Title: Testing in Microbiome Profiling Studies with the Microbiome Regression-based Kernel Association Test (MiRKAT)

Running Title: Mirobiome Association Test with MiRKAT

Ni Zhao [1], Jun Chen [2,*], Ian M. Carroll [3], Tamar Ringel-Kulka [4], Michael P. Epstein [5], Hua Zhou [6], Jin J. Zhou [7], Yehuda Ringel [3], Hongzhe Li [8], Michael C. Wu[1,*]

[1]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, 98109

[2]Division of Biomedical Statistics and Informatics and Center for Individualized Medicine, Mayo Clinic, Rochester, MN, 55905

[3]Department of Medicine, The University of North Carolina at Chapel Hill, Chapel Hill, NC, 27516

[4]Department of Maternal and Child Health, The University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599

[5]Department of Human Genetics, Emory University, Atlanta, GA, 30322

[6]Department of Statistics, North Carolina State University, Cary, Raleigh, NC, 27695

[7]Division of Epidemiology and Biostatistics, University of Arizona,Tucson, AZ, 85724

[8]Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA, 19014

March 31, 2015

Address for Correspondence:

Michael C. Wu

Public Health Sciences Division

Fred Hutchinson Cancer Research Center

1100 Fairview Avenue North, M3-C102

Seattle, WA 98109-1024

Phone: (206) 667-6603

Email: mcwu@fhcrc.org

and

Jun Chen

Division of Biomedical Statistics and Informatics and Center for Individualized Medicine

Mayo Clinic

200 First ST SW, Harwick Blg 3-19

Rochester, MN 55905

Phone: (507)293-2832

Email: chen.jun2@mayo.edu

Abstract

High-throughput sequencing technology has enabled population based studies into the role of the human microbiome in disease etiology and exposure response. Distance based analysis is a popular strategy to evaluate the overall association between microbiome diversity and outcome, wherein the phylogenetic distance between individuals' microbiome profiles is computed and tested for association via permutation. Despite their practical popularity, distance based approaches suffer from important challenges, especially in the difficulty in selecting the best distance and in extending the methods to alternative outcomes, such as survival outcomes. We propose the microbiome regression-based kernel association test (MiRKAT), which directly regresses the outcome on the microbiome profiles via the semi-parametric kernel machine regression framework. MiRKAT allows for easy covariate adjustment and extension to alternative outcomes while nonparametrically modeling the microbiome through a kernel which incorporates phylogenetic distance. A variance component score statistic is used to test for the association with analytical p-value calculation. The model also allows simultaneous examination of multiple distances, alleviating the problem of choosing the best distance. Simulations demonstrate that MiRKAT provides correctly controlled type I error and adequate power in detecting overall association. "Optimal" MiRKAT which considers multiple candidate distances is robust in that it suffers from little power loss compared to when the best distance is used and can achieve tremendous power gain compared to when a poor distance is chosen. Finally, we apply MiRKAT to real microbiome data sets to show that microbial communities are associated with smoking and with fecal protease levels

after controlling for confounders.

Key Words: Microbiome composition; Phylogenetic distance; Kernel machine regression; Multi-kernel testing.

Introduction

The advent of massively parallel sequencing has enabled high-throughput profiling of microbiota in a large number of samples via targeted sequencing of the 16S rDNA gene,[1–4] the sequence of which contains information about species identity. Knowledge on how microbial communities differ across individuals can provide key information on the role of communities in relation to variation in biological and clinical variables and is essential for gaining a broader understanding of biological mechanisms underlying disease and response to exposures.[5–9] Although considerable resources have been devoted to sequencing technologies and to quantifying individual taxa, successful application of microbial profiling to studying biomedical conditions requires novel statistical methods to efficiently test for associations with microbial diversity.

A popular strategy for evaluating the association between overall microbiome composition and outcomes of interest utilizes distance or dissimilarity based analysis, referred to as just distance based analysis for simplicity. Using standard methods, the 16S gene tags are clustered based on their sequence similarity to form Operational Taxonomic Units (OTUs), which can be essentially considered as surrogates for biological taxa. Distance metrics are then constructed to measure the phylogenetic or taxonomic dissimilarity between each pair of samples by exploring the phylogenetic relationship or the absolute and relative abundance of different taxa. Then to assess the association between the microbiome diversity and an outcome variable of interest, the pair-wise distance between each pairs of samples is compared to the distribution of the outcome variable. For categori-

cal outcome variables, this is essentially comparing the pair-wise distances within and between categories. Operationally, multivariate analysis[10] or the top principal coordinates (PCo)[11] of the matrix of pairwise distances are used to test for associations via permutation.

Among the many possible distances, the UniFrac distances are the most popular distances in the literature that are constructed based on a phylogenetic tree relating taxa to one another.[12,13] There are several different versions of UniFrac distances. The original, unweighted UniFrac distance between any pair of microbial communities is calculated as the proportion of the total branch length within the tree which leads to un-shared taxa (i.e. taxa in one community but not the other). Thus, the UniFrac distance primarily considers only the species presence and absence information and is most efficient in detecting abundance change in rare lineages since more prevalent species are likely to be present in all individuals. Weighted UniFrac distance uses species abundance information to weight the UniFrac distances, and thus is more powerful to detect changes in common lineages. Generalized UniFrac distances[14] were introduced as compromise between weighted and unweighted UniFrac distance, which down-weight their emphasis on either abundant or rare lineages and therefore are more powerful to detect changes in OTU clusters with modest abundance. Generalized UniFrac distance involves an additional parameter α that generalized UniFrac distance with $\alpha = 1$ is equivalent to weighted UniFrac distance. A range of other distances that do not incorporate phylogeny are also available. For example, Bray-Curtis dissimilarity, which is also commonly used, quantifies the taxonomic dissimilarity between two

different sites based on counts at each site. Similarly, Euclidean distance can also be used and is frequently thought to be similar to weighted UniFrac distance since abundance information from common taxa tends to dominate.

Despite successes, distance based analysis suffers from a number of limitations. First, as noted, many different distance metrics have been developed. While there are similarities, they are designed to capture distance differently leading to differential performance across different scenarios. This creates problems in choosing a particular metric to use as the best metric for any particular data set depends on the unknown true state of nature. A non-optimal distance metric will reduce power to discover true associations. Using multiple metrics and cherry picking the best result will result in inflated type I rates and lead to large numbers of spurious results. Beyond difficulties in choosing a particular distance metric, the need for permutation can be computationally expensive. Furthermore, the analysis framework is not easily interpretable or allow for easy covariate adjustment. Consequently, extending such approaches to accommodate more sophisticated outcomes such as survival or multivariate information is challenging.

We propose in this paper the microbiome regression-based kernel association test (MiRKAT), a flexible regression approach for testing the association between microbial community profiles and a continuous or dichotomous variable of interest such as an environmental exposure or disease. MiRKAT formalizes and extends the strategy of Chen and Li (2013)[15] to use the kernel machine regression framework, previously developed for genotyping data,[16–18] to directly regress the variable of interest on the covariates (including potential confounders) and the

microbiome compositional profiles. The kernel is a measure of similarity between samples' microbiome compositions and characterizes the relationship between the microbiome and the variable of interest. We propose to use kernels that incorporate phylogenetic relationships among taxa by transforming existing distance metrics into similarities. A variance component score test can be used to rapidly obtain a p-value for the association between microbial community profiles and the variable of interest.

In addition to fast computation, use of the kernel machine approach enables flexible modeling and testing, while still incorporating phylogenetic information and naturally accommodating covariates, under a well-studied, interpretable, and statistically rigorous framework. Beyond extensions to allow alternative types of outcomes, the framework allows for simultaneous examination of multiple distance metrics. This enables development of the "optimal" MiRKAT that has high power in the omnibus. We demonstrate through simulations and analysis of real data that MiRKAT and optimal MiRKAT are easy to apply and can be more robust than existing tests with well controlled type I error across a range of models for both continuous and dichotomous variables. We also explicitly establish connections between MiRKAT and existing distance based approaches.

The well-studied kernel machine framework forms the statistical underpinnings for our work, which is strength as this allows leverage of existing machinery within a rigorous framework. However, MiRKAT differs from previous, related kernel methods in the need to accommodate unique features of microbiome data. In particular, we tailor the approach to accommodate microbiome data by adopt-

ing kernels based on dissimilarity measures commonly used in microbiome compositional analysis. Furthermore, microbiome studies usually have more modest sample sizes, yet the kernels built on standard distance metrics are frequently of full rank with poor eigenvalue behavior. Consequently, in contrast to previous analytic[17–19] and perturbation based[20] p-value calculation approaches which do not control type I error well, we use alternative small sample corrections[21,22] and permutation methods. MiRKAT differs from our earlier conference manuscript[15] in that we formalize and fully flesh out the overall framework, we explicitly relate the approach to existing distance methods, we use alternative small sample corrections to control type I error, and we develop the optimal MiRKAT method for testing across choices of distance metrics.

Methods

Notationally, we assume that n samples have been collected and their microbial communities profiled. For the i^{th} subject, let y_i denote the outcome variable of interest, $\mathbf{Z}_i = (Z_{i1}, Z_{i2}, ..., Z_{ip})'$ denote the abundances of all OTUs for individual i and p is the total number of OTUs, and $\mathbf{X}_i = (X_{i1}, X_{i2}, ..., X_{im})'$ are covariates that we want to control for, such as age, gender, and other clinical and environmental variables which are suspected to influence microbial community diversity and related to outcomes. The goal is to test for association between the outcome and microbial profiles while adjusting for covariates $\mathbf{X}$. Note that we will refer to $\mathbf{y}$ as an "outcome" that depends on the microbiome composition while in some situa-

tions it may be a variable that is thought to influence microbial diversity; however, since our goal is association testing rather than causal modeling, the distinction does not affect the validity of our method given the duality.[23] We first consider the problem of testing under a single distance metric (kernel) and then extend the approach to optimally accommodate multiple distances simultaneously.

