


Exact variance component tests for longitudinal microbiome studies

Jing Zhai¹ | Kenneth Knox² | Homer L. Twigg III³ | Hua Zhou⁴ | Jin J. Zhou¹ 

¹Department of Epidemiology and Biostatistics, University of Arizona, Tucson, Arizona

²Division of Pulmonary, Allergy, Critical Care, Sleep Medicine, Department of Medicine, University of Arizona, Tucson, Arizona

³Division of Pulmonary, Critical Care, Sleep, and Occupational Medicine, Indiana University Medical Center, Indianapolis, Indiana

⁴Department of Biostatistics, University of California, Los Angeles, California

Correspondence

Jin J. Zhou, Department of Epidemiology and Biostatistics, University of Arizona, Tucson, AZ 85724.
Email: jzhou@email.arizona.edu

Funding information

National Science Foundation, Grant/Award Number: DMS-1645093; National Institute of General Medical Sciences, Grant/Award Numbers: GM105785, GM053275; National Human Genome Research Institute, Grant/Award Number: HG006139; National Institute of Diabetes and Digestive and Kidney Diseases, Grant/Award Number: K01DK106116; Arizona Biomedical Research Commission, Grant/Award Number: New Investigator Award

Abstract

In metagenomic studies, testing the association between microbiome composition and clinical outcomes translates to testing the nullity of variance components. Motivated by a lung human immunodeficiency virus (HIV) microbiome project, we study longitudinal microbiome data by using variance component models with more than two variance components. Current testing strategies only apply to models with exactly two variance components and when sample sizes are large. Therefore, they are not applicable to longitudinal microbiome studies. In this paper, we propose exact tests (score test, likelihood ratio test, and restricted likelihood ratio test) to (a) test the association of the overall microbiome composition in a longitudinal design and (b) detect the association of one specific microbiome cluster while adjusting for the effects from related clusters. Our approach combines the exact tests for null hypothesis with a single variance component with a strategy of reducing multiple variance components to a single one. Simulation studies demonstrate that our method has a correct type I error rate and superior power compared to existing methods at small sample sizes and weak signals. Finally, we apply our method to a longitudinal pulmonary microbiome study of HIV-infected patients and reveal two interesting genera *Prevotella* and *Veillonella* associated with forced vital capacity. Our findings shed light on the impact of the lung microbiome on HIV complexities. The method is implemented in the open-source, high-performance computing language Julia and is freely available at <https://github.com/JingZhai63/VCmicrobiome>.

KEYWORDS

human immunodeficiency virus, linear mixed effects models, longitudinal pulmonary microbiome, variance component models

1 | INTRODUCTION

Technology advances have led to a much deeper understanding of microbes and their link to human health (Eckburg et al., 2005; Haas et al., 2011; Hodkinson & Grice, 2015; Kuleshov et al., 2016; Wang & Jia, 2016). In particular, for the pulmonary microbiome, Rogers et al. (2010) hypothesized that a microbial lung community

might exist and can be considered as a unique, distinct pathogenic entity. The culture-independent microbial detection method, 16S ribosomal RNA (rRNA) gene sequencing, demonstrated the existence of the pulmonary microbiome, in both healthy (Erb-Downward et al., 2011; Morris et al., 2013; Twigg et al., 2013) and disease populations (Lozupone et al., 2013; Zemanick, Sagel, & Harris, 2011).

This paper is motivated by longitudinal microbiome studies. For instance, the lung human immunodeficiency virus (HIV) microbiome project studies the respiratory microbiome of HIV-infected patients and how the highly active antiretroviral therapy (HAART) may alter its construction (Twigg et al., 2016). A longitudinal cohort of HIV-infected subjects was collected before and up to 3 years after starting HAART. For a quantitative phenotype in a longitudinal design, we propose the model

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + h(\mathbf{G}) + \boldsymbol{\varepsilon}, \\ \mathbf{b} &\sim \mathcal{N}(0, \sigma_d^2 \mathbf{I}_n), \quad h(\mathbf{G}) \sim \mathcal{N}(0, \sigma_g^2 \mathbf{K}), \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma_e^2 \mathbf{I}_n), \end{aligned} \quad (1)$$

where \mathbf{y} , \mathbf{X} , \mathbf{G} , and $\boldsymbol{\varepsilon}$ are the vertically stacked vectors/matrices of individual-level \mathbf{y}_i , \mathbf{X}_i , \mathbf{G}_i , and $\boldsymbol{\varepsilon}_i$. \mathbf{y}_i is a vector of n_i repeated measures of a quantitative phenotype for individual i . \mathbf{X}_i is the $n_i \times p$ covariate matrix. \mathbf{G}_i is an $n_i \times u$ operational taxonomic unit (OTU) abundance matrix for individual i where u is the total number of OTUs. These OTUs are related by a known phylogenetic tree. $\boldsymbol{\varepsilon}_i$ is an $n_i \times 1$ vector of the random error. \mathbf{Z} is a block diagonal matrix with $\mathbf{1}_{n_i}$ on its diagonal. $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects and $\mathbf{b} = (b_i)$ is the subject-specific random effects. \mathbf{K} is a kernel matrix capturing distances between individuals, for example, the UniFrac distance (Lozupone & Knight, 2005) or the Bray–Curtis dissimilarity (Bray & Curtis, 1957; Web Appendix A). \mathbf{b} , $h(\mathbf{G})$ and $\boldsymbol{\varepsilon}$ are jointly independent; therefore,

$$\text{Var}(\mathbf{y}) = \sigma_d^2 \mathbf{Z}\mathbf{Z}' + \sigma_g^2 \mathbf{K} + \sigma_e^2 \mathbf{I}_n, \quad (2)$$

where σ_d^2 is the phenotypic variance due to the correlation of repeated measurements, σ_g^2 is the phenotypic variance explained by the microbiome, and σ_e^2 is the within-subject variance that cannot be explained by the microbiome and repeated measurements. The detection of overall microbiome association is to test $H_0: \sigma_g^2 = 0$ versus $H_A: \sigma_g^2 > 0$. When $\sigma_d^2 = 0$, Model (1) reduces to the microbiome regression-based kernel association test (MiRKAT; J. Chen, Chen, Zhao, Wu, & Schaid, 2016; Zhan et al., 2017; Zhao et al., 2015). In the longitudinal setting, the extra variance component σ_d^2 is necessary to capture the correlation between repeated measurements.

After the overall association is identified, localization of the signal to a specific component of the microbial community is essential for downstream mechanistic studies and drug discoveries. For instance, Jangi et al. (2016) found that multiple sclerosis patients had significantly increased abundance of the

phylum *Euryarchaeota*. However, such fine cluster effects can be tagged by other correlated microbials in the community (Gilbert et al., 2016), leading to false-positive discoveries. To detect association from specific taxonomic clusters, distances and kernel matrices can be formulated using abundances and tree information from specific clusters. Overall microbiome effects are then partitioned into different clusters at the same taxonomic level. That is,

$$\text{Var}(\mathbf{y}) = \sigma_d^2 \mathbf{Z}\mathbf{Z}' + \sum_i \sigma_{g_i}^2 \mathbf{K}_i + \sigma_e^2 \mathbf{I}_n, \quad (3)$$

where $\sum_i \sigma_{g_i}^2 \mathbf{K}_i$ is the summation of all microbiome clusters. We are now interested in testing effects from a specific taxonomic cluster: $H_0: \sigma_{g_i}^2 = 0$ versus $H_A: \sigma_{g_i}^2 > 0$.

