Testing in Microbiome-Profiling Studies with MiRKAT, the Microbiome Regression-Based Kernel Association Test

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] and Michael C. Wu[1,*]

High-throughput sequencing technology has enabled population-based studies of the role of the human microbiome in disease etiology and exposure response. Distance-based analysis is a popular strategy for evaluating 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 selecting the best distance and 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 non-parametrically modeling the microbiome through a kernel that incorporates phylogenetic distance. It uses a variance-component score statistic 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. Our simulations demonstrated 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 in comparison to when the best distance is used and can achieve tremendous power gain in comparison to when a poor distance is chosen. Finally, we applied MiRKAT to real microbiome datasets to show that microbial communities are associated with smoking and with fecal protease levels after confounders are controlled for.

Introduction

The advent of massively parallel sequencing has enabled high-throughput profiling of the microbiota in a large number of samples via targeted sequencing of the 16S rDNA sequence,[1–4] 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 for efficiently testing 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 here as just distance-based analysis for simplicity. Via standard methods, the 16S sequence tags are clustered on the basis of their sequence similarity to form operational taxonomic units (OTUs), which can essentially be considered surrogates for biological taxa. Distance metrics are then constructed to measure the phylogenetic or taxonomic dissimilarity between each pair of samples by incorporating the phylogenetic relationship or the absolute and relative abundance of different taxa. Then, for assessing the association between the microbiome diversity and an outcome variable of interest, the pairwise distance between each pair of samples is compared to the distribution of the outcome variable. For categorical outcome variables, this is essentially comparing the pairwise distances within and between categories. Operationally, multivariate analysis[10] or the top principal coordinates[11] of the matrix of pairwise distances are used for testing for associations via permutation.

Among the many possible distances, the UniFrac distances are the most popular in the literature and are constructed on the basis of 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 given that more prevalent species are likely to be present in all individuals. Weighted UniFrac distance uses species abundance information to weight the

[1]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA; [2]Division of Biomedical Statistics and Informatics and Center for Individualized Medicine, Mayo Clinic, Rochester, MN 55905, USA; [3]Division of Gastroenterology and Hepatology, Center for Gastrointestinal Biology and Disease, University of North Carolina at Chapel Hill, Chapel Hill, NC 27516, USA; [4]Department of Maternal and Child Health, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; [5]Department of Human Genetics, Emory University, Atlanta, GA 30322, USA; [6]Department of Statistics, North Carolina State University, Cary, Raleigh, NC 27695, USA; [7]Division of Epidemiology and Biostatistics, University of Arizona, Tucson, AZ 85724, USA; [8]Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA 19014, USA
*Correspondence: chen.jun2@mayo.edu (J.C.), mcwu@fhcrc.org (M.C.W.)
http://dx.doi.org/10.1016/j.ajhg.2015.04.003.

UniFrac distance and thus has more power to detect changes in common lineages. The generalized UniFrac distance[14] was introduced as a compromise between weighted and unweighted UniFrac distances; it down-weights its emphasis on either abundant or rare lineages and therefore has more power to detect changes in OTU clusters with modest abundance. Generalized UniFrac distance involves an additional parameter (α), such that the generalized UniFrac distance with $\alpha = 1$ is equivalent to the 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 on the basis of counts at each site. Similarly, Euclidean distance can also be used and is frequently thought to be similar to weighted UniFrac distance because 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. Although there are similarities, they are designed to capture distance differently, leading to differential performance across different scenarios. This creates problems in which choosing a particular metric to use as the best metric for any particular dataset 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 and does not 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[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 using 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 providing 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 providing extensions to allow alternative types of outcomes, the framework allows for simultaneous examination of multiple distance metrics. This enables development of the "optimal" MiRKAT, which has high power in the omnibus. We have demonstrated through simulations and analysis of real data that MiRKAT and optimal MiRKAT can be easily applied 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 a strength because 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 adopting kernels on the basis of 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 and have 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, our method uses alternative small-sample corrections[21] (unpublished data) and permutation methods. The present study differs from that detailed in our earlier conference manuscript[15] in that we formalize and fully flesh out the overall framework, explicitly relate the approach to existing distance methods, use alternative small-sample corrections to control type I error, and develop the optimal MiRKAT method for testing across choices of distance metrics.

Material and Methods

Notationally, we assume that n samples have been collected and that their microbial communities have been profiled. For the i^{th} subject, let y_i denote the outcome variable of interest, $\mathbf{Z}_i = (Z_{i1}, Z_{i2}, \ldots, Z_{ip})'$ denote the abundances of all OTUs for individual i (p is the total number of OTUs), and $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{im})'$ be the covariates—such as age, gender, and other clinical and environmental variables that are suspected to influence microbial community diversity and are related to outcomes—that we want to control for. 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, although in some situations it might be a variable that is thought to influence microbial diversity; however, because our goal is association testing rather than causal modeling, the distinction does not affect the validity of

our method given the duality.[22] 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 the kernel machine framework is that it compares pairwise similarity in the outcome variable to pairwise similarity in the microbiome profiles, and high correspondence is 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

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

and for a dichotomous outcome variable (e.g., $y = 1$ or 0 for case or control samples, respectively), we use the logistic kernel machine model

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

where β_0 is the intercept, $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_m]'$ is the vector of regression coefficients for the m covariates, and ε_i is an error term with mean 0 and variance σ^2 for continuous phenotypes. 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 the 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 $\alpha_1, \alpha_2, \ldots, \alpha_n$.