MiRKAT based on a single kernel

The intuition behind kernel machine framework is that it compares pairwise similarity in the outcome variable to pairwise similarity in the microbiome profiles, with high correspondence suggestive of association. MiRKAT exploits the kernel machine regression framework to relate the covariates and the microbiota profiles to the outcomes. Specifically, for a continuous outcome variable we use the linear kernel machine model:

$$y_i = \beta_0 + \boldsymbol{\beta}'\mathbf{X}_i + f(\mathbf{Z}_i) + \varepsilon_i \tag{1}$$

and for a dichotomous outcome variable (e.g. $\mathrm{y} = 1/0$ for case/control) we use the logistic kernel machine model:

$$\text{logit}(P(y_i = 1)) = \beta_0 + \boldsymbol{\beta}'\mathbf{X}_i + f(\mathbf{Z}_i) \tag{2}$$

where β_0 is the intercept, $\boldsymbol{\beta} = [\beta_1, ..., \beta_m]'$ is the vector of regression coefficients for the m covariates, and for continuous phenotypes ε_i is an error term with mean zero and variance σ^2. This regression framework can be easily extended to other more complicated outcomes, such as survival or multivariate outcomes.

The relationship between the microbiome profile and the outcome variable is fully characterized by function $f(\cdot)$: testing that there is no association between microbiome composition and the outcome is equivalent to testing that $f(\mathbf{Z}) = 0$. Under the kernel machine regression framework, $f(\mathbf{Z}_i)$ is assumed to be from a reproducing kernel Hilbert space $\mathcal{H}_k$ generated from a positive definite kernel function $K(\cdot, \cdot)$ such that $f(\mathbf{Z}_i) = \sum_{i'=1}^{n} \alpha_{i'} K(\mathbf{Z}_i, \mathbf{Z}_{i'})$ for some $\alpha_1, \alpha_2, ..., \alpha_n$.

The kernel measures the similarity between different individuals and different choices of $K(\mathbf{Z}_i, \mathbf{Z}_{i'})$ corresponds to different underlying models. For example, setting $K(\mathbf{Z}_i, \mathbf{Z}_{i'}) = \sum_{j=1}^{p} Z_{ij} Z_{i'j}$ implies $f(\mathbf{Z}_i) = \sum_{j=1}^{p} Z_{ij}\beta_j$, i.e. the model is linear. Therefore, by changing the kernel function, one is implicitly changing the model being used. Using more sophisticated kernels will result in more complex models which can allow for OTU interactions, nonlinear OTU effects or incorporation of phylogenetic relationships among OTUs. The matrix of pair-wise similarities between pairs of individuals is defined as the kernel matrix $\mathbf{K}$, where the (i, i')-th element of $\mathbf{K}$ is $K(\mathbf{Z}_i, \mathbf{Z}_{i'})$.

For microbiome composition data, the OTUs are related by a phylogenetic tree. Kernels that exploit the degree of divergence between different sequences can be much more powerful compared to similarity measures that ignore the phylogenetic tree information. We can construct the kernel matrix, which measures similarities between the microbiome composition among subjects, by exploiting the correspondence with the well-defined distance metrics, which measure dissimilarities between subjects. Specifically, we can construct the kernel matrix via the

following transformation of the phylogenetic or taxonomic distance metrics:

$$\mathbf{K} = -\frac{1}{2}(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n})\mathbf{D}^2(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n}) \tag{3}$$

where $\mathbf{D} = [d_{ij}]$ is the pair wise distance matrix, e.g. weighted or unweighted UniFrac distance or the Bray-Curtis dissimilarity, I is the identity matrix, $\mathbf{1}$ is a vector of 1's and $\mathbf{D}^2$ represents element wise square. For each distance metric, we can construct the corresponding kernel matrix, e.g., weighted or unweighted UniFrac kernels ($\mathbf{K}_W$ and $\mathbf{K}_U$) can be constructed based on weighted or unweighted distance metrics. This choice of kernel is in line with the relationship between kernel machine regression and distance based regression[24] in that it can recover the original distances using standard kernel operation: $d_{ij}^2 = K_{ii} + K_{jj} - 2K_{ij}$. Further, to ensure that the $\mathbf{K}$ to be a positive semi-definite matrix, we apply the same positive semi-definiteness correction procedure as in[15] that we first perform an eigen decomposition of $\mathbf{K} = \mathbf{U}\Lambda\mathbf{U}'$ with $\Lambda = diag(\lambda_1, \cdots, \lambda_n)$ being the eigen values and then reconstruct using the absolute eigen values $K^* = \mathbf{U}\Lambda^*\mathbf{U}'$ and $\Lambda^* = diag(|\lambda_1|, \cdots, |\lambda_n|)$.

When only a single kernel is considered, the estimation of the coeffcients β and $f(\mathbf{Z})$ are conducted by maximizing the following penalized log-likelihood:

$$\begin{aligned} pl(f, \beta) &= \sum_{i=1}^{n} \log L(f, \beta; y_i, x_i, z_i) - \frac{1}{2}\lambda||f||^2_{\mathcal{H}_k} \\ &= \sum_{i=1}^{n} \log L(f, \beta; y_i, x_i, z_i) - \frac{1}{2}\lambda\alpha' K\alpha \end{aligned}$$

Through an important relationship between kernel machine regression and mixed models,[25–27] $f(\mathbf{Z})$ can be viewed as a subject specific random effect which follows a distribution with mean 0 and variance $\tau\mathbf{K}$. Then testing for an association between the microbiome composition and the outcome is equivalent to testing the null hypothesis that $H_0 : \tau = 0$. Under the mixed model framework, this can be done using a standard variance component score test.[28]

In particular, the score statistic is computed as

$$Q = \frac{1}{2\phi}(\mathbf{y} - \hat{\mathbf{y}}_0)'\mathbf{K}(\mathbf{y} - \hat{\mathbf{y}}_0) \tag{4}$$

where $\hat{\mathbf{y}}_0$ is the predicted mean of $\mathbf{y}$ under H_0, i.e. $\hat{\mathbf{y}}_0 = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}'\mathbf{X}$ for continuous traits and $\hat{\mathbf{y}}_0 = \text{logit}^{-1}(\hat{\beta}_0 + \hat{\boldsymbol{\beta}}'\mathbf{X})$ for dichotomous traits, $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}$ are estimated under the null model by regressing $\mathbf{y}$ on only the covariates $\mathbf{X}$ and ϕ is the dispersion parameter. For the linear kernel machine regression case, $\phi = \hat{\sigma}_0^2$ where $\hat{\sigma}_0^2$ is the estimated residual variance under the null model. In the logistic kernel machine regression, $\phi = 1$.

Under the null hypothesis, Q asymptotically follows a weighted mixture of χ^2 distributions and p-value can be analytically obtained through higher order moment matching[29] or exact methods[30,31] with possible small sample adjustments via resampling.[19] However, the comparatively small sample sizes for many microbiome studies and the complexity of the kernels considered here (often of full rank with erratic eigenvalue behavior) lead to very conservative tests. Previously considered Satterthwaite methods[15] lead to type I error inflation. Thus, MiRKAT further considers the use of new, alternative small sample adjustments for both

continuous and dichotomous traits.[21,22]

A key advantage of the score test is that it only requires fitting the null model $y_i = \beta_0 + \boldsymbol{\beta}'\mathbf{X}_i + \varepsilon_i$ for continuous traits and $\text{logit}(P(y_i = 1)) = \beta_0 + \boldsymbol{\beta}'\mathbf{X}_i$ for dichotomous traits. Consequently, MiRKAT allows for fast, supervised, distance-based association testing under a regression framework that permits controls for potential confounding.

As the proposed test is a score test, all the parameters are estimated under the null model (linear regression or logistic regression), i.e. $f(\mathbf{Z})$ does not need to be estimated. This means that even if a poor kernel is chosen, the test is still statistically valid. Better choices of kernels simply improve power. From the perspective of testing, a metric that better reflects the true relationship between the microbiome compositional profiles and the outcome will result in substantially higher power.

Optimal MiRKAT based on multiple kernels

As noted, although MiRKAT is valid even if a poor kernel is chosen, better choices of kernel can lead to improved power. Unfortunately, the best kernel to use requires knowledge on how the microbiome influences the outcome. This is unknown *a priori* as knowledge of this would preclude need for analysis. Therefore, in this section, we develop the optimal MiRKAT which extends MiRKAT to simultaneously consider multiple possible kernels.

Suppose that we have a set of ℓ different candidate kernels $\mathbf{K}_1, ..., \mathbf{K}_\ell$, such as unweighted UniFrac, weighted UniFrac, Bray-Curtis kernels, etc., which are

constructed from corresponding distance metrices using equation (3).

The intuition behind the optimal MiRKAT is that it will consider testing using each individual kernel, obtain the p-value for each of the tests, select the minimum p-value and then adjust for having taken the minimum via a multiple comparison technique. If sample sizes are large, this can be accomplished via the perturbation based approach of Wu et al (2013).,[20] but when the sample size is more modest, we can apply a residual permutation approach to obtain the empirical null distribution of the test statistic. Specifically, we use the following procedure:

1. Fit the null linear or logistic regression model by regressing $\mathbf{y}$ on $\mathbf{X}$ and obtain the residuals $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}_0$ where $\hat{\mathbf{y}}_0$ is the estimated value of $\mathbf{y}$ based on the null model..

2. For each $\mathbf{K}_k$, calculate $Q_k = \frac{1}{2\phi}\mathbf{r}'\mathbf{K}_k\mathbf{r}$ and corresponding p values p_k through the asymptotically distribution of Q_k. Then the minimum p-value across all the ℓ kernels is $p_o = \min_{k\in(1,\ldots,\ell)} \mathbf{p}_k$.

3. Residual permutation is used to obtain the null distribution of p_o to accommodate the fact that we have considered multiple kernels.