Current methods for testing the null variance component in Models (2) and (3) are based on either asymptotics or parametric bootstrap. Under the assumption that the response variable vector can be partitioned into independent subvectors and the number of independent subvectors is sufficient, asymptotic null distribution of the likelihood ratio, Wald, and score tests is available (Self & Liang, 1987; Silvapulle & Sen, 2011; Stram & Lee, 1994). However, the asymptotic approximation deteriorates when the data are highly correlated without a sufficient number of independent blocks. Let m be the total number of phenotypic variance components except the error variance component. When $m = 1$, Crainiceanu and Ruppert (2004) developed a computational procedure for obtaining the approximate finite-sample distribution of the likelihood ratio and restricted likelihood ratio test (LRT) statistics. Greven, Crainiceanu, Küchenhoff, and Peters (2008) provided a pseudolikelihood-heuristic extension of this method to the $m > 1$ situation. Later Drikvandi, Verbeke, Khodadadi, and PartoviNia (2013) proposed a permutation test that does not depend on the distribution of the random effects and errors except for their mean and variance and can be applied to the $m > 1$ situation. However, the permutation test is computationally prohibitive for high dimensional tests. Qu, Guennel, and Marshall (2013) proposed a test statistic that is the weighted sum of the scores from the profile likelihood. Their method considered testing a subset of the variance components to be zero. When $m = 1$, Qu et al.'s (2013) method is exact; when $m > 1$, their test relies on asymptotic theory. Score-based tests can be less powerful than the LRTs, especially when sample sizes are limited as in most of the sequencing studies. Saville and Herring (2009) developed yet another type of test

based on the Bayes factors using the Laplace approximation. It cannot be easily extended to multiple random effects and relies on the subjective choice of the prior distribution of parameters. Others have suggested procedures based on Markov chain Monte Carlo methods (Z. Chen & Dunson, 2003; Kinney & Dunson, 2007), but they can be time-consuming, especially when the number of random effects is large.

In this study, we propose methods of performing the exact LRT (eLRT), the exact restricted LRT (eRLRT), and the exact score test (eScore) of a variance component being zero for the finite sample. Our approach combines the corresponding exact tests for the $m = 1$ case with a strategy of reducing the $m > 1$ case to the $m = 1$ case (Christensen, 1996; Ofversten, 1993). Our method is the first one that provides the eLRT, eRLRT, and eScore for testing the zero variance component when multiple variance components are present ($m > 1$).

2 | METHODS

2.1 | Exact tests with one variance component under H_0

We briefly review the three exact tests, eLRT, eRLRT, and eScore, for testing $H_0: \sigma_1^2 = 0$ in the model

$$\mathbf{V} = \sigma_e^2 \mathbf{I}_n + \sigma_1^2 \mathbf{V}_1. \quad (4)$$

Note the change of notation for general modeling. In the motivating microbiome example, $\sigma_1^2 = \sigma_g^2$ and $\mathbf{V}_1 = \mathbf{K}$, the kernel matrix calculated from microbiome abundances. A slight extension allows for testing the more general case $\mathbf{V} = \sigma_e^2 \mathbf{V}_0 + \sigma_1^2 \mathbf{V}_1$, where $\mathbf{V}_0 \in \mathbb{R}^{n \times n}$ is a known positive semidefinite matrix. Let $t = \text{rank}(\mathbf{V}_0)$. Given the thin eigen-decomposition $\mathbf{V}_0 = \mathbf{U}\mathbf{D}\mathbf{U}'$, define $\mathbf{T} = \mathbf{D}^{-1/2}\mathbf{U}' \in \mathbb{R}^{t \times n}$. (Only t column vectors of \mathbf{U} will be computed in thin eigen-decomposition.) Then $\mathbf{T}\mathbf{y} \sim \mathcal{N}(\mathbf{T}\mathbf{X}\boldsymbol{\beta}, \sigma_e^2 \mathbf{I}_t + \sigma_1^2 \mathbf{T}\mathbf{V}_1\mathbf{T}')$ and the eLRT and eRLRT (Crainiceanu & Ruppert, 2004) or the eScore test (Zhou, Hu, Qiao, Cho, & Zhou, 2016) can be applied to $\mathbf{T}\mathbf{y}$.

Let $\lambda = \sigma_1^2/\sigma_e^2$ be the signal-to-noise ratio, $s = \text{rank}(\mathbf{X})$, and write the covariance as $\mathbf{V} = \sigma_e^2(\mathbf{I}_n + \lambda\mathbf{V}_1) = \sigma_e^2\mathbf{V}_\lambda$. The model parameters are $(\boldsymbol{\beta}, \sigma_e^2, \lambda)$. Testing $H_0: \sigma_1^2 = 0$ is equivalent to testing $H_0: \lambda = 0$. The log-likelihood function is

$$L(\boldsymbol{\beta}, \sigma_e^2, \lambda) = -\frac{n}{2} \ln \sigma_e^2 - \frac{1}{2} \ln \det(\mathbf{V}_\lambda) - \frac{1}{2\sigma_e^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{V}_\lambda^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

The LRT statistic is

$$\begin{aligned} LRT &= 2 \sup_{H_1} L(\boldsymbol{\beta}, \sigma_e^2, \lambda) - 2 \sup_{H_0} L(\boldsymbol{\beta}, \sigma_e^2, \lambda) \\ &= \sup_{\lambda \geq 0} \{n \ln \mathbf{y}' \mathbf{A}_0 \mathbf{y} - n \ln \mathbf{y}' \mathbf{A}_\lambda \mathbf{y} \\ &\quad - \ln \det(\mathbf{V}_\lambda)\}, \end{aligned}$$

where $\mathbf{P}_X = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the projection matrix onto the column space $C(\mathbf{X})$, $\mathbf{A}_0 = \mathbf{I} - \mathbf{P}_X$, and $\mathbf{A}_\lambda = \mathbf{V}_\lambda^{-1} - \mathbf{V}_\lambda^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}_\lambda^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}_\lambda^{-1}$. Let $\{\xi_1, \dots, \xi_\ell\}$ be the positive eigenvalues of \mathbf{V}_1 and $\{\mu_1, \dots, \mu_k\}$ the positive eigenvalues of $\mathbf{A}_0\mathbf{V}_1\mathbf{A}_0$. Then

$$LRT \stackrel{\mathcal{D}}{=} \sup_{\lambda \geq 0} \left\{ n \ln \frac{\sum_{i=1}^{n-s} w_i^2}{\sum_{i=1}^k (w_i^2/(1 + \lambda\mu_i)) + \sum_{i=k+1}^{n-s} w_i^2} - \sum_{i=1}^l \ln(1 + \lambda\xi_i) \right\},$$

where, under the null, w_i are $(n - s)$ independent standard normals. Under the alternative, $w_i \sim \mathcal{N}(0, 1 + \lambda\mu_i)$ for $i = 1, \dots, k$, $w_i \sim \mathcal{N}(0, 1)$ for $i = k + 1, \dots, n - s$, and they are jointly independent. The null distribution can be obtained from computer simulation. A computationally efficient χ^2 approximation algorithm is given in the Supporting Information Material (Web Appendix B). The same derivation can be carried out for the eRLRT, in which case

$$\begin{aligned} RLRT & \stackrel{\mathcal{D}}{=} \sup_{\lambda \geq 0} \left\{ (n - s) \ln \frac{\sum_{i=1}^{n-s} w_i^2}{\sum_{i=1}^k (w_i^2/(1 + \lambda\mu_i)) + \sum_{i=k+1}^{n-s} w_i^2} \right. \\ & \quad \left. - \sum_{i=1}^k \ln(1 + \lambda\mu_i) \right\}. \end{aligned}$$

The null distribution generation for eRLRT is shown in Web Appendix B. Algorithms 1 and 2 in Web Appendix B contain a univariate optimization for each simulated point from the null distribution and can be computationally intensive for obtaining extremely small p -values. To further reduce the computational burden, we adopt the Satterthwaite method to approximate the null distributions (Zhou et al., 2016).

For eScore, it is easier to work with the original parameterization $\mathbf{V} = \sigma_e^2 \mathbf{I}_n + \sigma_1^2 \mathbf{V}_1$. The (Rao) score statistic is based on $\mathbf{I}_{\sigma_1^2, \sigma_e^2}^{-1} ((\partial/\partial\sigma_1^2)L)^2$, where the information matrix

$\mathbf{I}_{\sigma_1^2, \sigma_1^2} = E(-(\partial^2/\partial\sigma_1^2\partial\sigma_1^2)L)$ and score function $(\partial/\partial\sigma_1^2)L$ are evaluated at the maximum likelihood estimator under the null. The resultant test rejects the null when

$$S = \max \left\{ \frac{\mathbf{y}'(\mathbf{I} - \mathbf{P}_X)\mathbf{V}_1(\mathbf{I} - \mathbf{P}_X)\mathbf{y}}{\mathbf{y}'(\mathbf{I} - \mathbf{P}_X)\mathbf{y}}, \frac{\text{tr}(\mathbf{K})}{n} \right\}$$

is large. Let $\{\mu_1, \dots, \mu_k\}$ be the positive eigenvalues of $(\mathbf{I} - \mathbf{P}_X)\mathbf{V}_1(\mathbf{I} - \mathbf{P}_X)$. Then

$$S \stackrel{\mathcal{D}}{=} \max \left\{ \frac{\sum_{i=1}^k \mu_i w_i^2}{\sum_{i=1}^{n-s} w_i^2}, \frac{\text{tr}(\mathbf{K})}{n} \right\},$$

where w_i are $n - s$ independent standard normals. The null distribution can be obtained from computer simulation or inverting the characteristic function (Zhou et al., 2016). Both options, simulation and approximation of null distribution, are available in our program, <https://github.com/JingZhai63/VCmicrobiome>.

