Received: 30 April 2018 | Revised: 27 June 2018 | Accepted: 15 July 2018

DOI: 10.1002/gepi.22160

RESEARCH ARTICLE

WILEY Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

A small-sample kernel association test for correlated data with application to microbiome association studies

Xiang Zhan[1] | Lingzhou Xue[2] | Haotian Zheng[3] | Anna Plantinga[4] | Michael C. Wu[4,5] | Daniel J. Schaid[6] | Ni Zhao[7] | Jun Chen[6,8]

[1]Department of Public Health Sciences, Pennsylvania State University, Hershey, Pennsylvania

[2]Department of Statistics, Pennsylvania State University, University Park, Pennsylvania

[3]Department of Mathematical Sciences, Tsinghua University, Beijing, China

[4]Department of Biostatistics, University of Washington, Seattle, Washington

[5]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington

[6]Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, Minnesota

[7]Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland

[8]Center for Individualized Medicine, Mayo Clinic, Rochester, Minnesota

Correspondence
Ni Zhao, Department of Biostatistics, Johns Hopkins University, Baltimore 21202, MD.
Email: nzhao10@jhu.edu
Jun Chen, Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN.
Email: Chen.Jun2@mayo.edu

Funding information
US Public Health Service; National Institutes of Health, Grant/Award Number: GM065450; Mayo Clinic Center for Individualized Medicine

Abstract

Recent research has highlighted the importance of the human microbiome in many human disease and health conditions. Most current microbiome association analyses focus on unrelated samples; such methods are not appropriate for analysis of data collected from more advanced study designs such as longitudinal and pedigree studies, where outcomes can be correlated. Ignoring such correlations can sometimes lead to suboptimal results or even possibly biased conclusions. Thus, new methods to handle correlated outcome data in microbiome association studies are needed. In this paper, we propose the correlated sequence kernel association test (CSKAT) to address such correlations using the linear mixed model. Specifically, random effects are used to account for the outcome correlations and a variance component test is used to examine the microbiome effect. Compared to existing genetic association tests for longitudinal and family samples, we implement a correction procedure to better calibrate the null distribution of the score test statistic to accommodate the small sample size nature of data collected from a typical microbiome study. Comprehensive simulation studies are conducted to demonstrate the validity and efficiency of our method, and we show that CSKAT achieves a higher power than existing methods while correctly controlling the Type I error rate. We also apply our method to a microbiome data set collected from a UK twin study to illustrate its potential usefulness. A free implementation of our method in R software is available at https://github.com/jchen1981/SSKAT.

KEYWORDS

correlated outcomes, linear mixed model, microbiome association analysis, SKAT, small sample

1 | INTRODUCTION

The human microbiota is the collection of microorganisms that live in and on the human body, including the skin, oral mucosa and gastrointestinal tract, and the microbiome refers specifically to the collective genomes of these resident microorganisms (Turnbaugh et al., 2007). Recently, human microbiome research has received unprecedented attention from the scientific community, largely due to the advent of next-generation

sequencing technologies that enable measuring microbiome composition by direct DNA sequencing without the need for laborious cultivation (Lasken, 2012; Turnbaugh et al., 2007). Through analyses of large-scale microbiome studies, it has been found that the human microbiome contributes fundamentally to many human conditions, such as obesity (Turnbaugh et al., 2009), inflammatory bowel diseases (Morgan et al., 2015), colorectal cancer (Wang et al., 2012), Type II diabetes (Qin et al., 2012), and menopause-related symptoms (Mitchell et al., 2017).

Before understanding its impact on human health and disease conditions, a fundamental step in microbiome studies is to first quantify the taxonomic composition. There are two major approaches for characterizing microbiome composition: Metagenomic shotgun sequencing and 16S ribosomal RNA (rRNA) sequencing. Although the metagenomic shotgun approach is potentially more powerful and unbiased, 16S rRNA gene targeted sequencing is routinely performed to determine the taxonomic composition, mostly because of its reduced cost. The generated 16S rRNA sequence tags are usually clustered into operational taxonomic units (OTUs) with a specified amount of variability allowed within each OTU (Caporaso et al., 2010). At 97% similarity, these OTUs represent taxonomic species, and downstream analysis is often performed on the OTU abundances (Li, 2015). One such downstream analysis is to identify potential association between outcomes (e.g., disease status, treatment, and biological and environmental factors) and microbiome composition, which can lead to a better understanding of the underlying disease mechanism from a microbiome perspective, and potentially contribute to the development of new therapeutic strategies to modulate the microbiome composition and affect human health and disease conditions (Virgin & Todd, 2011).

Among existing methods for microbiome association analysis, one popular strategy is to use a distance/kernel-based multivariate test, which examines the global association between multiple OTUs and an outcome of interest (Chen et al., 2012; Koh, Blaser & Li, 2017; Plantinga et al., 2017; Tang, Chen & Alekseyenko, 2016, 2017; Zhao et al., 2015; Zhang, Han, Cox & Li, 2017; Zhan, 2017b) et al., 2017a). These distance/kernel-based multivariate tests can successfully address some common challenges in microbiome data (e.g., weak marginal OTU effect, large number of OTUs, and phylogenetic relatedness) by applying a meaningful ecological distance/dissimilarity metric and then evaluate the statistical significance of global association via a PERMANOVA-type test (McArdle & Anderson, 2001; Tang et al., 2016) or a kernel-based variance component test (Koh et al., 2017; Plantinga et al., 2017; Zhao et al., 2015; Zhan et al., 2017a). Using these methods, important insights regarding the association between microbiome composition and human health and disease conditions have been gained.

However, a common limitation of existing methods is that they are not directly applicable to microbiome association studies with related outcomes. These methods usually assume unrelated samples and can lead to incorrect inference when there exists correlation among the outcomes. Examples of related outcomes in microbiome studies include those from longitudinal and pedigree (family) studies (Borewicz et al., 2013; Goodrich et al., 2014; Lax et al., 2014; Morris et al., 2016; Turnbaugh et al., 2009). For example, in the UK twin studies (Goodrich et al., 2014), the body mass indices (BMIs) of the twins are highly correlated. If our interest is to test the effect of the microbiome on BMI using the twin data, existing association methods developed for independent data could potentially lead to exaggerated significance.

We propose to address the correlations among the outcomes through the linear mixed model (LMM), which uses random effects to capture the correlations. The effect of variables of interest (e.g., a set of OTUs in microbiome association studies or a single-nucleotide polymorphism [SNP]-set in genome-wide association studies) is also modeled via a random effect term in LMM. To account for high dimensionality, the effect of the variables is examined via a score statistic for testing whether the variance of the random effect is zero. Recently, many such variance component score tests have been proposed in genetic association studies, both for family samples (Chen, Meigs & Dupuis, 2013; Schaid, McDonnell, Sinnwell & Thibodeau, 2013; Schifano et al., 2012) and for longitudinal samples (Wang, Xu, Zhang, Wu & Wang, 2017). A common limitation of applying these existing score tests (Chen et al., 2013; Schaid et al., 2013; Schifano et al., 2012; Wang et al., 2017) to microbiome association analysis is that the null distribution of mixture of χ^2s for existing tests is uncalibrated for microbiome studies, due to the fact that microbiome studies usually have far fewer samples than genetic association studies (Chen, Chen, Zhao, Wu & Schaid, 2016; Schweiger et al., 2017). As a result, existing well-established genetic association tests tend to be conservative for examining microbiome associations (Chen et al., 2016; Zhan, 2017b et al., 2017a). To address this limitation, we propose a new correlated sequence kernel association test (CSKAT) with small sample correction. Instead of adopting the common null distribution of mixture of χ^2's as widely used in existing methods, we recalibrate the null distribution of CSKAT in a manner that can largely correct the conservativeness of those tests for a small sample size.