   (a) For a continuous outcome, the permutation approach of Freeman and Lane (1983)[32] is used. Specifically, for each permutation j,

      i. Reshuffle the residuals $\mathbf{r}$ to obtain the permuted residuals $\mathbf{r}^j$

      ii. Create new values of $\mathbf{y}^j$ as $\mathbf{y}^j = \hat{\mathbf{y}}_0 + \mathbf{r}^j$.

      iii. Consider $\mathbf{y}^j$ as new outcome. Refit the null linear regression model by regressing $\mathbf{y}^j$ on $\mathbf{X}$ to obtain the estimated residuals $\hat{\mathbf{r}}^j$ and $\hat{\phi}^j$

for calculation of the the score statistic $Q_k^j = \frac{1}{2\hat{\phi}^j}\hat{\mathbf{r}}^{j\prime}\mathbf{K}_k\hat{\mathbf{r}}^j$ using each kernel. Kernel specific p-value p_k^j can be obtained by comparing Q_k^j to the same asymptotic distribution as in step 2.

iv. Obtain $p_o^j = \min_{k\in(1,\ldots,\ell)} \mathbf{p}_k^j$

(b) For a dichotomous outcome, we use the permutation approach of Epistein *et al.* (2012),[33] which uses Fisher's non-central hypergeometric distribution to generate permuted $1/0$ outcome values. Specifically,

i. Obtain the estimated odds of being a case for each individual sample, i.e. $\exp(\hat{\beta}_0 + \hat{\beta}'\mathbf{X}_i)$ where $\hat{\beta}_0$ and $\hat{\beta}$ are the estimated coefficients under the null logistic regression model as in step 1.

ii. For each permutation j, generate new binary outcomes for each based on the estimated odds using the Fisher's non-central hypergeometric distribution (modified version BiasedUrn package[34] in R).

iii. The permuted outcome is subsequently used to calculate the score statistic Q_k^j as in step 2 for each kernel and the kernel specific p-value p_k^j by comparing Q_k^j to the same asymptotical mixture of χ^2 distribution.

iv. Obtain $p_o^j = \min_{k\in(1,\ldots,\ell)} \mathbf{p}_k^j$

4. Repeat step 3 for a large number of times B, to form the an empirical null distribution for p_o.

5. Calculate the final p-value as $p = \frac{1}{B}\sum_{b=1}^{B} I(p_o > p_o^b)$.

For each permutation j, $p_1^j, \cdots, p_\ell^j$ are calculated using the same set of permuted outcomes and thus correlated; taking the minimum p-value across different kernels accounts for this correlation. Although the optimal MiRKAT requires permutation for the final p-value calculation, it only estimates residuals under each permutated data using the null model, which essentially equates to finding the QR residuals for continuous outcomes or logistic regression for binary outcomes, and thus can be done very fast. Additionally, for each kernel, each Q_k^j follows the same weighted mixture of χ^2 distribution with the weights and degree of freedom need to be estimated only once.

Simulation study

We conducted simulation studies under a range of scenarios in order to verify that MiRKAT correctly controls type I error rate and to assess the relative power of MiRKAT using different kernels as well as the power of optimal MiRKAT.

We first simulated microbiome data sets following the general approach of Chen and Li (2013)[15] which has been shown to generate simulated data reflective of real OTU counts. In particular, we simulated data sets comprised of $n = 100$, 200 or 500 individuals. Then the OTU information for each individual in a simulated data set was generated from a dirichlet-multinomial distribution which accommodates the over-dispersion of OTU counts. To employ realistic parameter values for the dirichlet-multinomial distribution, we estimated the dispersion parameters and the proportion means from the real upper respiratory tract microbiome data set of Charlson *et al.*(2010)[35] which consists of 856 OTUs measured on

each of 60 samples. Then for each individual, we generated OTU counts on the same 856 OTUs using the estimated parameters and assumed 1000 total counts per sample. We considered two simulation scenarios for both continuous outcomes and dichotomous outcomes, which differ in how the OTUs are related to the outcome.

Under *Simulation Scenario 1*, the outcome is related to a group of taxa that depends on a phylogenetic tree. Specifically, we partitioned all the OTUs into 20 clusters (lineages) by performing PAM (Partition Around Medoid) based on the OTU distance matrix. The abundance of these OTU clusters varies greatly, with each OTU cluster corresponding to some possible bacterial lineage. We then chose a relatively abundant OTU cluster which constitutes of 19.4% of the total OTU reads to be related to the outcome using the model. For continuous outcomes, we simulated under the model:

$$y_i = 0.5X_{1i} + 0.5X_{2i} + \beta \text{scale}(\sum_{j \in \mathcal{A}} Z_{ij}) + \varepsilon_i \tag{5}$$

where $\varepsilon_i \sim N(0, 1)$.

For dichotomous outcomes, we simulated under the model

$$\text{logit}(E(y_i|\mathbf{X}_i, \mathbf{Z}_i)) = 0.5\text{scale}(X_{1i} + X_{2i}) + \beta \text{scale}(\sum_{j \in \mathcal{A}} Z_{ij})) \tag{6}$$

For both continuous and dichotomous outcomes, X_{1i} and X_{2i} are covariates to be adjusted for, and $\mathcal{A}$ denotes the indices of the OTU's the in the selected cluster. The "Scale" function standardizes the total OTU abundance in the asso-

ciated cluster to have mean 0 and standard deviation of 1. X_{1i} were simulated as Bernoulli random variables with success probability 0.5. For X_{2i}, we considered situations in which X_{2i} and microbiome profiles ($\mathbf{Z}$) were correlated and in which the X_{2i} and $\mathbf{Z}_i$ are independent. In the simulation that X_{2i} and $\mathbf{Z}_i$ are independent, X_{2i} is simulated as $N(0,1)$. For the case when X_{2i} and $\mathbf{Z}_i$ are correlated, we let $X_{2i} = \text{scale}\,(\sum_{j\in\mathcal{A}} Z_{ij}) + N(0,1)$.

Under *Simulation Scenario 2* the outcome is associated with the 10 most abundant OTUs in all the samples, without regard to the phylogeny. In particular, instead of clustering the OTUs based on the phylogenetic relationship, we simply selected the 10 OTUs with the largest average number of reads across all samples. Then the continuous outcome was simulated as

$$y_i = 0.5X_{1i} + 0.5X_{2i} + \beta\text{scale}(\sum_{j\in\mathcal{A}} \frac{Z_{i(j)}}{\bar{Z}_{(}j)}) + \varepsilon_i \tag{7}$$

The dichotomous outcome was simulated as

$$\text{logit}(E(y_i|\mathbf{X}_i, \mathbf{Z}_i)) = 0.5\text{scale}(X_{1i} + X_{2i}) + \beta\text{scale}(\sum_{j\in\mathcal{A}} \frac{Z_{i(j)}}{\bar{Z}_{(}j)}) \tag{8}$$

where $\varepsilon_i \sim N(0,1)$, X_{1i} and X_{2i} are defined as earlier but $\mathcal{A}$ note denotes the set of 10 most abundant OTU's and $\bar{Z}_{(}j)$ is the average reads for the j^{th} OTU across samples. The OTU reads were divided by its corresponding average to avoid the situation that a single or a few OTUs can dominate the total effect.

The additional covariates $\mathbf{X}$ were simulated as before and we again consider the scenario in which the covariates are associated with microbiome and the sce-

nario in which the covariates are independent of the microbiome.

For both simulation scenarios, we considered using the weighted and unweighted UniFrac kernel (K_W and K_U), Bray-Curtis kernel (K_{BC}), and four generalized UniFrac Kernels with α values chosen as $0, 0.25, 0.5$ and 0.75 , which are denoted as $K_0, K_{0.25}, K_{0.5}$,and $K_{0.75}$. All these kernels are computed from the corresponding distances. We selected these particular kernels (distances) for consideration since they represent a range of different classes of kernels: the UniFrac based methods utilize phylogenetic relationships while the Bray-Curtis does not, and the weighted and general UniFracs take into account the abundance information to differing degrees while the un-weighted UniFrac does not.

We applied MiRKAT using each single kernel to the simulated data sets to test for associations between the simulated OTUs (**Z**) and the outcome (**y**). Additionally, we also applied optimal MiRKAT. We applied tests with and without adjustment for the potential confounders **X**. For comparison, we further considered a naive Bonferroni adjusted test, which selects the minimum p-value across all the single kernel testing and uses $\ell * minP$ to be the final p-value where $minP$ is the smallest p-value across all the single kernel tests and ℓ is the total number of tests. For each choice of sample size n, simulation scenario, and correlation structure between microbiome and covariates, we conducted 5000 simulations with $\beta = 0$ to examine for type I error rate. To assess the statistical power of the tests across both simulation scenarios, we varied values of the coefficient β and conducted 2000 simulations for each choice of sample size, simulation scenario, correlation structure, and value of β.

Results

In this section, we present the simulation results examining the performance of our proposed MiRKAT and optimal methods as well as the results from applying our methods to two real data sets. We also consider the relationship between MiRKAT and existing methods and demonstrate a close connection.

Simulation results

The type I error rates of MiRKAT and optimal MiRKAT across different simulation scenarios for continuous outcome are shown in Table 1. The upper panel presents the type I error result for *Simulation Scenario 1* in which a single phylogenetic cluster of the OTUs is associated with the outcome and the bottom panel presents the type I error results for *Simulation Scenario 2* in which the 10 most abundant OTUs are associated with the outcome. Note that when the covariates are independent of the microbiome that both simulation scenarios are equivalent as there is no association between y and $\mathbf{Z}$. For both simulation scenarios, when covariates $\mathbf{X}$ and the microbiome composition $\mathbf{Z}$ are independent, MiRKAT was valid with or without adjusting for $\mathbf{X}$. However, when $\mathbf{X}$ and microbiome composition $\mathbf{Z}$ are correlated, adjusting for covariates $\mathbf{X}$ is necessary: the type I error is seriously inflated if the confounder $\mathbf{X}$ is not accounted for.