2.2 | Exact tests with more than one variance component under H_0

In this section, we consider the situation when $\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$ with $\mathbf{V} = \sigma_0^2 \mathbf{I} + \sigma_1^2 \mathbf{V}_1 + \dots + \sigma_m^2 \mathbf{V}_m$, $m > 1$. We are interested in testing $H_0: \sigma_m^2 = 0$ versus $H_A: \sigma_m^2 > 0$. We follow a strategy to reduce the problem to the $m = 1$ Case 2 (Christensen, 1996; Ofversten, 1993).

We first obtain an orthonormal basis $(\mathbf{Q}_0, \mathbf{Q}_1, \dots, \mathbf{Q}_m, \mathbf{Q}_{m+1})$ of \mathbb{R}^n such that \mathbf{Q}_0 is an orthonormal basis of $C(\mathbf{X})$, \mathbf{Q}_1 is an orthonormal basis of $C(\mathbf{X}, \mathbf{V}_1) - C(\mathbf{X})$, \mathbf{Q}_i is an orthonormal basis of $C(\mathbf{X}, \mathbf{V}_1, \dots, \mathbf{V}_i) - C(\mathbf{X}, \mathbf{V}_1, \dots, \mathbf{V}_{i-1})$ for $i = 2, \dots, m$, and \mathbf{Q}_{m+1} is an orthonormal basis of $\mathbb{R}^n - C(\mathbf{X}, \mathbf{V}_1, \dots, \mathbf{V}_m)$. Denoting their corresponding ranks by r_0, \dots, r_{m+1} . If $r_m > 0$, that is $C(\mathbf{X}, \mathbf{V}_1, \dots, \mathbf{V}_{m-1}) \subsetneq C(\mathbf{X}, \mathbf{V}_1, \dots, \mathbf{V}_{m-1}, \mathbf{V}_m)$, then $\mathbf{Q}'_m \mathbf{Y} \sim \mathcal{N}(\mathbf{0}, \sigma_e^2 \mathbf{I}_{r_m} + \sigma_m^2 \mathbf{Q}'_m \mathbf{V}_m \mathbf{Q}_m)$ and eLRT, eRLRT, and eScore can be applied to $\mathbf{Q}'_m \mathbf{Y}$. The order of $\mathbf{V}_1, \dots, \mathbf{V}_m$ does not matter. If $r_m = 0$, that is $C(\mathbf{X}, \mathbf{V}_1, \dots, \mathbf{V}_{m-1}) = C(\mathbf{X}, \mathbf{V}_1, \dots, \mathbf{V}_m)$, we construct a test based on the transformed data $\mathbf{Q}'_{m-1} \mathbf{Y} + \mathbf{CQ}'_{m+1} \mathbf{Y}$. Without loss of generality we assume \mathbf{Q}_{m-1} is nontrivial. If $r_{m-1} = 0$, we use \mathbf{Q}_{m-2} and so on. We consider the following cases:

1. If $\mathbf{Q}'_{m-1} \mathbf{V}_m = \mathbf{0}$, for example, when $C(\mathbf{V}_m) \subset C(\mathbf{X}, \mathbf{V}_1, \dots, \mathbf{V}_{m-2})$, then this test cannot be performed. Shifting the order of $\mathbf{X}, \mathbf{V}_1, \dots, \mathbf{V}_{m-1}$ might solve the issue.

2. If $\mathbf{Q}'_{m-1} \mathbf{V}_{m-1} \mathbf{Q}_{m-1} = \gamma \mathbf{I}_{r_{m-1}}$ and $\gamma \neq 0$, then

$$\begin{aligned} \mathbf{Q}'_{m-1} \mathbf{Y} &\sim \mathcal{N}(\mathbf{0}, \sigma_e^2 \mathbf{I}_{r_{m-1}} + \sigma_{m-1}^2 \mathbf{Q}'_{m-1} \mathbf{V}_{m-1} \mathbf{Q}_{m-1} \\ &\quad + \sigma_m^2 \mathbf{Q}'_{m-1} \mathbf{V}_m \mathbf{Q}_{m-1}) \\ &= \mathcal{N}(\mathbf{0}, (\sigma_e^2 + \gamma \sigma_{m-1}^2) \mathbf{I}_{r_{m-1}} \\ &\quad + \sigma_m^2 \mathbf{Q}'_{m-1} \mathbf{V}_m \mathbf{Q}_{m-1}), \end{aligned}$$

which is the case (4). The eLRT, eRLRT, and eScore can be applied without using the $\mathbf{CQ}'_{m+1} \mathbf{y}$ piece.

3. If $\mathbf{Q}'_{m-1} \mathbf{V}_{m-1} \mathbf{Q}_{m-1} \neq \gamma \mathbf{I}_{r_{m-1}}$, then the test requires the $\mathbf{CQ}'_{m+1} \mathbf{y}$ term. $\mathbf{CQ}'_{m+1} \mathbf{y}$ has the distribution $\mathbf{CQ}'_{m+1} \mathbf{Y} \sim \mathcal{N}(\mathbf{0}, \sigma_e^2 \mathbf{C}\mathbf{C}')$. Since $\mathbf{Q}'_{m-1} \mathbf{Y} \perp \mathbf{CQ}'_{m+1} \mathbf{Y}$, we pick \mathbf{C} such that

$$\mathbf{C}\mathbf{C}' = \zeta^{-1} \mathbf{Q}'_{m-1} \mathbf{V}_{m-1} \mathbf{Q}_{m-1} - \mathbf{I}_{r_{m-1}},$$

where the scalar ζ is chosen such that $\zeta^{-1} \mathbf{Q}'_{m-1} \mathbf{V}_{m-1} \mathbf{Q}_{m-1} - \mathbf{I}_{r_{m-1}}$ is positive semidefinite. Let $\mathbf{Q}'_{m-1} \mathbf{V}_{m-1} \mathbf{Q}_{m-1} = \mathbf{W}\boldsymbol{\Lambda}\mathbf{W}' = \mathbf{W}\text{diag}(\delta_i)\mathbf{W}'$ be the eigen-decomposition, ζ be the smallest positive eigenvalue, and $\mathbf{C} = \mathbf{W}\text{diag}(\sqrt{\delta_i/\zeta} - 1)$. Then the transformed data

$$\begin{aligned} \mathbf{Q}'_{m-1} \mathbf{Y} + \mathbf{CQ}'_{m+1} \mathbf{Y} &\sim \mathcal{N}(\mathbf{0}, (\sigma_{m-1}^2 \\ &\quad + \sigma_e^2/\zeta) \mathbf{Q}'_{m-1} \mathbf{V}_{m-1} \mathbf{Q}_{m-1} + \sigma_m^2 \mathbf{Q}'_{m-1} \mathbf{V}_m \mathbf{Q}_{m-1}) \end{aligned}$$

and the test for case (2.1) can be applied. A larger ζ leads to a higher signal-to-noise ratio $(\sigma_m^2/(\sigma_{m-1}^2 + \sigma_e^2/\zeta))$ and thus a more powerful test. Finally we test $H_0: \sigma_m^2 = 0$ using the eLRT, eRLRT, or eScore test on the transformed data:

$$\begin{aligned} \boldsymbol{\Lambda}^{-1/2} \mathbf{W}' (\mathbf{Q}'_{m-1} + \mathbf{CQ}'_{m+1}) \mathbf{Y} &\sim \mathcal{N}(\mathbf{0}, (\sigma_{m-1}^2 + \sigma_e^2/\zeta) \mathbf{I}_{r_{m-1}} \\ &\quad + \sigma_m^2 \boldsymbol{\Lambda}^{-1/2} \mathbf{W}' \mathbf{Q}'_{m-1} \mathbf{V}_m \mathbf{Q}_{m-1} \mathbf{W} \boldsymbol{\Lambda}^{-1/2}). \end{aligned}$$

We note that if in some applications that matrices have high or full rank, consuming most or all available degrees of freedom after the above reduction strategy. One could proceed with a low rank approximation. For example, if $m = 2$ and \mathbf{V}_1 has high or full rank, one could find the rank $r_{\mathbf{V}_1}$ approximation of \mathbf{V}_1 as follows: Let $r_{\mathbf{K}} = \text{rank}(\mathbf{V}_2)$, \mathbf{Q}_0 is an orthonormal basis of $C(\mathbf{X})$, and $r_0 = \text{rank}(\mathbf{Q}_0)$. A rank $r_{\mathbf{V}_1} \leq \lfloor (n - r_0 - r_{\mathbf{K}})/2 \rfloor$ approximation of \mathbf{V}_1 is enough to perform testing. Details can be found in the software's documentation (<http://vcmicrobiomejl.readthedocs.io/en/latest/>).