The kernel measures the similarity between different individuals, and different choices of $K(\mathbf{Z}_i, \mathbf{Z}_{i'})$ correspond to different underlying models. For example, setting $K(\mathbf{Z}_i, \mathbf{Z}_{i'}) = \sum_{j=1}^{p} Z_{ij} Z_{i'j}$ implies that $f(\mathbf{Z}_i) = \sum_{j=1}^{p} Z_{ij}\beta_j$, i.e., that 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 that can allow for OTU interactions, nonlinear OTU effects, or incorporation of phylogenetic relationships among OTUs. The matrix of pairwise similarities between pairs of individuals is defined as kernel matrix $\mathbf{K}$, where the $(i, i')^{\text{th}}$ 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 than 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:

$$\mathbf{K} = -\frac{1}{2}\left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n}\right)\mathbf{D}^2\left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n}\right), \qquad \text{(Equation 3)}$$

where $\mathbf{D} = [d_{ij}]$ is the pairwise distance matrix (e.g., weighted or unweighted UniFrac distance or the Bray-Curtis dissimilarity), $\mathbf{I}$ is the identity matrix, $\mathbf{1}$ in $(\mathbf{1}\mathbf{1}'/n)$ is a vector of ones, and $\mathbf{D}^2$ is the element-wise square. For each distance metric, we can construct the corresponding kernel matrix, e.g., weighted or unweighted UniFrac kernels ($\mathbf{K}_w$ or $\mathbf{K}_u$, respectively) can be constructed on the basis of weighted or unweighted distance metrics, respectively. This choice of kernel is in line with the relationship between kernel machine regression and distance-based regression[23] in that it can recover the original distances by using standard kernel operation: $d_{ij}^2 = K_{ii} + K_{jj} - 2K_{ij}$. Further, to ensure that $\mathbf{K}$ is a positive semi-definite matrix, we apply the same positive semi-definiteness correction procedure as in Chen and Li.[15] We first perform an eigenvalue decomposition of eigenvalues $\mathbf{K} = \mathbf{U}\Lambda\mathbf{U}'$, where $\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_n)$, and then reconstruct with the absolute eigenvalues $\mathbf{K}^* = \mathbf{U}\Lambda^*\mathbf{U}'$, where $\Lambda^* = \text{diag}(|\lambda_1|, \cdots, |\lambda_n|)$.

When only a single kernel is considered, we estimate the coefficients $\boldsymbol{\beta}$ and $f(\mathbf{Z})$ by maximizing the following penalized log-likelihood:

$$pl(f, \boldsymbol{\beta}) = \sum_{i=1}^{n} \log L\big(f, \boldsymbol{\beta}; y_i, x_i, z_i\big) - \frac{1}{2}\lambda\|f\|_{\mathcal{H}_k}^2$$

$$= \sum_{i=1}^{n} \log L\big(f, \boldsymbol{\beta}; y_i, x_i, z_i\big) - \frac{1}{2}\lambda\boldsymbol{\alpha}'K\boldsymbol{\alpha}.$$

Through an important relationship between kernel machine regression and mixed models,[24–26] $f(\mathbf{Z})$ can be viewed as a subject-specific random effect that 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 with a standard variance-component score test.[27]

In particular, the score statistic is computed as

$$Q = \frac{1}{2\phi}(\mathbf{y} - \widehat{\mathbf{y}}_0)'\mathbf{K}(\mathbf{y} - \widehat{\mathbf{y}}_0), \qquad \text{(Equation 4)}$$

where $\widehat{\mathbf{y}}_0$ is the predicted mean of $\mathbf{y}$ under H_0 (i.e., $\widehat{\mathbf{y}}_0 = \widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}'\mathbf{X}$ for continuous traits, and $\widehat{\mathbf{y}}_0 = \text{logit}^{-1}(\widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}'\mathbf{X})$ for dichotomous traits), $\widehat{\beta}_0$ and $\widehat{\boldsymbol{\beta}}$ are estimated under the null model by regression of $\mathbf{y}$ on only the covariates $\mathbf{X}$, and ϕ is the dispersion parameter. For the linear kernel machine regression, $\phi = \widehat{\sigma}_0^2$, where $\widehat{\sigma}_0^2$ 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 the p value can be analytically obtained through higher-order moment matching[28] or exact methods[29,30] 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 and with erratic eigenvalue behavior) lead to very conservative tests. Previously considered Satterthwaite methods[15] lead to inflation of type I error. Thus, MiRKAT further considers the use of new, alternative small-sample adjustments for both continuous and dichotomous traits[21] (unpublished data).

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.

Because 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 kernel choices can lead to improved power. Unfortunately, the best kernel requires knowledge of how the microbiome influences the outcome. This is unknown a priori given that 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, $\mathbf{K}_1, \ldots, \mathbf{K}_\ell$, such as unweighted UniFrac, weighted UniFrac, Bray-Curtis kernels, etc., which are constructed from corresponding distance matrices via Equation 3.

The intuition behind the optimal MiRKAT is that it will consider testing with 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.,[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 $\mathbf{r} = \mathbf{y} - \widehat{\mathbf{y}}_0$, where $\widehat{\mathbf{y}}_0$ is the estimated value of $\mathbf{y}$ based on the null model.
2. For each $\mathbf{K}_k$, calculate $Q_k = (1/2\phi)\mathbf{r}'\mathbf{K}_k\mathbf{r}$ and the corresponding p values, p_k, through the asymptotic distribution of Q_k. Then, the minimum p value across all the ℓ kernels is $\mathrm{p}_o = \min_{k\in(1,\ldots,\ell)}\mathbf{p}_k$.
3. Use residual permutation to obtain the null distribution of p_o to accommodate the fact that we have considered multiple kernels.
   a. For a continuous outcome, use the permutation approach of Freeman and Lane.[31] 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 $\mathbf{y}^j = \widehat{\mathbf{y}}_0 + \mathbf{r}^j$.
      iii. Consider $\mathbf{y}^j$ as the new outcome. Refit the null linear regression model by regressing $\mathbf{y}^j$ on $\mathbf{X}$ to obtain the estimated residuals $\widehat{\mathbf{r}}^j$ and $\widehat{\phi}^j$ for calculating the score statistic $Q_k^j = (1/2\widehat{\phi}^j)\widehat{\mathbf{r}}^{j\prime}\mathbf{K}_k\widehat{\mathbf{r}}^j$ with each kernel. Obtain the kernel-specific p value, p_k^j, by comparing Q_k^j to the same asymptotic distribution as in step 2.
      iv. Obtain $\mathrm{p}_o^j = \min_{k\in(1,\ldots,\ell)}\mathbf{p}_k^j$.
   b. For a dichotomous outcome, use the permutation approach of Epistein et al.,[32] 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., $\exp(\widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}'\mathbf{X}_i)$, where $\widehat{\beta}_0$ and $\widehat{\boldsymbol{\beta}}$ are the estimated coefficients under the null logistic regression model in step 1.
      ii. For each permutation j, generate new binary outcomes on the basis of the estimated odds by using the Fisher's non-central hypergeometric distribution (modified version of the BiasedUrn package[33] in R).
      iii. Use the permuted outcome 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 $\mathrm{p}_o^j = \min_{k\in(1,\ldots,\ell)}\mathbf{p}_k^j$.
4. Repeat step 3 for a large number of times B to form an empirical null distribution for p_o.
5. Calculate the final p value as $\mathrm{p} = (1/B)\sum_{b=1}^{B} I(\mathrm{p}_o > \mathrm{p}_o^b)$.

