

METHODOLOGY

Open Access



MiRKAT-S: a community-level test of association between the microbiota and survival times

Anna Plantinga¹, Xiang Zhan², Ni Zhao³, Jun Chen⁴, Robert R. Jenq⁵ and Michael C. Wu^{1,2*}

Abstract

Background: Community-level analysis of the human microbiota has culminated in the discovery of relationships between overall shifts in the microbiota and a wide range of diseases and conditions. However, existing work has primarily focused on analysis of relatively simple dichotomous or quantitative outcomes, for example, disease status or biomarker levels. Recently, there is also considerable interest in the relationship between the microbiota and censored survival outcomes, such as in clinical trials. How to conduct community-level analysis with censored survival outcomes is unclear, since standard dissimilarity-based tests cannot accommodate censored survival times and no alternative methods exist.

Methods: We develop a new approach, MiRKAT-S, for community-level analysis of microbiome data with censored survival times. MiRKAT-S uses ecologically informative distance metrics, such as the UniFrac distances, to generate matrices of pairwise distances between individuals' taxonomic profiles. The distance matrices are transformed into kernel (similarity) matrices, which are used to compare similarity in the microbiota to similarity in survival times between individuals.

Results: Simulation studies using synthetic microbial communities demonstrate correct control of type I error and adequate power. We also apply MiRKAT-S to examine the relationship between the gut microbiota and survival after allogeneic blood or bone marrow transplant.

Conclusions: We present MiRKAT-S, a method that facilitates community-level analysis of the association between the microbiota and survival outcomes and therefore provides a new approach to analysis of microbiome data arising from clinical trials.

Keywords: Survival data, Kernel machine regression, Distance-based analysis, Community-level analysis

Background

The human microbiota, or the collection of microorganisms that inhabits the human body, plays an important role in many areas of health and disease. The development of next-generation sequencing technologies allows culture-free profiling of entire microbial communities, often by sequencing the 16S ribosomal RNA (rRNA) gene [1–3]. Similar 16S sequences are clustered into

operational taxonomic units (OTUs); when clustered at the level of 97% similarity, these OTUs represent bacterial species [4]. Using these methods, the human microbiome has been studied at a variety of body sites including the gut [5], skin [6], and respiratory tract [7] and has been associated with many health conditions, such as inflammatory bowel disease, diabetes, psoriasis, and chronic obstructive pulmonary disease [8–10].

Associations between the human microbiota and health outcomes can be assessed by comparing individual OTU abundances or overall diversity metrics between samples or conditions [5, 11]. However, since taxonomic profiles are sparse and high-dimensional—hundreds to thousands of unique OTUs may be identified, many of which are

*Correspondence: mcwu@fredhutch.org

¹Department of Biostatistics, University of Washington, 1705 NE Pacific Street, Seattle, Washington, USA

²Public Health Sciences Division, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, Seattle, Washington, USA

Full list of author information is available at the end of the article

present in only a subset of samples—comparisons on the level of individual OTUs may have low power. An alternative to OTU-level analysis is to compare the microbiota at the community level, i.e., to compare overall taxonomic profiles between individuals [12–14]. This class of analyses is often performed by computing pairwise distances between communities (samples), where the distance metrics are ecologically relevant and may incorporate phylogenetic structure. The matrix of pairwise distances is summarized by its top principal coordinates for visualization, and distance-based multivariate methods coupled with permutation are used to determine if dissimilarity is related to an outcome. Community-level analyses may provide power gains by utilizing phylogenetic information, avoiding the multiple testing problem, and aggregating modest effects across multiple taxa [15].

Recently, as an alternative to distance-based approaches that use permutation analysis, Zhao et al. [14] proposed the microbiome regression-based kernel association test (MiRKAT). MiRKAT uses a kernel machine framework with a variety of ecologically informative kernels to test for associations between the human microbiota and either continuous or binary outcomes. Intuitively, MiRKAT compares similarity in taxonomic profiles between communities (where similarity is measured via a kernel, which can be obtained by transforming relevant distance matrices) to similarity in outcome measures. p values are obtained analytically using a variance-component score test. MiRKAT is equivalent to distance-based analysis but has the added advantages of flexible modeling of the relationship between the microbiota and outcome measures, natural incorporation of covariates, and efficient computation of p values.

A limitation of existing community-level analysis approaches is that they cannot accommodate censored survival outcomes. However, such outcomes are of tremendous interest as microbiome profiling studies move into the clinical arena. For example, the lung microbiota has been related to progression of idiopathic pulmonary fibrosis [16] and the gut microbiota to overall survival after allogeneic blood and bone marrow transplant [5]. Additional OTU-level studies with survival outcomes have shown associations between the intestinal microbiota and development of atopic dermatitis [17] and allergic rhinitis [18] in children.

To address this critical gap in the literature, in this paper, we propose a test for association between the microbiota and censored survival outcomes (MiRKAT-S), accounting for covariates and potential confounders. We perform a distance-based analysis using the kernel machine Cox regression framework, encoding taxonomic profiles into kernel matrices via a transformation of distance metrics appropriate for microbial communities. This allows the analysis to take into account phylogenetic information

and other features specific to biological communities. To formally test the association between the microbiota, as encoded in the kernel matrix, and censored survival times, we use a variance-component score test. However, when applied to microbial community profiles summarized by common kernels, the usual test statistic with p values calculated by resampling procedures is highly conservative [19, 20]. We therefore implement a small sample correction that provides proper control of type I error while maintaining adequate power, and we calculate p values analytically rather than by resampling. We demonstrate the performance of MiRKAT-S using real and simulated data summarized by a variety of kernels commonly used in microbial ecology.

This work represents the translation of existing methods in genetic studies with survival outcomes to applications in microbiome research. The first major contribution of this work is to allow survival outcomes in the kernel machine regression framework with kernels that appropriately encode microbiome data. Our small sample correction method provides proper control of type I error and improved power when using microbiota-appropriate kernels, whereas the kernel machine regression-based test as implemented for genetic studies has almost no power to detect relationships between microbial taxonomic profiles and survival. Secondly, the ability to perform the test using a variety of kernels provides robustness to the nature of the true association between the microbiota and survival. Therefore, although MiRKAT-S is technically similar to previous kernel machine regression methods, it enables microbiome analyses that are not possible using existing methods.

Methods

To associate the microbiota at the community level and censored survival times, we will relate censored survival times to taxonomic profiles using a flexible non-parametric modeling framework. We will assess significance via a variance-component score test which acknowledges the modest sample sizes of the most taxonomic profiling studies. In this section, we first describe the modeling framework, followed by the testing strategy and technical advances necessary to ensure proper control of type I error in this framework. Finally, we describe simulations encompassing a variety of true relationships between the microbiota and survival time.

Model specification

Suppose that for each of n subjects, we observe the microbial taxonomic profile, encoded by a q -vector of OTU counts Z_i , and a p -vector of other covariates X_i . Let T_i be the survival time and C_i the censoring time for the i th subject. We observe the bivariate vector (Δ_i, U_i) , where $U_i = \min(T_i, C_i)$ is the observed time and $\Delta_i = I(T_i \leq$

C_i) is the event indicator for subject i . We wish to test whether taxonomic profiles \mathbf{Z} are associated with survival time, adjusting for covariates \mathbf{X} .