TABLE 1 Simulation configurations

Sample size	Kernel type	Clustering	# Repeat	σ_g^2	Method
<i>Scenario 1: Testing the overall microbiome effect</i>					
100	$\mathbf{K}_W, \mathbf{K}_U, \mathbf{K}_{VAW}, \mathbf{K}_\alpha$	None	2	0–1.5	eRLRTeScore
100	$\mathbf{K}_W, \mathbf{K}_U, \mathbf{K}_{VAW}, \mathbf{K}_\alpha$	None	1	0–1.5	eRLRTeLRTeScore
<i>Scenario 2: Localizing fine microbiome cluster effects</i>					
100	\mathbf{K}_W	Yes	2	0–1.5	eRLRTeScore
100	\mathbf{K}_W	Yes	1	0–1.5	eRLRTeScore
<i>Scenario 3: Comparing with existing methods</i>					
20, 30, 50, 100	\mathbf{K}_W	None	2	0–1.5	eRLRTeScoreLinScore
20, 30, 50, 100	\mathbf{K}_W	None	1	0–1.5	eRLRTeLRTeScoreLinScoreMiRKAT

Note. For all simulations, $\sigma_e^2 = 1$ and $\sigma_d^2 = 0$ when the number of repeats (# Repeat) = 1 or $\sigma_e^2 = 1$ and $\sigma_d^2 = 0.6$ when the number of repeats > 1. There are 2,964 OTUs presented in the simulated count data. A phylogenetic tree generated using the real pulmonary microbiome data is used for kernel calculation and phenotype simulations. \mathbf{K}_W : weighted UniFrac kernel; \mathbf{K}_U : unweighted UniFrac kernel; \mathbf{K}_{VAW} : variance adjusted weighted UniFrac kernel; \mathbf{K}_α : generalized UniFrac kernels with $\alpha = 0$ and 0.5
OTU: operational taxonomic unit.

3 | SIMULATION

We evaluate the performance of the exact tests for the longitudinal microbiome study in three simulation scenarios (Table 1).

The longitudinal microbiome count data with two repeated measurements are simulated using the R package zero-inflated beta random effect (ZIBR) model (E. Z. Chen & Li, 2016). To mimic features of real microbiome datasets, the phylogenetic structure and average count information are extracted from the real HIV longitudinal pulmonary microbiome data. This microbiome data set contains 30 samples, each with 2–4 repeated measurements: baseline, 4 weeks, 1 year, and 3 years (Twigg et al., 2016). OTU alignment at the species level was produced by software Mothur (<https://www.mothur.org>; Schloss et al., 2009) and the Basic Local Alignment Search Tool (BLAST; <https://blast.ncbi.nlm.nih.gov/Blast.cgi>; Altschul, Gish, Miller, Myers, & Lipman, 1990) in the Ribosomal Database Project (RDP) 16S database release 11.4 (Maidak et al., 1996). The phylogenetic tree at the OTU level is generated using the RDP classifier (Twigg et al., 2016). We construct the higher taxon level, for example, phylum, using the phylogenetic tree generator phyloT (<http://phyloT.biobyte.de/>; Letunic & Bork, 2007, 2011) and the NCBI database taxonomy (<https://www.ncbi.nlm.nih.gov/taxonomy>; Federhen, 2012). There are 2,964 OTUs in total, 292 genera, and 24 phyla. Different distance measures are calculated using our Julia package PhylogeneticDistance (<https://github.com/JingZhai63/PhylogeneticDistance.jl>). The definition of different distance measures and the details of simulation of microbiome abundances are provided in Web Appendices A and C.

Phenotypes are generated under three different scenarios. For all three scenarios, two covariates are

included in the model. One of them is correlated with microbiome abundances. For individual i , $X_{1i} \sim \mathcal{N}(0, 1)$ and $X_{2i} = h(\mathbf{G}_i)_{baseline} + \mathcal{N}(0, 1)$. Their effects are $\beta_1 = \beta_2 = 0.1$. We set the within-individual variance to $\sigma_e^2 = 1$. For longitudinal data simulation, the between individual variance σ_d^2 is set to 0.6. This corresponds to 60% of the overall baseline phenotypic variance (Twigg et al., 2016).

3.1 | Scenario 1: Testing the overall microbiome effect

Longitudinal responses are generated using the model, $\mathbf{y} \sim \mathcal{N}(\mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2, \sigma_d^2\mathbf{Z}\mathbf{Z}' + \sigma_g^2\mathbf{K} + \sigma_e^2\mathbf{I})$, where $\sigma_g^2 = 0, 0.2, 0.5, 1.0,$ and 1.5 . We compare the performance of five different distance measures: unweighted UniFrac (Lozupone & Knight, 2005), weighted UniFrac distance (Lozupone, Hamady, Kelley, & Knight, 2007), variance adjusted weighted (VAW) UniFrac distance (Chang, Luan, & Sun, 2011), and generalized UniFrac distance with parameter $\alpha = 0.0$ and 0.5 (J. Chen et al., 2012).

3.2 | Scenario 2: Localizing fine microbiome cluster effects

We cluster OTUs into six phyla, *Actinobacteria*, *Bacteroidetes*, *Fusobacteria*, *Proteobacteria*, *Firmicutes*, and *other*. We assume that only cluster *other*, $h(\mathbf{G}_i)$, has effects. That is $\mathbf{y} \sim \mathcal{N}(\mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2, \sigma_d^2\mathbf{Z}\mathbf{Z}' + \sum_{l=1}^6 \sigma_{g_l}^2\mathbf{K}_l + \sigma_e^2\mathbf{I})$, where $\sigma_{g_l}^2 = 0, 0.5, 1.5$ and $\sigma_{g_1}^2 = 0$ for $l = 2, \dots, 6$. Due to the correlation between phyla, marginal tests of five individual phyla may show a false signal if we do not adjust for the effects of $h(\mathbf{G}_i)$. We show that testing of variance components in a joint model has the correct type I error.

3.3 | Scenario 3: Comparing with existing methods

We compare our method with MiRKAT (Zhao et al., 2015) and LinScore (Qu et al., 2013). As MiRKAT can only be used for testing overall microbiome effects for cross-sectional designs, we first compare the three methods when $\sigma_d^2 = 0$. Responses are generated according to simulation Scenario 1, where $\sigma_g^2 = 0, \dots, 1.5$.

In Scenarios 1 and 2, the sample size is fixed at $n = 100$. In Scenario 3, we compare the performance of the three methods under sample sizes 20, 30, 50, and 100. The performance of five different kernels is compared in Scenario 1. For Scenarios 2 and 3, we focus on the weighted UniFrac distance kernel only, which demonstrates a higher power than the other kernels in Scenario 1. Thousand Monte Carlo replicates are generated for all simulations and we use the nominal significance level 0.05 to evaluate Type 1 error and power.

4 | RESULTS

4.1 | Simulation results

4.1.1 | Scenario 1: Testing the overall microbiome effect

The type I error rate of eRLRT, eLRT, and eScore tests with various distance kernel matrices using real longitudinal OTU count data are shown in Table 2. Figure 1 shows the power comparison with different kernels. In Figure 1a,c, five different kernels are constructed using OTU count data directly. In Figure 1b,d, OTU counts are summarized at the phylum level for kernel calculations.