For each permutation j, $\mathrm{p}_1^j, \cdots, \mathrm{p}_\ell^j$ are calculated with the same set of permuted outcomes and are 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 by 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 the χ^2 distribution with the weights and degree of freedom needed 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 by using different kernels and the power of optimal MiRKAT.

We first simulated microbiome datasets according to Chen and Li's general approach,[15] which has been shown to generate simulated data reflective of real OTU counts. In particular, we simulated datasets composed of $n = 100$, 200, or 500 individuals. Then, we generated the OTU information for each individual in a simulated dataset 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 Charlson et al.'s real upper-respiratory-tract microbiome dataset,[34] which consists of 856 OTUs measured on each of 60 samples. Then, for each individual we generated OTU counts on the same 856 OTUs by using the estimated parameters and assumed 1,000 total counts per sample. For both continuous outcomes and dichotomous outcomes, we considered two simulation scenarios that differed in how the OTUs were related to the outcome.

Under simulation scenario 1, the outcome was related to a cluster of taxa that depend on a phylogenetic tree. Specifically, we partitioned all the OTUs into 20 clusters (lineages) by performing the partitioning-around-medoids algorithm on the basis of the OTU distance matrix. The abundance of these OTU clusters varied greatly, such that each OTU cluster corresponded to some possible bacterial lineage. We then used the model to choose a relatively abundant OTU cluster that constituted 19.4% of the total OTU reads to be related to the outcome. For continuous outcomes, we simulated under the model

$$y_i = 0.5X_{1i} + 0.5X_{2i} + \beta\mathrm{scale}\left(\sum_{j\in\mathcal{A}} Z_{ij}\right) + \varepsilon_i, \qquad \text{(Equation 5)}$$

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

For dichotomous outcomes, we simulated under the model

$$\mathrm{logit}\left(E(y_i \mid \mathbf{X}_i, \mathbf{Z}_i)\right) = 0.5\mathrm{scale}(X_{1i} + X_{2i}) + \beta\mathrm{scale}\left(\sum_{j\in\mathcal{A}} Z_{ij}\right). \qquad \text{(Equation 6)}$$

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 OTUs in the selected cluster. The "scale" function standardizes the total OTU abundance in the associated cluster to have mean 0 and SD 1. X_{1i} was simulated as a Bernoulli random variable with success probability 0.5. For X_{2i}, we considered situations in which X_{2i} and microbiome profiles ($\mathbf{Z}_i$) were correlated and in which X_{2i} and $\mathbf{Z}_i$ were independent. In the simulation wherein X_{2i} and $\mathbf{Z}_i$ were independent, X_{2i} was simulated as $N(0,1)$. For the case wherein X_{2i} and $\mathbf{Z}_i$ were correlated, we let $X_{2i} = \text{scale}\left(\sum_{j\in\mathcal{A}} Z_{ij}\right) + N(0,1)$.

Under simulation scenario 2, the outcome was associated with the ten most abundant OTUs in all samples, without regard for the phylogeny. In particular, instead of clustering the OTUs on the basis of the phylogenetic relationship, we simply selected the ten OTUs with the largest average number of reads across all samples. Then, we simulated the continuous outcome as

$$\gamma_i = 0.5X_{1i} + 0.5X_{2i} + \beta\text{scale}\left(\sum_{j\in\mathcal{A}} \frac{Z_{i(j)}}{\overline{Z}_{(j)}}\right) + \varepsilon_i. \qquad \text{(Equation 7)}$$

We simulated the dichotomous outcome as

$$\text{logit}\left(E(\gamma_i \mid \mathbf{X}_i, \mathbf{Z}_i)\right) = 0.5\text{scale}(X_{1i} + X_{2i}) + \beta\text{scale}\left(\sum_{j\in\mathcal{A}} \frac{Z_{i(j)}}{\overline{Z}_{(j)}}\right), \qquad \text{(Equation 8)}$$

where $\varepsilon_i \sim N(0,1)$, X_{1i} and X_{2i} are defined as earlier, $\mathcal{A}$ denotes the set of the ten most abundant OTUs, and $\overline{Z}_{(j)}$ is the average number of reads for the j^{th} OTU across samples. We divided the OTU reads by their corresponding average to avoid a situation in which a single or a few OTUs could dominate the total effect.

We simulated the additional covariates ($\mathbf{X}$) as before, and we again considered the scenario in which the covariates were associated with the microbiome and the scenario in which the covariates were independent of the microbiome.

For both simulation scenarios, we considered using the weighted and unweighted UniFrac kernels (K_w and K_u, respectively), the 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 $K_{0.75}$, respectively. All of these kernels were computed from the corresponding distances. We considered these particular kernels (distances) because they represent a range of different classes of kernels: the UniFrac-based methods utilize phylogenetic relationships, whereas the Bray-Curtis kernel does not, and the weighted and generalized UniFrac kernels account for abundance information to differing degrees, whereas the unweighted UniFrac kernel does not.

We used each kernel to apply MiRKAT to the simulated datasets to test for associations between the simulated OTUs ($\mathbf{Z}$) and the outcome ($\mathbf{y}$). Additionally, we also applied optimal MiRKAT. We applied tests with and without adjustment for the potential confounders ($\mathbf{X}$). For comparison, we further considered a naive Bonferroni-adjusted test, which selects the minimum p value across all the single-kernel testing and uses $\ell \times p_{\min}$, where $p_{\min}$ is the smallest p value across all single-kernel tests and ℓ is the total number of tests, as the final p value. For each choice of sample size n, simulation scenario, and correlation structure between the microbiome and covariates, we conducted 5,000 simulations with $\beta = 0$ to examine the 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 2,000 simulations for each choice of sample size, simulation scenario, correlation structure, and value of β.

