Title

Exact variance component tests for longitudinal microbiome studies

Journal

Genetic Epidemiology, 43(3)

ISSN

0741-0395

Authors

Zhai, Jing

Knox, Kenneth

Twigg, Homer L

et al.

Publication Date

2019-04-01

DOI

10.1002/gepi.22185

Peer reviewed

HHS Public Access

Author manuscript

Genet Epidemiol. Author manuscript; available in PMC 2020 April 01.

Published in final edited form as:

Genet Epidemiol. 2019 April ; 43(3): 250–262. doi:10.1002/gepi.22185.

Exact Variance Component Tests for Longitudinal Microbiome Studies

Jing Zhai[1], **Kenneth Knox**[2], **Homer L. Twigg III**[3], **Hua Zhou**[4], and **Jin J. Zhou**[1,*]

[1]Department of Epidemiology and Biostatistics, University of Arizona Tucson, AZ 85724

[2]Division of Pulmonary, Allergy, Critical Care, Sleep Medicine, Department of Medicine, University of Arizona Tucson, AZ 85724

[3]Division of Pulmonary, Critical Care, Sleep, and Occupational Medicine Indiana University Medical Center Indianapolis, IN 46202

[4]Department of Biostatistics University of California, Los Angeles Los Angeles, CA 90095

Abstract

In metagenomic studies, testing the association of microbiome composition and clinical outcomes translates to testing the nullity of variance components. Motivated by a lung HIV (human immunodeficiency virus) microbiome project, we study longitudinal microbiome data by variance component models with more than two variance components. Current testing strategies only apply to the 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 (1) test the association of the overall microbiome composition in a longitudinal design and (2) 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 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 human immunodeficiency virus (HIV) infected patients and reveal two interesting genera *Prevotella* and *Veillonella* associated with forced vital capacity. Our findings shed lights on the impact of lung microbiome to 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 (HIV); Linear mixed effects models; Longitudinal pulmonary microbiome; Variance component models

* jzhou@email.arizona.edu.

Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article.

Author Biography

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 and Grice, 2015; Kuleshov et al., 2016; Wang and Jia, 2016). In particular, for the pulmonary microbiome, Rogers et al. (2010) hypothesized that 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 pulmonary microbiome, both in healthy (Erb-Downward et al., 2011; Morris et al., 2013; Twigg et al., 2013) and disease populations (Zemanick et al., 2011; Lozupone et al., 2013).

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

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zb} + h(\boldsymbol{G}) + \boldsymbol{\varepsilon}, \quad (1)$$

$$\boldsymbol{b} \sim \mathcal{N}(0, \sigma_d^2 \boldsymbol{I}_n), \quad h(\boldsymbol{G}) \sim \mathcal{N}(0, \sigma_g^2 \boldsymbol{K}), \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma_e^2 \boldsymbol{I}_n),$$

where $\boldsymbol{y}$, $\boldsymbol{X}$, $\boldsymbol{G}$ and $\boldsymbol{\varepsilon}$ are the vertically stacked vectors/matrices of individual-level $\boldsymbol{y}_i$, $\boldsymbol{X}_i$, $\boldsymbol{G}_i$ and $\boldsymbol{\varepsilon}_i$. $\boldsymbol{y}_i$ is a vector of n_i repeated measures of a quantitative phenotype for individual i. $\boldsymbol{X}_i$ is the $n_i \times p$ covariate matrix. $\boldsymbol{G}_i$ 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. $\boldsymbol{Z}$ 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 $\boldsymbol{b} = (b_i)$ is the subject-specific random effects. $\boldsymbol{K}$ is a kernel matrix capturing distances between individuals, e.g., the UniFrac distance (Lozupone and Knight, 2005) or the Bray-Curtis dissimilarity (Bray and Curtis, 1957) (Web Appendix A). $\boldsymbol{b}$, $h(\boldsymbol{G})$ and $\boldsymbol{\varepsilon}$ are jointly independent; therefore,

$$\mathrm{Var}(\boldsymbol{y}) = \sigma_d^2 \boldsymbol{ZZ}' + \sigma_g^2 \boldsymbol{K} + \sigma_e^2 \boldsymbol{I}_n, \quad (2)$$

where σ_d^2 is the phenotypic variance due to the correlation of repeated measurements, σ_g^2 is the phenotypic variance explained by microbiome, and σ_e^2 is the within-subject variance that cannot be explained by microbiome and repeated measurements. 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) (Zhao et al.,

2015; Chen et al., 2016; Zhan et al., 2017). 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 microbial 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

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

where $\sum_i \sigma_{g_i}^2 \boldsymbol{K}_i$ 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 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 are available (Self and Liang, 1987; Stram and Lee, 1994; Silvapulle and Sen, 2011). 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 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 statistics. Greven et al. (2008) provided a pseudolikelihood-heuristic extension of this method to the $m > 1$ situation. Later Drikvandi et al. (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 et al. (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. (2013)'s method is exact; when $m > 1$, their test relies on asymptotic theory. Score-based tests can be less powerful than the likelihood ratio tests, 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 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

(Chen and Dunson, 2003; Kinney and Dunson, 2007), but they can be time-consuming, especially when the number of random effects is large.

