null are on the boundary of the parameter space defined by the alternative hypothesis[32, 33]. Stram and Lee proved that the likelihood ratio test for testing zero variance component has an asymptotic χ^2 mixture distribution ($0.5\chi^2_0 : 0.5\chi^2_1$) under the null[34]. This asymptotic distribution can hold only if the data are independent and identically distributed both under H_0 and H_1.

In many cases of LMMs, it is inconsistent due to the violation of independence assumption. Besides, the number of subjects may not be sufficient to ensure an accurate approximation[35]. In the case of LMM with one variance component, an appealing practical testing methodology was proposed by $Crainiceanu$. The finite sample distributions of the (restricted) likelihood ratio statistic were derived, in which the finite sample behaviour is explained, to overcome the inconsistency of independent and identically distributed (IID) data. The finite sample distribution of (R)LRT statistics can be obtained easily because it avoids the bootstrap; the eigenvalues need to be computed only once before simulation begins and a grid search over the variance parameter does not depend on the sample size n[36, 37].

$Greven$ provided a efficient mixture approximation to the parametric bootstrap. The reliable estimation of (1-α) quantiles based on parametric bootstrap is computationally expensive especially when α is small. A parametric approximation is proposed to reduce the computation time by reducing the necessary number of parametric bootstaps to determine the distribution of (R)LRT. The idea of approximation is to use the entire bootstrap sample to fit a flexible distribution with two parameters[35]. Therefore, the necessary number of simulations required for estimating tail quantiles is reduced.

1.8 Approach and Prospects

The current approaches, such as the distance-based test and the linear kernel machine framework, cannot be adopted in longitudinal microbiome studies. Besides, the biologists may be more interested in detecting which genus or phylum have effects on the bio-clinical conditions. Since there is a dependency structure (phylogenetic tree) in the microbiome profile, the abundance

change in one genus or other rank level cluster is not independent with others. Thus, the traditional approaches will give false positive results in detecting the association between one cluster without controlling the effect contributed by others. The microbiome association exact tests (MAETs) in linear mixed model proposed in this paper can overcome those problems. MAETs can test not only the effect of the overall microbiome in longitudinal studies but also the effect from specific cluster of the OTUs, while controlling for others by introducing more random effects in the model. Here, we consider a linear mixed effects model (LMM)[28] with multiple random effects in which the microbiome diversity is modeled as one or more random effects and th e correlation of repeated measurement is considered as another random effect which needs to be controlled. In this study, we propose a powerful and scalable strategy for testing zero variance components in presence of multiple variance components in LMM. The goal is achieved by reducing the multiple variance components to a single one using Ofversten's transformation strategy[38, 39]. Furthermore, we combined the recently developed exact likelihood ratio test(eLRT), exact restricted likelihood ratio test (eRLRT) and exact Score test for testing the variance component.

This thesis is organized as follows: Section 2 firstly reviews the UniFrac Distance, constructing kernel matrix and exact test for testing one variance component. In section 2, the five different scenario simulation study design is presented. Section 3 shows the results of the simulation study and application in longitudinal pulmonary microbiome study. We then discuss the advantages of our approach and future work to extend this approach to handle more sophisticated outcomes.

2

Study Design and Methods

2.1 UniFrac Distance

For each individual i we get n_i repeated microbiome profile measurements. Consider two microbiome communities A and B and suppose that we have a rooted phylogenetic tree with m branches(m is the total number of OTUs). Let b_i denote the length of branch i ($i = 1, 2, ...m$) and p_i^A, p_i^B are the taxa proportions descending from the branch i for community A and B respectively. T is the total reads of m OTUs and T_i is total reads of ith OTU from both community.

Unweighted Unifrac[16]

$$d^U = \sum_{i=1}^{m} \frac{b_i |I(p_i^A > 0) - I(p_i^B > 0)|}{\sum_{i=1}^{m} b_i} \tag{2.1}$$

Weighted Unifrac[17]

$$d^W = \frac{\sum_{i=1}^{m} b_i |p_i^A - p_i^B|}{\sum_{i=1}^{m} b_i (p_i^A - p_i^B)} \tag{2.2}$$

$$d^{(0)} = \frac{\sum_{i=1}^{m} b_i |\frac{p_i^A - p_i^B}{p_i^A + p_i^B}|}{\sum_{i=1}^{m} b_i} \tag{2.3}$$

VAW Unifrac[18]

$$d^{VAW} = \frac{\sum_{i=1}^{m} b_i \frac{|p_i^A - p_i^B|}{T(T - T_i)}}{\sum_{i=1}^{m} b_i \frac{|p_i^A - p_i^B|}{T(T - T_i)}} \tag{2.4}$$

General Unifrac[19]

$$d^{(\alpha)} = \frac{\sum_{i=1}^{m} b_i (p_i^A + p_i^B)^\alpha |\frac{p_i^A - p_i^B}{p_i^A + p_i^B}|}{\sum_{i=1}^{m} b_i (p_i^A + p_i^B)^\alpha} \tag{2.5}$$

2.2 Construct Kernels Based on UniFrac Distances

Let $d_{ii'}$ be the UniFrac distance between subject i and i'. In the longitudinal case, this denotes the UniFrac distance between each subject's repeated measurements.

The UniFrac distance matrix is generated using pacakge "GUniFrac" with two datasets: OTU count data, and rooted phylogenetic tree. In total, we have $\sum_{i=1}^{n} n_i = N$ observations.

The pairwise distance denotes the similarity of the microbiome composition. The larger distance between two observations means the microbiome compositions are more different. The diagonal of the matrix equals to 0 because it denotes the similarity of the microbiome in the same sample which means it couldn't have any difference in microbiome composition.

Then we define a kernel matrix by transforming the distance matrix through this equation to measure similarities between the microbiome composition among subjects.

$$\boldsymbol{K} \quad = \quad -\frac{1}{2}(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}'}{N})\boldsymbol{D}^2(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}'}{N}) \tag{2.6}$$

where $\boldsymbol{D} = (d_{ii'})$ is the pairwise distance matrix, $\boldsymbol{I}$ is the identity matrix and $\mathbf{1}$ is a vector of $\mathbf{1}$'s. It's easy to verify that the kernel matrix defined this way can recover the original distances using standard kernel operations: it could be UniFrac Distance or the Bray-Curtis dissimilarity. The i and i' here indicate the UniFrac distance between ith and i'th observation. The N denotes the total number of samples.

We then apply a positive semi-definite correction procedure to make sure the kernel matrix is psd. First, we perform an eigen-decomposition by $\boldsymbol{K} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$, where $\boldsymbol{\Lambda} = Diag(\lambda_1, \lambda_2, ..., \lambda_n)$. We then reconstruct $\boldsymbol{K}^*$ using the absolute eigenvalues by $\boldsymbol{K}^* = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$, where $\boldsymbol{\Lambda}^* = Diag(|\lambda_1|, |\lambda_2|, ..., |\lambda_n|)$.

2.3 Linear Mixed Model for Longitudinal Microbiome Studies

In experiments adopting longitudinal designs, phenotype measurements ($\boldsymbol{y}$, outcome variable of interest) are collected for each of n individuals at n_i (the total number of time points for ith individual) time points, where i indexes for subject and $\sum_{i=1}^{n} n_i = N$ (the sum of obs for n individuals). We expect that measurements acquired from the same individual will tend to be

more correlated than those obtained from different individuals. This correlation is due to both genetic and environmental effects shared between measurements. We consider a standard linear mixed model for the ith subjects,

$$\begin{aligned} \boldsymbol{y}_i &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + h(\boldsymbol{G}_i) + \boldsymbol{\varepsilon}_i \\ \boldsymbol{b}_i &\sim \mathcal{N}(0, \sigma_D^2) \\ \boldsymbol{\varepsilon}_i &\sim \mathcal{N}(0, \sigma_e^2\boldsymbol{I}_{n_i}) \end{aligned} \tag{2.7}$$

where $\boldsymbol{y}_i = (y_{i1}, y_{i2}, ...y_{in_i})'$ is an $n_i \times 1$ vector of n_i repeated measures of quantitative phenotype for individual i, $\boldsymbol{X}_i = (X_{ij1}, X_{ij2}, ..., X_{ijp})', j = 1, 2, .., n_i$ be the $n_i \times p$ covariates - such as age, gender, race and other clinical and environmental variables that are expected to influence microbial community diversity and are related to outcomes - that we want to control for. $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects. $\boldsymbol{G}_i = (G_{ij1}, G_{ij2}, ...Gijm)', j = 1, 2, .., n_i$ denote the abundances of all OTUs for individual i at the jth ($j = 1, 2, ..., n_i$) time point(m is the total number of OTUs). These OTUs are related by a known phylogenetic tree. $\boldsymbol{b}$ is a q vector of the subject -specific random effects. $\boldsymbol{b}_i$ is a one element vector random effects included to control the correlation in the repeated measurements. $\boldsymbol{Z}_i$ is a $n_i \times 1$ design matrix linking the vector of random effects $\boldsymbol{b}_i$ to $\boldsymbol{y}_i$. $\boldsymbol{\varepsilon}_i$ is a $n_i \times 1$ vector for the error term.

Typically, a random intercept or random intercept and random slope model is most often used. Now, we simply consider a random intercept model,

$$\boldsymbol{Z}_i = \mathbf{1}_{\mathbf{n_i}} \tag{2.8}$$

then, the covariance structure of $\boldsymbol{Z}_i\sigma_D^2\boldsymbol{Z}_i' + \sigma_e^2\boldsymbol{I}_{n_i}$ for ith subject is compound symmetry. The variances are homogeneous. There is a correlation between two measurements, but it is assumed that the correlation is constant regardless of how far apart the measurements are.

For n subjects, we have the following model,

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + h(\boldsymbol{G}) + \boldsymbol{\varepsilon} \\ \mathbf{h} &\sim N(\mathbf{0}, \sigma_{\mathbf{g}}^{\mathbf{2}}\mathbf{K}) \\ \boldsymbol{b} &\sim \mathcal{N}(0, \boldsymbol{Z}\sigma_D^2\boldsymbol{Z}') \\ \boldsymbol{\varepsilon} &\sim \mathcal{N}(0, \sigma_e^2\boldsymbol{I}) \end{aligned} \tag{2.9}$$

where $\boldsymbol{y}$ is a $N \times 1$ vector of repeated measures of quantitative phenotype, $\boldsymbol{X}$ is $N \times p$ covariate matrix (e.g., sex, smoking history, height, etc), $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects, p is the number of covariates and $h(\boldsymbol{G})$ is the random effects contributed by its microbiome profile. $\boldsymbol{Z}$ is a $N \times q$ the incident matrix for for random effects $\boldsymbol{b}$ where $\boldsymbol{b}$ is a $q \times q$ vector of the subject-specific random effects. $\mathbf{h}$ follows a distribution with mean zero and variance $\sigma_g^2 \boldsymbol{K}$. $\boldsymbol{b}$ follows a normal distribution with mean zero and $N \times N$ variance-covariance matrix $\boldsymbol{Z}\sigma_D^2\boldsymbol{Z}'$. It is also assumed that $\boldsymbol{b}$ and $\mathbf{h}$ are independent with each other.

$$\mathrm{Var}(\boldsymbol{y}) = \boldsymbol{V} = \sigma_g^2 \boldsymbol{K}_k + \boldsymbol{Z}\sigma_D^2\boldsymbol{Z}' + \sigma_e^2 \boldsymbol{I}_n. \tag{2.10}$$

where $\boldsymbol{K} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'$ is the kernel matrix capturing effects from the OTU set.