The most commonly used model for censored survival times is the Cox proportional hazards model [21] due to its flexibility and relative robustness. Therefore, to relate T and (\mathbf{X}, \mathbf{Z}) , we propose to use the kernel machine Cox proportional hazards model [19, 20], so that

$$\lambda(t; \mathbf{X}, \mathbf{Z}) = \lambda_0(t) \exp[\mathbf{X}\boldsymbol{\beta} + f(\mathbf{Z})] \quad (1)$$

where $\lambda_0(t)$ is the baseline hazard function. In the kernel machine regression framework, $f(\cdot)$ is generated by a positive definite kernel function $K(\cdot, \cdot)$, that is, $f(\cdot)$ lies in the reproducing kernel Hilbert space \mathcal{H}_K . Under the representer theorem [22], $f(\mathbf{Z}_i) = \sum_{i'=1}^n \alpha_{i'} K(\mathbf{Z}_i, \mathbf{Z}_{i'})$ for some $\alpha_1, \dots, \alpha_n$. Choosing different kernel functions $K(\cdot, \cdot)$ allows specification of a wide variety of models. For example, the kernel function $f(\mathbf{Z}_i) = \mathbf{Z}_i' \boldsymbol{\gamma}$, corresponding to the linear kernel $K(\mathbf{Z}_i, \mathbf{Z}_{i'}) = \mathbf{Z}_i \mathbf{Z}_{i'}'$, is used to specify a linear model. Kernels are similarity matrices, so each element $K_{j,k} = K(\mathbf{Z}_j, \mathbf{Z}_k)$ represents the pairwise similarity between samples j and k . Because we use a score test, which depends only on the null model, any kernel will result in a valid test; however, the choice of kernel does affect the power of the test.

To specify relevant models for microbial profiles, we use kernel functions that encode the similarity between the microbiota for two samples via a transformation of pairwise distance metrics. There are many commonly used ecological distance metrics, each with different features and strengths. For example, the UniFrac [23] and generalized UniFrac [24] distances take into account the organization of OTUs into phylogenetic trees, thereby gaining power when clusters of taxa are associated with the outcome. Other distances, such as the Bray-Curtis dissimilarity [25], look at the presence and relative abundance of each OTU regardless of phylogenetic structure. These and other commonly used distance metrics can be used to create distance matrices \mathbf{D} , where each element d_{ij} is a pairwise distance between the taxonomic profiles of two samples. The distance matrices are then transformed to kernels, or similarity matrices, via

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

as described in [14]. Here, \mathbf{I} is the identity matrix and $\mathbf{1}$ is an n -vector of ones. To ensure that \mathbf{K} is positive semi-definite, we replace negative eigenvalues with zero. That is, we perform an eigenvalue decomposition $\mathbf{K} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}$, where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_n)$, and then reconstruct the kernel matrix using the non-negative eigenvalues $\boldsymbol{\Lambda}^* = \text{diag}(\max(\lambda_1, 0), \dots, \max(\lambda_n, 0))$ so that $\mathbf{K} = \mathbf{U}\boldsymbol{\Lambda}^*\mathbf{U}$.

Score test

Testing whether taxonomic profiles are associated with the outcome in the kernel machine Cox model corresponds to testing the hypothesis $H_0 : f(\mathbf{Z}) = \mathbf{K}\boldsymbol{\alpha} = 0$. When the model is re-expressed using kernels as

$$\lambda(t; \mathbf{X}, \mathbf{Z}) = \lambda_0(t) \exp[\mathbf{X}\boldsymbol{\beta} + \mathbf{K}\boldsymbol{\alpha}]$$

where $\mathbf{K} = \{K(\mathbf{Z}_i, \mathbf{Z}_j)\}_{(i,j)}$, we can estimate $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ by maximizing the penalized log partial likelihood function

$$\log(PL) = \sum_{i=1}^n \int_0^\infty \log \left[\frac{e^{\boldsymbol{\beta}'\mathbf{X}_i + \boldsymbol{\alpha}'\mathbf{K}_i}}{\sum_{j=1}^n Y_j(s) e^{\boldsymbol{\beta}'\mathbf{X}_j + \boldsymbol{\alpha}'\mathbf{K}_j}} \right] dN_i(s) - \frac{c}{2} \boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha}$$

where $N_i(s) = I(U_i \leq s)\Delta_i$, $c \geq 0$ is the penalty parameter, and $Y_j(s) = I(U_j \geq s)$ is an indicator that subject j is at risk at time s . An important relationship between kernel regression and linear mixed models has been described for non-censored outcomes [26]; a similar relationship holds in the Cox model, as discussed in [20]. Therefore, solving the penalized log partial likelihood above is equivalent to fitting the random effects Cox model

$$\lambda(t; \mathbf{X}, \mathbf{Z}) = \lambda_0(t) \exp[\mathbf{X}\boldsymbol{\beta} + \mathbf{h}]$$

where $\mathbf{h} = (h_1, \dots, h_n)$ are random effects with mean 0 and variance $\tau\mathbf{K}$. Then, testing $H_0 : f(\mathbf{Z}) = \mathbf{K}\boldsymbol{\alpha} = 0$ is equivalent to testing $H_0 : \tau = 0$. This hypothesis can be tested using a variance-component score test. Since a score test only requires fitting the null model $\lambda(t; \mathbf{X}, \mathbf{Z}) = \lambda_0(t) \exp[\mathbf{X}\boldsymbol{\beta}]$, we do not need to estimate $f(\mathbf{Z})$, so the test is valid even if a poor kernel is chosen. However, choosing a kernel that accurately reflects the true relationship between the microbiota and survival time will provide higher power. Two factors determine how well the kernel reflects the true relationship: first, whether the abundance of the associated taxa matters (versus presence or absence), and second, whether the OTUs related to the outcome are clustered on a phylogenetic tree. For example, since the weighted UniFrac distance encodes both taxon abundance and phylogenetic information, a test based on the weighted UniFrac distance will have higher power when the true association is between the outcome and the abundance of a cluster of OTUs on a phylogenetic tree than when the true association is with the abundance of randomly selected OTUs (of similar frequencies) or with the presence or absence of a set of OTUs.

The variance-component score statistic is

$$Q = \hat{\mathbf{M}}' \mathbf{K} \hat{\mathbf{M}}$$

where $\hat{\mathbf{M}} = (\hat{M}_1, \dots, \hat{M}_n)$ is the vector of estimated martingale residuals under the null model, i.e., $\hat{M}_i = \Delta_i - \int_0^\infty Y_i(t) e^{\boldsymbol{\beta}'\mathbf{X}_i} d\hat{\Lambda}_0(t)$ [19, 20]. Here, $\hat{\Lambda}_0(u) = \sum_{i=1}^n \Delta_i I(U_i \leq u) / \hat{S}_0(U_i)$ is Breslow's estimator of the

baseline hazard function $\Lambda_0(u) = \int_0^u \lambda_0(t)dt$ under the null model and $\hat{S}_0(t) = \sum_{i=1}^n Y_i(t)e^{\hat{\beta}'X_i}$ is the estimator for the baseline survival function.

Q asymptotically follows a mixture of chi-square distributions under the null model. The distribution has been derived for a linear kernel [27] but can be written in general form: by the central limit theorem,

$$\mathbf{K}^{1/2}\mathbf{M} \sim \mathcal{N}\left(0, P_0^{1/2}\mathbf{K}P_0^{1/2}\right)$$

where $P_0 = V - VX(X'VX)^{-1}X'V$ with $V = \text{diag}\left(\int_0^\infty Y_i(t)e^{\hat{\beta}'X_i}d\hat{\Lambda}_0(t) - w_i(\beta, t_i)^2\right)$ and $w_i(\beta, t) = e^{\hat{\beta}'X_i}/\hat{S}_0(t)$. Therefore,

$$Q \sim \sum_{i=1}^n \tilde{\lambda}_i \chi_{1,i}^2$$

where $(\tilde{\lambda}_1, \dots, \tilde{\lambda}_n)$ are the eigenvalues of $P_0^{1/2}\mathbf{K}P_0^{1/2}$ and $\chi_{1,i}^2$ are independent χ_1^2 random variables. Note that \mathbf{K} need not be full rank for this distribution to hold, since $\lambda_i = 0$ for the terms associated with the singular components of \mathbf{K} , so those components of the distribution will have weight zero [28].

Thus far, we have assumed that there are no tied survival times. In practice, tied survival times are fairly common due to coarse time measurements resulting from specific visit schedules or study follow-up dates. We use the Efron approximation to accommodate tied survival times [29]. This approximation performs well even with relatively small sample sizes or a high proportion of ties [30].