Results

In this section, we present the simulation results from performing our proposed MiRKAT and optimal MiRKAT methods, as well as the results from applying our methods to two real datasets. 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 outcomes are shown in Table 1. In simulation scenario 1, a single phylogenetic cluster of OTUs was associated with the outcome, and in simulation scenario 2, the ten most abundant OTUs were associated with the outcome. Note that when the covariates were independent of the microbiome, both simulation scenarios were equivalent because there was no association between γ and $\mathbf{Z}$. For both simulation scenarios, when the covariates ($\mathbf{X}$) and the microbiome composition ($\mathbf{Z}$) were independent, MiRKAT was valid with or without adjusting for $\mathbf{X}$. However, when $\mathbf{X}$ and $\mathbf{Z}$ were correlated, adjusting for $\mathbf{X}$ was necessary: the type I error was seriously inflated if the confounder $\mathbf{X}$ was not accounted for.

Figures 1 and 2 show the statistical power for the tests with continuous outcomes in simulation scenario 1, in which a phylogenetic cluster of OTUs was associated with the outcome. Specifically, Figure 1 shows the power when $\mathbf{X}$ and $\mathbf{Z}$ were independent, and Figure 2 shows the power when $\mathbf{X}$ and $\mathbf{Z}$ were correlated. Note that for Figure 2, we only considered statistical tests that adjusted for $\mathbf{X}$ because the tests without $\mathbf{X}$ adjustment had inflated type I error and were invalid in such situations.

The power is presented for MiRKAT with each individual kernel, the optimal MiRKAT (which incorporates multiple kernels), and the naive Bonferroni-adjusted test. For all the kernels that were considered, the power increased when the association strength increased. Good kernel choices can greatly improve the statistical power of detecting association, whereas improper kernel choice 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, and the unweighted UniFrac kernel was the least powerful. Compared to the weighted UniFrac kernel, the optimal MiKRAT, which considers all metrics, lost some power but still maintained power considerably better than that of many other kernel choices. As expected, the optimal test was always more powerful than the naive Bonferroni-adjusted test.

Figures 3 and 4 show the statistical power for simulation scenario 2, where the top ten most abundant OTUs were associated with the outcome without regard for phylogeny.

Table 1. Empirical Type I Errors for MiRKAT and Optimal MiRKAT with Continuous Outcome

Simulation Setup	n	Empirical Type I Errors								
		K_w	K_u	K_{BC}	K_0	$K_{0.25}$	$K_{0.5}$	$K_{0.75}$	$K_{optimal}$	$K_{p_{min}}$
Simulation Scenario 1: Clustered OTUs										
$X \perp Z$, no adjustment for **X**	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
$X \perp Z$, adjustment 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
$X \not\perp Z$, no adjustment 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*
$X \not\perp Z$, adjustment 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 Ten OTUs										
$X \perp Z$, no adjustment for **X**	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
$X \perp Z$, adjustment 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
$X \not\perp Z$, no adjustment 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*
$X \not\perp Z$, adjustment 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

Type I error was evaluated for scenarios in which additional covariates were independent of the OTUs ($X \perp Z$) or related to the OTUs ($X \not\perp Z$) with the use of 5,000 simulated datasets. K_w, K_u, K_{BC}, K_0, $K_{0.25}$, $K_{0.5}$, and $K_{0.75}$ represent MiRKAT results for the weighted UniFrac kernel, unweighted UniFrac kernel, Bray-Curtis kernel, and generalized UniFrac kernels with $\alpha = 0$, 0.25, 0.5, and 0.75, respectively. $K_{optimal}$ represents the simulation results for optimal MiRKAT considering all seven kernels, and $K_{p_{min}}$ shows the results for a naive Bonferroni-adjusted test. The p values for optimal MiRKAT were obtained by 1,000 permutations. *Inflated type I error.

We again show the power when **X** and **Z** were independent (Figure 3) and when **X** and **Z** were correlated (Figure 4). Results were similar to those of simulation scenario 1, except that the Bray-Curtis distance metric gave the highest power. Optimal MiRKAT, which considers all distance metrics, had power that was smaller but comparable to that of the Bray-Curtis distance but much higher than that of 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 show that proper kernel choice is essential for being well powered to discover associations between microbiome composition and outcomes and that poor kernel choice leads to tremendous power loss. Optimal MiRKAT, however, alleviates the problem by considering different kernels and is more robust than single-distance-based analysis given that it hedges against different scenarios and works well in the omnibus.

The simulation results for dichotomous outcomes are quantitatively similar to the results obtained from continuous outcomes. The type I error results are summarized in Table S1, and power results are shown in Figures S1–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.[23] Consequently, MiRKAT with a single kernel can be viewed as a PERMANOVA generalization 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 without covariates to be adjusted for. For example, Figure S5 shows the p values for MiRKAT and the distance-based approach for 2,000 simulated datasets when a single distance or kernel was used. However, because it uses the asymptotic distribution, MiRKAT is considerably faster than corresponding distance-based approaches, especially with large sample sizes (Figure S6).

Analysis of Smoking Data

Recently, a microbiome-profiling study was conducted to examine the communities within the upper respiratory