The relationship between the microbiome profile and the outcome variable is fully characterized by the function $h(\cdot)$. Testing that there is no association between microbiome composition and the outcome is equivalent to testing that $h(\boldsymbol{G}) = 0$ ($\sigma_g^2 = 0$).

In model (2.9), we are interested in testing $H_0 : \sigma_g^2 = 0$ v.s. $H_1 : \sigma_g^2 > 0$.

Our method can handle the situation when there are two or more variance components present (model 2.9). Especially, when there are longitudinal microbiome profile data, our method is able to testing the microbiome effect while controlling for the correlation between repeated measurements. We provide one type of exact test and generate an accurate p-value for decision making.

2.4 Testing with one variance component

Here we give a brief review of eLRT, eRLRT and (Rao) score test for testing with one single variance component.

$$\boldsymbol{V} \;=\; \sigma_e^2 \boldsymbol{I}_n + \sigma_g^2 \boldsymbol{K} \tag{2.11}$$

A slight extension allows for testing the more general case

$$\boldsymbol{V} \;=\; \sigma_e^2 \boldsymbol{V}_0 + \sigma_g^2 \boldsymbol{K} \tag{2.12}$$

where $\boldsymbol{V}_0 \in \mathbb{R}^{n \times n}$ is a psd matrix (positive-definite matrix). Let r = rank($\boldsymbol{V}_0$). Given eigen-decomposition $\boldsymbol{V}_0 = \boldsymbol{U}_r \boldsymbol{D}_r \boldsymbol{U}_r^T$, define $\boldsymbol{T} = \boldsymbol{D}_r^{-1/2}\boldsymbol{U}_r^T \in \mathbb{R}^{r \times n}$. Then

$$\boldsymbol{T}\boldsymbol{Y} \sim N(\boldsymbol{T}\boldsymbol{X}\boldsymbol{\beta}, \sigma_e^2 \boldsymbol{I}_r + \sigma_g^2 \boldsymbol{T}\boldsymbol{K}\boldsymbol{T}^T) \tag{2.13}$$

and the eLRT, eRLRT or score test can be applied to $\boldsymbol{TY}$.

2.4.1 Exact LRT and RLRT

Let $\lambda = \sigma_g^2/\sigma_e^2$ be the signal-to-noise ratio, and write the covariance as

$$\boldsymbol{V} \;=\; \sigma_e^2(\boldsymbol{I}_n + \lambda \boldsymbol{K}) = \sigma_e^2 \boldsymbol{V}_\lambda \tag{2.14}$$

The model parameters are $(\boldsymbol{\beta}, \sigma_e^2, \lambda)$. Testing $H_0 : \sigma_g^2 = 0$ vs $H_A : \sigma_g^2 > 0$ is equivalent to testing $H_0 : \lambda = 0$ vs $H_A : \lambda > 0$. The Log-likelihood function is

$$l(\boldsymbol{\beta}, \sigma_e^2, \lambda) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma_e^2 - \frac{n}{2}\log|\boldsymbol{V}_\lambda| - \frac{1}{2\sigma_e^2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{V}_\lambda^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$$

The likelihood ratio test (LRT) statistic is

$$LRT = 2\underset{H_A}{sup} L(\boldsymbol{\beta}, \sigma_e^2, \lambda) - 2\underset{H_0}{sup} L(\boldsymbol{\beta}, \sigma_e^2, \lambda) = \underset{\lambda \geq 0}{sup} n \log \boldsymbol{Y}^T \boldsymbol{A}_0 \boldsymbol{Y} - n \log \boldsymbol{Y}^T \boldsymbol{A}_\lambda \boldsymbol{Y} - \log|\boldsymbol{V}_\lambda|$$

where projection matrix for X is

$$\boldsymbol{P}_X = \boldsymbol{X}(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}^T$$

$$\boldsymbol{A}_0 = \boldsymbol{I} - \boldsymbol{P}_X$$

Let $\{\xi_1, ..., \xi_l\}$ be positive eigenvalues of $\boldsymbol{K}$ and $\{\mu_1, ..., \mu_k\}$ be positive eigenvalues of $\boldsymbol{A}_0 \boldsymbol{K} \boldsymbol{A}_0$. Let s denote the $rank(\boldsymbol{X})$. Then,

$$\begin{aligned} LRT \;&=\; sup_{\lambda \geq 0} n \log \frac{\boldsymbol{y}\boldsymbol{A}_0\boldsymbol{y}^T}{\boldsymbol{y}\boldsymbol{A}_\lambda\boldsymbol{y}^T} - \log|\boldsymbol{I}_n + \lambda\boldsymbol{K}| \\ &\overset{D}{=}\; sup_{\lambda \geq 0} n \log \frac{\sum_{i=1}^{n-s} w_i^2}{\sum_{i=1}^{k} \frac{w_i^2}{1+\lambda\mu_i} + \sum_{i=k+1}^{n-s} w_i^2} - \sum_{i=1}^{l} \log(1 + \lambda\xi_i) \end{aligned} \tag{2.15}$$

$$\begin{aligned} RLRT \;&=\; sup_{\lambda \geq 0} (n-s) \log \frac{\boldsymbol{y}\boldsymbol{A}_0\boldsymbol{y}^T}{\boldsymbol{y}\boldsymbol{A}_\lambda\boldsymbol{y}^T} - \log|\boldsymbol{I}_n + \lambda\boldsymbol{K}| \\ &\overset{D}{=}\; sup_{\lambda \geq 0} (n-s) \log \frac{\sum_{i=1}^{n-s} w_i^2}{\sum_{i=1}^{k} \frac{w_i^2}{1+\lambda\mu_i} + \sum_{i=k+1}^{n-s} w_i^2} - \sum_{i=1}^{k} \log(1 + \lambda\xi_i) \end{aligned} \tag{2.16}$$

where under the null hypothesis that $\lambda = 0$, $w_i \overset{iid}{\sim} N(0,1)$, under the alternative hypothesis$\lambda > 0$, $w_i \sim N(0, 1+\lambda\mu_i)$, $i = 1, ...k$ and $w_i \sim N(0,1)$ when $i = k+1, ...n-s$.

2.4.2 (Rao) score test

The (Rao) score statistic is based on $S(\sigma_g^2) = \frac{U(\sigma_g^2)^2}{I(\sigma_g^2)}$ evaluated at the MLE under the null.

$$U(\sigma_g^2) = \frac{\partial l}{\partial \sigma_g^2} \tag{2.17}$$

$$I(\sigma_g^2) = E(-\frac{\partial^2}{\partial \sigma_g^2 \partial \sigma_g^2} l) \tag{2.18}$$

We evaluate the partial derivatives at the MLE under the null

$$\hat{\boldsymbol{\beta}} = \boldsymbol{y}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}, \hat{\sigma}_e^2 = \frac{1}{n}\boldsymbol{y}(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{y}^T \tag{2.19}$$

Equivalently the score test rejects when

$$max\{\frac{\boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{K}(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{y}}{(\boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{y})^2}, \frac{tr(\boldsymbol{K})}{n}\} \tag{2.20}$$

is large.

2.5 Testing microbiome effect in longitudinal study

2.5.1 Testing overall microbiome effect

We need to consider the subject specific random effect if we are testing overall microbiome association with outcomes in longitudinal studies. In this case, the mixed random intercept model can be written as

$$\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma_e^2\boldsymbol{I} + \sigma_g^2\boldsymbol{K} + \boldsymbol{Z}\sigma_D^2\boldsymbol{Z}') \tag{2.21}$$

where σ_g^2, σ_D^2, σ_e^2 are corresponding variance component parameters from microbiome, within individuals and environment effects.

The test provided by Ofversten[38] is valid regardless of the relationship of nested terms. A simpler and more general way presented in this section is using a single orthonormal basis Gram-Schmidt on the columns of $[X, K, Z, I_n]$ to obtain an orthonormal basis for $\boldsymbol{R}^n$: $C = [c_1, c_2, ...c_n]$. Partition C as

$$C = [C_1, C_2, C_3, C_4]$$

We have

C_1 : orthonormal basis for $\mathcal{C}(X)$

C_2 : orthonormal basis for the orthogonal complement of $\mathcal{C}(X)$ with respect to $\mathcal{C}(X, K)$

C_3 : orthonormal basis for the orthogonal complement of $\mathcal{C}(X, K)$ with respect to $\mathcal{C}(X, K, Z)$

C_4 : orthonormal basis for the orthogonal complement of $\mathcal{C}(X, K, Z)$

Executing test $H_0 : \sigma_g^2 = 0$

For special case $C(\boldsymbol{Z}) \subset C(\boldsymbol{X}, \boldsymbol{K}) : C_2'\boldsymbol{Z}\boldsymbol{Z}'C_2 = \lambda\boldsymbol{I}$, so that

$$C_2'\boldsymbol{Y} \sim N(0, (\lambda\sigma_D^2 + \sigma_e^2)\boldsymbol{I} + \sigma_g^2 C_2'\boldsymbol{K}C_2) \tag{2.22}$$

and an ordinary least-squares Wald test can be applied inmmediately without any use of $C_4'\boldsymbol{Y}$.

In general $C(Z) \subsetneq C(X, K) : C_2'\boldsymbol{Z}\boldsymbol{Z}'C_2 \neq \lambda I$

The goal is to choose a matrix Q so that the extended Wald's Test from

$$\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma_e^2\boldsymbol{I} + \sigma_g^2\boldsymbol{K} + \boldsymbol{Z}\sigma_D^2\boldsymbol{Z}') \tag{2.23}$$

to

$$C_2'\boldsymbol{Y} + \boldsymbol{Q}C_4'\boldsymbol{Y} \sim N(0, (\sigma_D^2 + \sigma_e^2/\lambda)C_2'\boldsymbol{Z}\boldsymbol{Z}'C_2 + \sigma_g^2 C_2'\boldsymbol{K}C_2) \tag{2.24}$$

We proceed with the following steps:

1. Obtain an orthonormal basis $\boldsymbol{C}_1$ of $\mathcal{C}(X)$ by QR decomposition. Let $r_0 = rank(\boldsymbol{C}_1)$.