Figure 1 shows that kernel type greatly impacts the power. The weighted UniFrac kernel yields the highest power and the unweighted UniFrac kernel has the least

TABLE 2 Scenario 1: Type I error of eLRT, eRLRT and eScore for detecting the overall microbiome effects

Simulation Design	Method	Kernel type				
		K_W	K_U	K_{VAW}	K_0	$K_{0.5}$
Cross-sectional	eRLRT	0.046	0.043	0.045	0.048	0.047
	eLRT	0.046	0.043	0.051	0.052	0.046
	eScore	0.039	0.031	0.047	0.045	0.042
Longitudinal	eRLRT	0.041	0.053	0.045	0.041	0.042
	eScore	0.034	0.048	0.048	0.050	0.045

Note. Five distance measures, weighted UniFrac kernel (K_W), unweighted UniFrac kernel (K_U), variance adjusted weighted UniFrac kernel (K_{VAW}), and generalized UniFrac kernels with $\alpha = 0$ (K_0) and 0.5 ($K_{0.5}$) are compared

eLRT: exact likelihood ratio test; eRLRT: exact restricted likelihood ratio test; eScore: exact score test; OTU: operational taxonomic unit.

power (Figure 1a,c). The pattern of the power increase with effect size differs according to which taxon-level count data are used to calculate the kernel. The power of five kernels became similar to each other in Figure 1b,d. Furthermore, the power of the unweighted UniFrac kernel K_{UW} , which is the least powerful kernel in Figure 1a,c, greatly improves in Figure 1b,d. The reason is when the reads are summarized at the higher phylum level, the difference of abundance between each phylum is less notable. The less variability of abundance between lineages, the more similar the power of detecting microbiome association. As expected, reducing variance components leads to reduced degrees of freedom for association testing and the test is slightly less powerful in the longitudinal study compared to the cross-sectional study given the same effect size in this simulation.

4.1.2 | Scenario 2. Localizing fine microbiome cluster effects

Table 3 shows the Type 1 error rates for testing the microbiome effect at the phylum level, with and without adjusting for the effect contributed by cluster, *other*. Most Type 1 error rates are inflated when not adjusting for cluster *other* effects. In the cross-sectional design, the Type I error rates of *Bacteroidetes* and *Proteobacteria* stay correct due to its weak correlation with cluster *other* (Pearson correlation = 0.04, 0.11 with $p = 0.70, 0.24$, respectively). After adjustment, Type 1 error rates stay correct even when confounding effects are large (Table 3).

In practice, symbiosis of bacteria causes correlation between them (Dickson, Erb-Downward, & Huffnagle, 2013; Xu et al., 2007; Zeng et al., 2016). Precise medication that targets specific pathogens can minimize the damage to essential symbiotic microbial species, and preserve community structure and function in the healthy (and developing) microbiome (Blaser, 2016; Hicks, Taylor, & Hunkler, 2013). Simulation Scenario 2 demonstrates that our method is capable of localizing fine microbiome cluster effects.

4.1.3 | Scenario 3: Comparing with existing methods MiRKAT and LinScore

Table 4 presents the Type 1 error rate and power for eRLRT, eLRT, eScore, MiRKAT, and LinScore tests in detecting overall microbiome effects. The power is shown for both cross-sectional and longitudinal studies with sample size from 20 to 100. eRLRT and eLRT outperform LinScore and MiRKAT in baseline simulation studies. For repeated measurements, eRLRT outperforms LinScore under small sample sizes (e.g., $n \leq 50$). Under sample size $n = 100$, eRLRT has similar or slightly higher power compared to LinScore when the association strength is weak. Microbiome

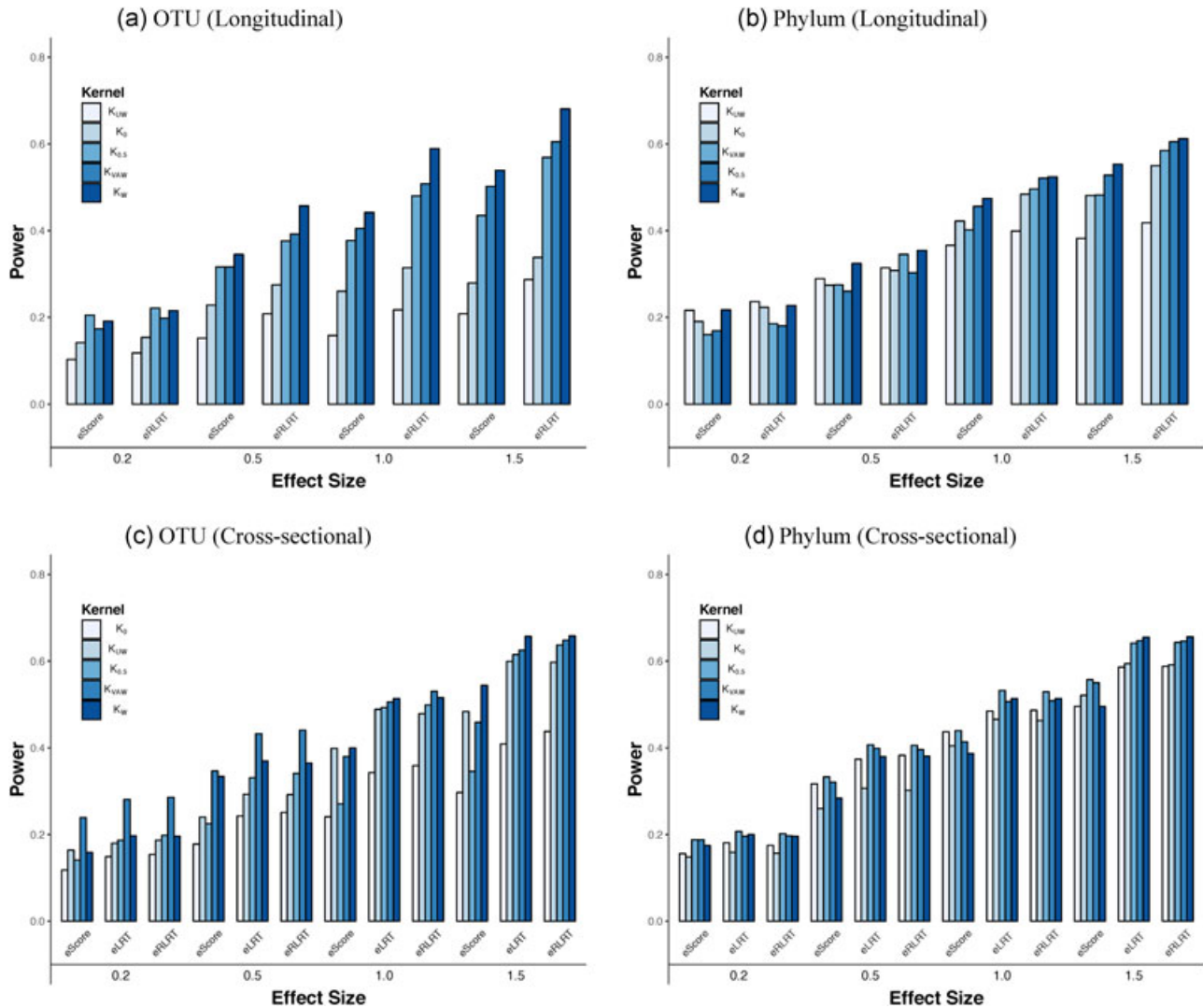


FIGURE 1 Scenario 1: Power of eRLRT, eLRT, and eScore using different distance measures. Figures to the left shows results where the OTU counts are used to calculate distances, and figures to the right shows that OTU counts are summarized at the phylum level to construct the distances. K_0 , $K_{0.5}$, K_W , K_U , and K_{VAW} represent the generalized UniFrac distance with $\alpha = 0, 0.5$, weighted UniFrac distance, unweighted UniFrac distance, and variance adjusted weighted UniFrac distance, respectively. eLRT: exact likelihood ratio test; eRLRT: exact restricted likelihood ratio test; eScore: exact score test; OTU: operational taxonomic unit

studies usually have a limited sample size due to the high cost. The higher power of the exact tests at small sample sizes will be particularly valuable for biologists and physicians for identifying the associated microbiome clusters.