The rest of the paper is organized as follows. We first introduce some notations and a LMM framework for correlated data. Then, we present a new variance component score test incorporating the small sample nature of microbiome data to examine the global association between the outcome of interest and microbiome composition. Both numerical simulation studies and an application example to a microbiome data set from a UK twin study were used to illustrate and evaluate the new method. The paper concludes with discussion.

2 | METHODS

2.1 | Notation

To establish notation used in the rest of this paper, we consider a total of N samples $(y_{ij}, \mathbf{X}_{ij}, \mathbf{Z}_{ij})$ for $i = 1, ..., n$ and $j = 1, ..., m_i$, where i denotes a cluster (e.g., a family), j denotes an observation within a cluster, and $N = \sum_{i=1}^{n} m_i$ is the total sample size. Here, y_{ij} is a continuous outcome variable, $\mathbf{X}_{ij} = (x_{ij0}, x_{ij1}, ..., x_{ijq})^T$ denotes a $(q + 1) \times 1$ vector of covariates, including $x_{ij0} \equiv 1$ as the intercept, and $\mathbf{Z}_{ij} = (z_{ij1}, ..., z_{ijp})^T$ is the composition vector of p taxa/OTUs (subject to appropriate normalization and transformation). The taxa/OTUs can be from a certain taxonomic rank, such as phylum, class, order, family, and genus. Depending on the rank, the number of OTUs p can range from the order of tens (family) to the order of thousands (species). By an abuse of notation, we simply the term "microbiome community" to refer the collection of all p OTUs being analyzed. An important feature of the data considered in this paper is that different samples $(y_{ij}, \mathbf{X}_{ij}, \mathbf{Z}_{ij})$ and $(y_{ij'}, \mathbf{X}_{ij'}, \mathbf{Z}_{ij'})$ from the same cluster are related (while we still assume that samples from different clusters are unrelated). Finally, we use $\mathbf{I}_n$ to denote the nth-order identity matrix and $\mathbf{1}_n$ to denote a $n \times 1$ vector of all ones.

2.2 | An LMM for correlated microbiome data

We model the association between the outcome and the microbiome via an LMM:

$$y_{ij} = \mathbf{X}_{ij}^T \boldsymbol{\alpha} + \mathbf{Z}_{ij}^T \boldsymbol{\beta} + \mathbf{U}_{ij}^T \boldsymbol{a}_i + \epsilon_{ij}, \quad (1)$$

where $\boldsymbol{\alpha} = (\alpha_0, ..., \alpha_q)^T$, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ are regression coefficients, $\boldsymbol{a}_i \sim MVN(0, \sigma_a^2 \Sigma_a)$ are the coefficients for cluster-specific random effects, $\mathbf{U}_{ij}$ are the covariates of the random effects, and $(\epsilon_{i1}, ..., \epsilon_{im_i})^T \sim MVN(\mathbf{0}, \sigma_e^2 \mathbf{I}_{m_i})$. Without loss of generality, we assume that there is only one random effect and $\mathbf{U}_{ij} = 1$, $\boldsymbol{a}_i \sim N(0, \sigma_a^2)$. That is, we only include the random intercept term in our LMM since it is often adequate to capture correlations within a cluster in practice (Chen & Li, 2016; Min & Agresti, 2005). Our primary interest is to test the null hypothesis H_0: $\boldsymbol{\beta} = 0$, that is, to examine whether the microbiome composition has an overall effect on the outcome variable after accounting for the effects of covariates.

Using matrix notation, the model can be rewritten as the following LMM:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\boldsymbol{\beta} + \mathbf{h}_a + \boldsymbol{\epsilon}. \quad (2)$$

Notationally, $\mathbf{y} = (y_{11}, ..., y_{1m_1}, ..., y_{n1}, ..., y_{nm_n})^T$ is a $N \times 1$ vector, $\mathbf{X} = (\mathbf{X}_{i1}, ..., \mathbf{X}_{nm_n})^T$ is $N \times (q + 1)$ covariates matrix with the first column being the intercept, $\mathbf{Z} = (\mathbf{Z}_{i1}, ..., \mathbf{Z}_{nm_n})^T$ is the $N \times p$ microbiome matrix with each row being the microbiome profile of one observation within a cluster, $\mathbf{h}_a = (\boldsymbol{a}_1^T, \boldsymbol{a}_2^T, ..., \boldsymbol{a}_n^T)^T = (a_1, ..., a_1, ..., a_n, ..., a_n)^T \sim MVN(\mathbf{0}, \sigma_a^2 \mathbf{A})$ with $\mathbf{A} = \text{diag}(\mathbf{1}_{m_1}\mathbf{1}_{m_1}^T, ..., \mathbf{1}_{m_n}\mathbf{1}_{m_n}^T)$ being a $N \times N$ block diagonal matrix, and $\boldsymbol{\epsilon} = (\epsilon_{11}, ..., \epsilon_{1m_1}, ..., \epsilon_{n1}, ..., \epsilon_{nm_n})^T \sim MVN(\mathbf{0}, \sigma_e^2 \mathbf{I}_N)$. We notice that positive semidefinite matrix $\mathbf{A}$ can take different forms depending on the context. For family-based studies, $\mathbf{A}$ can be set as the kinship matrix (Chen et al., 2013; Schifano et al., 2012). In longitudinal studies, some common choices for $\mathbf{A}$ are compound symmetry matrix and autoregressive correlation matrix (Wang et al., 2017).

When the dimension of microbiome features p is large (e.g., all OTUs in the gut microbiome community), testing $\boldsymbol{\beta}$ as a fixed-effect is model (2) can be challenging. An alternative approach is to treat $\boldsymbol{\beta}$ as a random effect distributed as $\boldsymbol{\beta} \sim N(0, \tau \mathbf{I}_p)$, and then testing the variance component $\tau = 0$ in LMM (2) using a score-type test (Chen et al., 2013; Wang et al., 2017). Another advantage of the variance component score test is that phylogenetic structure can be incorporated, which usually leads to statistically more powerful analysis (Chen et al., 2012; Tang et al., 2017; Zhao et al., 2015). In particular, rather than assuming a linear relationship between microbiome and outcome, one can use a general random effect $\mathbf{h}_Z$ to replace $\mathbf{Z}\boldsymbol{\beta}$ in the original LMM (2). That is,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{h}_Z + \mathbf{h}_a + \boldsymbol{\epsilon}, \quad (3)$$

where $\mathbf{h}_Z \sim N(\mathbf{0}, \tau \mathbf{K}_Z)$ and covariance matrix $\mathbf{K}_Z$ can be viewed as a $N \times N$ similarity (kernel) measure of microbiome profiles. Phylogenetic relationship has been successfully incorporated in $\mathbf{K}_Z$ via UniFrac-type beta-diversities (Chen et al., 2012; Zhao et al., 2015). Under model representation (3), to test no microbiome effect on the outcome after adjusting covariates, it suffices to

examine H_0: $\tau = 0$. Score tests for the variance component $\tau = 0$ are presented in the next section.