2. Obtain an orthonormal basis $\boldsymbol{C}_2$ of $\mathcal{C}(X, Z) - \mathcal{C}(X)$ by QR decomposition. And,

$$\boldsymbol{C}_2'\boldsymbol{Y} \sim N(0, \sigma_D^2\boldsymbol{C}_2'\boldsymbol{Z}\boldsymbol{Z}'\boldsymbol{C}_2 + \sigma_g^2\boldsymbol{C}_2'\boldsymbol{K}\boldsymbol{C}_2 + \sigma_e^2\boldsymbol{I}) \tag{2.25}$$

3. Eigen-decomposition $\boldsymbol{C}_2'\boldsymbol{Z}\boldsymbol{Z}'\boldsymbol{C}_2 = \boldsymbol{W}\Lambda\boldsymbol{W}^T = \boldsymbol{W}diag(\lambda_i)\boldsymbol{W}^T$. Let λ be the smallest positive-eigenvalue and set

$$\boldsymbol{Q} = \boldsymbol{W}\mathrm{diag}(\sqrt{\lambda_i/\lambda - 1}) \in \mathbb{R}^{\mathrm{rank}(\boldsymbol{C}_1)\times\mathrm{rank}(\boldsymbol{C}_1)}. \tag{2.26}$$

4. Obtain any $\mathrm{rank}(\boldsymbol{C}_1)$ orthonormal basis vectors $\boldsymbol{C}_4 \in \mathbb{R}^{n\times\mathrm{rank}(\boldsymbol{Q}_1)}$ of the space $\mathbb{R}^n - \mathcal{C}(\boldsymbol{C}_1, \boldsymbol{C}_2, \boldsymbol{K})$. Note,

$$\boldsymbol{C}_4^T\boldsymbol{Y} \sim N(\boldsymbol{0}, \sigma_e^2\boldsymbol{Q}\boldsymbol{Q}^T) = N(\boldsymbol{0}, (\sigma_e^2/\lambda)\boldsymbol{C}_2'\boldsymbol{Z}\boldsymbol{Z}'\boldsymbol{C}_2 - \sigma_e^2\boldsymbol{I}) \tag{2.27}$$

and thus

$$(\boldsymbol{C}_2' + \boldsymbol{Q}\boldsymbol{C}_4')\boldsymbol{Y} \sim N(0, (\sigma_D^2 + \sigma_e^2/\lambda)\boldsymbol{C}_2'\boldsymbol{Z}\boldsymbol{Z}'\boldsymbol{C}_2 + \sigma_g^2\boldsymbol{C}_2'\boldsymbol{K}\boldsymbol{C}_2) \tag{2.28}$$

Test $H_0 : \sigma_g^2 > 0$ using eRLRT or eScore test on the transformed data

$$\boldsymbol{\Lambda}^{-1/2}\boldsymbol{W}^T(\boldsymbol{C}_2' + \boldsymbol{Q}\boldsymbol{C}_4')\boldsymbol{Y} \tag{2.29}$$

$$\sim \quad N(\boldsymbol{0}, (\sigma_D^2 + \sigma_e^2/\lambda)\boldsymbol{I} + \sigma_g^2\boldsymbol{\Lambda}^{-1/2}\boldsymbol{W}^T\boldsymbol{C}_2'\boldsymbol{K}\boldsymbol{C}_2\boldsymbol{W}\boldsymbol{\Lambda}^{-1/2}). \tag{2.30}$$

2.5.2 Testing multiple microbiome effects

In this section, we consider to test microbiome effect from a specific part of the microbial composition. We can cluster the OTUs by their lineage information at higher rank such as genus, order or phylum. For instance, the count of OTUs belong to one phylum may be correlated with another phylum. Therefore the other phylum may contribute effect when we test the microbiome effect of the phylum of interest. The linear mixed model with multiple variance components can be written as

$$\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma_e^2\boldsymbol{I} + \boldsymbol{Z}\sigma_D^2\boldsymbol{Z}' + \sigma_{g1}^2\boldsymbol{K}_1 + \sigma_{g2}^2\boldsymbol{K}_2) \tag{2.31}$$

where σ_D^2 and σ_e^2 are corresponding variance component parameters from within individuals and environment effects. σ_{m1}^2 and σ_{m2}^2 are corresponding parameters of microbiome effect which need to be adjusted and the microbiome effect being tested.

We are interested in testing $H_0 : \sigma_{m2}^2 = 0$, vs $H_A : \sigma_{m2}^2 > 0$. Since there are more than one variance components in the above model, the idea is to reduce the multiple variance components to the single variance component cases by performing Gram-Schmidt on the matrix $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{K}_1, \boldsymbol{K}_2, \boldsymbol{I}_n)$ to obtain an orthonormal basis of $\boldsymbol{R}^n$

$$C \quad = \quad [C_1, C_2, C_3, C_4, C_5]$$

where C_1 is an orthonormal basis for $\mathcal{C}(X)$, C_2 is orthonormal basis for the orthogonal complement of $\mathcal{C}(X)$ with respect to $\mathcal{C}(X, Z)$, C_3 is orthonormal basis for the orthogonal complement of $\mathcal{C}(X, Z, K_1)$ with respect to $\mathcal{C}(X, Z)$, C_4 is orthonormal basis for the orthogonal complement of $\mathcal{C}(X, Z, K_1, K_2)$ with respect to $\mathcal{C}(X, Z, K_1)$ and C_5 is orthonormal basis of $\mathcal{C}(X, Z, K_1, K_2)$.

Similarly, we consider the following situations:

1. If rank$(C_4) > 0$, that is $\mathcal{C}(X, Z, K_1) \subsetneq \mathcal{C}(X, Z, K_1, K_2)$, then

$$C_4'\boldsymbol{Y} \sim N(0, \sigma_e^2\boldsymbol{I} + \sigma_{g2}^2 C_4'\boldsymbol{K}_2 C_4) \tag{2.32}$$

and eLRT, eRLRT and eScore can be applied to $C_4'\boldsymbol{Y}$. The order of $\boldsymbol{Z}, \boldsymbol{K}_1, \boldsymbol{K}_2, \boldsymbol{I}$ does not matter.

2. If rank$(C_4) = 0$, that is $\mathcal{C}(X, Z, K_1) = \mathcal{C}(X, Z, K_1, K_2)$, then we construct a test based on transformed data $C_3'\boldsymbol{Y} + \boldsymbol{Q}C_5'\boldsymbol{Y}$, where the matrix $\boldsymbol{Q}$ is chosen such that

$$C_3'\boldsymbol{Y} + \boldsymbol{Q}C_5'\boldsymbol{Y} \sim N(0, (\sigma_{g1}^2 + \sigma_e^2/\lambda)C_3'\boldsymbol{K}_1 C_3 + \sigma_{g2}^2 C_3'\boldsymbol{K}_2 C_3) \tag{2.33}$$

When $C_3'\boldsymbol{K}_1' C_3 \neq \lambda I$, the test requires $\boldsymbol{Q}C_5'\boldsymbol{Y}$ which follows distribution $N(\boldsymbol{0}, \sigma_e^2\boldsymbol{Q}\boldsymbol{Q}^T)$.

Since $C_3'\boldsymbol{V}C_5 = \boldsymbol{0}$, $C_3'\boldsymbol{Y} \perp \boldsymbol{Q}C_5'\boldsymbol{Y}$. The $\boldsymbol{Q}$ is chosen via eigen-decomposition as 2.26, such that

$$\boldsymbol{Q}\boldsymbol{Q}^T = \lambda^{-1}C_3'\boldsymbol{K}_1 C_3 - \boldsymbol{I} \tag{2.34}$$

Therefore, the test for single variance component can be applied. Test $H_0 : \sigma_{g2}^2 > 0$ using eRLRT or eScore test on the transformed data

$$\boldsymbol{\Lambda}^{-1/2}\boldsymbol{W}^T(\boldsymbol{C}_3' + \boldsymbol{Q}\boldsymbol{C}_5')\boldsymbol{Y} \tag{2.35}$$

$$\sim \quad N(\boldsymbol{0}, (\sigma_{g1}^2 + \sigma_e^2/\lambda)\boldsymbol{I} + \sigma_{g2}^2\boldsymbol{\Lambda}^{-1/2}\boldsymbol{W}^T\boldsymbol{C}_3'\boldsymbol{K}_2\boldsymbol{C}_3\boldsymbol{W}\boldsymbol{\Lambda}^{-1/2}). \tag{2.36}$$

2.6 Monte Carlo Simulation Study Design

We conducted simulation studies under a range of scenarios in order to verify that our method correctly controls type I error rate and to assess the power by using different kernels.

The tree based simulation strategy simulates microbiome datasets by utilizing a real longitudinal pulmonary microbiome profile set consisting of 100 samples and 2964 OTUs. There are two parts of the microbiome profile: the OTU counts data and the phylogenetic tree. The phylogenetic tree contains the lineage information of the OTUs. It can be constructed by using the FastTree algorithm in the Qiime pipeline or by using the lineage information from the NCBI Taxonomy database.

2.6.1 Scenario 1: Simulation for Overall Microbiome Effect (without Clustering)

In particular, for simulating microbiome effect, we constructed seven different types of kernel matrix K based on the real longitudinal microbiome profile (2964 OTUs without clustering), which is composed of $n = 30$ individuals and each contains 2-4 repeated measurements. Then, we applied this kernel matrix which captures the overall microbiome effect to simulate the outcome. In this case, we changed the effect size from 0 to 30 of this kernel matrix to study the type I error rate and power of the test.

In addition, we assumed that there are two covariates X_{1i} and X_{2i} and assume that the covariates for each individual do not change during the follow-up. Therefore, we first simulated two covariates for 30 individuals at baseline and assigned the same covariates to the following repeated measurements to make sure the covariates for each individual do not change with time. Moveover, the covariates were simulated in three scenarios: no covariates, covariates are independent with microbiome effect ($X \perp K$), and dependent with the baseline microbiome effect ($X \not\perp K_{baseline}$). It can be shown as follows:

$$\begin{aligned} X \perp K &: X_{1i}, X_{2i} \sim \mathcal{N}(0,1) \\ X \not\perp K &: X_{1i} \sim \mathcal{N}(0,1) \\ & \quad X_{2i} = h(G)_{baseline} + \mathcal{N}(0,1) \end{aligned}$$

We can display the model as:

$$\begin{aligned} \boldsymbol{y}_i &= \beta_1 X_{1i} + \beta_2 X_{2i} + h(\boldsymbol{G}_i) + \boldsymbol{Z}_i \boldsymbol{b}_i + \boldsymbol{\varepsilon}_i \\ \boldsymbol{V}_i &= \sigma^2 \boldsymbol{Z}_i \boldsymbol{Z}_i' + \sigma_g^2 \boldsymbol{K} + \sigma_e^2 \boldsymbol{I}_i \\ \boldsymbol{y}_i &\sim \mathcal{N}(\boldsymbol{X}_i \boldsymbol{\beta}, \boldsymbol{V}_i) \end{aligned}$$

where $\sigma_e^2 = 1$, $\sigma^2 = 4$, $\beta_1 = \beta_2 = 0$ for no covariates and $\beta_1 = \beta_2 = 0.1$ for existing covariates.