4.2 | Analysis of longitudinal pulmonary microbiome data

It is well known that HIV infection is associated with alterations in the respiratory microbiome (Twigg et al., 2016). However, due to limited investigation, the clinical implications of lung microbial dysbiosis are currently unknown. As an initial step to reveal the connection of the respiratory microbiome to pulmonary complications in HIV-infected individuals, we investigate the relationship

between pulmonary function and the respiratory microbiota profiles in the bronchoalveolar lavage (BAL) fluid of 30 HIV-infected patients at the advanced stage (baseline mean CD4 count, 262 cells/mm³). Their acellular BAL fluid was sampled at baseline, 4 weeks, 1 year, and 3 years. 16S rRNA gene sequencing technology was used to quantify pulmonary microbiota. The details of microbiome composition have been discussed in Section 2.3. Pulmonary function is measured by spirometry and diffusion capacity tests. Spirometry tests measure how much and how quickly air can move out of lung. Typical spirometry tests include forced vital capacity, forced expiratory volume in 1 s (FEV1), and average forced expiratory flow (FEF). Diffusion capacity of the lungs for carbon monoxide (DLCO) measures how much oxygen travels from the lung alveoli

TABLE 3 Scenario 2: Type I error rate of localizing fine microbiome cluster effects

Longitudinal design							
Phylum	No adjustment for Other			Adjustment for Other			
	Effect size σ_g^2			Effect size σ_g^2			
	0	0.5	1.5	0	0.5	1.5	
	eRLRT, eScore	eRLRT, eScore	eRLRT, eScore	eRLRT, eScore	eRLRT, eScore	eRLRT, eScore	
<i>Actinobacteria</i>	0.050, 0.038	0.108, 0.075	0.151, 0.100	0.049, 0.038	0.051, 0.048	0.033, 0.040	
<i>Bacteroidetes</i>	0.045, 0.040	0.060, 0.055	0.061, 0.055	0.041, 0.040	0.047, 0.042	0.042, 0.037	
<i>Firmicutes</i>	0.043, 0.043	0.049, 0.044	0.063, 0.067	0.042, 0.043	0.041, 0.043	0.052, 0.051	
<i>Fusobacteria</i>	0.052, 0.048	0.038, 0.041	0.060, 0.048	0.052, 0.048	0.045, 0.044	0.048, 0.037	
<i>Proteobacteria</i>	0.051, 0.046	0.041, 0.048	0.056, 0.050	0.049, 0.042	0.040, 0.035	0.053, 0.036	
Cross-sectional design							
<i>Actinobacteria</i>	0.041, 0.036	0.117, 0.065	0.111, 0.083	0.050, 0.040	0.052, 0.043	0.048, 0.035	
<i>Bacteroidetes</i>	0.051, 0.047	0.048, 0.049	0.051, 0.041	0.051, 0.041	0.048, 0.043	0.048, 0.037	
<i>Firmicutes</i>	0.037, 0.038	0.059, 0.052	0.068, 0.062	0.044, 0.038	0.051, 0.045	0.052, 0.048	
<i>Fusobacteria</i>	0.053, 0.050	0.070, 0.060	0.078, 0.065	0.052, 0.033	0.051, 0.041	0.048, 0.040	
<i>Proteobacteria</i>	0.042, 0.035	0.038, 0.042	0.053, 0.047	0.048, 0.047	0.049, 0.050	0.041, 0.033	

Note. Only cluster *Other* contains effects, 0, 0.5 and 1.5. Type 1 error rates are evaluated with or without adjustment for effect from cluster *Other*. The weighted UniFrac kernel is used. Top panel shows results from simulation using longitudinal data while the bottom panel shows results using cross-sectional data only eLRT: exact likelihood ratio test; eRLRT: exact restricted likelihood ratio test; eScore: exact score test.

to the blood stream. DLCO corrected for hemoglobin (DsbHb) and diffusion capacity corrected for alveolar volume and hemoglobin (DVAsbHb) are evaluated. Descriptive statistics of these measures are summarized in Web Appendix Table 1.

Exact tests and LinScore are used to study the association. Associations with p values less than 0.05 are reported to be significant. Covariates include gender, race, smoking status, CD4 counts, and HIV virus load (Table 5). The missing covariate is imputed by its mean. For the overall microbiome association test, no tests find significant associations. However at the phylum level, *Bacteroidetes* shows a significant association with spirometry while *Firmicutes* shows a significant association with diffusing capacity measures. Similar results have been reported by Hewitt and Molyneaux (2017) and Tunney et al. (2013). We then focus on analyzing genera from both phyla *Bacteroidetes* and *Firmicutes* given their important status in normal lungs (Cui et al., 2014). Only by eRLRT and eScore, genus *Prevotella*, *Porphyromonas*, and *Parvimonas* show significant effects on FEF and FEV1 (Table 5). Genus *Veillonella* shows a significant association with FEF. It appears that both *Parvimonas* and *Veillonella* in phylum *Firmicutes* are significantly associated with FEF and both genus *Prevotella* and *Porphyromonas* in phylum *Bacteroidetes* are significantly associated with FEF and FEV1. We therefore

perform the test in a joint model to localize the fine cluster effect. Interestingly, by eRLRT the significant association between genus *Parvimonas* and FEF still remains after adjusting for the effects from genus *Veillonella*. But the opposite is not true. This supports the findings of the previous studies that *Parvimonas* abundance changed in subjects with pulmonary disease (e.g., asthma or COPD) compared to the control group (Kim et al., 2018; Pragman, Kim, Reilly, Wendt, & Isaacson, 2012). However, either *Prevotella* or *Porphyromonas* lost its significance when adjusting for the other. This likely suggests that *Prevotella* and *Porphyromonas* are correlated and both tag effects to lung function. In comparison, LinScore only detects the significant microbiome effect of *Bacteroidetes* with FEF. Our results further support the conclusions from previous studies and sheds light for future clinical causality research (Segal et al., 2017; Twigg et al., 2016; Weiden et al., 2017). None of the tests (exact tests, LinScore, and MiRKAT) identify significant associations using only baseline data (results not shown). In conclusion, our exact tests provides innovative association evidence of pulmonary microbiome and lung function in the HIV-infected population, which have not been reported before. While the modeling is compelling, interpretation of the data and how correlations translate to meaningful clinical outcomes needs further study.

TABLE 4 Scenario 3: Comparing with existing methods

n	#Repeat	Method	Effect size (σ_g^2)								
			0	0.10	0.2	0.5	0.8	1.0	1.5		
20	1	eScore	0.045	0.059	0.050	0.074	0.078	0.079	0.104		
		eLRT	0.051	0.089	0.095	0.111	0.118	0.141	0.152		
		eRLRT	0.050	0.097	0.088	0.108	0.122	0.142	0.160		
		MiRKAT	0.048	0.056	0.046	0.071	0.069	0.077	0.104		
		LinScore	0.050	0.060	0.046	0.075	0.072	0.078	0.106		
	2	eScore	0.050	0.055	0.040	0.057	0.068	0.077	0.088		
		eRLRT	0.051	0.055	0.074	0.081	0.092	0.085	0.118		
		LinScore	0.049	0.057	0.063	0.055	0.072	0.078	0.090		
		30	1	eScore	0.043	0.059	0.050	0.074	0.078	0.079	0.104
				eLRT	0.046	0.089	0.095	0.111	0.118	0.141	0.152
eRLRT	0.052			0.097	0.088	0.108	0.122	0.142	0.160		
MiRKAT	0.055			0.056	0.046	0.071	0.069	0.077	0.104		
LinScore	0.054			0.060	0.046	0.075	0.072	0.078	0.106		
2	eScore		0.045	0.058	0.067	0.093	0.114	0.127	0.151		
	eRLRT		0.052	0.063	0.081	0.105	0.127	0.145	0.178		
	LinScore		0.046	0.054	0.061	0.076	0.088	0.132	0.134		
	50		1	eScore	0.036	0.070	0.071	0.118	0.151	0.164	0.240
				eLRT	0.048	0.084	0.094	0.135	0.188	0.214	0.306
eRLRT		0.049		0.086	0.088	0.127	0.192	0.201	0.307		
MiRKAT		0.047		0.065	0.069	0.114	0.156	0.183	0.257		
LinScore		0.045		0.070	0.077	0.124	0.176	0.189	0.267		
2		eScore	0.047	0.069	0.084	0.110	0.148	0.177	0.257		
		eRLRT	0.041	0.074	0.097	0.134	0.188	0.217	0.315		
		LinScore	0.051	0.063	0.096	0.156	0.205	0.261	0.333		
		100	1	eScore	0.050	0.096	0.165	0.304	0.383	0.390	0.532
				eLRT	0.052	0.114	0.191	0.377	0.472	0.516	0.664
eRLRT	0.049			0.105	0.195	0.375	0.460	0.510	0.661		
MiRKAT	0.051			0.093	0.181	0.329	0.427	0.483	0.622		
LinScore	0.048			0.106	0.194	0.347	0.439	0.507	0.630		
2	eScore		0.037	0.140	0.205	0.277	0.378	0.411	0.525		
	eRLRT		0.041	0.161	0.244	0.327	0.447	0.498	0.626		
	LinScore		0.046	0.121	0.214	0.347	0.451	0.545	0.652		

Note. Type 1 error rate and power from eLRT, eRLRT, eScore, LinScore, and MiRKAT at baseline when #Repeat = 1. When #Repeat = 2, only LinScore is compared with eRLRT and eScore. Sample sizes (n) range from 20 to 100 and effect sizes (σ_g^2) range from 0 to 1.5.

eLRT: exact likelihood ratio test; eRLRT: exact restricted likelihood ratio test; eScore: exact score test.