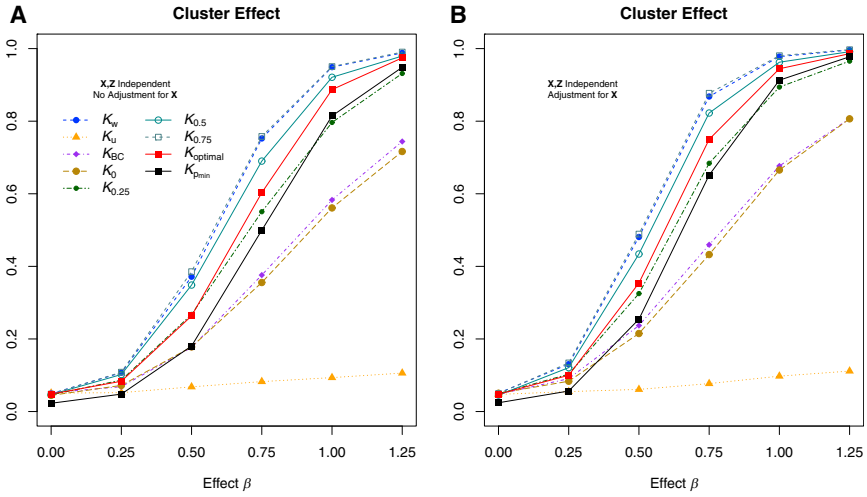

Figure 1. Type I Error and Power of MiRKAT Based on Different Kernels for Simulation Scenario 1 with Continuous Outcome when X and Z Are Independent
A selected phylogenetic cluster of the OTUs were associated with the outcome, and covariates (**X**) and the microbiome profiles (**Z**) were simulated independently. Results are shown for tests that did (A) or did not (B) adjust for **X**. K_w, K_u, K_{BC}, K_0, $K_{0.25}$, $K_{0.5}$, and $K_{0.75}$ represent MiRKAT results from the weighted UniFrac kernel, unweighted UniFrac kernel, Bray-Curtis kernel, and generalized UniFrac kernels with $\alpha = 0$, 0.25, 0.5, and 0.75, respectively. $K_{optimal}$ represents the simulation results for optimal MiRKAT considering all seven kernels, and $K_{p_{min}}$ shows the results for a naive Bonferroni-adjusted test. Sample size $n = 100$.

tract[34] in order to explain the effect of cigarette smoking on the orpharyngeal and nospharyngeal microbiome. Although details can be found in the original manuscript and subsequent re-analyses,[14] in brief, swab samples were collected from the right and left nasopharynx and oropharynx of 29 smoking and 33 non-smoking adults. The variable region 1–2 (V1–V2) of the bacterial gene 16S rRNA was PCR amplified and subjected to multiplexed pyrosequencing. OTUs were constructed with the QIIME pipeline. Samples with fewer than 500 reads and OTUs with only one read were removed, resulting in an OTU table with 60 samples (28 smokers and 32 nonsmokers) and 856 OTUs. Additional covariates in these data included gender and antibiotic use within the last 3 months.

Distance-based analysis of the oropharyngeal samples via 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, 75% of smokers were male, yet only 56% of non-smokers were male. The odds ratio of smoking between males and females was 2.33 within the dataset. 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 by the gender imbalance. Additionally, the tests were conducted with 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 by using MiRKAT. Specifically, we applied MiKRAT to analyze the association between smoking and microbial community composition by using weighted and unweighted UniFrac distance matrices and the Bray-Curtis distance, except that here we transformed them to be similarity metrics to form the kernels and 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} to be 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 the potential confounders are controlled for, 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 accommodating covariates and multiple kernels.

Analysis of Fecal Protease Data

Fecal proteases (FPs) are enteric enzymes that are elevated in subsets of individuals with irritable bowel syndrome (IBS) and inflammatory bowel disease (MIM: 266600). It was demonstrated that FPs from IBS-affected individuals have a profound impact on intestinal physiology, including visceral sensitivity and colonic permeability in mice.[35] 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.[36] conducted a study to examine the relationship between FP levels and microbiota in human fecal samples from 30 individuals affected by IBS and 24 healthy adults. 454 pyrosequencing of the gene 16S rRNA was again used for profiling 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 dataset

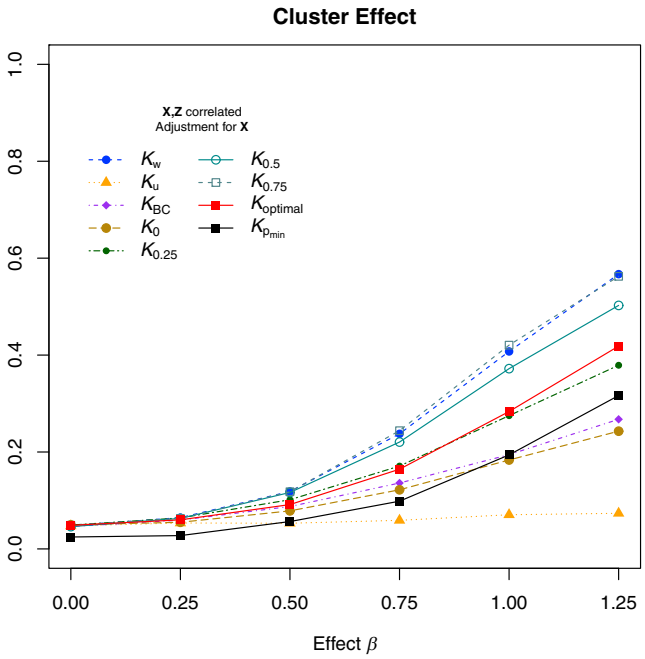

Figure 2. Type I Error and Power of MiRKAT Based on Different Kernels for Simulation Scenario 1 with Continuous Outcome when X and Z Are Correlated

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

(limiting to the 23 diarrhea-predominant IBS-affected subjects and 23 healthy control subjects) 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 by 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 unweighted UniFrac kernel and Bray-Curtis kernel yielded significant p values ($p = 0.0046$ and 0.039, respectively), whereas the weighted UniFrac kernel gave a non-significant result ($p = 0.124$). The unweighted UniFrac kernel is primarily based on the presence or absence of an OTU, whereas the weighted UniFrac kernel 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, gave a single p value of 0.0116 after covariate adjustment. This further demonstrates the advantages of optimal MiRKAT to be able to consider multiple kernels given that using individual distance metrics yielded disparate results and is difficult to interpret.

Discussion

We propose 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 non-parametrically. The kernel matrix, which defines the functional form of the microbiome effect, is constructed via the exploitation of its correspondence with the popular distance metric designed to convey phylogenetic or taxonomic information among different OTUs. Additionally, the proposed method allows the 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 kernel choice. 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 because 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 the false-positive rate, but the trade-off is that one might lose power if there are too many overly disparate kernels under consideration—use of highly correlated kernels will not affect power very much. In the most extreme cases, optimal MiRKAT from multiple perfectly correlated kernels will generate the same p value as will each of the individual kernel tests. Furthermore, we note that the tests using each of the individual kernels are constructed on the basis of the same datasets 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 likelihood-ratio tests, recently have been the focus of considerable statistical research.[37,38] However, this is frequently much more computationally intensive than the score