To be more specific, for simulating $X_2 \not\perp K$, we first simulated X_2^* following $\mathcal{N}(0,1)$ and X_2^{**} following $\mathcal{N}(0, 0.25K_{baseline})$. Then the $X_2 = X_2^* + X_2^{**}$.

Under this scenario, we changed the effect size of the microbiome random effect $h(\boldsymbol{G})$ to study the type I error rate and the power of the test.

For study of the type I error rate, we set the effect size to be 0 and prepare three sets of response and covarites data under these three situations: no covariates, $X \perp K$ and $X \not\perp K$. Then we performed the eRLRT and eScore test with or without adjusting covariates. We also perform the test ignoring the correlation between repeated measurements and using MiRKAT and optional MiRKAT which could not take the correlation into account. Since the optional MiRKAT is computationally expensive, we only perform this test on the outcome simulated from $K_{0.25}$ (see table 1).

For the power study, we set up the effect size to be 0, 2, 5, 10, 15, 20, 25 and 30. The power studies were performed under the same situations as above to see whether different kernel types have different power of detecting the microbiome effect.

2.6.2 Scenario 2: Simulate Longitudinal Microbiome Count Data

Since the sample size has great impact on power, we simulated the longitudinal microbiome counts data under different sample sizes to see how sample size affects the power of the exact tests. The count data simulation is based on the real pulmonary microbiome profile. In this case, we use sample size 10, 20 50, 100 and 200. Then, we picked a sample size which was relatively powerful and has reasonable computation speed to repeat the simulation scenario 1. Since the MiRKAT and ignoring correlation between repeated measurement show the inflated type I error in scenario 1, we only assessed the type I error and power for outcomes $\boldsymbol{y}$ associated without covariates and with independent or dependent covariates.

Several methods or models were proposed to simulate sparse count data. To realistically simulate the data, it is important to model extra-variation or overdispersion of the OTU counts due to high degree of heterogeneity among the samples. For this reason, eventually we chose the two-part zero-inflated Beta regression model which exhibits more variance than expected from a model using Poisson marginal distribution and Pearson correlation.

Pearson correlation and Poisson margin distribution

The first one is sampling high-dimensional correlated count random variables with a prespecified Pearson correlation and exact Poisson marginal distribution. For real pulmonary microbiome data, we only have similar sample size at the first three time points, so three repeated microbiome

data were simulated for each individual. Since we needed to specify the mean parameters λ of Poisson, we calculated the mean of counts for each OTU at each time point based on the real data. For the correlation matrix, we first estimated the mean proportions $\boldsymbol{\pi}_t$ of 2964 OTUs at each time points $t = 1, 2, 3$ using the maximum likelihood method ("dirmult" package in R). Then, we computed the correlation within each of the three time points using the mean proportion $\boldsymbol{\pi}_{2964\times 3}$. Therefore, each of the 2964 OTUs was sampled at each time points by adopting the unstructured correlation matrix.

Two-part zero-inflated Beta regression model

The longitudinal microbiome compositional data are highly skewed, bounded in [0,1), and often sparse with many zeros. In addition, the observations from repeated measures are correlated. $H\ Li$ (in submission) proposed a two-part zero-inflated Beta regression model with random effects (ZIBR) for simulating the microbial abundance for longitudinal microbiome data. The model includes a logistic component to model presence/absence of the microbe in samples and a Beta component to model non-zero microbial abundance. Each component includes a random effect to take into account the correlation among repeated measurements on the same subject. We used this two-part zero-inflated Beta regression model to simulate the abundance for 2964 OTUs and generated counts for each measurement via multiple the average counts based on the real microbiome data.

2.6.3 Scenario 3: Simulation for Baseline Overall Microbiome Effect (without clustering)

Since the exact tests proposed in this paper can analyze both longtitudinal and cross-sectional microbiome data, for comparing with the MiRKAT and optional MiRKAT proposed by $N\ Zhao\ et\ al$[27], we simulated outcome using microbiome data at baseline with and without covariates. The linear mixed model, microbiome effect size, and coefficients of covariates was the same as above.

2.6.4 Scenario 4: Simulation for Overall Microbiome Effect (Clustering by Phylum)

The unweighted UniFrac distance is most efficient in detecting abundance change in rare lineages and the weighted is more sensitive for common lineage. In addition, the generalized UniFrac distance was designed for detecting change in moderate abundant lineage. When the difference of lineage abundance becomes smaller, the power for detecting the abundance change may become smaller. Since the cluster of taxa depend on a phylogenetic tree, we partitioned all OTUs into 24 clusters by adopting the lineage information at rank of phylum of all the OTUs. In other words, we constructed the kernel matrix which captures the overall microbiome effect via the sum of the counts of OTUs in each phylum. The abundance of each lineage is more even by clustering.

Similarly, we constructed seven different types of kernel matrix K based on the new count data after clustering. Then, the outcomes was simulated using the kernel matrix which captured the overall microbiome effect at higher level of rank. In this case, we simulated the outcome without covariates and changed the microbiome effect size from 0 to 30 to assess the type I error rate and power of the test using different types of kernel matrix.

2.6.5 Scenario 5: Simulation for outcome associated with two microbiome clusters

We clustered the OTU by their rank phylum and then we constructed the kernel matrix based on the OTU counts for each phylum group. Under this simulation scenario, the outcome was associated with the two microbiome clusters $h(\boldsymbol{G}_{1i}), h(\boldsymbol{G}_{2i})$ in all samples related to the phylogenetic information of each OTU. Then the continuous phenotype was simulated as

$$\boldsymbol{y}_i = 0.1X_{1i} + 0.1X_{2i} + h(\boldsymbol{G}_{1i}) + h(\boldsymbol{G}_{2i}) + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i$$

where $h(\boldsymbol{G}_{1i})$ and $h(\boldsymbol{G}_{2i})$ denotes the random effect from each cluster. Covariates X_{1i} and X_{2i} are defined and simulated as earlier.

In this scenario, suppose we are interested if $h(\boldsymbol{G}_{2i})$ has significant association with outcome $\boldsymbol{y}$. If two clusters are correlated, cluster $h(\boldsymbol{G}_{1i})$ may contribute false effect when testing the signal from the cluster of interest $h(\boldsymbol{G}_{2i})$. Therefore we studied whether there is correlation between the clusters or not in this section.

First, we simulated outcome $\boldsymbol{y}$ associated with $h(\boldsymbol{G}_{1i})$ under different effect size while not associated with $h(\boldsymbol{G}_{2i})$. Then we tested if the other cluster $h(\boldsymbol{G}_{2i})$ shows association on the response variable. The type I error rate will inflate if $h(\boldsymbol{G}_{1i})$ does contribute effect. Moreover, we studied whether the type I error inflated with adjustment for the $h(\boldsymbol{G}_{1i})$.

Based on the simulation results of scenario 1, we chose the most powerful kernel type K_W to assess the rate of type I error. We first clustered OTUs into 6 phyla based on their lineage information and constructed K_W for each phylum using real OTU counts data. Secondly, we asscessed the power of phylum $Firmicutes$ under different effect strengths in order to guide setting the effect size for the next step. Then, the outcome $\boldsymbol{y}$ was simulated to associate with phylum $Firmicutes$ under effect size $0, 5, 30$ to get 0%, 20% and 40% power. Due to the small sample size, 40% was the highest power. Finally, we tested the microbiome effect of the other 5 phyla to assess if they have correlation with phylum $Firmicutes$. Additionally, we assessed the correlation using the baseline real micriobiome data. The procedure was similar as above except setting the effect size as $0, 2.5, 5$ to get 0%, 48% and 86% power.

We considered adjusting for the effect of $Firmicutes$ which contributed the effects on outcomes. The idea is to reduce the multiple variance components to the single variance components case using Ofvensten's method. In our algorithm, we did a low rank approximation of the kernel matrix which needed to be adjusted because a high rank leading to a small signal-to-noise ratio will decrease the power of test. The type I error rate was reassessed after adjusting for the the effect of $Firmicutes$ as above.

3

Results

In this section, we display the simulation results from performing our proposed exact tests in the linear mixed model with and without adjusting the correlation across time in longitudinal study, as well as the results from applying our methods to real pulmonary longitudinal microbiome datasets. Additionally, we compared our method with existing methods in simulation study and real data analysis.

3.1 Simulation Results

3.1.1 Scenario 1: Simulation for Overall Microbiome Effect (without Clustering)

The type I error rates of eRLRT and eScore tests across different simulation design are shown in Table 3.1. In simulation scenario 1, we tested the overall microbiome effect associated with outcome as one random effect. Notably, when the covariates were independent with the microbiome, both scenarios are equivalent since there's no association between $h(\boldsymbol{G})$ and $\boldsymbol{y}$. Therefore, our methods were valid with or without adjusting for the covariates $\boldsymbol{X}$ in this case. However, the type I error was inflated if the $\boldsymbol{X}$ dependent of $h(\boldsymbol{G})$ was not adjusted for. Moreover, the type I error was also seriously inflated if we ignore the correlation within the repeated measurements on same individuals. Since the MiRKAT and optimal MiRKAT couldn't handle this correlation structure, the type I error was also inflated.

Figure 3.1 and Figure 3.2 shows the statistical power for the tests with continuous outcomes in simulation scenario 1, in which the overall microbiome effect without clustering was associated with outcome. The power is presented with each kernel type. Specifically, Figure 3.1 shows

Table 3.1: Type I Error for Testing Overall Microbiome Effect under Different Simulation Design

		Kernel Type						
Simulation Design	Test	K_W	K_U	K_{VAW}	K_0	$K_{0.25}$	$K_{0.5}$	$K_{0.75}$
No Covariates $\boldsymbol{X}$	eRLRT	0.064	0.051	0.060	0.065	0.065	0.058	0.066
	eScore	0.053	0.045	0.045	0.044	0.051	0.052	0.046
No adjustment for $\boldsymbol{X}$ ($X \perp K$)	eRLRT	0.054	0.063	0.067	0.065	0.069	0.059	0.066
	eScore	0.040	0.047	0.043	0.045	0.042	0.050	0.051
Adjustment for $\boldsymbol{X}$ ($X \perp K$)	eRLRT	0.056	0.057	0.067	0.057	0.063	0.062	0.057
	eScore	0.049	0.047	0.046	0.049	0.052	0.046	0.051
No adjustment for $\boldsymbol{X}$ ($X \not\perp K$)	eRLRT	0.060	0.068	0.060	0.083	0.070	0.060	0.056
	eScore	0.046	0.046	0.040	0.052	0.053	0.051	0.044
Adjustment for $\boldsymbol{X}$ ($X \not\perp K$)	eRLRT	0.049	0.068	0.053	0.065	0.065	0.059	0.060
	eScore	0.037	0.044	0.041	0.046	0.045	0.044	0.042
No adjustment for $\boldsymbol{Z}$	eRLRT	0.166	0.236	0.211	0.306	0.270	0.222	0.181
	eScore	0.133	0.114	0.124	0.137	0.113	0.112	0.130
MiRKAT and optimal MiRKAT	MiRKAT	0.177	0.201	0.211	0.288	0.267	0.232	0.208
	optimal				0.226			

Type I error of eRLRT (eScore) is evaluated in which no covariates, additional covariates were independent of the microbiome ($X \perp K$) or dependent of the microbiome ($X \not\perp K$) with the use of 1,000 simulated datasets. It's also evaluated in which no adjustment for correlation within subject ($\boldsymbol{Z}$) and MiRKAT while adjusting for the additional independent covariates. K_W, K_U, K_{VAW}, K_0, $K_{0.25}$, $K_{0.5}$, and $K_{0.75}$ represent results for the weighted UniFrac kernel, unweighted UniFrac kernel, variance adjusted weighted UniFrac kernel, and generalized UniFrac kernels with = 0, 0.25, 0.5, and 0.75, respectively. Sample size is 100. There are 30 subject and 2-4 repeated measurements for each individual.

the power when outcome was not associated with covariates, and Figure 3.2 shows the power when covariates $\boldsymbol{X}$ and $\boldsymbol{K}$ were independent. Note that for Figure 3.2, we only considered tests that adjusted for covariate since it's better to consider the covariate when it can be included. We didn't perform power simulation using MiRKAT and optional MiRKAT since both tests produced inflated type I error and were invalid in such situation.