Small sample correction

The test outlined above is highly conservative for modest sample sizes and complicated kernels, such as kernels commonly used for the microbiota (see Additional file 1: Table S1). We therefore propose an approximate test using a modified score statistic that accounts for overdispersion. Analogous small-sample corrections have been proposed for quantitative and binary traits [31]. Specifically, we propose the modified score statistic

$$Q^* = \frac{\hat{M}'\mathbf{K}\hat{M}}{\hat{M}'\hat{M}}.$$

To derive the distribution of this statistic, we need the covariance of the residuals \hat{M} . We use a diagonal small-sample approximation to $\text{Cov}(\hat{M})$ that is motivated by the corresponding weighted linear model at convergence. This approximation is justified both in existing literature (e.g., [32, 33]) and through empirical evidence, namely the rapid convergence of the iteratively reweighted least squares (IRLS) algorithm using this weight matrix to the correct coefficients and the proper empirical type 1 error of our method.

Specifically, the fitted kernel machine Cox model (Eq. 1) is equivalent to a weighted linear model at convergence with weight matrix estimated using an iteratively reweighted least squares (IRLS) algorithm, as described in Appendix 1. We use a diagonal approximation for both the covariance of the residuals \hat{M} and the weight matrix W for IRLS. Several versions of W have been used for weighted partial least squares in the literature (e.g., [34] and [32]); we use an intermediate version that is diagonal as in [32] and [33] but whose elements are defined by the negative Hessian with respect to β as in [34].

To express this mathematically, let $\tilde{y} = X\beta + W^{-1}\hat{M}$ be the working response. Again, although the covariance matrix of the residuals \hat{M} is non-diagonal, we approximate $\text{Cov}(\hat{M})$ using a diagonal form proportional to the weight matrix W . Then, by defining $\tilde{y}^* = W^{1/2}\tilde{y}$, $X^* = W^{1/2}X$, and $\epsilon^* = W^{1/2}\epsilon$, the weighted linear model can be written as

$$\tilde{y}^* = X^*\beta + \epsilon^*, \epsilon^* \sim \mathcal{N}(0, \sigma^2\mathbf{I})$$

with projection matrix $P_0^* = I - X^*(X^{*'}X^*)^{-1}X^{*'}$. This model can be reframed using $\tilde{y}^* = W^{1/2}\tilde{y}$, $X^* = W^{1/2}X$, and $\epsilon^* = W^{1/2}\epsilon$ and fit as a weighted linear model. Then, at convergence, $\text{Var}(\epsilon^*) = W^{1/2}\text{Var}(\epsilon)W^{1/2} = W^{1/2}\sigma^2W^{-1}W^{1/2} = \sigma^2\mathbf{I}$.

Based on this, the distribution of Q^* satisfies

$$\begin{aligned} P(Q^* > q) &= P(\hat{M}'\mathbf{K}\hat{M} - \hat{M}'q\hat{M} > 0) \\ &= P((P_0^*\mathbf{M})'(\mathbf{K} - q\mathbf{I})(P_0^*\mathbf{M}) > 0) \\ &= P(\epsilon'P_0^{1/2}P_0^*(\mathbf{K} - q\mathbf{I})P_0^*P_0^{1/2}\epsilon > 0) \quad (2) \end{aligned}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. For the second equality, it is fairly straightforward to show that $\hat{M} = P_0^*\mathbf{M}$; this derivation is included in Appendix 2. The third equality uses the distribution of \mathbf{M} stated in the previous section. Then, under the null hypothesis,

$$Q^* \sim \sum_{i=1}^n \lambda_i^* \chi_{1,i}^2$$

where $(\lambda_1^*, \dots, \lambda_n^*)$ are the eigenvalues of $P_0^{1/2}P_0^*(\mathbf{K} - q\mathbf{I})P_0^*P_0^{1/2}$ and, as before, $\chi_{1,i}^2$ are independent χ_1^2 random variables. p values can be calculated efficiently using Davies' exact method [35]. For very small samples (e.g., $n \leq 50$), Davies p values may be anticonservative and permutation p values may be used instead.

Simulation scenarios

We carried out simulation studies in a range of settings to confirm that MiRKAT-S properly controls type I error and to assess its power using a variety of kernels. Microbiome OTU counts were generated using the

same approach as [14]. Specifically, for each individual, we simulated OTU counts from a Dirichlet-multinomial distribution with dispersion parameters and proportions estimated from Charlson et al.'s real upper respiratory tract microbiome dataset, in which 856 OTUs were measured on each of 60 individuals [7]. The data for each simulated individual consists of 1000 total OTU counts distributed among the 856 OTUs of Charlson et al. We also simulated two covariates for each individual, X_{1i} and X_{2i} , from a standard normal and a Bernoulli (0.5) distribution independently of taxonomic profiles. We considered sample sizes ranging from $n = 25$ to $n = 500$ individuals. For all simulation scenarios, we generated datasets with approximately 25 and 50% censoring. Four simulation settings were considered, varying (1) whether OTU abundance or the presence/absence of OTUs was associated with the outcome and (2) whether phylogenetically clustered or unclustered OTUs were associated with the outcome.

In setting 1, the abundances of OTUs in one cluster on a phylogenetic tree were associated with survival time. We partitioned all of the OTUs into 20 clusters using the partitioning-around-medoids algorithm based on the cophenetic distances of OTUs in the phylogenetic tree. The abundance of clusters ranged from 0.05 to 19.7% of all OTU reads. We selected an abundant cluster, containing 19.7% of all reads, and a rare cluster, containing 0.9% of all reads, to be associated with exponentially distributed survival times through the model

$$T_i = \frac{-\log(U_i)}{\lambda \exp\left(X'_i \beta + \gamma \text{scale}\left(\sum_{j \in \mathcal{A}} Z_{ij}\right)\right)} \quad (3)$$

where γ is the true effect size for the cluster, λ is the scale parameter, $U_i \sim \text{Uniform}(0, 1)$, \mathcal{A} is the set of indices of OTUs in the selected cluster, and the "scale" function standardizes the total OTU abundance in the cluster to have mean 0 and standard deviation 1:

$$\text{scale}\left(\sum_{j \in \mathcal{A}} Z_{ij}\right) = \frac{\sum_{j \in \mathcal{A}} Z_{ij} - \frac{1}{n} \sum_i \left(\sum_{j \in \mathcal{A}} Z_{ij}\right)}{SD_i\left(\sum_{j \in \mathcal{A}} Z_{ij}\right)}.$$

Censoring times are simulated independently as $C_i \sim \text{Exp}(\mu)$, and λ and μ are selected to give approximately 25% or approximately 50% censoring.

In setting 2, the ten most abundant OTUs overall, accounting for 31.5% of all OTU reads, were associated with survival time regardless of cluster membership. In this setting, we simulated survival times as

$$T_i = \frac{-\log(U_i)}{\lambda \exp\left(X'_i \beta + \gamma \text{scale}\left(\sum_{j \in \mathcal{A}} \frac{Z_{i(j)}}{\bar{Z}_{(j)}}\right)\right)} \quad (4)$$

where $\bar{Z}_{(j)}$ is the average across samples of the counts for the j th OTU. This limits the ability of a single OTU

to dominate the communal effect of the microbiota. Setting 2 is comparable to setting 1 when the associated cluster is common, since in both cases the abundance of common OTUs is associated with survival times, but it lacks setting 1's close phylogenetic relationship between associated OTUs.

In setting 3, the presence or absence of each OTU in a cluster was associated with survival time. OTUs were clustered as in setting 1, but in this case, were associated with survival time via the model

$$T_i = \frac{-\log(U_i)}{\lambda \exp\left(X'_i \beta + \gamma \text{scale}\left(\sum_{j \in \mathcal{A}} I(Z_{ij} > 0)\right)\right)} \quad (5)$$

As in setting 1, we simulated situations where an abundant cluster, containing 19.7% of all reads, was associated with the outcome and where a rare cluster, containing 0.9% of all reads, was associated with the outcome.