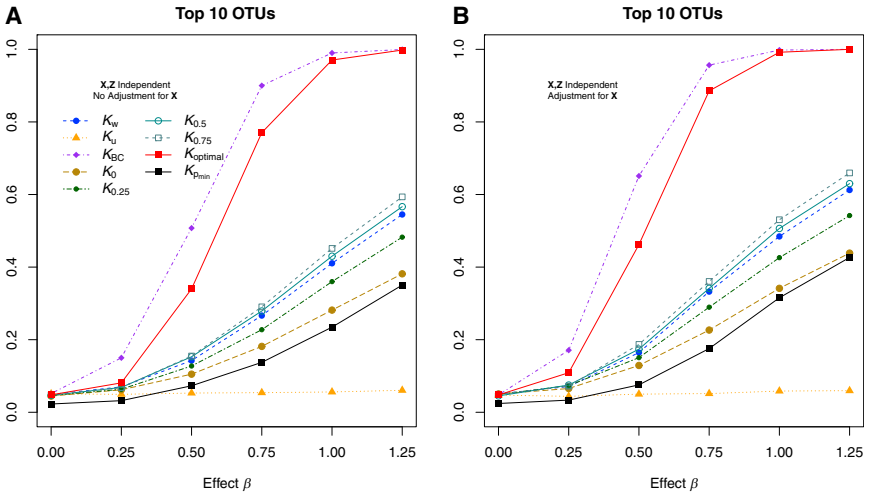

Figure 3. Type I Error and Power of MiRKAT Based on Different Kernels for Simulation Scenario 2 with Continuous Outcome when X and Z Are Independent
The ten most abundant OTUs were associated with the outcome. Additional covariates (**X**) and the microbiome profiles (**Z**) were simulated independently. Results are shown for tests that did (A) or did not (B) adjust for **X**. K_w, K_u, K_{BC}, K_0, $K_{0.25}$, $K_{0.5}$, and $K_{0.75}$ represent MiRKAT results for the weighted UniFrac kernel, unweighted UniFrac kernel, Bray-Curtis kernel, and generalized UniFrac kernels with $\alpha = 0$, 0.25, 0.5, and 0.75, respectively. K_{optimal} represents the simulation results for optimal MiRKAT considering all seven kernels, and $K_{p_{\min}}$ shows the results for a naive Bonferroni-adjusted test. Sample size $n = 100$.

test, especially when many kernels are under consideration. Furthermore, very limited work has been conducted on the 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 to 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 as a result of 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 composition percentages 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 relatively fast computation, MiRKAT could enable a stepwise variable selection approach with the Akaike information criterion or the Bayesian information criterion. Such an approach could be applied post hoc to identify the variables most strongly driving apparent associations. It might also be possible to use a penalized regression approach within the kernel framework,[39] 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, current microbiome studies consider issues such as confounding, covariate adjustment, and accommodation of more-sophisticated outcomes to be increasingly important. 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 the 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 six figures and one table and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2015.04.003.

Acknowledgments

This study was supported in part by NIH grants K01DK092330, R01HG007508, R01HG006139, and R01GM097505; Center for Gastrointestinal Biology and Disease pilot feasibility grant P30DK03498; the Hope Foundation; and the Gerstner Family Career Development Award in Individualized Medicine.

Received: January 21, 2015
Accepted: April 7, 2015
Published: May 7, 2015

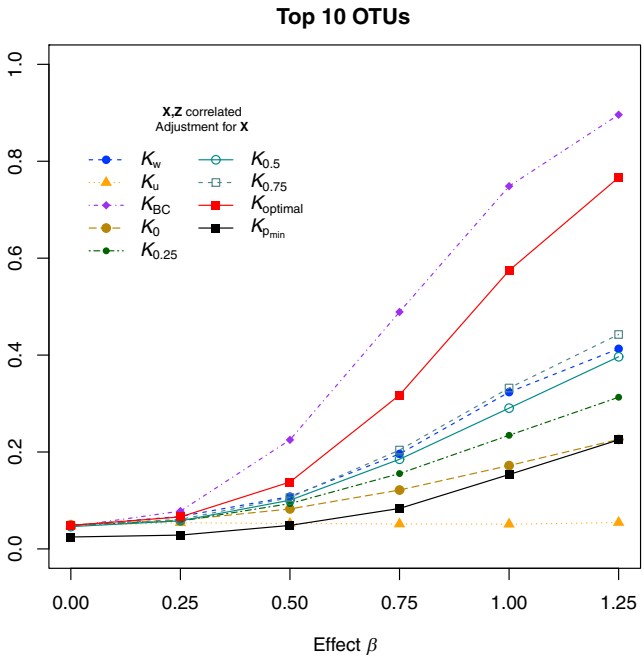

Figure 4. Type I Error and Power of MiRKAT Based on Different Kernels for Simulation Scenario 2 with Continuous Outcome when X and Z Are Correlated

The ten most abundant OTUs were associated with the outcome. Additional covariates (**X**) and the microbiome profiles (**Z**) were correlated such that $X_{2i} = \text{scale}(\sum_{j \in \mathcal{A}} Z_{ij}) + N(0,1)$, where $\mathcal{A}$ represents the top ten most abundant OTUs. Results are presented only for MiRKAT with **X** adjustment because unadjusted tests gave seriously inflated type I error. K_w, K_u, K_{BC}, K_0, $K_{0.25}$, $K_{0.5}$, and $K_{0.75}$ represent MiRKAT results for the weighted UniFrac kernel, unweighted UniFrac kernel, Bray-Curtis kernel, and generalized UniFrac kernels with $\alpha = 0$, 0.25, 0.5, and 0.75, respectively. K_{optimal} represents the simulation results for optimal MiRKAT considering all seven kernels, and $K_{p_{\min}}$ shows the results for a naive Bonferroni-adjusted test. Sample size $n = 100$.

Web Resources

The URLs for data presented herein are as follows:

Implementation of MiRKAT in the R language, http://research.fhcrc.org/wu/en.html

MiRKAT R package and manual, http://research.fhcrc.org/wu/en.html

OMIM, http://www.omim.org

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., Bergström, G., Behre, C.J., Fagerberg, B., Nielsen, J., and Bäckhed, 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.; MetaHIT Consortium (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*, 2106–2113.
15. Chen, J., and Li, H. (2013). Kernel Methods for Regression Analysis of Microbiome Compositional Data. In Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association, M. Hu, Y. Liu, and J. Lin, eds. (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. Am. J. Hum. Genet. *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.; NHLBI GO Exome Sequencing Project—ESP Lung Project Team (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*, 224–237.