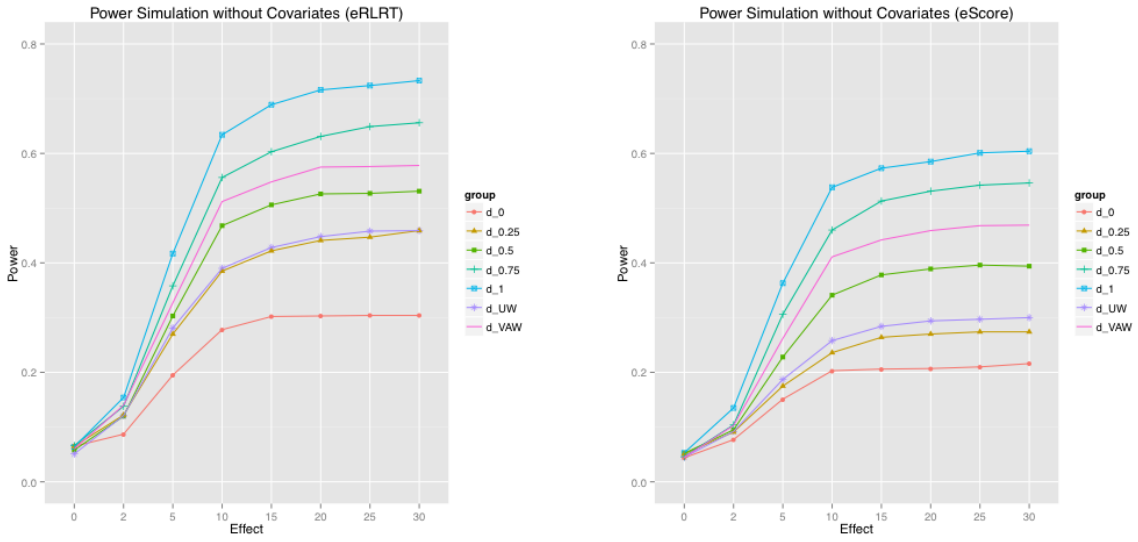

Figure 3.1: **Type I Error and Power of eRLRT and eScore Based on Different Kernels for Simulation Scenario 1 with Continuous Outcome (No Covariates)** The outcome associated with the overall microbiome were simulated using real OTU count data. Results are shown for eRLRT and eScore test. d_0, $d\_0.25$, $d\_0.5$, $d\_0.75$,d_1, d_UW and d_VAW represent test results from generalized UniFrac kernels with $\alpha = 0, 0.25, 0.5, 0.75$, weighted UniFrac kernel, unweighted UniFrac kernel and variance adjusted weighted UniFrac kernel, respectively. Sample size is 30. There are 30 individuals with $2 - 4$ repeated measurements each.

For all the kernel types that were considered, the power increased when the association strength increased. However, the choice of kernel types can greatly affect the statistical power of detecting the association. A proper choice of kernels is able to improve the statistical power compared with the improper one. For scenario 1, the weighted UniFrac kernel gave the highest power and the gerneralized UniFrac kernel with $\alpha = 0$ was the least powerful. Note that the highest power of eRLRT test is lower than 0.8 since the sample size was too small. In addition,

the eScore is more conservative than the eRLRT based on Figure 3.1 and Figure 3.2.

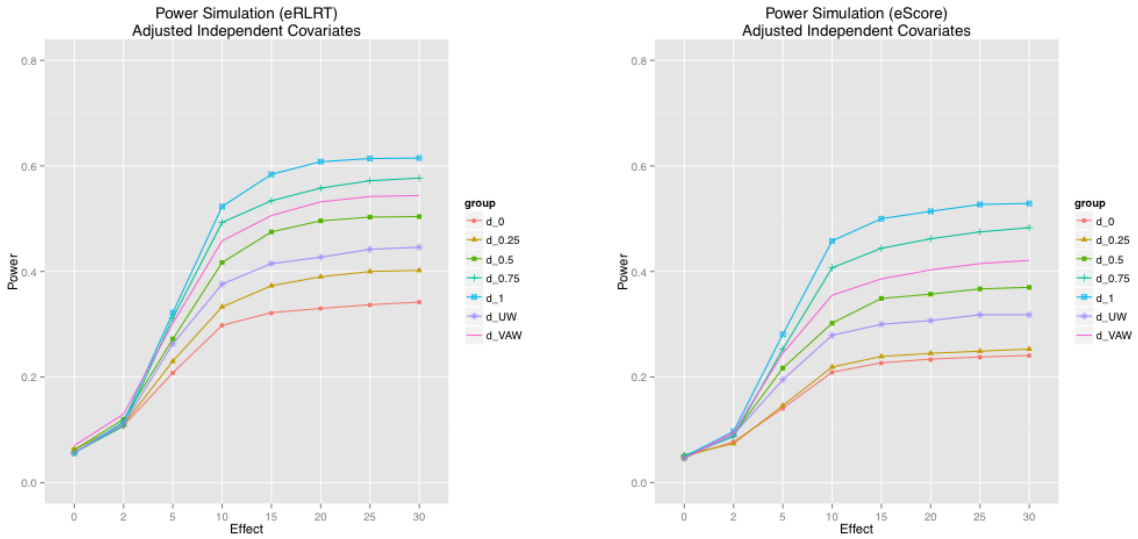

Figure 3.2: **Type I Error and Power of eRLRT and eScore Based on Different Kernels for Simulation Scenario 1 with Continuous Outcome (Adjusted $X \perp K$)** The outcome associated with the overall micriobiome were simulated using real OTU count data, and covariates $\boldsymbol{X}$ and microbiome profile $\boldsymbol{K}$ were simulated independently. Results are shown for eRLRT and eScore test. d_0, $d\_0.25$, $d\_0.5$, $d\_0.75$,d_1, d_UW and d_VAW represent test results from generalized UniFrac kernels with $\alpha = 0, 0.25, 0.5, 0.75$, weighted UniFrac kernel, unweighted UniFrac kernel and variance adjusted weighted UniFrac kernel, respectively. Sample size is 30. There are 30 individuals with $2-4$ repeated measurements each.

3.1.2 Scenario 2: Simulation Study using Simulated Longitudinal Data

Figure 3.3 presents the type I error rate and power for eScore (left) and eRLRT association tests using simulated longitudial counts data based on the two-part zero inflated Beta regression model with different sample sizes. The weighted UniFrac kernel was used in this simulation design. The type I error rate was well controlled under different sample sizes. Also, the power of detecting association increased with the sample size. Additionally, Figure 3.3 shows that the power of eRLRT is higher than eScore when sample size is small under the same association strength.

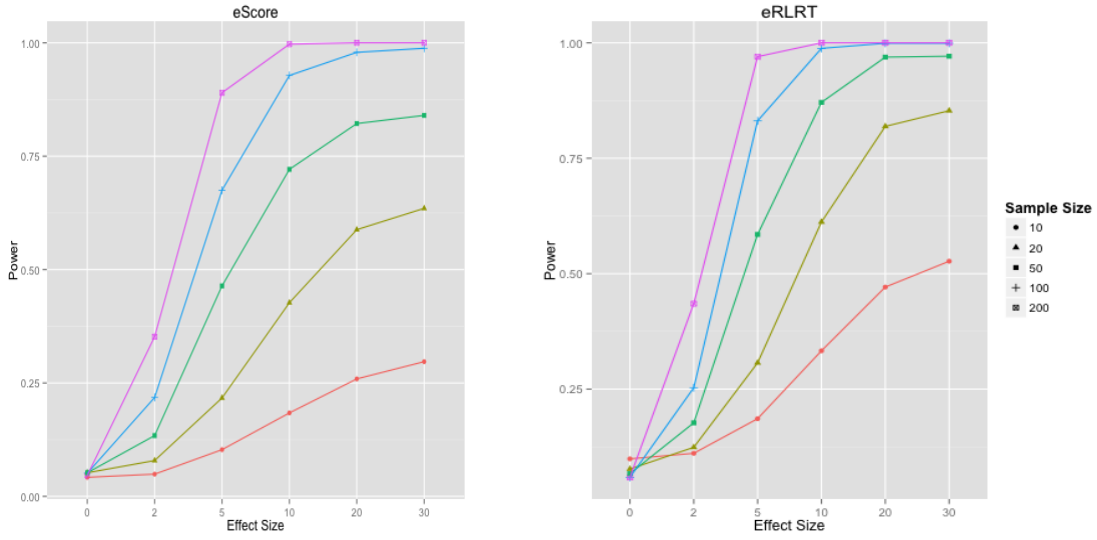

Figure 3.3: **Type I Error Rate and Power for Different Sample Size using Simulated Longitudinal Counts Data** The counts data was simulated using two-part zero inflated beta regression model. The association strength between microbiome and outcomes is ranging from 0 to 30. The longitudinal OTU count data were simulated for sample size equal to 10, 20, 50, 100 and 200. Each sample has 3 repeated microbiome profile measurements. (Left: eScore, Right: eRLRT)

Type I error rates of eRLRT and eScore across different type of covariates and kernel type are shown in Table 3.2. The type I error rates were well controlled under larger sample size and simulated longitudinal count data using ZIBR model.