5 | DISCUSSION

In this report, motivated by a longitudinal pulmonary microbiome study, we develop and implement three computationally efficient exact variance component tests (eScore, eLRT, and eRLRT). Our method extends previous exact variance component tests to the case when the null hypothesis contains more than one variance component (Zhou et al., 2016). They can be applied to longitudinal

studies testing the overall microbiome effects, as well as cross-sectional studies identifying microbiome associations at the fine-grained level. The latter has been emerging as the focus of many current microbiome studies (Lloyd-Price et al., 2017; Nayfach, Rodriguez-Mueller, Garud, & Pollard, 2016; Truong, Tett, Pasolli, Huttenhower, & Segata, 2017). Unlike Qu et al. (2013) and Zhao et al.'s (2015) score test that uses moment-matching to approximate null distribution, our tests are exact in finite samples, therefore

TABLE 5 Application to the longitudinal pulmonary microbiome studies

PFT	Overall			Bacteroidetes			Firmicutes		
	eRLRT	eScore	LinScore	eRLRT	eScore	LinScore	eRLRT	eScore	LinScore
DVA	1.0	1.0	0.53	1.0	1.0	0.07	1.0	1.0	0.47
DsbHb	0.10	1.0	0.18	0.32	1.0	0.43	< 0.01	0.12	0.42
FEV1	0.26	1.0	0.63	0.02	0.05	0.79	1.0	1.0	0.22
FVC	0.11	1.0	0.22	0.01	0.03	0.53	1.0	1.0	0.23
FEF	0.17	0.34	0.82	0.21	0.20	0.04	0.07	0.07	0.47

PFT	Firmicutes			Veillonella			Parvimonas		
	eRLRT	eScore	LinScore	eRLRT	eScore	LinScore	eRLRT	eScore	LinScore
DVA	0.35	0.34	0.12	1.0	1.0	0.12	1.0	1.0	0.70
DsbHb	0.33	0.32	0.87	1.0	1.0	0.07	0.26	0.25	0.62
FEV1	0.05 (1.0)	0.05 (1.0)	0.75	0.03 (1.0)	0.03 (1.0)	0.16	0.30	0.31	0.25
FVC	0.12	0.10	0.50	0.17	0.16	0.25	1.0	1.0	0.90
FEF	0.05 (0.34)	0.05 (0.35)	0.37	< 0.01 (1.0)	< 0.01 (1.0)	0.06	0.03 (1.0)	0.04 (1.0)	0.09

Note. eRLRT, eScore, and LinScore are used to detect association. Genus *Porphyromonas* and *Prevotella* belong to phylum Bacteroidetes while genus *Veillonella* and *Parvimonas* belong to phylum Firmicutes. Upper panel shows the testing results at the phylum level, while the lower panel shows the results at the genus level. *p* Values less than 0.05 are highlighted in bold font. *p* Values in parenthesis show the results from a joint model where a significant genus in the same phylum is included

eRLRT: exact restricted likelihood ratio test; eScore: exact score test; FEF: average forced expiratory flow; FEV1: forced expiratory volume in 1 s; FVC: forced vital capacity.

beneficial to studies with a limited sample size. Compared to the score test, our eLRT and eRLRT tests can further boost power when the microbiome effects are weak. Simulation studies verify that our exact tests have the correct size and many innovative utilizations. In the application to the real longitudinal pulmonary microbiome study, only our exact tests detect multiple interesting genera associated with lung function. We then further demonstrated the ability of our exact tests to differentiate associated genera by using two real data examples. Although the derivation of eLRT and eRLRT require the normality assumption, a sensitivity simulation shows that even with a misspecified phenotypic distribution, like the t -distribution, our tests still preserve the correct Type I error rate (Web Appendix E, Table 2). The software package is implemented in an open-source, high-performance computing language Julia and is freely available. We offer unweighted, weighted, VAW, and generalized UniFrac distance calculation to further ease the computation and advance microbiome studies.

There are a few directions for future work. First, there are linear mixed effects models not of form (3), for example, those including both random intercepts and random slopes (Drikvandi et al., 2013). Our methods extend to these cases naturally and we defer them to future research. The second direction is to incorporate multiple types of kernels into exact tests. Finally, we consider extension to the generalized linear mixed effects models, although it can be challenging especially for LRT and RLRT. Score-based tests may be possible through penalized quasi-likelihood (H. Chen et al., 2016; Lin, 1997).

6 | SOFTWARE

The Julia package is freely available at <https://github.com/JingZhai63/VCmicrobiome>. In the real longitudinal data analysis with sample size 30 and 2,964 OTUs, the elapsed CPU times are 0.1 and 0.04 s for eRLRT and eScore, respectively. The analysis was performed by a MacBook Pro with 2.3 GHz Intel Core i7 processor and 8 GB 1600 MHz DDR3 memory.

ACKNOWLEDGMENTS

J. J. Z. is supported by NIH grant K01DK106116 and the Arizona Biomedical Research Commission (ABRC) grant. H. Z. is partially supported by NIH grants HG006139, GM105785, GM053275 and the NSF grant DMS-1645093.

ORCID

Jin J. Zhou  <http://orcid.org/0000-0001-7983-0274>

REFERENCES

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215, 403–410.
- Blaser, M. J. (2016). Antibiotic use and its consequences for the normal microbiome. *Science*, 352, 544–545.
- Bray, J. R., & Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, 27, 325–349.
- Chang, Q., Luan, Y., & Sun, F. (2011). Variance adjusted weighted UniFrac: A powerful beta diversity measure for comparing communities based on phylogeny. *BMC Bioinformatics*, 12, 118.
- Chen, E. Z., & Li, H. (2016). A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics*, 32, 2611–2617.
- Chen, H., Wang, C., Conomos, M. P., Stilp, A. M., Li, Z., Sofer, T., ... Celedón, J. C. (2016). Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *The American Journal of Human Genetics*, 98, 653–666.
- 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, 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, 5–19.
- Chen, Z., & Dunson, D. B. (2003). Random effects selection in linear mixed models. *Biometrics*, 59, 762–769.
- Christensen, R. (1996). Exact tests for variance components. *Biometrics*, 52, 309–314.
- Crainiceanu, C. M., & Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66, 165–185.
- Cui, L., Morris, A., Huang, L., Beck, J. M., Twigg, H. L., III, VonMutius, E., & Ghedin, E. (2014). The microbiome and the lung. *Annals of the American Thoracic Society*, 11, S227–S232.
- Dickson, R. P., Erb-Downward, J. R., & Huffnagle, G. B. (2013). The role of the bacterial microbiome in lung disease. *Expert Review of Respiratory Medicine*, 7, 245–257.
- Drikvandi, R., Verbeke, G., Khodadadi, A., & PartoviNia, V. (2013). Testing multiple variance components in linear mixed-effects models. *Biostatistics*, 14, 144–59.
- Eckburg, P. B., Bik, E. M., Bernstein, C. N., Purdom, E., Dethlefsen, L., Sargent, M., & Relman, D. A. (2005). Diversity of the human intestinal microbial flora. *Science*, 308, 1635–1638.
- Erb-Downward, J. R., Thompson, D. L., Han, M. K., Freeman, C. M., McCloskey, L., Schmidt, L. A., ... Sundaram, B. (2011). Analysis of the lung microbiome in the “healthy” smoker and in COPD. *PloS One*, 6, e16384.
- Federhen, S. (2012). The NCBI taxonomy database. *Nucleic Acids Research*, 40, D136–D143.
- Gilbert, J. A., Quinn, R. A., Debelius, J., Xu, Z. Z., Morton, J., Garg, N., & Knight, R. (2016). Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature*, 535, 94–103.
- Greven, S., Crainiceanu, C. M., Küchenhoff, H., & Peters, A. (2008). Restricted likelihood ratio testing for zero variance components