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. 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).
22. 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.
23. Pan, W. (2011). Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing. Genet. Epidemiol. *35*, 211–216.
24. 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*, 1079–1088.
25. 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.
26. Gianola, D., and van Kaam, J.B. (2008). Reproducing kernel hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics *178*, 2289–2303.
27. Lin, X. (1997). Variance component testing in generalised linear models with random effects. Biometrika *84*, 309–326.
28. 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. Comput. Stat. Data Anal. *53*, 853–856.
29. Davies, R. (1980). The distribution of a linear combination of chi-2 random variables. J. R. Stat. Soc. Ser. C Appl. Stat. *29*, 323–333.
30. 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. Comput. Stat. Data Anal. *54*, 858–862.
31. Freedman, D., and Lane, D. (1983). A nonstochastic interpretation of reported significance levels. J. Bus. Econ. Stat. *1*, 292–298.
32. 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.
33. Fog, A. (2008). Sampling methods for wallenius' and fisher's noncentral hypergeometric distributions. Commun. Stat. Simul. Comput. *37*, 241–257.
34. 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.
35. Annaházi, A., Gecse, K., Dabek, M., Ait-Belgnaoui, A., Rosztóczy, A., Róka, R., Molnár, T., Theodorou, V., Wittmann, T., Bueno, L., and Eutamene, H. (2009). Fecal proteases from diarrheic-IBS and ulcerative colitis patients exert opposite effect on visceral sensitivity in mice. Pain *144*, 209–217.
36. 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.
37. Crainiceanu, C.M., and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. J. R. Stat. Soc. Series B Stat. Methodol. *66*, 165–185.
38. Greven, S., Crainiceanu, C.M., Kchenhoff, H., and Peters, A. (2008). Restricted likelihood ratio testing for zero variance components in linear mixed models. J. Comput. Graph. Stat. *17*, 870–891.
39. Allen, G.I. (2013). Automatic feature selection via weighted kernels and regularization. J. Comput. Graph. Stat. *22*, 284–299.

The American Journal of Human Genetics

Supplemental Data

Testing in Microbiome-Profiling Studies with MiRKAT, the Microbiome Regression-Based Kernel Association Test

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

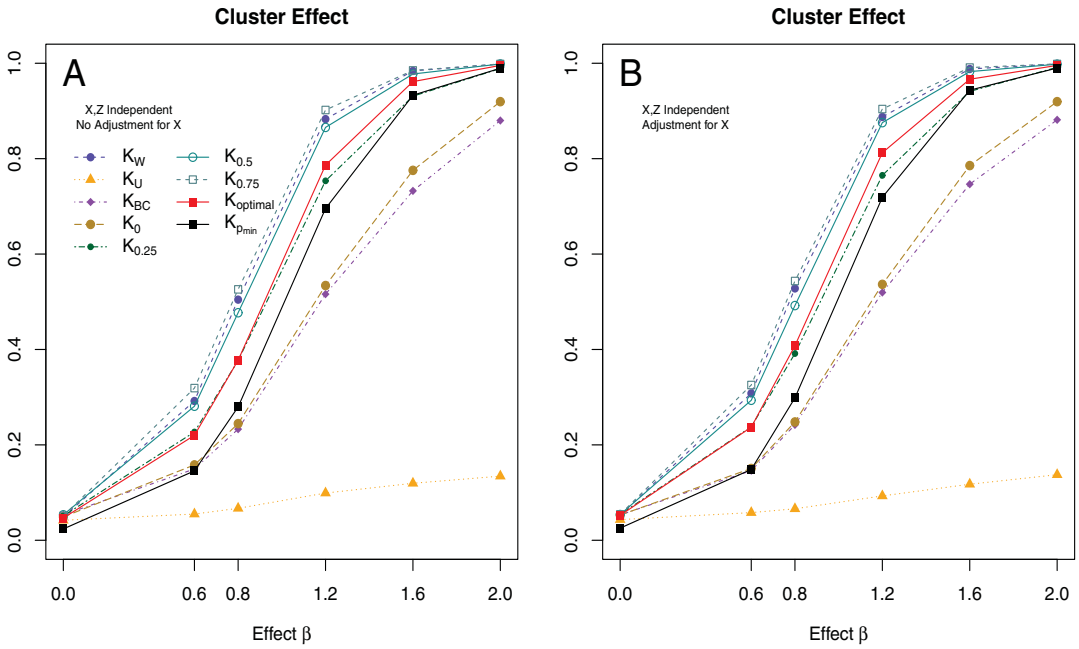

Figure S1: **Type I error and Power of MiRKAT Based on Different Kernels for *Simulation Scenario 1* with Dichotomous Outcome when X and Z are Independent**: a selected phylogenetic cluster of the OTUs are associated with the outcome. Additional covariates X and microbiome effect Z were simulated independently. 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 $n = 200$.

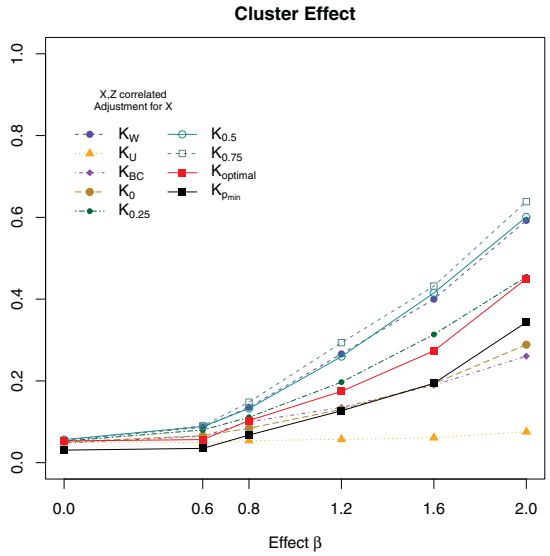

Figure S2: **Type I error and Power of MiRKAT Based on Different Kernels for *Simulation Scenario 1* with Dichotomous Outcome when X and Z are Correlated**: a selected phylogenetic cluster of the OTUs are associated with the outcome. Additional covariates X and microbiome composition Z are correlated 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 $n = 200$.