Table 3.2: Type I Error for Testing Overall Microbiome Effect under Large Sample Size

		Kernel Type						
Simulation Design	Test	K_W	K_U	K_{VAW}	K_0	$K_{0.25}$	$K_{0.5}$	$K_{0.75}$
No Covariates $\boldsymbol{X}$	eRLRT	0.059	0.062	0.046	0.053	0.050	0.052	0.048
	eScore	0.051	0.050	0.047	0.048	0.045	0.047	0.049
Adjustment for $\boldsymbol{X}$ ($X \perp K$)	eRLRT	0.064	0.055	0.052	0.055	0.049	0.053	0.053
	eScore	0.047	0.044	0.046	0.049	0.044	0.055	0.045
Adjustment for $\boldsymbol{X}$ ($X \not\perp K$)	eRLRT	0.065	0.063	0.054	0.059	0.058	0.054	0.061
	eScore	0.053	0.054	0.052	0.056	0.050	0.056	0.057

Type I error of eRLRT (eScore) was evaluated in which no covariates, additional covariates were independent of the microbiome ($X \perp K$) or dependent of the microbiome ($X \not\perp K$) with the use of 1,000 simulated datasets. K_W, K_U, K_{VAW}, K_0, $K_{0.25}$, $K_{0.5}$, and $K_{0.75}$ represent results for the weighted UniFrac kernel, unweighted UniFrac kernel, variance adjusted weighted UniFrac kernel, and generalized UniFrac kernels with = 0, 0.25, 0.5, and 0.75, respectively. Sample size is 100. There are 100 individual with 3 repeated measurements for each individual.

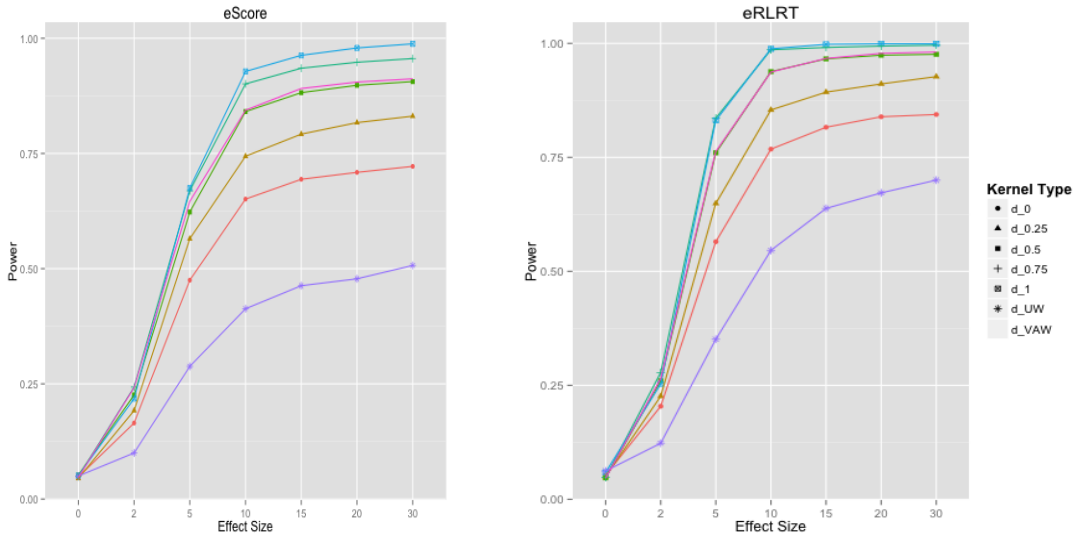

Figure 3.4: **Type I Error Rate and Power under Different Kernel Type using Simulated Longitudinal Counts Data** The counts data was simulated using two-part zero inflated beta regression model. The association strength between microbiome and outcomes ranged from 0 to 30. The longitudinal OTU count data were simulated for sample size equal to 100. Seven kernel types are shown as above. Each sample has 3 repeated microbiome profile measurements. (Left: eScore, Right: eRLRT)

Figure 3.4 shows the Type I error and power across seven kernel types using simulated longitudinal counts data (sample size = 100). Similarly, the power increased when the association strength increased. The choice of kernel types can greatly affect the statistical power of detecting association strength. The weighted UniFrac kernel provided the highest power for both eScore and eRLRT. Also, the power of exact tests was greatly increased when the sample size was large enough. The eRLRT was more powerful and more sensitive to the microbiome effect than eScore.

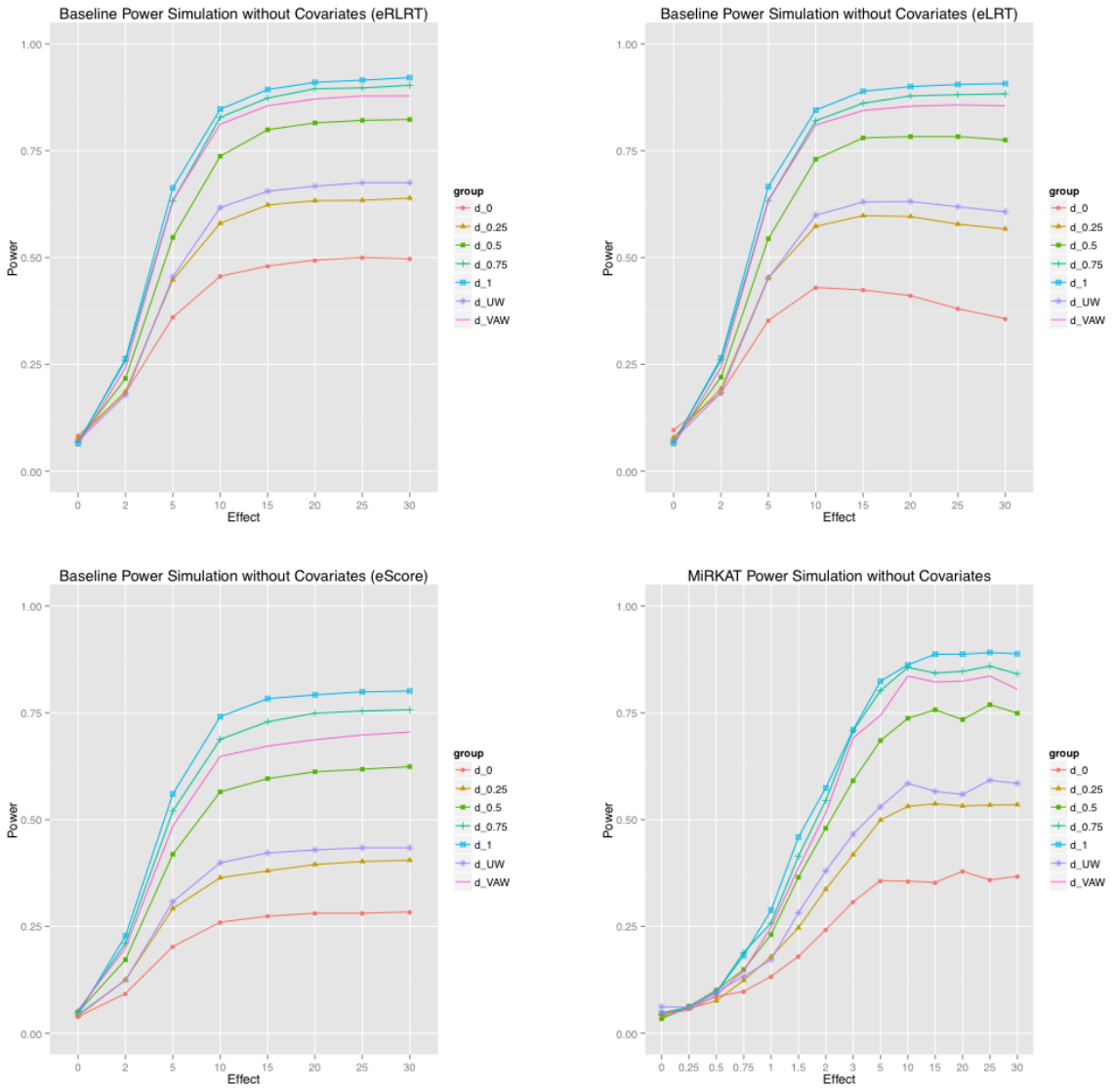

Figure 3.5: **Type I Error and Power Comparison for Baseline Microbiome Effect without Clustering** The outcome associated with the micriobiome were simulated using real OTU count data. Results are shown for exact test (eRLRT, eLRT eScore) and MiRKAT. d_0, $d\_0.25$, $d\_0.5$, $d\_0.75$,d_1, d_UW and d_VAW represent test results from generalized UniFrac kernels with $\alpha = 0, 0.25, 0.5, 0.75$, weighted UniFrac kernel, unweighted UniFrac kernel and variance adjusted weighted UniFrac kernel, respectively. Sample size is 30.

3.1.3 Scenario 3: Simulation for Baseline Overall Microbiome Effect (without clustering)

Figure 3.5 shows the type I error rate and power for the eRLRT, eLRT, eScore and MiRKAT tests with continuous outcomes in simulation scenario 3, in which the baseline overall microbiome effect without clustering was associated with outcome. The power is presented with each kernel types as in scenario 1. We also performed power simulation using MiRKAT since it's valid when there is no correlation between each sample. The trend of power was similar with results of scenario 1 for all the kernel types. However, even when the outcomes were simulated without covariates, the software still introduced one covariate equal to 1. The pattern of eRLRT and eLRT are not exactly the same. The power was higher than the longitudinal microbiome effect test in scenario 1. In addition, the eScore is more conservative than eRLRT and eLRT when the sample size was small. The MiRKAT test gave similar power to eact tests except it was more sensitive when the association strength was small.

3.1.4 Scenario 4: Simulation for Overall Microbiome Effect (Clustering by Phylum)

Figure 3.6 presents statistical power for scenario 4, where the clustered counts at the phylum level were associated with the outcome. We again show the power that when outcome has no association with covariates. Results were notably different with the scenario 1. The statistical power of the seven kernels became similar and the unweighted UniFrac kernel produced the highest power. Moreover, the power of the generalized UniFrac kernel, the least powerful one in scenario 1, improved in scenario 4. The weighted and generalized UniFrac kernel with parameter α greater than 0.5 gave the same statistical power. Since the OTU counts were clustered at the phylum level, the difference of abundance between each phylum cluster is less notable. As we mentioned earlier, detecting power of different kernels is mainly based on the abundance in each lineage. In other words, the less variation of abundance for each lineage, resulted in similar power of detecting microbiome association for each kernel type. This kind of trend was more obvious using eRLRT compared with eScore.

The power depended on the choice of kernel type, clustering or not and the sample size.