2.3 | A small-sample adjusted kernel score statistic for association testing

Without loss of generality, we derive the association test based on model (3), while model (2) can be viewed as a special case of (3) with $\mathbf{K}_Z = \mathbf{Z}\mathbf{Z}^T$. Define $\Sigma_{\text{full}} = \tau\mathbf{K}_Z + \sigma_a^2\mathbf{A} + \sigma_e^2\mathbf{I}_N$ and $\Sigma = \sigma_a^2\mathbf{A} + \sigma_e^2\mathbf{I}_N$ be the covariance matrix of $\mathbf{y}$ under the full model and null model, respectively. To derive the score test statistic for H_0: $\tau = 0$, one can calculate the partial derivative of the log-likelihood with respect to τ:

$$\left.\frac{\partial l(\boldsymbol{\alpha}, \Sigma_{\text{full}})}{\partial \tau}\right|_{H_0} \propto -\text{tr}[\Sigma^{-1}\mathbf{K}_Z] + (\mathbf{y} - \mathbf{X}\boldsymbol{\alpha})^T\Sigma^{-1}\mathbf{K}_Z\Sigma^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}).$$

Following the rationale used in existing SKAT-type statistics (Chen et al., 2013; Schifano et al., 2012; Wang et al., 2017; Wu et al., 2011), we take the second term of partial derivative of the log-likelihood evaluated under H_0 as the score statistic for $\tau = 0$:

$$Q = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\alpha}})^T\Sigma^{-1}\mathbf{K}_Z\Sigma^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\alpha}}), \tag{4}$$

where $\hat{\boldsymbol{\alpha}} = (\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^T\Sigma^{-1}\mathbf{y}$ and Σ can be replaced by a consistent estimator $\hat{\Sigma}$ under the null LMM. Finally, to calculate the p-value of Q, existing tests usually use a mixture of χ^2's as the approximate null distribution of Q and further use Davies method on the mixture of χ^2's to calculate the p-value (Duchesne & De Micheaux, 2010). Examples of such a test include the family-based genetic association test (famSKAT; Chen et al., 2013) and the longitudinal SNP-set test (LSKAT; Wang et al., 2017). For simplicity, we term all this type of tests as a Q test in the rest of this paper.

Increasing evidence indicates that the performance of a Q test is largely affected by the accuracy of variance component estimation (Chen et al., 2016; Schweiger et al., 2017; Zhan et al., 2017a). The mixture of χ^2's null distribution is exact if Σ is known, and is only approximate if it is estimated. Asymptotically, when Σ can be accurately estimated by $\hat{\Sigma}$, the mixture of χ^2's distribution provides a good approximation to the null distribution of Q. In practice, estimating error in $\hat{\Sigma}$ could cause large discrepancy between the Q's approximate distribution and its exact distribution under certain scenarios (Schweiger et al., 2017). One example is the small-sample scenario, where variance components (σ_a^2 and σ_e^2) cannot be accurately estimated due to a small sample size (Chen et al., 2016; Zhan et al., 2017a). Furthermore, it is also possible to have a large discrepancy even with very large samples, and the coefficient of variation (CV) of the kernel's spectrum is a good indicator to quantify the discrepancy (Schweiger et al., 2017). For kernels considered in this paper, we defer details on the CV of some common kernels for microbiome data to Section 3.2.1.

Recently, the discrepancy has been addressed for independent samples with univariate or multivariate outcomes using an adjusted score statistic (Chen et al., 2016; Zhan et al., 2017a). Along the same line, we introduce the new CSKAT statistic with the denominator offsetting the impact of variance component estimation error. That is, we propose

$$Q_1 = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\alpha}})^T\Sigma^{-1}\mathbf{K}_Z\Sigma^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\alpha}})}{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\alpha}})^T\Sigma^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\alpha}})} \tag{5}$$

as our new association test statistic for related microbiome samples. Compared with Q, the new statistic Q_1 is standardized and scale-free with respect to the error variance scale. The detailed procedure to calculate the p-value based on the Q_1 statistic is presented in the Appendix.

3 | RESULTS

3.1 | Simulation studies

Following simulation designs used in previous microbiome association studies (Zhao et al., 2015), we simulated the outcome through

$$y_{ij} = 0.5\cdot(1 + X_{i1} + X_{ij2}) + \beta \cdot scale\left(\sum_{l\in\mathcal{A}} Z_{ijl}\right) + \boldsymbol{a}_i + \epsilon_{ij}, \tag{6}$$

where $i = 1, ..., n$, $j = 1, ..., m$. To mimic the UK twin study microbiome data analyzed later in this paper, we set $m = 2$, and $n = 30$ or 100 in this simulation. X_{i1} was an Bernoulli variable with success probability 0.5 that did not depend on j, and X_{ij2} was a standard normal variable. Z_{ijl} is the relative abundance of the lth OTU in jth microbiome sample from cluster i, and function $scale(\cdot)$ standardized the mean of the vector $\{\sum_{l\in\mathcal{A}} Z_{ijl}\}_{i,j}$ to zero and standard deviation to one. Index set $\mathcal{A}$ was a collection of OTUs or a subset of the whole microbiome community, which will be described in detail in the next paragraph. We simulated the random terms $\boldsymbol{a}_i \sim N(0, \sigma_a^2)$, $\epsilon_{ij} \sim N(0, \sigma_e^2)$ where $\sigma_e^2 = 1$. We picked $\sigma_a^2 = 1/4, 1, 4$ so that the intraclass correlation coefficient $\rho = \sigma_a^2/(\sigma_a^2 + \sigma_e^2) = 0.2, 0.5, 0.8$ representing different levels of within-cluster correlation strength.

Compared to the relatively straightforward clustered outcome variables, simulating related microbiome samples was quite difficult. In particular, the multidimensionality and the compositionality (sum to one) constraint posed grand challenges for simulating $\mathbf{Z}_{ij}$'s. To capture potential within-cluster relatedness, the following random perturbation was considered in our simulations (Zhao, Zhan, Guthrie, Mitchell & Larson, 2018). Let $(a_1, ..., a_p)$ and $(b_1, ..., b_p)$ be two compositional vectors, that is, $\sum_{i=1}^{p} a_i = \sum_{i=1}^{p} b_i = 1$. Then, the perturbation between $(a_1, ..., a_p)$ and $(b_1, ..., b_p)$ is defined as $a \otimes b \equiv (a_1 b_1 / \sum a_i b_i, ..., a_p b_p / \sum a_i b_i)$, which is also a compositional vector. Assuming the first microbiome composition sample $\mathbf{Z}_{i1} = (Z_{i11}, ..., Z_{i1p})$, then using the perturbation, the next sample within the same cluster was simulated as a random perturbation of the previous microbiome composition. That is, $\mathbf{Z}_{ij} = \mathbf{Z}_{i(j-1)} \otimes \mathbf{Z}, j = 2, ..., m$, where $\mathbf{Z}$ is a random vector with each coordinate nonnegative and all coordinates sum to 1. To simulate $\mathbf{Z}_{i1}$, we followed an existing strategy that generated OTU counts using a Dirichlet-multinomial distribution estimated from real respiratory-tract microbiome data (Charlson et al., 2010), and then normalized the OTU counts to relative abundances by dividing the total count for each sample. In total, $p = 856$ OTUs were simulated. To select an active set $\mathcal{A}$ of OTUs that affected the outcome, two schemes were used: (a) a random 10% of OTUs were selected to be active, and (b) we partitioned the 856 OTUs into 20 clades (lineages) using the partition around medoids algorithm and selected one of the clades as active. The second scheme represented a scenario in which associated OTUs were phylogenetically related. For simplicity, we denote two schemes as $\mathcal{A}_1$ and $\mathcal{A}_2$, respectively, in the rest of this section.