Finally, in setting 4, the presence or absence of 40 randomly selected OTUs was associated with the survival time. This mimics the size of an average cluster, since the mean number of OTUs assigned to a cluster was 42.8, with cluster sizes ranging from 3 to 118 OTUs. Since the majority of OTUs are rare, the overall number of OTU reads associated with the outcome is low in this setting. The model for T_i was the same as in setting 3. Setting 4 is comparable to setting 3 when the associated cluster is rare: in both cases, the presence or absence of rare OTUs is associated with survival times. However, it lacks setting 3's close phylogenetic relationship between associated OTUs.

In all simulation settings, we considered the weighted (K_w) and unweighted (K_u) UniFrac kernels, the Bray-Curtis kernel (K_{BC}), and the generalized UniFrac kernel with $\alpha = 0.5$ ($K_{0.5}$). These kernels are expected to have high power in different simulation settings. All of the UniFrac kernels take phylogenetic information into account. The unweighted UniFrac kernel does not account for OTU abundance, whereas the weighted UniFrac kernel does; the generalized UniFrac kernel is intermediate between the weighted and unweighted. The Bray-Curtis kernel does not account for phylogenetic structure or overall abundance of an OTU but does compare both presence/absence and relative abundance between samples of each OTU. Each kernel will have highest power when its measure of distance (and therefore similarity) accurately reflects the true relationship between the microbiome and the outcome.

For each simulation setting, sample size n , and censoring proportion, and using each kernel, we applied the test described above to test for associations between OTU counts and survival time. We used 5000 simulations with $\gamma = 0$ to estimate the empirical type I error rate with a nominal significance level of 0.05 and estimated empirical power across a range of γ values using 1000 simulations.

We also compare MiRKAT-S to two alternative approaches sometimes used for community-level analysis. First, we performed OTU-level tests of all OTUs. For each of the 856 OTUs in the dataset, we ran a marginal Cox regression model. The minimum p value from the 856 marginal models was compared to the null distribution to produce an overall p value for any association of the microbiota with survival times. In practice, the null distribution would be generated for an individual study using permutation; however, in the interest of computational efficiency, we generated this distribution using the minimum p values from 5000 simulations where survival times were not associated with the microbiota. Second, we performed principal coordinates analysis (PCoA) on a relevant distance matrix (see, e.g., [16]). Since it is not clear how to make PCoA plots with censored time-to-event outcomes, we followed PCoA by Cox proportional hazards regression. Specifically, we generated the UniFrac and Bray-Curtis distance matrices as above, then took the top two principal coordinates as our covariates for a Cox regression analysis. We tested the two principal coordinates jointly by using a chi-square test to compare nested models with and without the microbiota-related predictors. The covariates X_1 and X_2 were included in all models exactly as in the MiRKAT-S simulations.

Results

Power and type I error in simulated datasets

Empirical type I error rates with 25% censoring are reported in Table 1. Note that Eqs. 3, 4, and 5 are identical when $\gamma = 0$ (i.e., there is no true association between the microbiota and survival time), so settings 1–4 all have the same type I error. From the table, we see that MiRKAT-S is valid for all kernels and sample sizes of at least 100 individuals. For comparison, empirical estimates of type I error without the small-sample correction are reported in (Additional file 1: Table S1), demonstrating the uncorrected test is highly conservative. To further describe the behavior of p values based on the uncorrected test statistic, we have generated Q-Q plots (see Additional file 1: Figure S1). These plots demonstrate that p values based on the corrected statistic do not deviate significantly from the theoretical distribution, whereas p values based on the

uncorrected statistic are far from the theoretical distribution. They also show that, more specifically, p values based on the uncorrected statistic tend to be less extreme than they should be: p values that are truly smaller than 0.2 tend to be overestimated, whereas p values larger than 0.2 tend to be underestimated. All type I error results are similar with 50% censoring (see Additional file 1: Table S2). For sample sizes smaller than 100, the size of MiRKAT-S is close to correct, though it may be slightly anticonservative. Empirical type I errors for small sample sizes are reported in Table 2. If the sample size is smaller than $n = 50$ and the p value from MiRKAT-S is borderline, it may be preferable to report p values obtained using permutation.

To interpret the simulation results evaluating the power of the test, recall that two aspects of the relationship between the microbiome and survival are important for understanding which kernel will provide the highest power: the relationship between associated OTUs (whether or not they cluster on a phylogenetic tree) and the importance of taxon abundance (whether OTU count or presence/absence matters). All of the UniFrac distances account for phylogeny, while the Bray-Curtis dissimilarity does not. The weighted UniFrac distance and Bray-Curtis dissimilarity both utilize taxon abundance (OTU counts), whereas the unweighted UniFrac distance only incorporates presence/absence of taxa, and the generalized UniFrac distance is a compromise between the weighted and unweighted UniFrac distances.

Figure 1 shows the estimated power under all simulation settings. As expected, in all settings, power increases with increasing true effect size (γ). We first consider settings 1 and 3, in which a cluster of OTUs is associated with the outcome. When the OTU counts of an abundant cluster are associated with survival times (panel A), the weighted UniFrac kernel and the generalized UniFrac kernel with $\alpha = 0.5$ provide the highest power, since the corresponding distance metrics take both abundance and phylogeny into consideration. Since the associated cluster is common, nearly all individuals have at least one read for each

Table 1 Empirical type I errors for $n = 100, 200,$ or 500

Number	K_w	K_u	$K_{0.5}$	K_{BC}
100	0.0544	0.0540	0.0530	0.0542
200	0.0494	0.0480	0.0470	0.0462
500	0.0506	0.0478	0.0536	0.0442

Empirical type I errors for sample sizes $n = 100, 200,$ and 500 with approximately 25% censoring. Results are based on 5000 simulated datasets. $K_w, K_u, K_{BC},$ and $K_{0.5}$ represent results for the weighted UniFrac kernel, unweighted UniFrac kernel, Bray-Curtis kernel, and generalized UniFrac kernel with $\alpha = 0.5,$ respectively

Table 2 Empirical type I errors for $n < 100$

Number	Method	K_w	K_u	$K_{0.5}$	K_{BC}
25	MiRKAT-S	0.054	0.055	0.055	0.056
	Permutation	0.046	0.048	0.046	0.049
50	MiRKAT-S	0.045	0.058	0.051	0.051
	Permutation	0.041	0.052	0.045	0.045
75	MiRKAT-S	0.054	0.058	0.051	0.053
	Permutation	0.051	0.053	0.048	0.049

Empirical type I errors for small sample sizes ($n < 100$) with approximately 25% censoring. Results are based on 5000 simulated datasets, and permutation p values were obtained using 1000 permutations. $K_w, K_u, K_{0.5},$ and K_{BC} represent results for the weighted UniFrac kernel, unweighted UniFrac kernel, Bray-Curtis kernel, and generalized UniFrac kernel with $\alpha = 0.5,$ respectively