Figures 1 and 2 shows the statistical power for the tests with continuous outcomes in *Simulation Scenario 1* in which a phylogenetic cluster of OTUs is associated with the outcome. Specifically, Figure 1 shows the power when $\mathbf{X}$ and $\mathbf{Z}$ are

independent and Figure 2 shows the power when **X** and **Z** are correlated. Note that we only considered statistical tests that adjust for covariates **X** because the tests without **X** adjustment have inflated type I error and are invalid.

The power is presented for MiRKAT using each individual kernel, the optimal MiRKAT which incorporates multiple kernels, as well as the naive Bonferroni adjusted test. For all the kernels that were considered, the power increases when the association strength increases. Good choices of kernel can greatly improve the statistical power in detecting association while improper choice of kernel leads to little power to detect the association. For this simulation scenario, the weighted UniFrac kernel and the generalized UniFrac Kernel with $\alpha = 0.75$ produced the highest power as opposed to other kernels and the unweighted UniFrac kernel was the least powerful. The optimal MiKRAT considering all metrics has power close to the weighted UniFrac kernel, losing some power relative to the weighted UniFrac kernel but still maintaining considerably better power than many other choices of kernel. As expected, the optimal test is always more powerful than the naive Bonferroni adjusted test.

Figures 3 and 4 show the statistical power for *Simulation Scenario 2* where the top 10 most abundant OTUs were associated with the outcome without regard for phylogeny. We again show the power when **X** and **Z** are independent (Figure 3) and when **X** and **Z** are correlated (Figure 4). Results were similar to *Simulation Scenario 1* except that the Bray-Curtis distance metric gave the highest power. MiRKAT, which considers all distance metrics, had smaller but comparable power as the Bray-Curtis distance, and much higher power than the naive Bonferroni

corrected test. The unweighted UniFrac kernel provided the least power.

In practice, the optimal kernel depends on the true state of nature and can vary from case to case. The two simulation scenarios shows that proper choice of kernel is essential in being well powered to discover associations between microbiome composition and outcomes, and poor choices of kernels leads to tremendous power loss. Optimal MiRKAT, however, alleviates the problem by considering different kernels and is more robust compared to single-distance based analysis as it hedges against different scenarios and works well in the omnibus.

The simulation results for dichotomous outcome are quanlitatively similar to the results obtained from continuous outcome. The type I error results are summarized in Table S1 and power results are shown in Figures S1, S2, S3 and S4.

Relationship Between MiRKAT and Existing Methods

A key advantage of MiRKAT is that it is already closely related to existing approaches for analyzing the association between microbiome composition and an outcome. In particular, with large sample size, the PERMANOVA method[10] can be shown to be a special case of the kernel machine testing framework under the scenario in which there are no confounding variables.[24] Consequently, the MiRKAT with a single kernel can be viewed as a generalization of PERMANOVA that accommodates additional covariates. In numerical simulations, the correlation between p-values obtained from single kernel MiRKAT and the corresponding distance based method is usually more than 0.99 in scenarios when there are no covariates to adjust for. For example, Figure S5 showed the p-values for

MiRKAT and the distance based approach using 2000 simulated data sets when a single distance/kernel was used. However, by using the asymptotic distribution, MiRKAT is considerably faster than corresponding distance based approach, especially with large sample size (Figure S6).

Analysis of Smoking Data

Recently, a microbiome profiling study was conducted to examine the communities within the upper respiratory tract[35] in order to understand the effect of cigarette smoking on the orpharyngeal and nospharyngeal microbiome. While details can be found in the original manuscript and subsequent re-analyses,[14] briefly, swab samples were collected from right and left nasopharynx and oropharynx of 29 smoking and 33 nonsmoking adults. The variable region 1-2 (V1-V2) of the bacterial 16S rRNA gene was PCR amplified and subject to multiplexed pryosequencing. OTUs were constructed using the QIIME pipeline. Samples with less than 500 reads and OTUs with only one read were removed, resulting in an OTU table with 60 samples (28 smokers vs 32 nonsmokers) and 856 OTUs. Additional covariates in this data included gender and antibiotic use within 3 months.

Distance based analysis of the oropharyngeal samples using permutation based distance analysis (PERMANOVA) with both weighted and unweighted UniFrac distances identified significant association between microbiome profiles and smoking status. However, the analyses did not take into account potential confounders: within the collected study sample, among the smokers 75% were male, yet among the non-smokers, only 56% were male. The odds ratio of smoking between males

and females is 2.33 within the data set. The imbalance in the proportion of male and female subjects indicates strong potential for confounding: it is unclear whether the differences in microbiome profiles between smokers and non-smokers is driven by smoking or driven by the gender imbalance. Additionally, the tests were conducted using either weighted or unweighted UniFrac distance; it is practically attractive to consider multiple possible distance measurements while controlling for possible confounding effects. MiRKAT represents a natural analysis approach.

Therefore, we re-analyzed the data from the oropharyngeal samples using MiRKAT. Specifically, we applied MiKRAT method to analyze the association between smoking and microbial community composition, using weighted and unweighted UniFrac distance matrices and the Bray-Curtis distance, except here we transformed them to be similarity metrics to form the kernels, but we further adjusted for gender, and antibiotic use. We also applied the optimal MiRKAT. Using MiRKAT under individual distance metrics, we found the p-values from K_W, K_U and K_{BC} are 0.0048, 0.014 and 0.002 respectively. The optimal MiRKAT generated a p-value of 0.0031. Thus, despite the potential for confounding, our results show that the association between microbiome profiles and smoking status remains significant after controlling for the potential confounders, reaffirming and providing greater confidence in the earlier results. In addition to validating a previous analysis, this result also demonstrates the utility and importance of MiRKAT with regard to accommodation of covariates and multiple kernels.

Analysis of Fecal Protease Data

Fecal proteases (FP) are enteric enzymes that are elevated in subsets of individuals with irritable bowel syndrome (IBS) and inflammatory bowel disease (IBD, MIM 266600). It was demonstrated that FP from IBS affected individuals have a profound impact on intestinal physiology including visceral sensitivity and colonic permeability in mice.[36] Although there is evidence that elevated FP levels can alter intestinal physiology by activating proteinase activated receptors, it remains unclear whether the FP levels are of human or microbial origin. Consequently, Carroll *et al.*[37] conducted a study to examine the relationship between FP levels and microbiota in human fecal samples from 30 individuals affected with IBS and 24 healthy adults. 454 pyrosequencing of the 16S rRNA gene was again used to profile the microbiomes and QIIME was again applied to quantify the composition and diversity of each community.

The original study identified a significant association between microbiome composition and FP levels. However, analyses were restricted to the subjects with the highest and lowest FP levels. Thus, we applied MiRKAT to the data set (limiting to the 23 diarrhea-predominant IBS affected subjects and 23 healthy controls) to test for an association between FP levels and microbiome composition, except that we treated FP levels as continuous (so as to use all subjects) and we further adjusted for additional potential confounders, including age, body mass index, gender, race and functional bowel disorder. We considered MiRKAT using the weighted UniFrac, unweighted UniFrac and Bray-Curtis kernels as well as the optimal MiRKAT.

Interestingly, the three distances gave discordant conclusions in that the un-weighted UniFrac kernel and Bray-Curtis kernel yielded significant p values (p =0.0046 and 0.039 respectively) while the weighted UniFrac kernel gave non-significant result (p = 0.124). Un-weighted UniFrac is primarily based on the presence or absence of an OTU, while weighted UniFrac distances further incorporates abundance which could account for the differences but the difference in association results makes it difficult to draw a single conclusion. The optimal MiRKAT which simultaneously considers the three candidate kernels gives a single p-value of 0.0116 after covariate adjustment. This further demonstrates the advantages of MiRKAT to be able to consider multiple kernels since using individual distance metrics yielded disparate results and is difficult to interpret.

Discussion

We propose a kernel machine regression based method (MiRKAT) to test for the association between microbial community composition and a continuous or dichotomous outcome of interest in which covariate effects are modeled parametrically and the microbiome effect is modeled nonparametrically. The kernel matrix, which defines the functional form of the microbiome effect, is constructed by exploiting its correspondence with the popular distance metric designed to convey phylogenetic or taxonomic information among different OTUs. Additionally, the proposed method allows incorporation of multiple candidate kernels simultaneously, enabling development of the optimal MiRKAT. Simulations and real data

analyses indicate that the approach has reasonable power and that the optimal MiRKAT is robust to poor choices of kernels. Close connections between MiRKAT and existing analysis frameworks ensure that the approach is a natural addition to the currently available methodology.

The optimal MiRKAT enables researchers to consider multiple distance and dissimilarity metrics, simultaneously. Here, we focused primarily on the UniFrac, weighted UniFrac, generalized UniFrac and Bray-Curtis metrics as our experiences have shown that these tend to work well in practice. In principle, one can include a wide range of other metrics with little penalty with regard to false positive rate, but the trade-off is that one may lose power if there are too many kernels under consideration that are too disparate – use of highly correlated kernels will not impact power very much. In the most extreme cases, optimal MiRKAT from multiple perfectly correlated kernels will generate the same p-value as from each of the individual kernel tests. Furthermore, we note that the tests using each of the individual kernels are constructed based on the same data sets and are non-negatively correlated (i.e. not competitive). Thus, the optimal MiRKAT should always have higher power than the naive Bonferroni adjusted test.

A reasonable alternative to the proposed omnibus test approach is to construct, as a kernel, a weighted combination of multiple kernels. In practice, the optimal "weight" is unknown and needs to be estimated from data or selected via other approaches, such as a grid search. From the mixed model point of view, estimating the weights is equivalent to estimating a variance component that disappears when the null hypothesis is true; this violates the common regularity conditions in

the standard asymptotic tests. Statistical methods for such problems, such as the likelihood ratio tests, recently have been the focus of considerable statistical research.[38,39] However, this is frequently much more computational intensive than the score test, especially when there are many kernels under consideration. Furthermore, there is very limited work on likelihood ratio test for variance components when some parameters disappear under the null AND when the null values are on the boundary of the parameter space. On the other hand, selecting the best "weight" through a grid search can be conducted similarly as the optimal MiRKAT in which each of the weighted combination of candidate kernels is treated as a new kernel. However, when the number of kernels under consideration increases or when a finer grid is used, the computation burden increases quickly due to the large search space and rapidly becomes computationally prohibitive. Therefore, if prior evidence is available to suggest that a single kernel is the best kernel, then using that single kernel or using a smaller set of kernels will be more powerful. In the absence of prior knowledge, then we suggest using a modest range of kernels with differing characteristics, e.g. a combination of phylogeny based and non-phylogeny based kernels as in our simulations.