Under the null model, we simulated data according to (6) by setting $\beta = 0$. Because $\mathcal{A}$ would not affect y_{ij} when $\beta = 0$ according to (6), no active set $\mathcal{A}$ was formed under the null model. Under each simulation scenario, 10, 000 replicates were simulated to assess the Type I error rates. Under the alternative model, both $\mathcal{A}_1$ and $\mathcal{A}_2$ were evaluated, and 1,000 replicates were generated to calculate the power of different tests. To avoid power saturation (all powers either too close to 0 or to 1), we used $\beta = 1, 2, 3$ for $n = 30$ and $\beta = 0.5, 1, 1.5$ for $n = 100$. After the data were simulated, we applied both the Q test and the Q_1 test to examine the overall association between the related microbiome profiles and outcome. For the kernel matrix $\mathbf{K}_Z$ in Q/Q_1, the Bray–Curtis kernel K_{BC} and UniFrac-type kernels (Zhao et al., 2015) including the weighted UniFrac kernel K_w, unweighted UniFrac kernel K_u and generalized UniFrac kernel K_5 with the parameter set as 0.5 were used. The Bray–Curtis kernel only depends on the microbial abundance. On the other hand, the UniFrac kernels incorporate both abundance and phylogenetic information, and are expected to be more powerful under scheme $\mathcal{A}_2$.

3.2 | Simulation results

3.2.1 | Spectrum of microbiome kernels

For independent samples, the discrepancy between a Q test and an exact test has been shown to be related to the CV of the kernel matrix's nonzero eigenvalues (Schweiger et al., 2017). In particular, let $\lambda_1, ..., \lambda_L (L \leq N)$ be the nonzero eigenvalues of $\mathbf{K}_Z$, then CV is the ratio of the standard deviation of $(\lambda_1, ..., \lambda_L)$ to the mean. Schweiger et al. (2017) suggest that: (a) If CV $\gg 1$, discrepancy between Q test and its exact counterpart is negligible; (b) The larger the CV, the smaller the discrepancy.

Since the Q_1 test can be viewed as a more exact version of the Q test by addressing the estimation error in the scale of error variance, it is interesting to observe the CVs of various microbiome kernels, which may provide insights of the discrepancy between the Q test and the proposed Q_1 test. We calculated the sample CV of some common microbiome kernels using 1,000 replicates. The summary of these CVs is presented in Table 1, based on which the CVs of these microbiome kernels are all around 1 when the sample size is relatively small (i.e., $n = 30$ and $n = 100$). Thus, there is plenty room to improve the performance of the Q test in a small-sample size scenario, which motivates our Q_1 test.

3.2.2 | Type I error results

The empirical Type I error rates are presented in Table 2. Overall, the Q test is conservative under all simulation scenarios. With independent samples, it has been observed that p values of genetic association score tests (e.g., the Q test considered in this paper) can be conservative when sample size is small or moderate (e.g., less than a few thousand; Chen et al., 2016). The same phenomenon is also observed in our simulations for related microbiome samples. On the other hand, the proposed Q_1 test has

TABLE 1 CV of microbiome kernels

n	K_w	K_u	K_5	K_{BC}
30	1.399 (0.049)	0.469 (0.018)	1.088 (0.013)	1.165 (0.022)
100	1.978 (0.050)	0.858 (0.011)	1.374 (0.013)	1.536 (0.022)

Note. Each cell is the mean value of CVs based on 1,000 replicates and the number in parentheses is the standard deviation.
CV: coefficient of variation.

TABLE 2 Type I errors of Q test and Q_1 test

		$\alpha = 0.05$			$\alpha = 0.005$		
Test	**Kernel**	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$
	K_w	0.0136	0.0180	0.0244	0.0002	0.0002	0.0011
	K_u	0.0001	0	0	0	0	0
Q	K_5	0.0004	0.0012	0.0060	0	0	0
	K_{BC}	0.0028	0.0068	0.0135	0	0	0.0003
	K_w	0.0311	0.0391	0.0507	0.0020	0.0042	0.0053
	K_u	0.0017	0.0028	0.0019	0.0001	0	0
Q_1	K_5	0.0073	0.0153	0.0428	0	0.0008	0.0051
	K_{BC}	0.0176	0.0273	0.0490	0.0006	0.0024	0.0058
	K_w	0.0321	0.0332	0.0324	0.0019	0.0020	0.0024
	K_u	0.0064	0.0067	0.0057	0	0	0
Q	K_5	0.0083	0.0126	0.0183	0.0001	0.0003	0.0002
	K_{BC}	0.0196	0.0234	0.0204	0.0003	0.0011	0.0012
	K_w	0.0441	0.0494	0.0471	0.0047	0.0049	0.0047
	K_u	0.0179	0.0176	0.0163	0.0003	0.0006	0.0004
Q_1	K_5	0.0282	0.0367	0.0454	0.0015	0.0029	0.0048
	K_{BC}	0.0362	0.0476	0.0444	0.0032	0.0035	0.0035

Note. The upper half corresponds to $n = 30$ and and the lower half corresponds to $n = 100$.