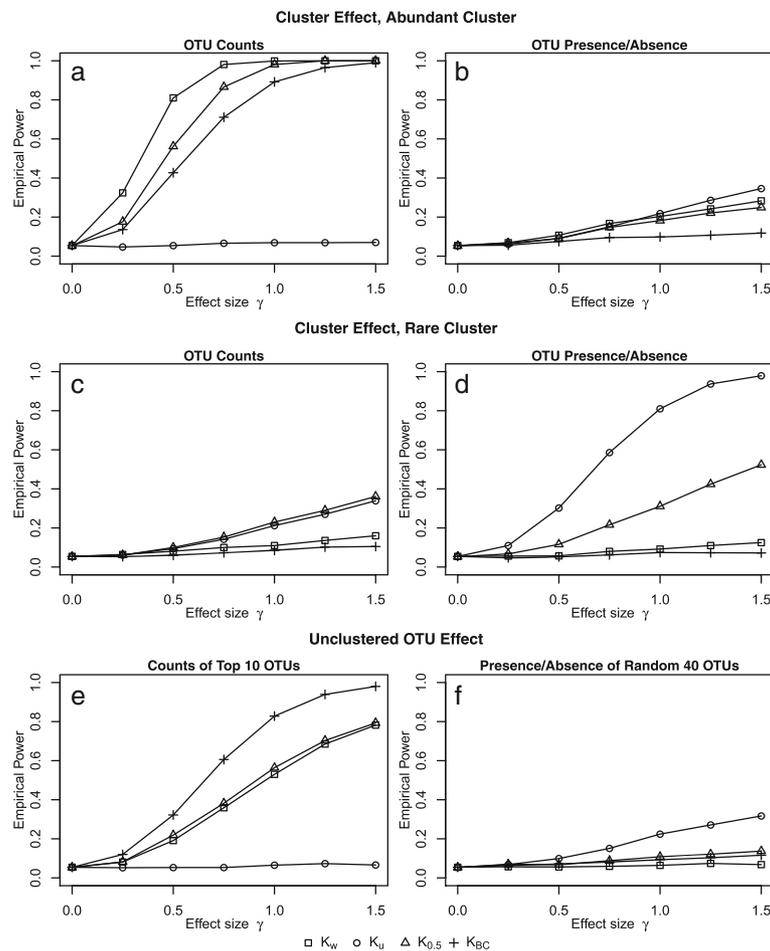


Fig. 1 Empirical power. Empirical power was evaluated in all simulation settings, using a sample size of $n = 100$ and 25% censoring. K_w , K_u , K_{BC} , and $K_{0.5}$ represent results for the weighted UniFrac kernel, unweighted UniFrac kernel, Bray-Curtis kernel, and generalized UniFrac kernel with $\alpha = 0.5$, respectively. γ is the true effect size for the associated cluster or OTUs. **a** Setting 1; survival is associated with OTU counts in a common cluster containing 19.7% of all reads. **b** Setting 3; survival is associated with the presence or absence of each taxon in a common cluster. **c** Setting 1; survival is associated with OTU counts in a rare cluster containing 0.9% of all reads. **d** Setting 3; survival is associated with the presence or absence of each taxon in a rare cluster. **e** Setting 2; survival is associated with the ten most common OTUs, regardless of cluster membership. **f** Setting 4; survival is associated with 40 OTUs selected at random, regardless of cluster membership

OTU in the cluster, so individuals cannot be distinguished based on OTU presence/absence. Therefore, in this setting, the unweighted UniFrac kernel has almost no power to detect the association. In contrast, when OTU presence/absence in a rare cluster is associated with survival time (panel D), the unweighted UniFrac kernel has highest power, since this distance metric is based on the presence and absence of OTUs. Here, the weighted UniFrac kernel has very low power because OTU counts of a rare cluster do not vary much between individuals. When the OTU counts of a rare cluster are associated with survival time (panel C) or when the presence or absence of an abundant cluster is associated with survival time (panel B), the OTU effect is small and similar across most individuals. Therefore, in these settings, no kernel provided power comparable to settings with larger effect sizes.

The power under settings 2 and 4, in which unclustered OTUs are associated with the outcome, is reported in panels e and f of Fig. 1. When the OTU counts of the ten most common OTUs are associated with survival time (panel e), the Bray-Curtis kernel has highest power, followed by the weighted UniFrac kernel and generalized UniFrac kernel with $\alpha = 0.5$. Since the Bray-Curtis dissimilarity metric does not incorporate phylogenetic information, this distance is designed for unclustered rather than clustered OTUs. However, since it takes abundance into account, the Bray-Curtis kernel performs better when OTU counts are associated with survival (e.g., panel a) rather than OTU presence/absence (e.g., panel d) and when the associated cluster is at least moderately abundant. When the presence or absence of a random 40 OTUs were associated with survival time (panel f), no kernel had high power, but

the unweighted UniFrac kernel had higher power than any other tested kernel. The low power is likely due to the rarity of most randomly selected clusters and inability to gain power by utilizing phylogenetic information.

Kernel choice has a strong effect on the power of the test, and different kernels are optimal depending on the nature of the true relationship between the microbiota and survival time. In practice, a kernel representing relationships of particular scientific interest could be selected. For example, if a healthy microbiome at a certain body site has relatively few dominant taxa at high frequencies, and changes in the relative abundance of these taxa is hypothesized to be associated with the time to a disease outcome or death, choosing a kernel that accounts for taxon abundance will have the highest power to detect the hypothesized changes. If there is no specific hypothesized relationship, multiple kernels can be tested and then the resulting p values adjusted for multiple comparisons. Testing the four kernels discussed here is a reasonable starting point, and the limited number of tests reduces the power loss due to adjusting for multiple comparisons. In these simulations, if the analysis is performed using all four kernels included in Fig. 1 ($K_U, K_{0.5}, K_W, K_{BC}$) and then the minimum p value after an FDR adjustment is used for testing, the power does not quite match the best kernel but is comparable to or better than the remaining three kernels (see Additional file 1: Figure S2).

We also compared the power of MiRKAT-S to two approaches used in current practice: performing a marginal, or OTU-level, analysis for all OTUs and including the top principal coordinates of the distance matrix as the covariates of interest in a regression model (see Additional file 1: Figure S2). We find that for most simulation settings, MiRKAT-S has substantially better power than the marginal analysis or PCoA, and in the remainder, the power is comparable between methods. In particular, the marginal analysis has power to detect an association between counts of OTUs in a cluster and survival times, but virtually no power to detect associations between presence/absence of clustered OTUs and survival or associations involving unclustered OTUs. The power of the marginal test for detecting an association with clustered OTU counts has similar power regardless of how common the cluster is. Therefore, since MiRKAT-S is more powerful for OTU counts of relatively common clusters, the marginal analysis and MiRKAT-S have similar power for rare clusters, and for very large effect sizes in this simulation setting, the marginal analysis outperforms MiRKAT-S slightly (Additional file 1: Figure S2 (panel C)). However, in all other cases, MiRKAT-S is substantially more powerful than the OTU-level analysis. PCoA regression analysis performs similarly to MiRKAT-S for each kernel when OTU counts in a common cluster are associated with survival times (Additional file 1: Figure S2 (panel A)). In most

other simulation settings, PCoA matches or approaches the power of MiRKAT-S in only the case of the best kernel. That is, MiRKAT-S is more robust to kernel choice than PCoA. In addition, in settings where clustering information does not matter (Additional file 1: Figure S2 (panels e and f)), PCoA has very low power, while MiRKAT-S has moderate power provided that the associated OTUs are not too rare.

Analysis of blood and bone marrow transplant data

Acute graft-versus-host disease (aGVHD) is a leading cause of death after allogeneic blood or bone marrow transplantation. There is a suspected relationship between the intestinal microbiome and aGVHD, but previous studies in mice and humans have yielded mixed results about the presence and nature of this relationship. Therefore, Jenq et al. recently studied the association of a particular bacterial species (intestinal *Blautia*) and of intestinal microbiome diversity indices with time to each of aGVHD onset, aGVHD-related mortality, and adverse outcomes unrelated to aGVHD [5].

In the original study, subjects were stratified into two cohorts depending on sequencing platform. The combined dataset used here results from resequencing of the first cohort of patients using the Illumina MiSeq platform; unfortunately, four patients did not have additional DNA available for MiSeq sequencing and were excluded from our analysis. Therefore, 481 stool samples were available for 111 unique subjects, and for each sample, the Illumina MiSeq platform was used to sequence the V4–V5 regions of the 16S rRNA gene. OTUs were generated as described in [5]. Briefly, mothur version 1.34 was used to compile and process sequence data [36], and quality filters were applied as in [37]. This procedure yielded OTU counts for 2436 OTUs. As in the original paper, for each subject, we only included the sample collected closest to 12 days post-transplant in our analysis, and we excluded subjects for whom no samples were collected between 8 and 16 days post-transplant, so that 94 subjects were included. We used QIIME with default settings to align the sequences and generate a rooted phylogenetic tree. The 109 OTUs that failed to be placed on the tree were excluded from our analysis, leaving 2327 OTUs. We performed the test using the unweighted and weighted UniFrac kernels, the generalized UniFrac kernel with $\alpha = 0.5$, and the Bray-Curtis kernel, adjusting for age and gender. The outcomes considered were overall survival and time to stage III aGVHD.