Beyond assessing the association with overall composition, there is considerable interest in identifying the individual taxa that are driving the apparent associations. This approach for analyzing microbiome data is frequently complementary and parallel to methods for testing overall composition and diversity. One common approach for doing this is to assess the marginal association between each OTU and the outcome. However, in addition to difficulties in determining

the scale of the analysis, i.e. whether to use percent composition or raw OTU counts, a problem of considerable interest lies in using distance metrics to inform the identification of individual taxa related to the outcome. To this end, as a regression based approach combined with the relatively fast computation, MiRKAT could enable a step-wise variable selection approach with AIC or BIC. Such an approach could be applied post-hoc to identify the variables most strongly driving apparent associations. It may also be possible to use a penalized regression approach within the kernel framework,[40] but this remains a topic for future research.

Microbiome studies are now being included within epidemiological, population based, and clinical studies. In contrast to early microbiome studies with modest sample sizes and relatively controlled experimental conditions, issues such as confounding, covariate adjustment, and accommodation of more sophisticated outcomes are increasingly important in such studies. MiRKAT's ability to control for confounders within a principled regression based framework while maintaining type I error and adequate power make it an attractive alternative to currently available methods. Furthermore, although we focused on dichotomous and continuous variables of interest, the framework can be generalized to alternative types of outcomes such as multivariate, longitudinal, and survival data. Thus, with growing interest in applying microbiome to complex clinical and population based studies, MiRKAT can be extended to open new avenues of research by enabling analysis of data from the emerging studies with more sophisticated outcomes.

Supplemental Data

Supplemental Data include 6 figures and 1 table.

Acknowledgements

This study was supported in part by a NIH grants K01 DK092330 and R01 HG007508; CGIBD pilot feasibility grant P30 DK03498; the Hope Foundation and the Gerstner Family Career Development Award in Individualized Medicine. The authors declare that there are no conflicts of interest.

Web Resources

The URLs for data presented herein are as follows:

Online Mendelian Inheritance in Man (OMIM): `http://www.omim.org`
An implementation of MiRKAT in the R language can be found at: `http://research.fhcrc.org/wu/en.html`.
MiRKAT R package with manual can be found at: `http://research.fhcrc.org/wu/en.html`.

References

[1] Woese, C. R., Fox, G. E., Zablen, L., Uchida, T., Bonen, L., Pechman, K., Lewis, B. J., and Stahl, D. (1975). Conservation of primary structure in 16S ribosomal RNA. Nature, *254*, 83–86.

[2] Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S., and Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature, *428*, 37–43.

[3] Wooley, J. C., Godzik, A., and Friedberg, I. (2010). A primer on metagenomics. PLoS Comput. Biol., *6*, e1000667.

[4] Lasken, R. S. (2012). Genomic sequencing of uncultured microorganisms from single cells. Nat. Rev. Microbiol., *10*, 631–640.

[5] Willing, B. P., Russell, S. L., and Finlay, B. B. (2011). Shifting the balance: antibiotic effects on host-microbiota mutualism. Nat. Rev. Microbiol., *9*, 233–243.

[6] Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., Sogin, M. L., Jones, W. J., Roe, B. A., Affourtit, J. P., et al. (2009). A core gut microbiome in obese and lean twins. Nature, *457*, 480–484.

[7] Larsen, N., Vogensen, F. K., van den Berg, F. W., Nielsen, D. S., Andreasen, A. S., Pedersen, B. K., Al-Soud, W. A., Sørensen, S. J., Hansen, L. H., and

Jakobsen, M. (2010). Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. PLoS ONE, *5*, e9085.

[8] Peterson, D. A., Frank, D. N., Pace, N. R., and Gordon, J. I. (2008). Metagenomic approaches for defining the pathogenesis of inflammatory bowel diseases. Cell Host Microbe, *3*, 417–427.

[9] Karlsson, F. H., Tremaroli, V., Nookaew, I., Bergstrom, G., Behre, C. J., Fagerberg, B., Nielsen, J., and Backhed, F. (2013). Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature, *498*, 99–103.

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

[11] Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., Fernandes, G. R., Tap, J., Bruls, T., Batto, J. M., et al. (2011). Enterotypes of the human gut microbiome. Nature, *473*, 174–180.

[12] Lozupone, C. and Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. Appl. Environ. Microbiol., *71*, 8228–8235.

[13] Lozupone, C. A., Hamady, M., Kelley, S. T., and Knight, R. (2007). Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. Appl. Environ. Microbiol., *73*, 1576–1585.

[14] Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., Wu, G. D., Collman, R. G., Bushman, F. D., and Li, H. (2012). Associating microbiome composition with environmental covariates using generalized unifrac distances. Bioinformatics, *28*, 2106–13.

[15] Chen, J. and Li, H. (2013). Kernel methods for regression analysis of microbiome compositional data. In Hu, M., Liu, Y., and Lin, J., eds., Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association. (Springer), pp. 191–201.

[16] Kwee, L. C., Liu, D., Lin, X., Ghosh, D., and Epstein, M. P. (2008). A powerful and flexible multilocus association test for quantitative traits. Am. J. Hum. Genet., *82*, 386–397.

[17] Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J., and Lin, X. (2010). Powerful SNP-set analysis for case-control genome-wide association studies. The American Journal of Human Genetics, *86*, 929–942.

[18] Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet., *89*, 82–93.

[19] Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., Christiani, D. C., Wurfel, M. M., and Lin, X. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet, *91*, 224–37.

[20] Wu, M. C., Maity, A., Lee, S., Simmons, E. M., Harmon, Q. E., Lin, X., Engel, S. M., Molldrem, J. J., and Armistead, P. M. (2013). Kernel machine SNP-set testing under multiple candidate kernels. Genet. Epidemiol., *37*, 267–275.

[21] Zhou, J. J. and Zhou, H. (2015). Powerful exact variance component tests for the small sample next generation sequencing studies (evc test). Submitted.

[22] Chen, W., Zhao, N., Wu, M. C., Schaid, D. J., and Chen, J. (2015). Small sample kernel association test for genetic association studies. Technical report, Mayo Clinic.

[23] Goeman, J. J., Van De Geer, S. A., De Kort, F., and Van Houwelingen, H. C. (2004). A global test for groups of genes: testing association with a clinical outcome. Bioinformatics, *20*, 93–99.

[24] Pan, W. (2011). Relationship between genomic distance based regression and kernel machine regression for multi-marker association testing. Genetic epidemiology, *35*, 211–216.

[25] Liu, D., Lin, X., and Ghosh, D. (2007). Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. Biometrics, *63*, 1079–88.

[26] Liu, D., Ghosh, D., and Lin, X. (2008). Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. BMC Bioinformatics, *9*, 292.

[27] Gianola, D. and van Kaam, J. B. (2008). Reproducing Kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics, *178*, 2289–303.

[28] Lin, X. (1997). Variance component testing in generalised linear models with random effects. Biometrika, *84*, 309–326.

[29] Liu, H., Tang, Y., and Zhang, H. H. (2009). A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. Computational Statistics & Data Analysis, *53*, 853–856.

[30] Davies, R. (1980). The distribution of a linear combination of chi-2 random variables. *29*, 323–333.

[31] Duchesne, P. and Lafaye de Micheaux, P. (2010). Computing the distribution of quadratic forms: Further comparisons between the liu-tang-zhang approximation and exact methods. Computational Statistics and Data Analysis, *54*, 858–862.

[32] Freedman, D. and Lane, D. (1983). A nonstochastic interpretation of reported significance levels. Journal of Business and Economic Statistics, *1*, 292–298.

[33] Epstein, M. P., Duncan, R., Jiang, Y., Conneely, K. N., Allen, A. S., and Satten, G. A. (2012). A permutation procedure to correct for confounders in case-control studies, including tests of rare variation. Am. J. Hum. Genet., *91*, 215–223.

[34] Fog, A. (2008). Sampling methods for wallenius' and fisher's noncentral hypergeometric distributions. Communications in Statistics, Simulation and Computation, *37*, 241–257.

[35] Charlson, E. S., Chen, J., Custers-Allen, R., Bittinger, K., Li, H., Sinha, R., Hwang, J., Bushman, F. D., and Collman, R. G. (2010). Disordered microbial communities in the upper respiratory tract of cigarette smokers. PLoS One, *5*, e15216.

[36] Annaházi, A., Gecse, K., Dabek, M., Ait-Belgnaoui, A., Rosztóczy, A., Róka, R., Molnár, T., Theodorou, V., Wittmann, T., Bueno, L., et al. (2009). Fecal proteases from diarrheic-IBS and ulcerative colitis patients exert opposite effect on visceral sensitivity in mice. Pain, *144*, 209–217.

[37] Carroll, I. M., Ringel-Kulka, T., Ferrier, L., Wu, M. C., Siddle, J. P., Bueno, L., and Ringel, Y. (2013). Fecal protease activity is associated with compositional alterations in the intestinal microbiota. PloS one, *8*, e78017.

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

[39] Greven, S., Crainiceanu, C. M., Kchenhoff, H., and Peters, A. (2008). Restricted likelihood ratio testing for zero variance components in linear mixed models. Journal of Computational and Graphical Statistics, *17*, 870–891.

[40] Allen, G. I. (2013). Automatic feature selection via weighted kernels and regularization. Journal of Computational and Graphical Statistics, *22*, 284–299.