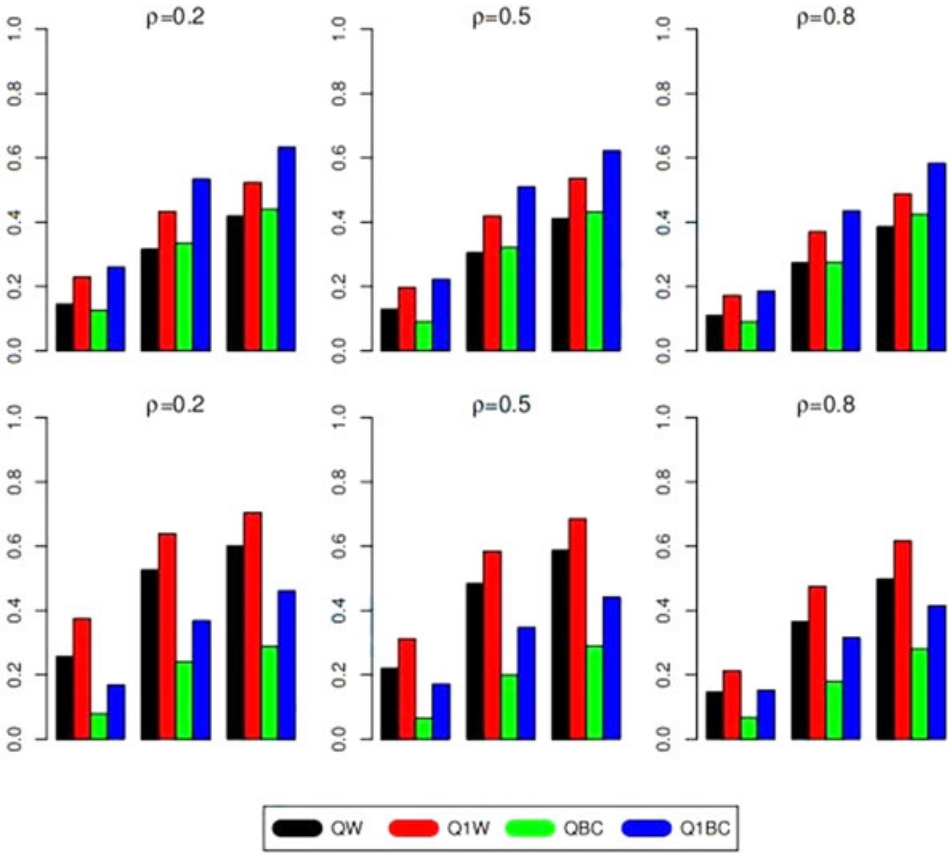

FIGURE 1 Power under $n = 30$. The Y-axis represents the power and the X-axis represents the effect size β. The first row corresponds to scheme $\mathscr{A}_1$, and the second row corresponds to scheme $\mathscr{A}_2$. QW/Q1W represents K_w-based Q/Q_1 test, and QBC/Q1BC represents K_{BC}-based Q/Q_1 test

much more accurate null p values than the Q test. For each kernel, the Q_1 test is much less conservative than the corresponding Q test on the same kernel. We notice that some Q_1 tests (especially K_u-based) may still be a little conservative. This is because our Q_1 test is still not an exact test in the sense that it is affected by the estimation error in $\hat{\Sigma}$ (see Appendix for details). For an approximate test, its relative performance (compared to the corresponding exact test) typically is indicated by the CV of eigenvalues of the underlying kernel (Schweiger et al., 2017). As seen in Table 1, K_u has much smaller CV compared with the other three which explains that a K_u-based Q test is much more conservative than others. The K_u-based Q_1 performs much better, although it is still conservative given the nature of the kernel.

To summarize, existing Q tests as widely used in genetic association analysis (Chen et al., 2013; Wang et al., 2017) can be very conservative in a typical microbiome association studies where the sample size is usually small or moderate. Despite not being perfect, the newly proposed Q_1 test consistently corrects the conservativeness of Q test to a large extent in the small sample setting typical of microbiome studies. In the Type I error simulations presented here, we considered the scenario in which covariates ($\mathbf{X}'s$) are independent of OTUs. The Type I error performance of Q and Q_1 are similar under other more complicated scenarios. Additional simulations are presented in supplementary materials.

3.2.3 | Power results

For ease of presentation, we evaluated power of tests based on K_{BC} and K_w (as a representative of the UniFrac kernels). The empirical power under $n = 30$ and $n = 100$ are presented in Figures 1 and 2, respectively. Under scheme $\mathcal{A}_1$, the outcome-associated OTUs are randomly distributed. Thus, K_{BC}-based tests are more powerful than UniFrac kernels-based tests (K_w, K_u, and K_5). In contrast, under scheme $\mathcal{A}_2$, the outcome-associated OTUs belong to the same cluster and hence K_w-based tests are more powerful. Based on these two figures, the Q_1 test is consistently more powerful than the corresponding Q test under each scenario, which is not surprising given the conservativeness of the Q test as observed in Type I error simulations. The power gain from Q to Q_1 is usually larger for K_{BC} than for K_w. According to Table 1, K_{BC} has smaller CV than K_w. Thus, the discrepancy between a K_{BC}-based Q test and a K_{BC}-based exact test is larger than that of K_w(Schweiger

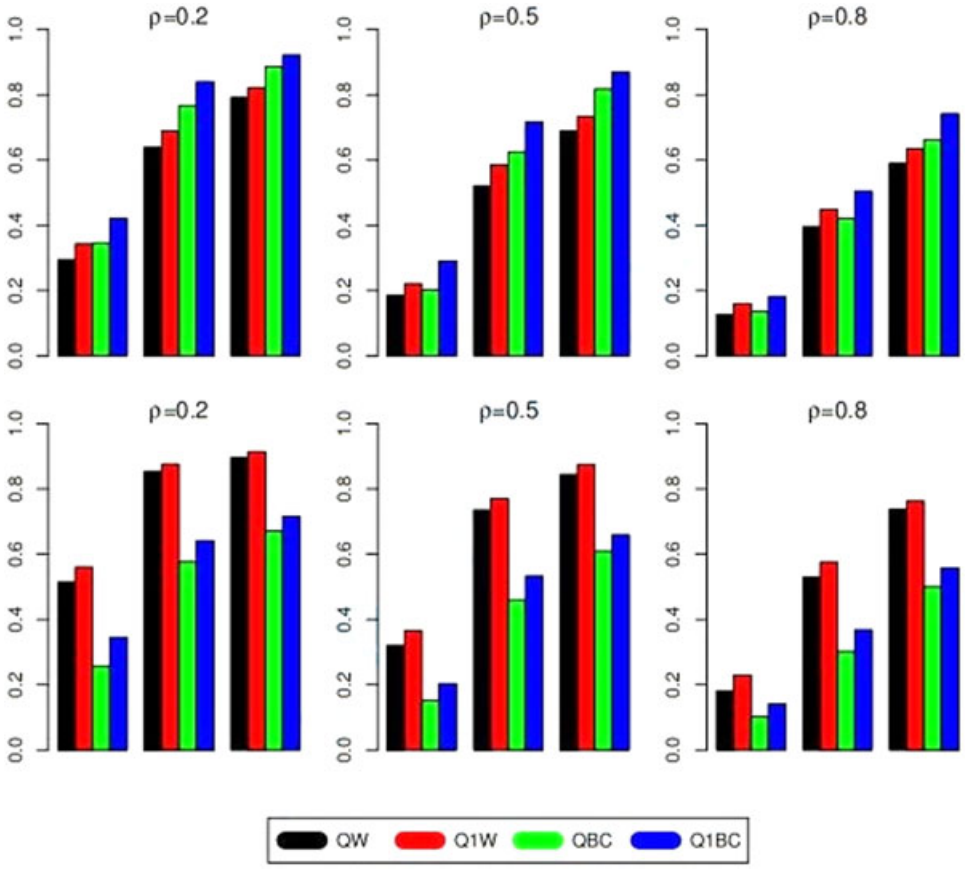

FIGURE 2 Power under $n = 100$. The Y-axis represents the power and the X-axis represents the effect size β. The first row corresponds to scheme $\mathcal{A}_1$ and the second row corresponds to scheme $\mathcal{A}_2$. QW/Q1W represents K_w-based Q/Q_1 test, and QBC/Q1BC represents K_{BC}-based Q/Q_1 test

et al., 2017). In other words, the improvement from a K_{BC}-based Q test to a K_{BC}-based Q_1 test is larger than that for K_w. The same reason can explain why the power gain from the Q test to the Q_1 test is larger at $n = 30$ than $n = 100$.