The results of applying MiRKAT-S to these data with and without the small-sample correction are reported in Table 3. We do not find a significant association between the microbiota and time until development of stage III aGVHD using any kernel. However, the association between overall survival and the microbiota is

Table 3 Analysis of gut microbiome after allogeneic transplant

Outcome	Method	K_u	$K_{0.5}$	K_w	K_{BC}
Overall Survival	Uncorrected	0.049	0.008	0.065	0.029
	Corrected	0.046	0.007	0.065	0.022
Grade III aGVHD	Uncorrected	0.496	0.514	0.472	0.849
	Corrected	0.560	0.575	0.518	0.933

p values from MiRKAT-S using the weighted (K_w) and unweighted (K_u) UniFrac kernels, the generalized UniFrac kernel with $\alpha = 0.5$ ($K_{0.5}$), and the Bray-Curtis kernel (K_{BC}) with outcomes of overall survival and severe (at least grade 3) graft-versus-host disease. "Corrected" indicates the p values are based on the modified score statistic with proper type I error; "uncorrected" indicates the p values are based on the original score statistic

significant at $\alpha = 0.05$ using the unweighted UniFrac kernel K_u , generalized UniFrac kernel $K_{0.5}$, and Bray-Curtis kernel K_{BC} , but not using the weighted UniFrac kernel K_w (Table 3). The association remains significant after we adjust for multiple comparisons (multiple kernels) using either the false discovery rate method or the Bonferroni correction. The differences between the corrected and uncorrected p values are fairly small. However, they are in the direction we would expect based on simulation results. In particular, we saw that low and high p values are less frequent than would be expected for a null distribution of p values (see Additional file 1: Figure S1). This is consistent with seeing higher p values for the uncorrected statistic in the overall survival case, where the p values based on the corrected statistic are fairly small, and conversely, seeing lower p values for the uncorrected statistic in the grade III aGVHD case, where the p values based on the corrected statistic are large.

To visualize this association, we clustered individuals using Ward's agglomerative hierarchical clustering method [38] based on the generalized UniFrac distance with $\alpha = 0.5$. Ward's method is a generic clustering method that can be used for many data types. Generally speaking, the goal is to divide samples into clusters (groups) that tend to be similar in the ways that we care about; here, clusters should reflect similarity of taxonomic profiles. Operationally, Ward's method begins by assigning each sample to its own cluster and sequentially merges pairs of clusters that are most similar into larger clusters until all samples are merged into a single cluster. Which clusters to merge is decided by minimizing the increase in the sum of within-cluster squared distances (when Euclidean distances are used, this is the within-cluster variance). Through this process, a hierarchical tree is created. The tree can be cut at different levels to create the desired number of final clusters used for analysis. Although Euclidean distances are often used for Ward's method, other squared distances (in this case, ecologically relevant metrics such as the UniFrac distances) can be substituted while still using the same form of criterion

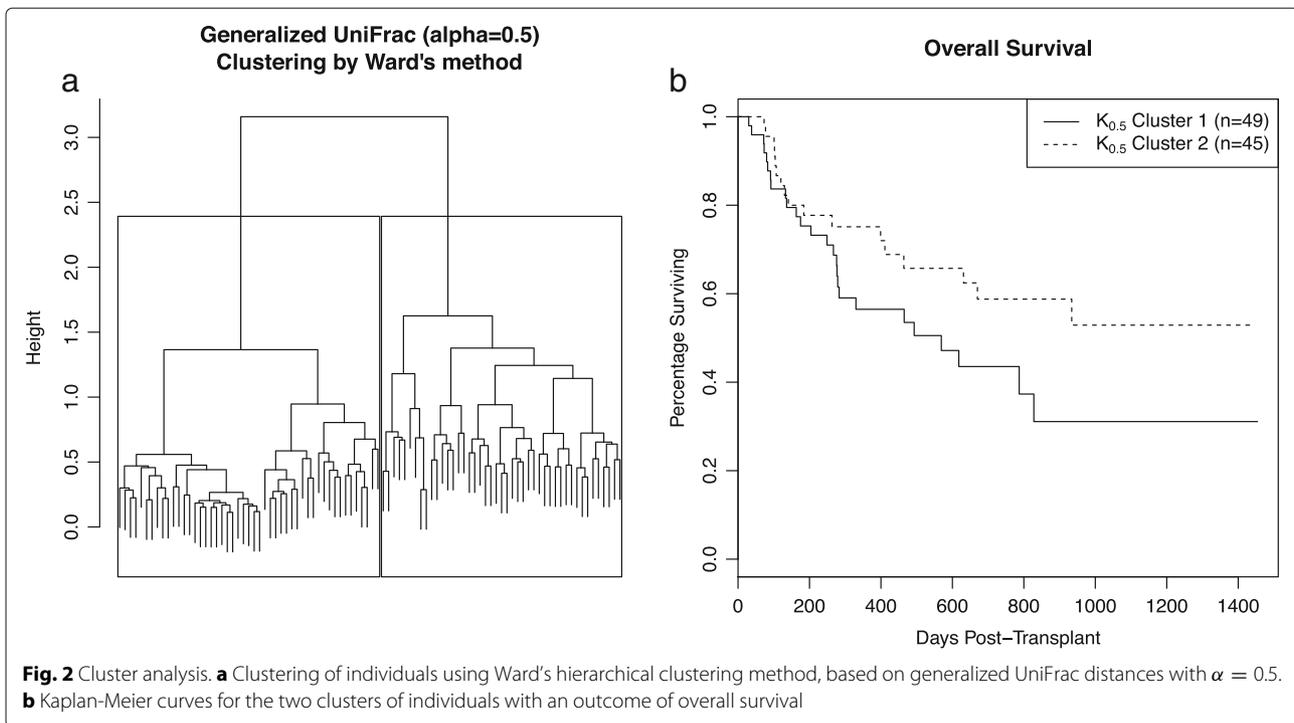
and algorithm [39]. For our analysis, we used the generalized UniFrac distance to measure dissimilarity between individuals to ensure that clusters are similar with regard to the presence and abundance of taxa, accounting for phylogenetic relationships. We chose to cut the tree to create two clusters; a clear separation into clusters of sizes $n = 45$ and $n = 49$ can be seen in Fig. 2a.

Kaplan-Meier curves for overall survival in the two clusters are shown in Fig. 2b. However, the simple Cox regression p value is not significant ($p = 0.09$). That is, the similarity between individuals was measured the same way in both analyses. MiRKAT-S yielded a highly significant p value for the association of the microbiota with overall survival, whereas the analysis based on clustering individuals gave a non-significant p value. This demonstrates that MiRKAT-S has higher power to detect this association than a simple clustering analysis based on the same distance metric.

The highly significant result using MiRKAT-S with $K_{0.5}$ may also provide information about the form of the association between the gut microbiota and survival post-transplant. The generalized UniFrac kernel incorporates phylogenetic information and represents a compromise between abundance and presence or absence of OTUs. Therefore, this kernel has highest power to detect relationships between taxonomic profiles and overall survival that occur through moderately rare clusters of OTUs or through a combination of common and rare clusters of OTUs. Accordingly, the high significance of MiRKAT-S using $K_{0.5}$ may indicate that one of those settings holds: either moderately rare clusters of OTUs are driving the relationship between the microbiota and overall survival, or multiple clusters of OTUs, some of which are abundant and some of which are rare, are associated with overall survival. However, without further analysis, we cannot determine which OTUs or clusters are associated with survival times in aGVHD patients.