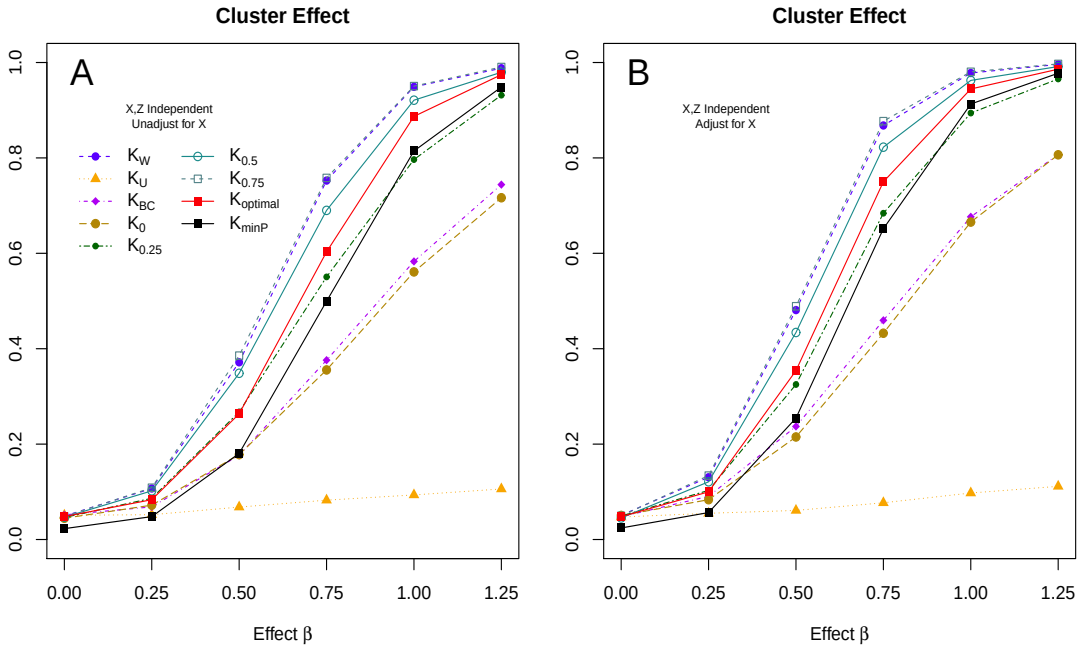

Figure 1: Type I error and Power of MiRKAT based on different kernels for *Simulation Scenario 1* with continuous outcome: A selected phylogenetic cluster of the OTUs are associated with the outcome and covariates **X** and the microbiome profiles **Z** were simulated independently. Panel A shows the results for tests that do not adjust for **X** and panel B shows results that adjust for **X**. K_w, K_u, K_{BC}, K_0, $K_{0.25}$, $K_{0.5}$ and $K_{0.75}$ represents MiRKAT results using different individual kernels respectively: weighted UniFrac, unweighted UniFrac, Bray-Curtis, and generalized UniFrac kernels with $\alpha = 0$, 0.25, 0.5 and 0.75. $K_{optimal}$ represents the simulation results for optimal MiRKAT considering all seven kernels and K_{minP} shows the results using a naive Bonferroni adjusted test. Sample size $n = 100$.

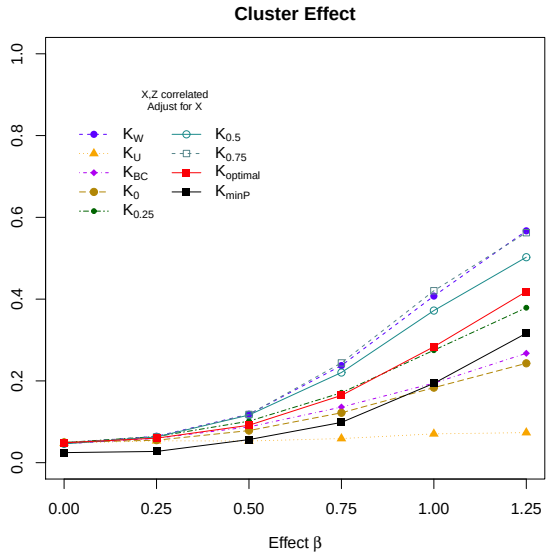

Figure 2: Type I error and Power of MiRKAT based on different kernels for *Simulation Scenario 1* with continuous outcome: A selected phylogenetic cluster of the OTUs are associated with the outcome and covariates **X** and microbiome composition **Z** are correlated through $X_{2i} = \text{scale}\left(\sum_{j\in\mathcal{A}} Z_{ij}\right) + N(0,1)$ where $\mathcal{A}$ represents the selected cluster. Results are presented only for MiRKAT with **X** adjustment because unadjusted tests give seriously inflated type I error. K_w, K_u, K_{BC}, K_0, $K_{0.25}$, $K_{0.5}$ and $K_{0.75}$ represents MiRKAT results using different individual kernels respectively: weighted UniFrac, unweighted UniFrac, Bray-Curtis, and generalized UniFrac kernels with $\alpha = 0$, 0.25, 0.5 and 0.75. $K_{optimal}$ represents the simulation results for optimal MiRKAT considering all seven kernels and K_{minP} shows the results using a naive Bonferroni adjusted test. Sample size $n = 100$.

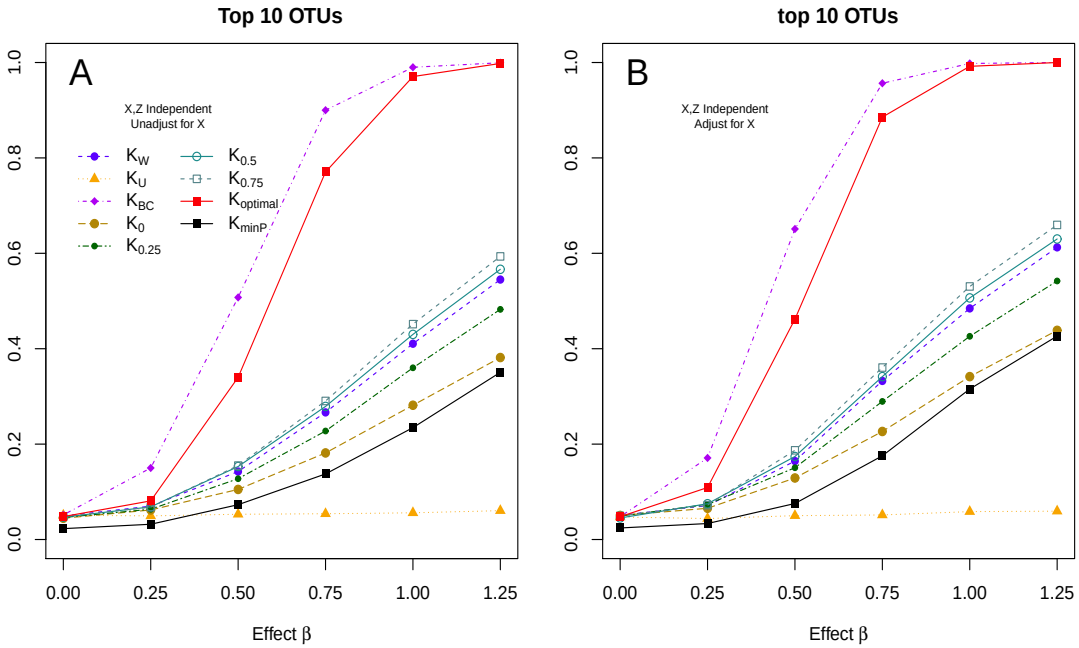

Figure 3: Type I error and Power of MiRKAT based on different kernels for *Simulation Scenario 2* with continuous outcome: The 10 most abundant OTUs are associated with the outcome. Additional covariates **X** and the microbiome profiles **Z** were simulated independently. Panel A shows the results for tests that do not adjust for **X** and panel B shows results that adjust for **X**. K_w, K_u, K_{BC}, K_0, $K_{0.25}$, $K_{0.5}$ and $K_{0.75}$ represents MiRKAT results using different individual kernels respectively: weighted UniFrac, unweighted UniFrac, Bray-Curtis, and generalized UniFrac kernels with $\alpha = 0$, 0.25, 0.5 and 0.75. $K_{optimal}$ represents the simulation results for optimal MiRKAT considering all seven kernels and K_{minP} shows the results using a naive Bonferroni adjusted test. Sample size $n = 100$.

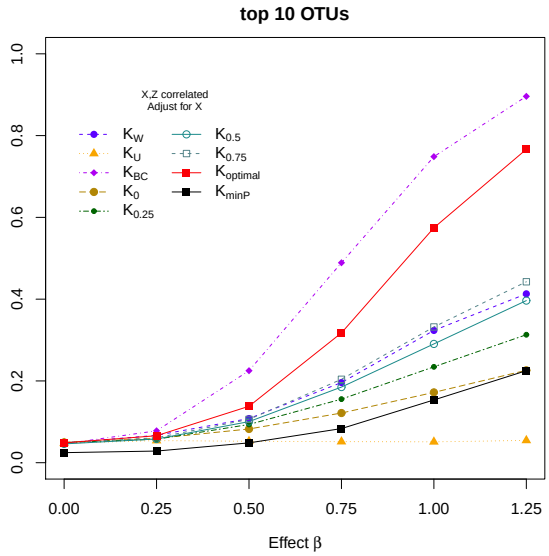

Figure 4: Type I error and Power of MiRKAT based on different kernels for *Simulation Scenario 1* with continuous outcome: The 10 most abundant OTUs are associated with the outcome. Additional covariates **X** and the microbiome profiles **Z** are correlated in that $X_{2i} = \text{scale}\left(\sum_{j\in\mathcal{A}} Z_{ij}\right) + N(0,1)$ where $\mathcal{A}$ represents the top 10 most abundant OTUs. Results are presented only for MiRKAT with **X** adjustment because unadjusted tests give seriously inflated type I error. K_w, K_u, K_{BC}, K_0, $K_{0.25}$, $K_{0.5}$ and $K_{0.75}$ represents MiRKAT results using different individual kernels respectively: weighted UniFrac, unweighted UniFrac, Bray-Curtis, and generalized UniFrac kernels with $\alpha = 0$, 0.25, 0.5 and 0.75. $K_{optimal}$ represents the simulation results for optimal MiRKAT considering all seven kernels and K_{minP} shows the results using a naive Bonferroni adjusted test. Sample size $n = 100$.