3.3 | Application to UK twins data

We applied the proposed method to a microbiome data set from a UK twin study (Goodrich et al., 2014). The data set was produced by targeted sequencing of the V4 region of the 16S rRNA gene in fecal samples and was originally used to investigate the genetic effects on the human gut microbiome. In this application, we used it to test for the association between BMI and the gut microbiome. We retrieved the microbiome data set ($n = 1024$) from Qiita database with the study ID 2014 (https://qiita.ucsd.edu/). The raw sequencing data were processed by QIIME (version 1.9.0-dev) using reference-based OTU picking method, which clustered the 16S sequencing tags at 97% similarity level into 7,365 nonsingleton OTUs. We focused our analysis on the monozygotic twins and rarefied the OTU counts to an equal depth of 10,000 counts per sample. Distance matrices (Bray–Curtis, unweighted, generalized, and weighted UniFrac) were constructed based on the rarefied counts and the phylogenetic tree (GUniFrac R package; Chen et al., 2012), and converted into kernels as in Zhao et al. (2015). A total of 311 samples from 145 families were finally included in the analysis. The age range for the data set was 27–83 with a median of 63, and age was weakly correlated with BMI ($p = 0.016$ and $p = 0.58$ for linear/quadratic age terms, LMM). The BMIs of twins were highly correlated (intraclass correlation coefficient: 0.78) and thus we need to take into account the correlation when testing the association between BMI and the gut microbiome. We thus calculated the Q/Q_1 statistics by fitting a linear mixed effects model with a family-level random intercept to capture the shared genetic and environment effects within each family. A linear age term was included in the null model. Table 3 presents the association p values based on Q/Q_1 statistics under four kernels. We first observe that the association is very significant when the unweighted UniFrac or Bray–Curtis kernels are used, indicating a potential diverse community change associated with BMI. The test based on Q_1 produces consistently smaller p values than that based on Q, although the difference is moderate. Since many microbiome studies have a much smaller sample size, it is informative to compare the performance of Q/Q_1 on a smaller sample size. To achieve this end, we randomly sampled 50 families from the 145 families and performed the same association tests based on the subsampled data set. We then repeated the random sampling 200 times and compared the distribution of p values over the 200 random samples. In Figure 3, we observe much lower p values for our small sample adjusted approach, especially for unweighted UniFrac and Bray–Curtis kernels. Using a p-value cutoff of 0.05, our approach invariably declares more significant associations than the unadjusted approach (Table 4). When the unweighted UniFrac kernel is used, our approach achieves significance in 53.5% of the cases, compared to 36.5% for the unadjusted approach. Therefore, our method has a clear advantage in real data sets with a small sample size.

TABLE 3 Q/Q_1-based p values for body mass indes–microbiome associations on the UK twin data set

Test	K_w	K_u	K_5	K_{BC}
Q	0.0838	0.0002	0.0068	0.0004
Q_1	0.0823	0.0001	0.0057	0.0002

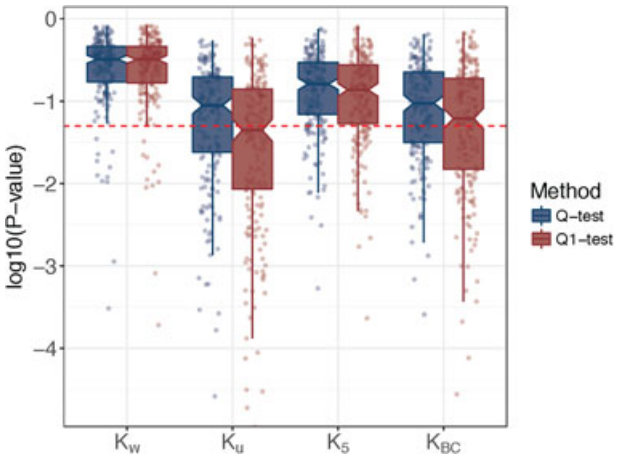

FIGURE 3 Power comparison on the subsampled UK twin data set. Q/Q_1-based tests were performed on 200 random samplings of 50 families from the original 145 families. Four types of kernels (K_w, K_u, K_5, K_{BC}) were used. The association p values on the log scale are shown. Red dashed line indicates the significance level of 0.05

TABLE 4 Percentage of subsamples with p values less than 0.05

Test	K_w (%)	K_u (%)	K_5 (%)	K_{BC} (%)
Q	5.5	36.5	17.0	33.0
Q_1	5.5	53.5	24.0	44.5

4 | DISCUSSION

Family-based and longitudinal study designs have several advantages compared to population-based studies with unrelated individuals, and have become increasingly

popular in the scientific community. However, along with the benefits of these complex designs come statistical challenges, particularly regarding how to address the correlation among related samples. Previous studies have shown that statistical analysis without considering such correlations can lead to suboptimal or even invalid inference (Chen et al., 2013; Wang et al., 2017). In this paper, we propose CSKAT to test the association between microbiome compositions and an outcome of interest, where microbiome samples and outcome measurements within the same cluster are related to each other. Like famSKAT (Chen et al., 2013) and LSKAT (Wang et al., 2017), CSKAT can be viewed as an extension of SKAT (Wu et al., 2011) to consider correlated data. Through extensive numerical studies, we have seen that CSKAT can protect the correct Type I error and achieve higher power than these existing methods, especially when the sample size is small or moderate. Besides the improved power using small-sample correction, CSKAT is more flexible than famSKAT and LSKAT in that it can test many types of correlated data. The null LMM in CSKAT can be fitted by many softwares such as the lme4 R package (Bates, Mächler, Bolker, & Walker, 2015) and lme4qtl R package (Ziyatdinov et al., 2018), allowing for many different types of correlated data including longitudinal data and pedigree data by specifying the kinship matrix. Although we illustrate CSKAT using microbiome data as an example, the methodology we developed is general and would have substantial applicability to many types of genetic and genomic data (Chen et al., 2013; Wang et al., 2017; Zhan, Girirajan, Zhao, Wu & Ghosh, 2016).

We mainly formulate CSKAT as a tool to examine the overall association between correlated microbiome compositions and an outcome of interest. But the method can be applied to an arbitrary set of OTUs. To detect biologically meaningful results, we can form the OTU-set according to a particular taxonomic rank (e.g., kingdom, phylum, class, order, family, genus, or species). Similar to other methods (Koh et al., 2017; Wu, Chen, Kim, & Pan, 2016), our CSKAT framework can also be used as a taxon association mapping tool by shifting the analysis unit to a lower taxonomic rank (e.g., genus or family). For example, by targeting OTUs within each genus at a time, our method can identify associated genera. In this manner, our method can fine-map the outcome-associated lineage on the phylogenetic tree. To account for multiple testing correction needed for each rank (except for kingdom), many procedures are available to control the false discovery rate of multiple testing (Xiao, Cao & Chen, 2017).