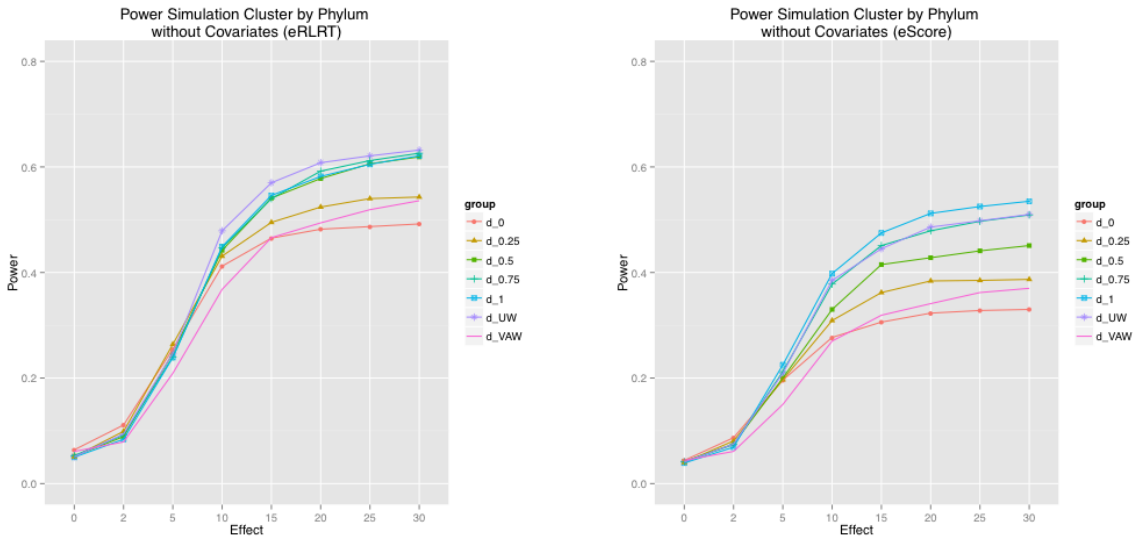

Figure 3.6: **Type I Error and Power of eRLRT and eScore Based on Different Kernels with Clustering (No X)** The OTUs were clustered by phylum to reconstructed the kernels. The outcome associated with the overall microbiome were simulated using the reconstructed kernels. Results are shown for eRLRT and eScore test. d_0, $d\_0.25$, $d\_0.5$, $d\_0.75$,d_1, d_UW and d_VAW represent test results from generalized UniFrac kernels with $\alpha = 0, 0.25, 0.5, 0.75$, weighted UniFrac kernel, unweighted UniFrac kernel and variance adjusted weighted UniFrac kernel, respectively. Sample size is 82. There are 30 individuals with 2-4 repeated measurements each.

These two simulation scenarios showed that the choice of kernel is important for being well powered to discover the associations between overall microbiome effect and outcomes. However, the choice of proper kernel was less essential when the counts data were clustered at the higher phylogenetic level. In practice, it's a good option for researchers to cluster the count when sample size is small and there is not enough information to choose the proper kernel type.

3.1.5 Scenario 5: Simulation for outcome was associated with two microbiome clusters

In simulation scenario 5, we performed the exact test on the association of one cluster while the outcome is associated with another OTU cluster. Table 3.3 and Table 3.4 show the type I error

Table 3.3: Type I Error of Testing Signal with or without Adjustment for Effect of $Firmicutes$ (Longitudial Data)

		Without Adjustment for Noise		With Adjustment for Noise	
Strength of Noise	0 (0 %)	5 (20 %)	30 (40 %)	5 (20 %)	30 (40 %)
Actinobacteria	0.066, 0.061	0.071, 0.050	0.077, 0.044	0.056, 0.039	0.043, 0.039
Bacteroidetes	0.063, 0.050	0.088, 0.066	0.101, 0.076	0.045, 0.052	0.051, 0.044
Fusobacteria	0.063, 0.051	0.081, 0.065	0.093, 0.085	0.035, 0.049	0.050, 0.044
Proteobacteria	0.069, 0.062	0.087, 0.072	0.089, 0.067	0.045, 0.043	0.042, 0.035
Other	0.055, 0.052	0.065, 0.046	0.117, 0.078	0.044, 0.058	0.049, 0.039

Type I error was evaluated via eRLRT and eScore test with no covariates using 1,000 simulated datasets. It was also evaluated with no adjustment for effect of $Firmicutes$ while adjusting for this effect. The outcomes were simulated with effect of $Firmicutes$ in which the effect size was set up as 0, 5 and 30. 0%, 20% and 40% indicates the power when testing the effect of $Firmicutes$. K_W, the weighted UniFrac kernel, was used to simulate the longitudinal outcomes. There are 30 individuals in total and $2-4$ repeated measurements for each individual.

rate of testing signal with and without adjustment for effect contributed by cluster $Firmicutes$. Specifically, Table 3.3 presents the result using the longitudinal microbiome data and Table 3.4 shows the result of baseline. The type I error rates in both tables without adjustment were higher with greater association between outcomes and phylum $Firmicutes$ effect. The type I error rate of $Actinobacteria$ didn't change as much as the others in Table 3.3, however, it is greatly inflated in Table 3.4. It may indicate that the correlation between phylum may change with time or treatment. Additionally, the difference in the type I error rates for eRLRT and eScore was greater with the association strength of $Firmicutes$ getting larger. After adjusting for the effect from $Firmicutes$, the type I error rate was no longer inflated.

Table 3.4: Type I Error of Testing Signal with or without Adjustment for Effect of $Firmicutes$ (Baseline Data)

		Without Adjustment for Noise		With Adjustment for Noise	
Strength of Noise	0 (0 %)	2.5 (48 %)	30 (86 %)	2.5 (48 %)	30 (86 %)
Actinobacteria	0.057, 0.051	0.129, 0.067	0.259, 0.091	0.064, 0.031	0.058, 0.042
Bacteroidetes	0.064, 0.044	0.105, 0.100	0.212, 0.207	0.040, 0.049	0.063, 0.051
Fusobacteria	0.066, 0.042	0.105, 0.073	0.145, 0.093	0.059, 0.035	0.050, 0.029
Proteobacteria	0.072, 0.062	0.103, 0.089	0.159, 0.134	0.034, 0.022	0.050, 0.018
Other	0.058, 0.050	0.099, 0.064	0.177, 0.096	0.064, 0.039	0.049, 0.047

Type I error was evaluated via eRLRT and eScore test with no covariates using 1,000 simulated baseline outcomes. It was also evaluated with no adjustment for effect of $Firmicutes$ while adjusting for this effect. The outcomes were simulated with effect of $Firmicutes$ in which the effect size was set up as 0, 5 and 30. 0%, 20% and 40% indicates the power when testing the effect of $Firmicutes$. K_W, the weighted UniFrac kernel, was used to simulate the outcomes. The sample size was 24.

3.2 Application to Longitudinal Pulmonary Microbiome Data

Pulmonary infection is a common and frequently fatal complication of individuals living with HIV infection, though little is known regarding the lung microbiome composition of this population. The analysis objectives were to determine relationships with clinical, immunological, and microbiological variables, to characterize the lung microbiome effect of HIV-infected patients and to further study which phyla have significant effects on lung function or the immune system. Pulmonary Function Tests (PFTs), including spirometry and diffusion capacity are a group of tests that measure how well the lungs take in and release air and how well they move gases such as oxygen from the atmosphere into the body's circulation. Spirometry is the test to measure static lung volumes, such as slow vital capacity (sVC), forced vital capacity (FVC) and dynamic volumes, such as forced expiratory volume in one second (FEV1), flow-volume loops. Diffusing capacity (or DLCO) is the carbon monoxide uptake from a single inspiration in a standard time (usually 10 seconds). The exhaled gas is tested to determine how much of the tracer gas was

absorbed during the breath. This will pick up diffusion impairments, for instance in pulmonary fibrosis. Researchers conducted a longitudinal study to examine the relationship of pulmonary function and immunological factor level with microbiota in human pulmonary samples from 30 individuals with HIV infection. 454 pyrosequencing of the gene 16S rRNA was used for profiling the microbiomes at baseline, 4 week, 1 year and 3 years for each individual. The total number of measurements is 100.

The previous studies about the relation between microbiome composition and pulmonary function identified that the number of bacteria belonging to $Firmicute$ and $Bacteroidetes$ phyla increased in chronic obstructive pulmonary disease (COPD). Intuitively, the composition of these two phyla might affect the lung function. However, analyses were restriceted to the change of numbers or ratios of microbiota rather than their effect on pulmonary function. Moreover, the previous analyses failed to adjusting for the covariates such as sex, race $etc.$ Thus, we applied our method to the dataset (microbiome profiles, pulmonary function test, confounders) to test for an association between pulmonary function and microbiome composition. We further adjusted for additional potential confounders, including age, sex, race and smoking status. We considered generalized UniFrac Distance which can detect a wider range of changes in composition to test the association.

Interestingly, the overall microbiome effect test using longitudinal and baseline data gave non-significant results while the $Bacteroidetes$ and $Firmicutes$ showed significant effects on FEV and DVA respectively when tested by phylum (p=0.01 and p=0.05 repectively). $Tunney$ and colleagues detected the compositional change of genus $Prevotella$ and $Veillonella$ belonging to $Bacteroidetes$ and $Firmicutes$ respectively in sputum samples from patients with clinically stable bronchiectasis. Therefore, we analyzed the effect of genus $Prevotella$ and $Veillonella$ on pulmonary function and found that these two genera $Prevotella$ (p=0.05 at baseline) and $Veillonella$ (p=0.03 in 3 years follow-up) have a significant effect on forced vital capacity. Thus, considering potential confounding, our results showed significant association between particular microbiome profiles and pulmonary function after the potential covariates were controlled for, providing new evidence to reaffirm the discoveries in the earlier studies. In addition to validating the former findings, these results demonstrates the utility and importance of our method with regard to accommodating covariates in a longitudinal study.

4

Discussion and Conclusion

Microbiome studies are now being extensively involved in individualized medicine, epidemiological, clinical and population-based studies. Most of previous microbiome studies focused on evaluating the variant microbe populations among the exposed or unexposed groups, while current research on microbiome association also consider that the microbial composition depends on a number of factors including, but not limited to, the age, ethncity, gender and geographic information of an individual. The method proposed in this paper is therefore more attractive than the previous studies due to it's ability of handling the issues together while controlling type I error and adequate power.

4.1 Advantages between Exact Tests and Existing Methods

First of all, the key advantage of the methods we propose is that it enables researchers to analyze the relationship between microbiome composition and outcome of interest with or without covariates in longitudinal studies. In particular, the distance-based PERMANOVA method cannot easily adjust for covariates and correlation structure within repeated measurements when evaluating the microbiome. Based on the simulation study results, it is necessary to adjust for the confounders especially when they are corrlated with the microbiome. Besides, the intensive computation of permutation procedure is another problem that cannot be neglected.

Even though the MiRKAT with one single kernel can accommodate confounding variables and the optimal MiRKAT can deal with different features of microbiome data, MiRKAT could not be adopted in the longitudinal case since it's not able to adjust for the correlation between

repeated measurements across time. The linear mixed model we proposed in this paper provides researchers one powerful way to deal with the longitudinal microbiome profile and to learn more about the cause-effect relationships, because it has the ability to show the patterns of microbiome effect over time and to discover the sleeper effects or connections over a long period of time. In contrast with the previous microbiome studies, it's not only a powerful method for testing microbiome association in longitudinal studies, but also an attractive and more accurate strategy for detecting the potential relationship up to genus level by adjusting for the effect from others.