- in linear mixed models. *Journal of Computational and Graphical Statistics*, 17, 870–891.
- Haas, B. J., Gevers, D., Earl, A. M., Feldgarden, M., Ward, D. V., Giannoukos, G., ... Sodergren, E. (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Research*, 21, 494–504.
- Hewitt, R. J. & Molyneaux, P. L. (2017). The respiratory microbiome in idiopathic pulmonary fibrosis. *Annals of Translational Medicine*, 5, <http://atm.amegroups.com/article/view/14049>
- Hicks, L. A., Taylor, T. H., Jr, & Hunkler, R. J. (2013). US outpatient antibiotic prescribing, 2010. *New England Journal of Medicine*, 368, 1461–1462.
- Hodkinson, B. P., & Grice, E. A. (2015). Next-generation sequencing: A review of technologies and tools for wound microbiome research. *Advances in Wound Care*, 4, 50–58.
- Jangi, S., Gandhi, R., Cox, L. M., Li, N., VonGlehn, F., Yan, R., ... Glanz, B. L. (2016). Alterations of the human gut microbiome in multiple sclerosis. *Nature Communications*, 7, 12015.
- Kim, B.-S., Lee, E., Lee, M.-J., Kang, M.-J., Yoon, J., Cho, H.-J., & Hong, S.-J. (2018). Different functional genes of upper airway microbiome associated with natural course of childhood asthma. *Allergy*, 73, 644–652.
- Kinney, S. K., & Dunson, D. B. (2007). Fixed and random effects selection in linear and logistic models. *Biometrics*, 63, 690–698.
- Kuleshov, V., Jiang, C., Zhou, W., Jahanbani, F., Batzoglou, S., & Snyder, M. (2016). Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. *Nature Biotechnology*, 34, 64–69.
- Letunic, I., & Bork, P. (2007). Interactive Tree of Life (iTOL): An online tool for phylogenetic tree display and annotation. *Bioinformatics*, 23, 127–128.
- Letunic, I., & Bork, P. (2011). Interactive Tree of Life v2: Online annotation and display of phylogenetic trees made easy. *Nucleic Acids Research*, 39, W475–W478.
- Lin, X. (1997). Variance component testing in generalised linear models with random effects. *Biometrika*, 84, 309–326.
- Lloyd-Price, J., Mahurkar, A., Rahnavard, G., Crabtree, J., Orvis, J., Hall, A. B., ... Giglio, M. G. (2017). Strains, functions and dynamics in the expanded human microbiome project. *Nature*, 550, 61.
- Lozupone, C., Cota-Gomez, A., Palmer, B. E., Linderman, D. J., Charlson, E. S., Sodergren, E., ... Yao, G. (2013). Widespread colonization of the lung by *Tropheryma whipplei* in HIV infection. *American Journal of Respiratory and Critical Care Medicine*, 187, 1110–1117.
- Lozupone, C., & Knight, R. (2005). UniFrac: A new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*, 71, 8228–8235.
- Lozupone, C. A., Hamady, M., Kelley, S. T., & Knight, R. (2007). Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Applied and Environmental Microbiology*, 73, 1576–1585.
- Maidak, B. L., Olsen, G. J., Larsen, N., Overbeek, R., McCaughey, M. J., & Woese, C. R. (1996). The ribosomal database project (RDP). *Nucleic Acids Research*, 24, 82–85.
- Morris, A., Beck, J. M., Schloss, P. D., Campbell, T. B., Crothers, K., Curtis, J. L., ... Huang, L. (2013). Comparison of the respiratory microbiome in healthy nonsmokers and smokers. *American Journal of Respiratory and Critical Care Medicine*, 187, 1067–1075.
- Nayfach, S., Rodriguez-Mueller, B., Garud, N., & Pollard, K. S. (2016). An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Research*, 26, 1612–1625.
- Ofversten, J. (1993). Exact tests for variance components in unbalanced mixed linear models. *Biometrics*, 49, 45–57.
- Pragman, A. A., Kim, H. B., Reilly, C. S., Wendt, C., & Isaacson, R. E. (2012). The lung microbiome in moderate and severe chronic obstructive pulmonary disease. *PLoS one*, 7, e47305.
- Qu, L., Guennel, T., & Marshall, S. L. (2013). Linear score tests for variance components in linear mixed models and applications to genetic association studies. *Biometrics*, 69, 883–892.
- Rogers, G. B., Carroll, M., Hoffman, L., Walker, A., Fine, D., & Bruce, K. (2010). Comparing the microbiota of the cystic fibrosis lung and human gut. *Gut Microbes*, 1, 85–93.
- Saville, B. R., & Herring, A. H. (2009). Testing random effects in the linear mixed model using approximate bayes factors. *Biometrics*, 65, 369–376.
- Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., ... Robinson, C. J. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, 75, 7537–7541.
- Segal, L. N., Clemente, J. C., Li, Y., Ruan, C., Cao, J., Danckers, M., ... Diaz, P. (2017). Anaerobic bacterial fermentation products increase tuberculosis risk in antiretroviral-drug-treated HIV patients. *Cell Host and Microbe*, 21, 530–537.
- Self, S. G., & Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82, 605–610.
- Silvapulle, M. J., & Sen, P. K. (2011). *Constrained statistical inference: Order, inequality, and shape constraints*. Vol. 912. John Wiley & Sons.
- Stram, D. O., & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, 50, 1171–1177.
- Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C., & Segata, N. (2017). Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Research*, 27, 626–638.
- Tunney, M. M., Einarsson, G. G., Wei, L., Drain, M., Klem, E. R., Cardwell, C., & Elborn, J. S. (2013). Lung microbiota and bacterial abundance in patients with bronchiectasis when clinically stable and during exacerbation. *American Journal of Respiratory and Critical Care Medicine*, 187, 1118–1126.
- Twigg, H. L., Morris, A., Ghedin, E., Curtis, J. L., Huffnagle, G. B., Crothers, K., Beck, J. M., & Beck, J. M. (2013). Use of bronchoalveolar lavage to assess the respiratory microbiome: Signal in the noise. *The Lancet Respiratory Medicine*, 1, 354–356.
- Twigg, H. L., III, Knox, K. S., Zhou, J., Crothers, K. A., Nelson, D. E., Toh, E., ... Dong, Q. (2016). Effect of advanced HIV infection on the respiratory microbiome. *American Journal of Respiratory and Critical Care Medicine*, 194, 226–235.
- Wang, J., & Jia, H. (2016). Metagenome-wide association studies: Fine-tuning the microbiome. *Nature Reviews Microbiology*, 14, 508–522.
- Weiden, M. D., Segal, L. N., Clemente, J., Li, Y., Danckers-Degregory, M., Morris, A. M., ... Van Zyl-Smit, R. (2017). Lung microbiome dysbiosis is a risk factor for pulmonary diffusion abnormalities in

- antiretroviral treated HIV-infection. *American Journal of Respiratory and Critical Care Medicine*, 195, A1002. https://www.atsjournals.org/doi/abs/10.1164/ajrccm-conference.2017.195.1_MeetingAbstracts.A1002
- Xu, J., Mahowald, M. A., Ley, R. E., Lozupone, C. A., Hamady, M., Martens, E. C., ... Latreille, P. (2007). Evolution of symbiotic bacteria in the distal human intestine. *PLoS Biology*, 5, 1–13.
- Zemanick, E. T., Sagel, S. D., & Harris, J. K. (2011). The airway microbiome in cystic fibrosis and implications for treatment. *Current Opinion in Pediatrics*, 23, 319–324.
- Zeng, M. Y., Cisalpino, D., Varadarajan, S., Hellman, J., Warren, H. S., Cascalho, M., & Núñez, G. (2016). Gut microbiota-induced immunoglobulin G controls systemic infection by symbiotic bacteria and pathogens. *Immunity*, 44, 647–658.
- Zhan, X., Tong, X., Zhao, N., Maity, A., Wu, M. C., & Chen, J. (2017). A small-sample multivariate kernel machine test for microbiome association studies. *Genetic Epidemiology*, 41, 210–220.
- 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. *American Journal of Human Genetics*, 96, 797–807.
- Zhou, J. J., Hu, T., Qiao, D., Cho, M. H., & Zhou, H. (2016). Boosting gene mapping power and efficiency with efficient exact variance component tests of single nucleotide polymorphism sets. *Genetics*, 204, 921–931.

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: Zhai J, Knox K, Twigg III HL, Zhou H, Zhou JJ. Exact variance component tests for longitudinal microbiome studies. *Genet. Epidemiol.* 2019;43:250–262.
<https://doi.org/10.1002/gepi.22185>