Despite the large improvement of CSKAT over famSKAT/LSKAT in small sample scenarios by correcting the estimation error in scale of variance, the p-value calculation procedure implemented in CSKAT depends on estimator $\hat{\Sigma}^{-1}$, as can be seen in the procedure detailed in the Appendix. Thus, CSKAT is still not a perfect exact test in the sense that a bad estimator $\hat{\Sigma}^{-1}$ may still lead to a conservative test (especially for a K_u-based CSKAT). It is of future interest to further improve CSKAT by correcting for the remaining estimation error. The current method fits a LMM for a continuous outcome. Further developments are needed to accommodate other types of outcomes, such as a generalized LMM framework for a dichotomous outcome and mixed-effect Cox model framework for a survival time outcome. We will leave these extensions in future investigation.

ACKNOWLEDGMENTS

We thank the editor and two reviewers for their constructive comments that have greatly improved the paper. The research of D. J. S. was supported by the US Public Health Service, National Institutes of Health (contract Grant No. GM065450). The research of J. C. was supported by Mayo Clinic Center for Individualized Medicine. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

CONFLICTS OF INTEREST

The authors declare that there are no conflicts of interest.

ORCID

Xiang Zhan http://orcid.org/0000-0001-9650-143X
Ni Zhao http://orcid.org/0000-0002-7762-3949

REFERENCES

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Borewicz, K., Pragman, A. A., Kim, H. B., Hertz, M., Wendt, C., & Isaacson, R. E. (2013). Longitudinal analysis of the lung microbiome in lung transplantation. *FEMS Microbiology Letters*, *339*(1), 57–65.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., & Gordon, J. I., et al. (2010). Qiime allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*(5), 335–336.

Charlson, E. S., Chen, J., Custers-Allen, R., Bittinger, K., Li, H., Sinha, R., & Collman, R. G. (2010). Disordered microbial communities in the upper respiratory tract of cigarette smokers. *PLOS One*, *5*(12), e15216.

Chen, E. Z., & Li, H. (2016). A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics*, *32*(17), 2611–2617.

Chen, H., Meigs, J. B., & Dupuis, J. (2013). Sequence kernel association test for quantitative traits in family samples. *Genetic Epidemiology*, *37*(2), 196–204.

Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., Wu, G. D., & Li, H. (2012). Associating microbiome composition with environmental covariates using generalized unifrac distances. *Bioinformatics*, *28*(16), 2106–2113.

Chen, J., Chen, W., Zhao, N., Wu, M. C., & Schaid, D. J. (2016). Small sample kernel association tests for human genetic and microbiome association studies. *Genetic Epidemiology*, *40*(1), 5–19.

Duchesne, P., & DeMicheaux, P. L. (2010). Computing the distribution of quadratic forms: Further comparisons between the liu-tang-zhang approximation and exact methods. *Computational Statistics and Data Analysis*, *54*(4), 858–862.

Goodrich, J. K., Waters, J. L., Poole, A. C., Sutter, J. L., Koren, O., Blekhman, R., & Bell, J. T. (2014). Human genetics shape the gut microbiome. *Cell*, *159*(4), 789–799.

Koh, H., Blaser, M. J., & Li, H. (2017). A powerful microbiome-based association test and a microbial taxa discovery framework for comprehensive association mapping. *Microbiome*, *5*(1), 45.

Lasken, R. S. (2012). Genomic sequencing of uncultured microorganisms from single cells. *Nature Reviews Microbiology*, *10*(9), 631.

Lax, S., Smith, D. P., Hampton-Marcell, J., Owens, S. M., Handley, K. M., Scott, N. M., & Weiss, S. (2014). Longitudinal analysis of microbial interaction between humans and the indoor environment. *Science*, *345*(6200), 1048–1052.

Li, H. (2015). Microbiome, metagenomics and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application*, *2*(1), 73–94.

McArdle, B. H., & Anderson, M. J. (2001). Fitting multivariate models to community data: A comment on distance-based redundancy analysis. *Ecology*, *82*(1), 290–297.

Min, Y., & Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling*, *5*(1), 1–19.

Mitchell, C. M., Srinivasan, S., Zhan, X., Wu, M. C., Reed, S. D., Guthrie, K. A., & Liu, C. (2017). Vaginal microbiota and genitourinary menopausal symptoms: A cross-sectional analysis. *Menopause*, *24*(10), 1160–1166.

Morgan, X. C., Kabakchiev, B., Waldron, L., Tyler, A. D., Tickle, T. L., Milgrom, R., & Silverberg, M. S. (2015). Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease. *Genome Biology*, *16*(1), 67.

Morris, A., Paulson, J. N., Talukder, H., Tipton, L., Kling, H., Cui, L., & Ghedin, E. (2016). Longitudinal analysis of the lung microbiota of cynomolgous macaques during long-term shiv infection. *Microbiome*, *4*(1), 38.

Plantinga, A., Zhan, X., Zhao, N., Chen, J., Jenq, R. R., & Wu, M. C. (2017). Mirkat-s: A community-level test of association between the microbiota and survival times. *Microbiome*, *5*(1), 17.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., & Shen, D. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, *490*(7418), 55–60.

Schaid, D. J., McDonnell, S. K., Sinnwell, J. P., & Thibodeau, S. N. (2013). Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. *Genetic Epidemiology*, *37*(5), 409–418.

Schifano, E. D., Epstein, M. P., Bielak, L. F., Jhun, M. A., Kardia, S. L., Peyser, P. A., & Lin, X. (2012). Snp set association analysis for familial data. *Genetic Epidemiology*, *36*(8), 797–810.

Schweiger, R., Weissbrod, O., Rahmani, E., Müller-Nurasyid, M., Kunze, S., Gieger, C., & Halperin, E. (2017). Rl-skat: An exact and efficient score test for heritability and set tests. *Genetics*, *207*(4), 1275–1283.

Tang, Z.-Z., Chen, G., & Alekseyenko, A. V. (2016). Permanova-s: Association test for microbial community composition that accommodates confounders and multiple distances. *Bioinformatics*, *32*(17), 2618–2625.

Tang, Z.-Z., Chen, G., Alekseyenko, A. V., & Li, H. (2017). A general framework for association analysis of microbial communities on a taxonomic tree. *Bioinformatics*, *33*(9), 1278–1285.

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., & Affourtit, J. P. (2009). A core gut microbiome in obese and lean twins. *Nature*, *457*(7228), 480–484.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C., Knight, R., & Gordon, J. I. (2007). The human microbiome project. *Nature*, *449*(7164), 804–810.

Virgin, H. W., & Todd, J. A. (2011). Metagenomics and personalized medicine. *Cell*, *147*(1), 44–56.

Wang, T., Cai, G., Qiu, Y., Fei, N., Zhang, M., Pang, X., & Zhao, L. (2012). Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. *The ISME Journal*, *6*(2), 320.

Wang, Z., Xu, K., Zhang, X., Wu, X., & Wang, Z. (2017). Longitudinal snp-set association analysis of quantitative phenotypes. *Genetic Epidemiology*, *41*(1), 81–93.

Wu, C., Chen, J., Kim, J., & Pan, W. (2016). An adaptive association test for microbiome data. *Genome Medicine*, *8*(1), 56.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, *89*(1), 82–93.

Xiao, J., Cao, H., & Chen, J. (2017). False discovery rate control incorporating phylogenetic tree increases detection power in microbiome-wide multiple testing. *Bioinformatics*, *33*(18), 2873–2881.