Discussion

We propose MiRKAT-S for testing the association between the human microbiota and survival outcomes. In the kernel machine Cox model framework, taxonomic profiles are modeled through a kernel function. This allows comparison of microbial community profiles using microbiome-specific distance metrics such as the UniFrac distances or Bray-Curtis dissimilarity. The kernel machine regression framework also allows linear (or, more generally, parametric) adjustment for covariates and potential confounders. We test the significance of the association between the microbiota and survival times using a variance-component score test, and we develop a small-sample correction to account for the modest sample sizes and sparse, high-dimensional data that often result from microbiome studies. In contrast to existing methods that



use resampling, p values are obtained analytically using the Davies approximation.

Like other distance-based analyses, MiRKAT-S is limited to detecting the presence of an association between the microbiota and survival times. It cannot identify individual taxa that are associated with the outcome and does not provide information about relationships among taxa within a microbial community. MiRKAT-S is therefore designed to be used when the question of interest is whether an entire microbial community is associated with the outcome. Alternative ways to answer this question include testing the association of each OTU individually with the outcome of interest or using a dimension reduction technique such as PCoA and testing the top few principal coordinates. Our simulation studies show that MiRKAT-S has power at least comparable to, and often substantially greater than, either of these methods for community-level association testing. Community-level tests can be used in combination with other methods that identify taxa of interest. These include marginal tests for particular OTUs of interest, identification of OTUs with high loadings from PCA or PCoA, or penalized regression methods that account for the structure and compositional nature of the data.

Our simulation results show that MiRKAT-S correctly controls type I error. However, under conditions of extreme censoring or very small sample sizes, the analytic p values provided by MiRKAT-S may be slightly anticonservative. In these cases, obtaining p values by

permutation may be preferable. Type I error is accurate regardless of the choice of kernel; that is, the test is valid even when a poor choice of kernel is made. The power of the test depends heavily on how well the selected kernel encodes the true relationship between the microbiome and the outcome of interest. For example, when the abundance of an OTU or set of OTUs is related to the outcome, a kernel that encodes abundance information, such as the weighted UniFrac or Bray-Curtis kernel, will have higher power than a kernel that encodes only taxon presence or absence. There are situations in which MiRKAT-S has low power regardless of kernel choice, but any method would have low power in those settings because the true effect size is very small. For example, if the presence or absence of a common OTU is associated with the outcome, nearly all subjects will have the OTU present in the sample, so the association will be very difficult to detect using any method.

If there is no a priori hypothesis about which kernel will best represent relationships of scientific interest, the analysis can be performed using multiple kernels and an overall p value can be obtained by permutation or adjustment for multiple comparisons. This analysis approach can provide information not only about the presence of a relationship but also about its form, depending on the distance metrics considered and their relative power for different forms of the true association. That is, if the metric with the lowest p value has the highest power to detect associations with abundance of common clusters,

that may be the form of the true association. Furthermore, weighted combinations of kernels could be used to simultaneously detect different types of shifts in the microbiota. Specific combinations or kernel weights could either be selected a priori or via a grid search, again using permutation to test overall significance. As the field of microbiome analysis matures and new distance metrics are proposed, our approach will continue to increase in power.

Conclusions

We present MiRKAT-S, a method for testing the association between the microbiota, assessed on the community level, and survival (time-to-event) outcomes. Similar methods exist for binary and continuous outcomes; however, MiRKAT-S is the first community-level test for microbiome data that allows analysis of censored survival outcomes. Community-level analyses have several benefits: they often provide higher statistical power to detect associations, and they allow investigators to address additional scientific questions, such as whether the entire microbiome is collectively associated with survival time or time to development of a disease. We use the kernel machine regression framework, encoding microbiome data in ecologically relevant kernels. With judicious choice of kernels, the test can detect a wide range of true forms of association, including association of the outcome with OTU presence/absence or abundance and with either phylogenetically clustered or unclustered sets of taxa. Therefore, MiRKAT-S facilitates a robust community-level analysis of the association between the microbiota and censored survival outcomes that is not possible using existing methods.

Appendix 1

An iteratively reweighted least squares (IRLS) algorithm can be used to fit the linear model at convergence that is equivalent to the Cox PH model of interest. At the k th step of the IRLS algorithm, we solve

$$\tilde{y}^k = X\beta + \epsilon, \quad \epsilon \sim \mathcal{N}\left(0, \sigma^2(\tilde{W}^k)^{-1}\right)$$

with weight matrix

$$\tilde{W}^k = \text{diag}\left(\int_0^\infty I(T_i \geq t) e^{X_i^T \tilde{\beta}^k} d\hat{\Lambda}_0(t) - \int_0^\infty I(T_i \geq t) w_i(\tilde{\beta}^k, t) e^{X_i^T \tilde{\beta}^k} d\hat{\Lambda}_0(t)\right)$$

where $w_i(t) = \frac{e^{X_i^T \beta}}{\sum_{l=1}^n Y_l(t) e^{X_l^T \beta}}$ and working response

$$\tilde{y} = X\tilde{\beta}^{k-1} + \left(\tilde{W}^{k-1}\right)^{-1} \hat{M}^{k-1}.$$

The corresponding quantities without the superscript k refer to the model at convergence. Then, the modified score statistic is equivalent to

$$Q^* = \frac{(\tilde{y} - X\tilde{\beta})' \tilde{W} K \tilde{W} (\tilde{y} - X\tilde{\beta})}{\hat{\sigma}^2}$$

which is analogous to the linear and logistic cases considered in [31]. Multiplying both sides of the equation by $\tilde{W}^{1/2}$ and defining $\tilde{y}^* = \tilde{W}^{1/2} \tilde{y}$, $X^* = \tilde{W}^{1/2} X$, and $\epsilon^* = \tilde{W}^{1/2} \epsilon$, the model can be expressed as

$$\tilde{y}^* = X^* \beta + \epsilon^*, \quad \epsilon^* \sim \mathcal{N}(0, \sigma^2 I)$$

with projection matrix $P_0^* = I - X^* (X^{*'} X^*)^{-1} X^{*}$.

Appendix 2

To derive the relationship between \hat{M} and M , recall that

$$\begin{aligned} \hat{z} &= X\hat{\beta} + W^{-1}\hat{M} \\ z &= X\beta + W^{-1}M \\ \hat{\beta} &= (X'WX)^{-1}X'Wz \end{aligned}$$

Then, solving the first equation for \hat{M} gives

$$\begin{aligned} \hat{M} &= W(z - X\hat{\beta}) = W[X\beta + W^{-1}M - X\hat{\beta}] \\ &= W[I - X(X'WX)^{-1}X'W]z \\ &= W[I - X(X'WX)^{-1}X'W](X\beta + W^{-1}M) \\ &= [W^{-1}M - WX(X'WX)^{-1}X'W^{-1}M] \\ &= [I - WX(X'WX)^{-1}X']M \\ &= [I - X^*(X^{*'}X^*)^{-1}X^{*}]M = P_0^*M \end{aligned}$$

so that $\hat{M} = P_0^*M$, as claimed.

Additional file

Additional file 1: PDF file includes supplemental tables (Tables S1–S3) and figures (Figures S1–S2). (PDF 933 kb)

Acknowledgements

Not applicable.

Funding

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1256082 (AP). This research was additionally supported by the National Institutes of Health (R01-HL124112, RRJ; U10-CA180819, MCW; R01-HG007508, MCW) and The Hope Foundation (MCW).

Availability of data and materials

The data used in this article are available in the European Bioinformatics Institute Nucleotide Archive under accession number PRJEB15259 (<http://www.ebi.ac.uk/ena/data/view/PRJEB15259>). The R package MiRKAT-S is available at <http://research.fhcrc.org/wu/en/software.html> or on GitHub at <https://github.com/aplantin/MiRKATS> in formats appropriate for Macintosh or Windows systems. In addition, each of these sites includes a vignette demonstrating use of the package.

Authors' contributions

AP developed the method, analyzed the data, and wrote the manuscript. RRJ collected the data and contributed to the data analysis. JC and MCW conceived the study and critically reviewed the manuscript. XZ and NZ contributed to development and testing of the method. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable. Because this study only involved secondary analysis of an existing, de-identified dataset, it is not considered human subject research.