Table 1: Empirical type I errors for MiRKAT and "optimal" MiRKAT with continuous outcome. Type I error was evaluated for scenarios when additional covariates are independent with the OTUs ($X \perp Z$) and scenarios when covariates are related to the OTUs ($X \not\perp Z$) using 5000 simulated data sets. K_w, K_u, K_{BC}, K_0, $K_{0.25}$, $K_{0.5}$ and $K_{0.75}$ represents MiRKAT results using different individual kernels respectively: weighted UniFrac, unweighted UniFrac, Bray-Curtis, and generalized UniFrac kernels with $\alpha = 0$, 0.25, 0.5 and 0.75. $K_{optimal}$ represents the simulation results for optimal MiRKAT considering all seven kernels and K_{minP} shows the results using a naive Bonferroni adjusted test. P-values for "optimal" MiRKAT were obtained by 1000 permutations. Numbers in **bold** show inflated type I error.

Simulation scenario 1: Clustered OTUs									
$X \perp Z$	Unadjust for **X**								
n	K_W	K_U	K_{BC}	K_0	$K_{0.25}$	$K_{0.5}$	$K_{0.75}$	K_{opt}	K_{minP}
100	0.053	0.050	0.050	0.046	0.047	0.048	0.052	0.050	0.023
200	0.052	0.047	0.051	0.053	0.049	0.048	0.051	0.051	0.026
$X \perp Z$	Adjust for **X**								
100	0.056	0.048	0.047	0.049	0.045	0.050	0.048	0.046	0.024
200	0.051	0.050	0.053	0.048	0.047	0.052	0.049	0.050	0.027
$X \not\perp Z$	Unadjust for **X**								
100	**0.389**	**0.062**	**0.172**	**0.268**	**0.345**	**0.384**	**0.182**	**0.268**	**0.183**
200	**0.790**	**0.080**	**0.398**	**0.587**	**0.732**	**0.791**	**0.387**	**0.651**	**0.547**
$X \not\perp Z$	Adjust for **X**								
100	0.055	0.047	0.047	0.049	0.046	0.049	0.046	0.049	0.024
200	0.052	0.049	0.051	0.047	0.047	0.052	0.050	0.049	0.026
Simulation scenario 2: top 10 OTUs									
$X \perp Z$	Unadjust for **X**								
n	K_W	K_U	K_{BC}	K_0	$K_{0.25}$	$K_{0.5}$	$K_{0.75}$	K_{opt}	K_{minP}
100	0.053	0.050	0.050	0.045	0.048	0.049	0.053	0.050	0.025
200	0.051	0.047	0.050	0.053	0.050	0.047	0.051	0.050	0.026
$X \perp Z$	Adjust for **X**								
100	0.056	0.048	0.047	0.050	0.046	0.051	0.047	0.049	0.021
200	0.051	0.049	0.053	0.047	0.047	0.052	0.050	0.051	0.023
$X \not\perp Z$	Unadjust for **X**								
100	**0.153**	**0.048**	**0.669**	**0.105**	**0.124**	**0.147**	**0.157**	**0.516**	**0.067**
200	**0.307**	**0.048**	**0.976**	**0.194**	**0.239**	**0.293**	**0.320**	**0.932**	**0.151**
$X \not\perp Z$	Adjust for **X**								
100	0.056	0.048	0.047	0.049	0.046	0.050	0.047	0.049	0.020
200	0.052	0.049	0.051	0.048	0.048	0.051	0.049	0.049	0.024

Supplemental Material

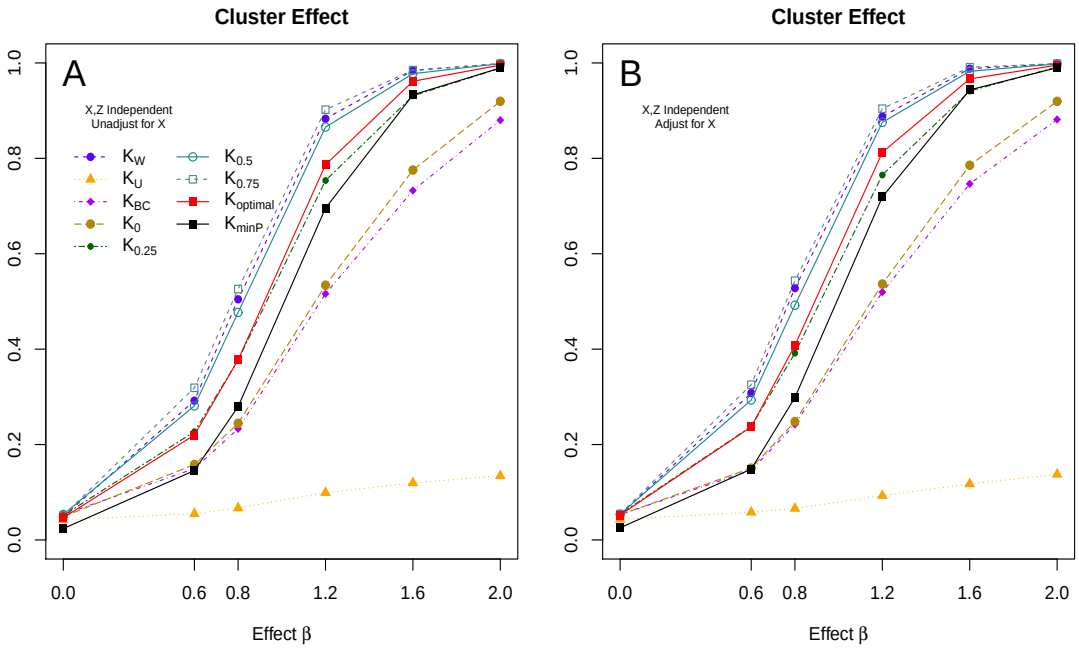

Figure S1: **Type I error and Power of MiRKAT based on different kernels for *Simulation Scenario 1* with dichotomous outcome**: a selected phylogenetic cluster of the OTUs are associated with the outcome. Additional covariates X and microbiome effect Z were simulated independently. Panel A shows the results for tests that do not adjust for **X** and panel B shows results that adjust for **X**. K_w, K_u, K_{BC}, K_0, $K_{0.25}$, $K_{0.5}$ and $K_{0.75}$ represents MiRKAT results using different individual kernels respectively: weighted UniFrac, unweighted UniFrac, Bray-Curtis, and generalized UniFrac kernels with $\alpha = 0$, 0.25, 0.5 and 0.75. $K_{optimal}$ represents the simulation results for optimal MiRKAT considering all seven kernels and K_{minP} shows the results using a naive Bonferroni adjusted test. Results were presented at $n = 200$.

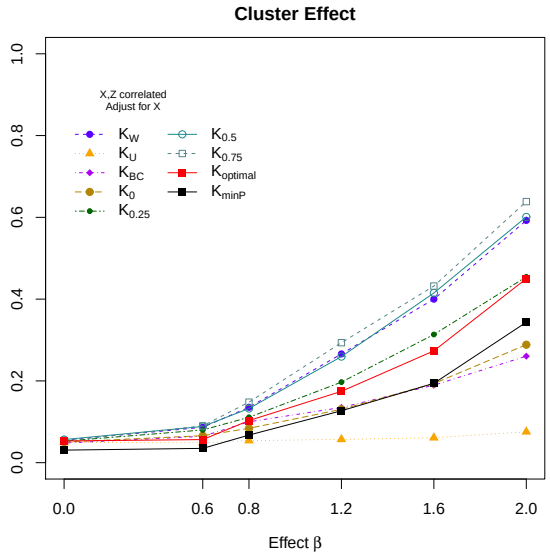

Figure S2: **Type I error and Power of MiRKAT based on different kernels for for *Simulation Scenario 1* with dichotomous outcome**:a selected phylogenetic cluster of the OTUs are associated with the outcome. Additional covariates X and microbiome composition Z are correlated through $X_{2i} = \text{scale}\left(\sum_{j\in\mathcal{A}} Z_{ij}\right) + N(0,1)$. We only considered MiRKAT with **X** adjustment because unadjusted tests give seriously inflated type I error. K_w, K_u, K_{BC}, K_0, $K_{0.25}$, $K_{0.5}$ and $K_{0.75}$ represents MiRKAT results using different individual kernels respectively: weighted UniFrac, unweighted UniFrac, Bray-Curtis, and generalized UniFrac kernels with $\alpha = 0$, 0.25, 0.5 and 0.75. $K_{optimal}$ represents the simulation results for optimal MiRKAT considering all seven kernels and K_{minP} shows the results using a naive Bonferroni adjusted test. Sample Size $n = 200$.