Zhan, X., Girirajan, S., Zhao, N., Wu, M. C., & Ghosh, D. (2016). A novel copy number variants kernel association test with application to autism spectrum disorders studies. *Bioinformatics*, *32*(23), 3603–3610.

Zhan, X., Plantinga, A., Zhao, N., & Wu, M. C. (2017b). A fast small-sample kernel independence test for microbiome community-level association analysis. *Biometrics*, *73*(4), 1453–1463.

Zhan, X., Tong, X., Zhao, N., Maity, A., Wu, M. C., & Chen, J. (2017a). A small-sample multivariate kernel machine test for microbiome association studies. *Genetic Epidemiology*, *41*(3), 210–220.

Zhang, Y., Han, S. W., Cox, L. M., & Li, H. (2017). A multivariate distance-based analytic framework for microbial interdependence

association test in longitudinal study. *Genetic Epidemiology*, *41*(8), 769–778.

Zhao, N., Chen, J., Carroll, I. M., Ringel-Kulka, T., Epstein, M. P., Zhou, H., & Wu, M. C. (2015). Testing in microbiome-profiling studies with mirkat, the microbiome regression-based kernel association test. *The American Journal of Human Genetics*, *96*(5), 797–807.

Zhao, N., Zhan, X., Guthrie, K. A., Mitchell, C. M., & Larson, J. (2018). Generalized hotelling's test for paired compositional data with application to human microbiome studies. *Genetic Epidemiology*, *42*(5), 459–469.

Ziyatdinov, A., Vázquez-Santiago, M., Brunel, H., Martinez-Perez, A., Aschard, H., & Soria, J. M. (2018). lme4qtl: Linear mixed models with flexible covariance structure for genetic studies of related individuals. *BMC Bioinformatics*, *19*(1), 68.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Zhan X, Xue L, Zheng H, et al. A small-sample kernel association test for correlated data with application to microbiome association studies. *Genet. Epidemiol*. 2018;42: 772–782. https://doi.org/10.1002/gepi.22160

APPENDIX

p-Value calculation

Let $\mathbf{P} \equiv \Sigma - \mathbf{X}(\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^T$, then $\mathbf{P}^{1/2} = \Sigma^{1/2} - \mathbf{X}(\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^{*T}\Sigma^{-1/2}$ since $\mathbf{P}^{1/2}(\mathbf{P}^{1/2})^T = \mathbf{P}$. Let $\mathbf{P}^{-1/2}$ be the generalized inverse of $\mathbf{P}^{1/2}$ and define $\boldsymbol{\zeta} \equiv \mathbf{P}^{-1/2}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\alpha}})$. Under the null model, $\boldsymbol{\zeta} \sim MVN(\mathbf{0}, \mathbf{I}_N)$. Correspondingly,

$$Q_1 = \frac{\boldsymbol{\zeta}^T\mathbf{P}^{1/2}\Sigma^{-1}\mathbf{K}_Z\Sigma^{-1}\mathbf{P}^{1/2}\boldsymbol{\zeta}}{\boldsymbol{\zeta}^T\mathbf{P}^{1/2}\Sigma^{-1}\mathbf{P}^{1/2}\boldsymbol{\zeta}}. \tag{A1}$$

Let q_1 be the observed Q_1-value calculated from the sample. Then, CSKAT p value $\Pr(Q_1 \geq q_1)$ is calculated as:

$$\Pr(Q_1 \geq q_1) = \Pr(\boldsymbol{\zeta}^T[\mathbf{P}^{1/2}\Sigma^{-1}\mathbf{K}_Z\Sigma^{-1}\mathbf{P}^{1/2} - q_1\mathbf{P}^{1/2}\Sigma^{-1}\mathbf{P}^{1/2}]\boldsymbol{\zeta} \geq 0). \tag{A2}$$

Since $\boldsymbol{\zeta} \sim MVN(\mathbf{0}, \mathbf{I}_N)$, $\boldsymbol{\zeta}^T[\mathbf{P}^{1/2}\Sigma^{-1}\mathbf{K}_Z\Sigma^{-1}\mathbf{P}^{1/2} - q_1\mathbf{P}^{1/2}\Sigma^{-1}\mathbf{P}^{1/2}]\boldsymbol{\zeta} \sim \sum_{i=1}^{N}\lambda_i\chi_i^2(1)$, where $\lambda_1, ..., \lambda_N$ are eigenvalues of matrix $[\mathbf{P}^{1/2}\Sigma^{-1}\mathbf{K}_Z\Sigma^{-1}\mathbf{P}^{1/2} - q_1\mathbf{P}^{1/2}\boldsymbol{\Sigma}^{-1}\mathbf{P}^{1/2}]$ and $\chi_i^2(1)$, $i = 1, ..., N$ are independent and identically distributed χ^2 random variable with one degree of freedom. Finally, we can apply the Davies method (Duchesne & De Micheaux, 2010) to compute the CSKAT p-value from the mixture of χ^2's distribution $\sum_{i=1}^{N}\lambda_i\chi_i^2(1)$. For comparison, the mixture of χ^2's distribution $\sum_{i=1}^{N}\psi_i\chi_i^2(1)$ used in a Q test (Chen et al., 2013; Wang et al., 2017) have eigenvalues $\psi_1, ..., \psi_N$ of the matrix $\mathbf{P}^{1/2}\Sigma^{-1}\mathbf{K}_Z\Sigma^{-1}\mathbf{P}^{1/2}$ as its weights.

Note that in all the formulas above, we assume Σ is known. In practice, many softwares can be used to estimate Σ under the null LMM $\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{h}_a + \boldsymbol{\epsilon}$. Let $\hat{\Sigma}$ be the estimate of Σ, then eigenvalues $\hat{\lambda}_1, ..., \hat{\lambda}_N$ (or $\hat{\psi}_1, ..., \hat{\psi}_N$) can be calculated by plugging $\hat{\Sigma}$ into matrix $[\mathbf{P}^{1/2}\Sigma^{-1}\mathbf{K}_Z\Sigma^{-1}\mathbf{P}^{1/2} - q_1\mathbf{P}^{1/2}\Sigma^{-1}\mathbf{P}^{1/2}]$ (or $\mathbf{P}^{1/2}\Sigma^{-1}\mathbf{K}_Z\Sigma^{-1}\mathbf{P}^{1/2}$). The denominator of Q_1 statistic is introduced to account for the variability in estimating σ_e^2 in Σ ($\Sigma = \sigma_a^2\mathbf{A} + \sigma_e^2\mathbf{I}_N$). The Q_1 test is not perfect in the sense that its performance can still be affected by the estimating error of σ_a^2. In other words, it can depart from the exact distribution if $\hat{\lambda}_1, ..., \hat{\lambda}_N$ depart from $\lambda_1, ..., \lambda_N$. In general, as observed for independent samples (Chen et al., 2016; Zhan et al., 2017a), the small-sample corrected $\sum_{i=1}^{N}\hat{\lambda}_i\chi_i^2(1)$ distribution usually provides a more accurate Q_1 test than the $\sum_{i=1}^{N}\hat{\psi}_i\chi_i^2(1)$ distribution-based Q test.