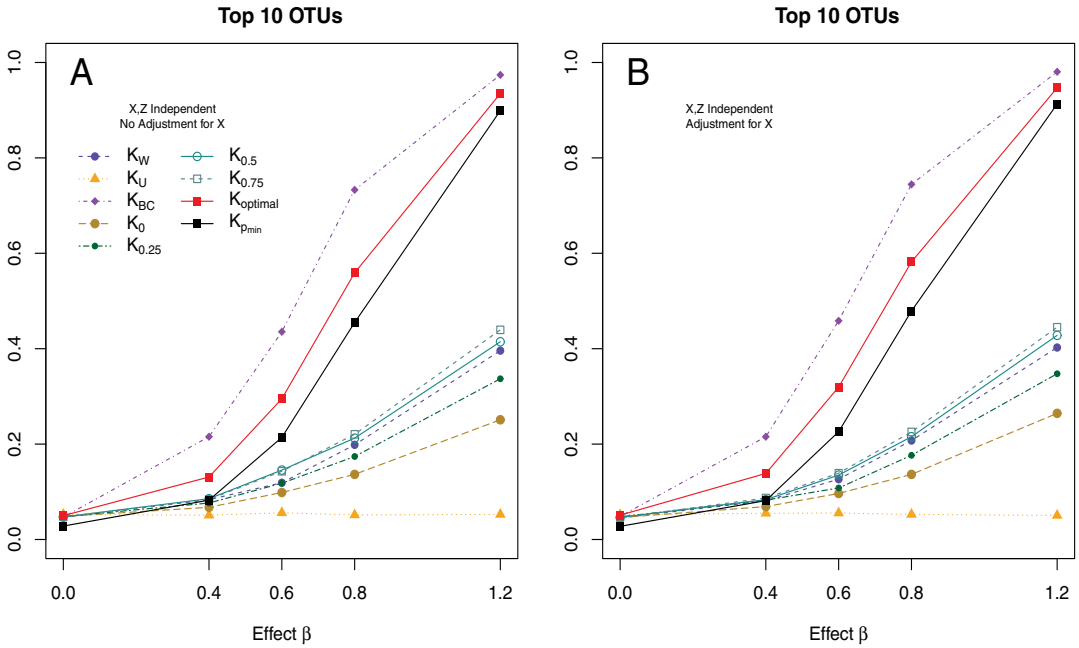

Figure S3: **Type I error and Power of MiRKAT Based on Different Kernels for *Simulation Scenario 2* with Dichotomous Outcome when X and Z are Independent**: the 10 most abundant OTUs are associated with the outcome. Additional covariates X and microbiome effect Z were simulated independently. Panel 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 $n = 200$.

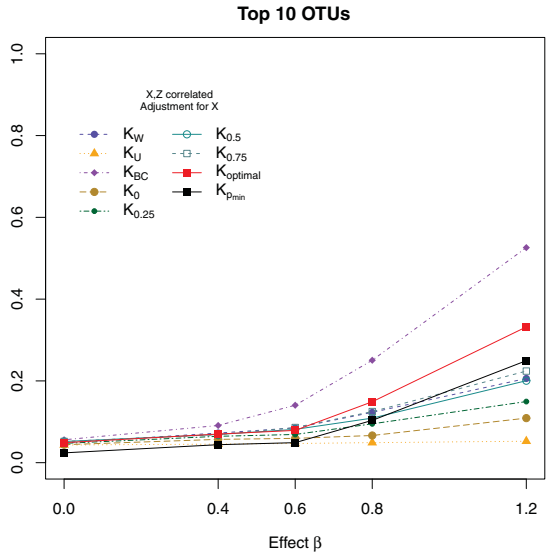

Figure S4: **Type I error and Power of MiRKAT Based on Different Kernels for *Simulation Scenario* 2 with Dichotomous Outcome when X and Z are Correlated**: the 10 most abundant OTUs are associated with the outcome. Additional covariates X and microbiome composition Z are correlated through $X_{2i} =$ scale $(\sum_{j\in\mathcal{A}} Z_{ij}) + N(0,1)$. 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 $n = 200$.

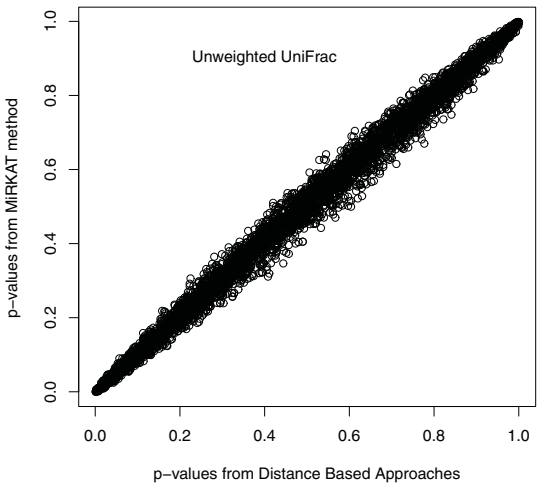

Figure S5: Example plot of the p-value correlation using distance based approach and MiRKAT when no additional covariates are considered. 5000 simulations are plotted at sample size n = 200 for continuous outcome. Unweighted UniFrac distance and kernel were used for the distance based approach and MiRKAT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

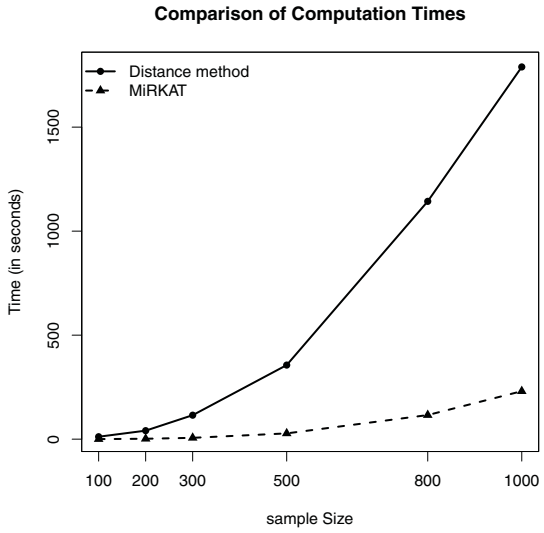

Figure S6: Computation times of MiRKAT and distance based test as a function of the sample size for continuous outcome. The figure presents the total computation time for 100 repeated tests with each sample size. 999 permutations (the default number) were used in distance based approaches.

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