Author details

¹Department of Biostatistics, University of Washington, 1705 NE Pacific Street, Seattle, Washington, USA. ²Public Health Sciences Division, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, Seattle, Washington, USA.

³Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, 615 N Wolfe St, Baltimore, Maryland, USA. ⁴Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, 200 First Street SW, Rochester, Minnesota, USA.

⁵Departments of Genomic Medicine and Stem Cell Transplantation, Division of Cancer Medicine, The University of Texas MD Anderson Cancer Center, Unit 1954, 1515 Holcombe Blvd, Houston, TX, USA.

Received: 6 September 2016 Accepted: 31 January 2017

Published online: 08 February 2017

References

- Morgan XC, Huttenhower C. Human microbiome analysis. *PLoS Comput Biol*. 2012;8(12):1002808.
- Hamady M, Knight R. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Res*. 2009;19(7):1141–1152.
- Wooley JC, Godzik A, Friedberg I. A primer on metagenomics. *PLoS Comput Biol*. 2010;6(2):1000667.
- Stackebrandt E, Goebel B. Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *Int J Syst Evol Microbiol*. 1994;44(4):846–9.
- Jenq RR, Taur Y, Devlin SM, Ponce DM, Goldberg JD, Ahr KF, Littmann ER, Ling L, Gobourne AC, Miller LC, et al. Intestinal *Blautia* is associated with reduced death from graft-versus-host disease. *Biol Blood Marrow Transplants*. 2015;21(8):1373–1383.
- Fitz-Gibbon S, Tomida S, Chiu BH, Nguyen L, Du C, Liu M, Elashoff D, Erfe MC, Loncaric A, Kim J, et al. Propionibacterium acnes strain populations in the human skin microbiome associated with acne. *J Invest Dermatol*. 2013;133(9):2152–160.
- Charlson ES, Chen J, Custers-Allen R, Bittinger K, Li H, Sinha R, Hwang J, Bushman FD, Collman RG. Disordered microbial communities in the upper respiratory tract of cigarette smokers. *PLoS ONE*. 2010;5(12):15216.
- Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. *Nature Reviews Genetics*. 2012;13(4):260–70.
- Erb-Downward JR, Thompson DL, Han MK, Freeman CM, McCloskey L, Schmidt LA, Young VB, Toews GB, Curtis JL, Sundaram B, et al. Analysis of the lung microbiome in the “healthy” smoker and in COPD. *PLoS ONE*. 2011;6(2):16384.
- Giongo A, Gano KA, Crabb DB, Mukherjee N, Novelo LL, Casella G, Drew JC, Ilonen J, Knip M, Hyöty H, et al. Toward defining the autoimmune microbiome for type 1 diabetes. *The ISME J*. 2011;5(1):82–91.
- Morris A, Beck JM, Schloss PD, Campbell TB, Crothers K, Curtis JL, Flores SC, Fontenot AP, Ghedin E, Huang L, et al. Comparison of the respiratory microbiome in healthy nonsmokers and smokers. *Am J Respiratory Critical Care Med*. 2013;187(10):1067–1075.
- Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, et al. A core gut microbiome in obese and lean twins. *Nature*. 2009;457(7228):480–4.
- Karlsson FH, Tremaroli V, Nookaew I, Bergström G, Behre CJ, Fagerberg B, Nielsen J, Bäckhed F. Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature*. 2013;498(7452):99–103.
- Zhao N, Chen J, Carroll IM, Ringel-Kulka T, Epstein MP, Zhou H, Zhou JJ, Ringel Y, Li H, Wu MC. Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. *Am J Human Genet*. 2015;96(5):797–807.
- Chen J, Li H. Kernel Methods for Regression Analysis of Microbiome Compositional Data In: Hu M, Liu Y, Lin J, editors. *Topics in Applied Statistics*, vol 55. New York: Springer; 2013. p. 191–201.
- Han MK, Zhou Y, Murray S, Tayob N, Noth I, Lama VN, Moore BB, White ES, Flaherty KR, Huffnagle GB, et al. Lung microbiome and disease progression in idiopathic pulmonary fibrosis: an analysis of the COMET study. *Lancet Respir Med*. 2014;2(7):548–56.
- Penders J, Gerhold K, Stobberingh EE, Thijs C, Zimmermann K, Lau S, Hamelmann E. Establishment of the intestinal microbiota and its role for atopic dermatitis in early childhood. *J Allergy Clin Immunol*. 2013;132(3):601–7.
- Bisgaard H, Li N, Bonnelykke K, Chawes BLK, Skov T, Paludan-Müller G, Stokholm J, Smith B, Krogfelt KA. Reduced diversity of the intestinal microbiota during infancy is associated with increased risk of allergic disease at school age. *J Allergy Clin Immunol*. 2011;128(3):646–52.
- Lin X, Cai T, Wu MC, Zhou Q, Liu G, Christiani DC, Lin X. Kernel machine SNP-set analysis for censored survival outcomes in genome-wide association studies. *Genet Epidemiol*. 2011;35(7):620–31.
- Cai T, Tonini G, Lin X. Kernel machine approach to testing the significance of multiple genetic markers for risk prediction. *Biometrics*. 2011;67(3):975–86.
- Cox DR. Regression models and life tables (with discussion). *J Royal Stat Soc*. 1972;34:187–220.
- Kimeldorf GS, Wahba G. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Annals Math Stat*. 1970;41(2):495–502.
- Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol*. 2005;71(12):8228–235.
- Chen J, Bittinger K, Charlson ES, Hoffmann C, Lewis J, Wu GD, Collman RG, Bushman FD, Li H. Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics*. 2012;28(16):2106–113.
- Bray JR, Curtis JT. An ordination of the upland forest communities of southern Wisconsin. *Ecol Monogr*. 1957;27(4):325–49.
- Liu D, Lin X, Ghosh D. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics*. 2007;63(4):1079–1088.
- Chen H, Lumley T, Brody J, Heard-Costa NL, Fox CS, Cupples LA, Dupuis J. Sequence kernel association test for survival traits. *Genet Epidemiol*. 2014;38(3):191–7.
- Styan GP. Notes on the distribution of quadratic forms in singular normal variables. *Biometrika*. 1970;57(3):567–72.
- Efron B. The efficiency of Cox's likelihood function for censored data. *J Am Stat Assoc*. 1977;72(359):557–65.
- Hertz-Picciotto I, Rockhill B. Validity and efficiency of approximation methods for tied survival times in Cox regression. *Biometrics*. 1997;53(3):1151–1156.
- Chen J, Chen W, Zhao N, Wu MC, Schaid DJ. Small sample kernel association tests for human genetic and microbiome association studies. *Genet Epidemiol*. 2016;40(1):5–19.
- Nygård S, Borgan Ø, Lingjærde OC, Størvold HL. Partial least squares Cox regression for genome-wide data. *Lifetime Data Anal*. 2008;14(2):179–95.
- Park PJ, Tian L, Kohane IS. Linking gene expression data with patient survival times using partial least squares. *Bioinformatics*. 2002;18(suppl 1):120–7.
- Lambert-Lacroix S, Letué F, et al. Partial least squares and Cox model with application to gene expression. Technical report. 2011. http://sites.uclouvain.be/IAP-Stat-Phase-V-VI/PhaseVI/publications_2011/TR/Lambert-LacroixLetuelAP.pdf.
- Davies RB. The distribution of a linear combination of chi-2 random variables. *J R Stat Soc Ser C (Appl Stat)*. 1980;29(3):323–33.
- Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, et al. Introducing

mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol.* 2009;75(23):7537–541.

37. Schloss PD, Gevers D, Westcott SL. Reducing the effects of pcr amplification and sequencing artifacts on 16s rRNA-based studies. *PLoS ONE.* 2011;6(12):27310.
38. Ward Jr JH. Hierarchical grouping to optimize an objective function. *J Am Stat Assoc.* 1963;58(301):236–44.
39. Murtagh F, Legendre P. Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? *J Classif.* 2014;31(3):274–95.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