The linear mixed model we proposed in this paper introduced multiple random effects in order to model the correlation structure and the microbiome effect. The current methods for multiple variance component tests are based on either asymptotic distribution or parametric bootstrap. Even in testing one variance component, the usual asymptotic chi-square distribution of the likelihood ratio and score statistics under the null leads to poor performance. The reason is that the null variance component parameter lies on the boundary of the parameter space. A novel approach introduced in this paper combines the recent development of a series exact (restrictive) likelihood ratio test and a strategy for reducing multiple variance components to a single case in LMM. Since the exact test is defined without parametric assumptions and evaluated without using approximate algorithms, we can be free of the extensive computational issue of parametric bootstrap and the parameter that lies on the boundary of the null hypothesis.

4.2 Conclusion

We propose microbiome association exact tests (MAETs) in the linear mixed model to detect the association between microbial community composition and a continuous outcome of interest in a longitudinal study. The covariates are modeled parametrically and the microbiome effects are modeled as non-parametric random effects. The kernel matrix, which captures the microbiome effect, is constructed via transformation from the UniFrac distance metric including the phylogenetic dependency information of OTUs. The proposed exact tests allow the incorporation of testing multiple variance components, enabling development of longitudinal microbiome association test. The MAETs package also enable researchers to test association of OTU clusters separately via constructing multiple kernel matrix as variance components in order to capture the microbiome effect separately. Additionally, the proposed MAETs allows detection of the

potential relationship at the more interesting genus level while controlling for the effect from the other part of microbiome composition. Our method is a natural extension to the currently available analysis frameworks. Simulations show that this approach has controlled type I error and superior power to test the microbiome effect in a longitudinal study. In real-data analysis, the MAETs provides new evidence to re-affirm that $Prevotella$ and $Veillonella$ have significant effect on pulmonary function. This application demonstrates the importance and utility of our method in longitudinal microbiome studies.

Even though our method merely focused on continuous outcomes with correlation and covariates, the linear mixed model can be extended to analyze dichotomous outcomes of interest in future research. Therefore, with increasing interest in detecting the microbiome association and identifying the variant microbial composition in clinical and epidemiological studies, the microbiome association exact tests in the linear mixed model can be a powerful tool to detect how changes in the human microbiome are associated with human health or disease by enabling analysis of more sophisticated microbiome profiles.

A

Principal Abbreviations

OTU: operational taxonomic units

VAW-UniFrac: variance adjusted weighted UniFrac distance

MiRKAT: microbiome reggression based kernel association test

LMM: Linear Mixed Effects Model

eLRT: exact likelihood ratio test

eRLRT: exact restricted likelihood ratio test

ZIBR: zero-inflated Beta regression model with random effects

PFTs: Pulmonary Function Tests

DLCO: diffusing capacity

MAETs: microbiome association exact tests

References

[1] R MacDougall. Nih human microbiome project defines normal bacterial makeup of the body. *Website: http://www. nih. gov/news/health/jun2012/nhgri-13. htm*, 2012.

[2] John Penders, Ellen E Stobberingh, Piet A van den Brandt, and Carel Thijs. The role of the intestinal microbiota in the development of atopic disorders. *Allergy*, 62(11):1223–1236, 2007.

[3] Ann M O'Hara and Fergus Shanahan. The gut flora as a forgotten organ. *EMBO reports*, 7(7):688–693, 2006.

[4] Peter J Turnbaugh, Micah Hamady, Tanya Yatsunenko, Brandi L Cantarel, Alexis Duncan, Ruth E Ley, Mitchell L Sogin, William J Jones, Bruce A Roe, Jason P Affourtit, et al. A core gut microbiome in obese and lean twins. *nature*, 457(7228):480–484, 2009.

[5] Aleksandar D Kostic, Dirk Gevers, Chandra Sekhar Pedamallu, Monia Michaud, Fujiko Duke, Ashlee M Earl, Akinyemi I Ojesina, Joonil Jung, Adam J Bass, Josep Tabernero, et al. Genomic analysis identifies association of fusobacterium with colorectal carcinoma. *Genome research*, 22(2):292–298, 2012.

[6] Ruth E Ley, Fredrik Bäckhed, Peter Turnbaugh, Catherine A Lozupone, Robin D Knight, and Jeffrey I Gordon. Obesity alters gut microbial ecology. *Proceedings of the National Academy of Sciences of the United States of America*, 102(31):11070–11075, 2005.

[7] Martin J Blaser, Yu Chen, and Joan Reibman. Does helicobacter pylori protect against asthma and allergy? *Gut*, 57(5):561–567, 2008.

[8] Timothy L Cover and Martin J Blaser. Helicobacter pylori in health and disease. *Gastroenterology*, 136(6):1863–1873, 2009.

[9] Paul Baumann and Nancy A Moran. Non-cultivable microorganisms from symbiotic associations of insects and other hosts. *Antonie van Leeuwenhoek*, 72(1):39–48, 1997.

[10] Gene W Tyson, Jarrod Chapman, Philip Hugenholtz, Eric E Allen, Rachna J Ram, Paul M Richardson, Victor V Solovyev, Edward M Rubin, Daniel S Rokhsar, and Jillian F Banfield. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37–43, 2004.

[11] Roger S Lasken. Genomic sequencing of uncultured microorganisms from single cells. *Nature Reviews Microbiology*, 10(9):631–640, 2012.

[12] Antonia Suau, Régis Bonnet, Malène Sutren, Jean-Jacques Godon, Glenn R Gibson, Matthew D Collins, and Joel Doré. Direct analysis of genes encoding 16s rrna from complex communities reveals many novel molecular species within the human gut. *Applied and environmental microbiology*, 65(11):4799–4807, 1999.

[13] Mark Blaxter, Jenna Mann, Tom Chapman, Fran Thomas, Claire Whitton, Robin Floyd, and Eyualem Abebe. Defining operational taxonomic units using dna barcode data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1462):1935–1943, 2005.

[14] Patrick D Schloss, Sarah L Westcott, Thomas Ryabin, Justine R Hall, Martin Hartmann, Emily B Hollister, Ryan A Lesniewski, Brian B Oakley, Donovan H Parks, Courtney J Robinson, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*, 75(23):7537–7541, 2009.

[15] Nadja Larsen, Finn K Vogensen, FW Van Den Berg, Dennis Sandris Nielsen, Anne Sofie Andreasen, Bente K Pedersen, Waleed Abu Al-Soud, Soren J Sorensen, Lars H Hansen, and Mogens Jakobsen. Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PloS one*, 5(2):e9085, 2010.

[16] Catherine Lozupone and Rob Knight. Unifrac: a new phylogenetic method for comparing microbial communities. *Applied and environmental microbiology*, 71(12):8228–8235, 2005.

[17] Catherine A Lozupone, Micah Hamady, Scott T Kelley, and Rob Knight. Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Applied and environmental microbiology*, 73(5):1576–1585, 2007.

[18] Qin Chang, Yihui Luan, and Fengzhu Sun. Variance adjusted weighted unifrac: a powerful beta diversity measure for comparing communities based on phylogeny. *BMC bioinformatics*, 12(1):118, 2011.

[19] Jun Chen, Kyle Bittinger, Emily S Charlson, Christian Hoffmann, James Lewis, Gary D Wu, Ronald G Collman, Frederic D Bushman, and Hongzhe Li. Associating microbiome composition with environmental covariates using generalized unifrac distances. *Bioinformatics*, 28(16):2106–2113, 2012.

[20] Zhenqiu Liu, William Hsiao, Brandi L Cantarel, Elliott Franco Drábek, and Claire Fraser-Liggett. Sparse distance-based learning for simultaneous multiclass classification and feature selection of metagenomic data. *Bioinformatics*, 27(23):3242–3249, 2011.

[21] Eric E Schadt, Michael D Linderman, Jon Sorenson, Lawrence Lee, and Garry P Nolan. Computational solutions to large-scale data management and analysis. *Nature Reviews Genetics*, 11(9):647–657, 2010.

[22] JULIA Fukuyama, PAUL J McMurdie, Les Dethlefsen, David A Relman, and Susan Holmes. Comparisons of distance methods for combining covariates and abundances in microbiome studies. In *Pac Symp Biocomput*. World Scientific, 2012.

[23] Justin Kuczynski, Elizabeth K Costello, Diana R Nemergut, Jesse Zaneveld, Christian L Lauber, Dan Knights, Omry Koren, Noah Fierer, Scott T Kelley, Ruth E Ley, et al. Direct sequencing of the human microbiome readily reveals community differences. *Genome Biol*, 11(5):210, 2010.

[24] Gary D Wu, Jun Chen, Christian Hoffmann, Kyle Bittinger, Ying-Yu Chen, Sue A Keilbaugh, Meenakshi Bewtra, Dan Knights, William A Walters, Rob Knight, et al. Linking

long-term dietary patterns with gut microbial enterotypes. *Science*, 334(6052):105–108, 2011.

[25] Brian H McArdle and Marti J Anderson. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology*, 82(1):290–297, 2001.

[26] Mingxiu Hu, Yi Liu, and Jianchang Lin. Topics in applied statistics.

[27] Ni Zhao, Jun Chen, Ian M Carroll, Tamar Ringel-Kulka, Michael P Epstein, Hua Zhou, Jin J Zhou, Yehuda Ringel, Hongzhe Li, and Michael C Wu. Testing in microbiome-profiling studies with mirkat, the microbiome regression-based kernel association test. *The American Journal of Human Genetics*, 96(5):797–807, 2015.

[28] Nan M Laird and James H Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.

[29] Harald Cramér. *Mathematical methods of statistics*, volume 9. Princeton university press, 1945.

[30] Gunnar Kulldorff. On the conditions for consistency and asymptotic efficiency of maximum likelihood estimates. *Scandinavian Actuarial Journal*, 1957(3-4):129–144, 1957.

[31] Herman Chernoff. On the distribution of the likelihood ratio. *The Annals of Mathematical Statistics*, pages 573–578, 1954.

[32] Patrick AP Moran. Maximum-likelihood estimation in non-standard conditions. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 70, pages 441–450. Cambridge Univ Press, 1971.

[33] Steven G Self and Kung-Yee Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610, 1987.

[34] Daniel O Stram and Jae Won Lee. Variance components testing in the longitudinal mixed effects model. *Biometrics*, pages 1171–1177, 1994.

[35] Sonja Greven, Ciprian M Crainiceanu, Helmut Küchenhoff, and Annette Peters. Restricted likelihood ratio testing for zero variance components in linear mixed models. *Journal of Computational and Graphical Statistics*, 2012.

[36] Ciprian M Crainiceanu and David Ruppert. Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):165–185, 2004.

[37] Ciprian Crainiceanu, David Ruppert, Gerda Claeskens, and Matthew P Wand. Exact likelihood ratio tests for penalised splines. *Biometrika*, 92(1):91–103, 2005.

[38] J Ofversten. Exact tests for variance components in unbalanced mixed linear models. *Biometrics*, pages 45–57, 1993.

[39] Ronald Christensen. Exact tests for variance components. *Biometrics*, pages 309–314, 1996.