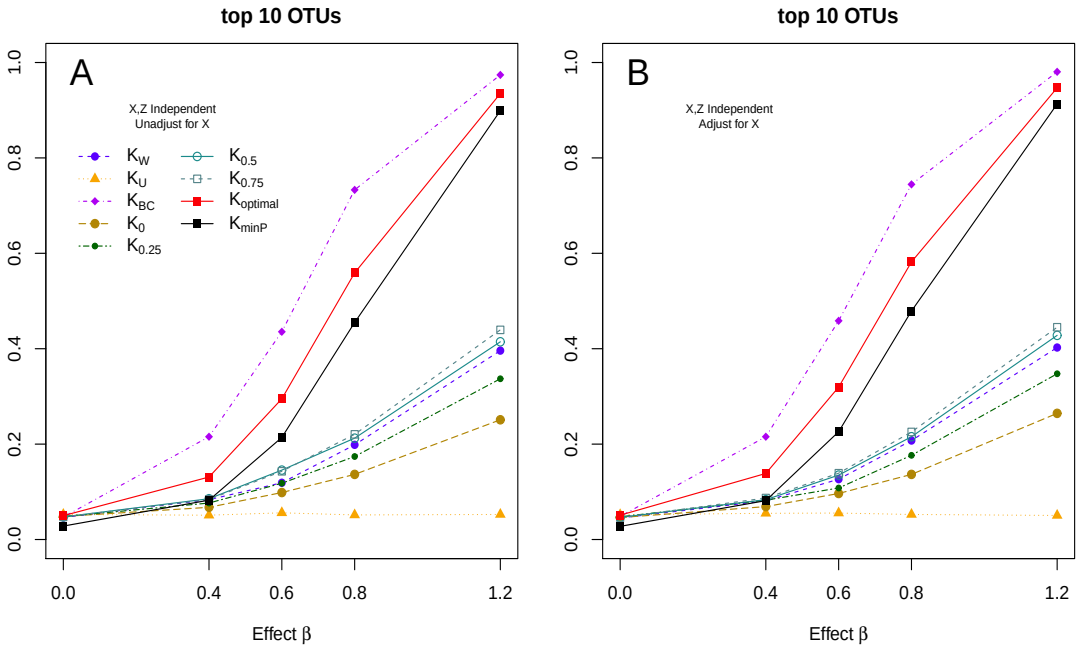

Figure S3: **Type I error and Power of MiRKAT based on different kernels for *Simulation Scenario 2* with dichotomous outcome**: the 10 most abundant OTUs are associated with the outcome. Additional covariates X and microbiome effect Z were simulated independently. Panel A shows the results for tests that do not adjust for **X** and panel B shows results that adjust for **X**. K_w, K_u, K_{BC}, K_0, $K_{0.25}$, $K_{0.5}$ and $K_{0.75}$ represents MiRKAT results using different individual kernels respectively: weighted UniFrac, unweighted UniFrac, Bray-Curtis, and generalized UniFrac kernels with $\alpha = 0$, 0.25, 0.5 and 0.75. $K_{optimal}$ represents the simulation results for optimal MiRKAT considering all seven kernels and K_{minP} shows the results using a naive Bonferroni adjusted test. Results were presented at $n = 200$.

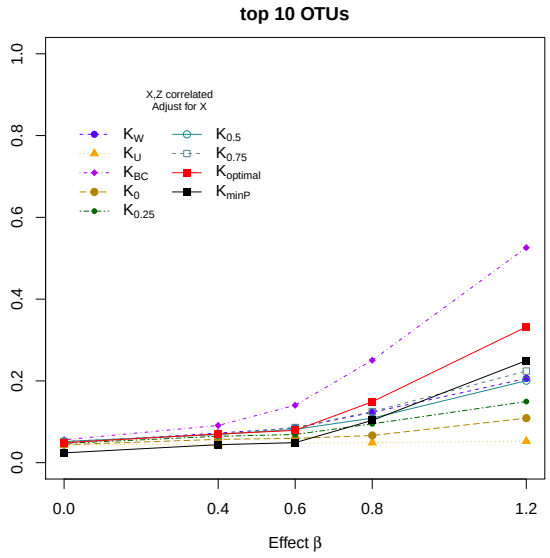

Figure S4: **Type I error and Power of MiRKAT based on different kernels for for *Simulation Scenario 2* with dichotomous outcome**:the 10 most abundant OTUs are associated with the outcome. Additional covariates X and microbiome composition Z are correlated through $X_{2i} = \text{scale}\left(\sum_{j\in\mathcal{A}} Z_{ij}\right) + N(0,1)$. We only considered MiRKAT with **X** adjustment because unadjusted tests give seriously inflated type I error. K_w, K_u, K_{BC}, K_0, $K_{0.25}$, $K_{0.5}$ and $K_{0.75}$ represents MiRKAT results using different individual kernels respectively: weighted UniFrac, unweighted UniFrac, Bray-Curtis, and generalized UniFrac kernels with $\alpha = 0$, 0.25, 0.5 and 0.75. $K_{optimal}$ represents the simulation results for optimal MiRKAT considering all seven kernels and K_{minP} shows the results using a naive Bonferroni adjusted test. Results were presented at $n = 200$.

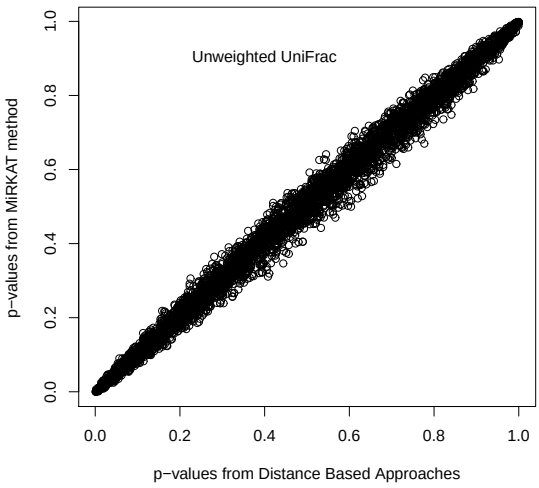

Figure S5: Example plot of the p-value correlation using distance based approach and MiRKAT when no additional covariates are considered. 5000 simulations are plotted at sample size n = 200 for continuous outcome. Unweighted UniFrac distance and kernel were used for the distance based approach and MiRKAT respectively.

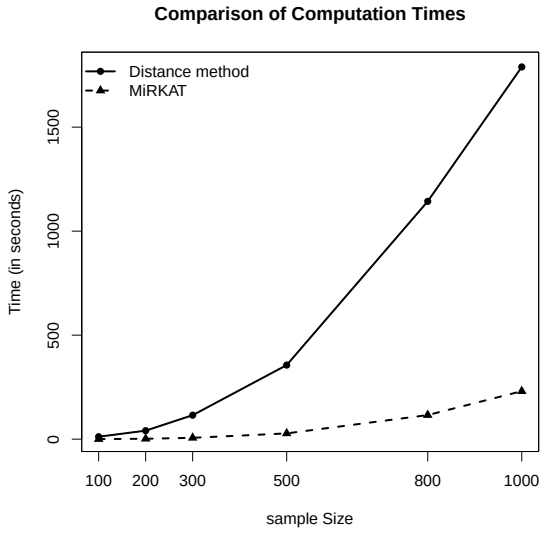

Figure S6: Computation times of MiRKAT and distance based test as a function of the sample size for continuous outcome. The figure presents the total computation time for 100 repeated tests with each sample size. 999 permutations (the default number) were used in distance based approaches.

Table S1: Empirical type I errors for MiRKAT and "optimal" MiRKAT with dichotomous outcome. Type I error was evaluated for scenarios when additional covariates are independent with the OTUs ($X \perp Z$) and scenarios when covariates are related to the OTUs ($X \not\perp Z$) using 5000 simulated data sets. K_w, K_u, K_{BC}, K_0, $K_{0.25}$, $K_{0.5}$ and $K_{0.75}$ represents MiRKAT results using different individual kernels respectively: weighted UniFrac, unweighted UniFrac, Bray-Curtis, and generalized UniFrac kernels with $\alpha = 0$, 0.25, 0.5 and 0.75. $K_{optimal}$ represents the simulation results for optimal MiRKAT considering all seven kernels and K_{minP} shows the results using a naive Bonferroni adjusted test. P-values for "optimal" MiRKAT were obtained by 1000 permutations. Numbers in **bold** show inflated type I error.

Simulation scenario 1: Clustered OTUs									
$X \perp Z$	Unadjust for **X**								
n	K_W	K_U	K_{BC}	K_0	$K_{0.25}$	$K_{0.5}$	$K_{0.75}$	K_{opt}	K_{minP}
200	0.051	0.049	0.049	0.051	0.052	0.054	0.051	0.049	0.025
500	0.046	0.049	0.054	0.056	0.053	0.054	0.053	0.053	0.028
$X \perp Z$	Adjust for **X**								
200	0.054	0.051	0.050	0.051	0.053	0.054	0.054	0.053	0.028
500	0.047	0.048	0.051	0.053	0.055	0.051	0.049	0.055	0.029
$X \not\perp Z$	Unadjust for **X**								
200	**0.105**	**0.054**	**0.075**	**0.081**	**0.099**	**0.116**	**0.123**	**0.092**	**0.057**
500	**0.156**	**0.056**	**0.092**	**0.149**	**0.210**	**0.260**	**0.285**	**0.214**	**0.138**
$X \not\perp Z$	Adjust for **X**								
200	0.048	0.054	0.049	0.050	0.050	0.053	0.052	0.051	0.028
500	0.045	0.051	0.050	0.051	0.048	0.049	0.049	0.048	0.024
Simulation scenario 2: top 10 OTUs									
$X \perp Z$	Unadjust for **X**								
n	K_W	K_U	K_{BC}	K_0	$K_{0.25}$	$K_{0.5}$	$K_{0.75}$	K_{opt}	K_{minP}
200	0.046	0.052	0.047	0.048	0.048	0.047	0.047	0.050	0.028
500	0.058	0.044	0.045	0.051	0.050	0.052	0.053	0.048	0.025
$X \perp Z$	Adjust for **X**								
200	0.045	0.052	0.048	0.046	0.048	0.046	0.046	0.051	0.028
500	0.052	0.045	0.040	0.048	0.052	0.052	0.050	0.042	0.022
$X \not\perp Z$	Unadjust for **X**								
200	**0.066**	**0.051**	**0.201**	**0.064**	**0.069**	**0.070**	**0.073**	**0.125**	**0.077**
500	**0.123**	**0.049**	**0.544**	**0.101**	**0.104**	**0.123**	**0.126**	**0.378**	**0.307**
$X \not\perp Z$	Adjust for **X**								
200	0.047	0.056	0.052	0.044	0.047	0.052	0.052	0.049	0.024
500	0.051	0.047	0.056	0.051	0.050	0.046	0.049	0.054	0.024