In this article, we propose methods of performing exact Likelihood Ratio Test (`eLRT`), exact Restricted Likelihood Ratio Test (`eRLRT`), and 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 (Ofversten, 1993; Christensen, 1996). Our method is the first one that provides `eLRT`, `eRLRT`, and `eScore` for testing 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 model

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

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

Let $\lambda = \sigma_1^2/\sigma_e^2$ be the signal-to-noise ratio, $s = \text{rank}(\boldsymbol{X})$, and write the covariance as $\boldsymbol{V} = \sigma_e^2(\boldsymbol{I}_n + \lambda \boldsymbol{V}_1) = \sigma_e^2 \boldsymbol{V}_\lambda$. 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 $L(\boldsymbol{\beta}, \sigma_e^2, \lambda) = -\frac{n}{2}\ln\sigma_e^2 - \frac{1}{2}\text{lndet}(\boldsymbol{V}_\lambda) - \frac{1}{2\sigma_e^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{V}_\lambda^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$. The likelihood ratio test (LRT) statistic is

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

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

$$\mathrm{LRT} \overset{\mathscr{D}}{=} \sup_{\lambda \geq 0}\left\{ n\ln\frac{\sum_{i=1}^{n-s}\omega_i^2}{\sum_{i=1}^{k}\frac{\omega_i^2}{1+\lambda\mu_i} + \sum_{i=k+1}^{n-s}\omega_i^2} - \sum_{i=1}^{l}\ln(1+\lambda\xi_i)\right\},$$

where, under the null, w_i are $(n-s)$ independent standard normals. Under the alternative, $\omega_i \sim \mathcal{N}(0, 1+\lambda\mu_i)$ for $i = 1, \ldots, k$, $w_i \sim \mathcal{N}(0, 1)$ for $i = k+1, \ldots, n-s$, 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 Supplementary Material (Web Appendix B). The same derivation can be carried out for the `eRLRT`, in which case

$$\mathrm{RLRT} \overset{D}{=} \sup_{\lambda \geq 0}\left\{ (n-s)\ln\frac{\sum_{i=1}^{n-s}\omega_i^2}{\sum_{i=1}^{k}\frac{\omega_i^2}{1+\lambda\mu_i} + \sum_{i=k+1}^{n-s}\omega_i^2} - \sum_{i=1}^{k}\ln(1+\lambda\mu_i)\right\}.$$

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 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 $\boldsymbol{V} = \sigma_e^2\boldsymbol{I}_n + \sigma_1^2\boldsymbol{V}_1$. The (Rao) score statistic is based on $\boldsymbol{I}^{-1}_{\sigma_1^2,\sigma_1^2}\left(\frac{\partial}{\partial\sigma_1^2}L\right)^2$, where the information matrix $\boldsymbol{I}_{\sigma_1^2,\sigma_1^2} = E\left(-\frac{\partial^2}{\partial\sigma_1^2\partial\sigma_1^2}L\right)$ and score function $\frac{\partial}{\partial\sigma_1^2}L$ are evaluated at the maximum likelihood estimator (MLE) under the null. The resultant test rejects the null when

$$S = \max\left\{\frac{\boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{V}_1(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}}{\boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}}, \frac{\mathrm{tr}(\boldsymbol{K})}{n}\right\}$$

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

$$S \overset{D}{=} \max\left\{\frac{\sum_{i=1}^{k}\mu_i\omega_i^2}{\sum_{i=1}^{n-s}\omega_i^2}, \frac{\mathrm{tr}(\boldsymbol{K})}{n}\right\},$$

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 components under H_0

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

We first obtain an orthonormal basis $(\boldsymbol{Q}_0, \boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_m, \boldsymbol{Q}_{m+1})$ of $\mathbb{R}^n$ such that $\boldsymbol{Q}_0$ is an orthonormal basis of $\mathrm{C}(\boldsymbol{X})$, $\boldsymbol{Q}_1$ is an orthonormal basis of $\mathrm{C}(\boldsymbol{X}, \boldsymbol{V}_1) - \mathrm{C}(\boldsymbol{X})$, $\boldsymbol{Q}_i$ is an orthonormal basis of $\mathrm{C}(\boldsymbol{X}, \boldsymbol{V}_1, \ldots, \boldsymbol{V}_i) - \mathrm{C}(\boldsymbol{X}, \boldsymbol{V}_1, \ldots, \boldsymbol{V}_{i-1})$ for $i = 2, \ldots, m$, and $\boldsymbol{Q}_{m+1}$ is an orthonormal basis of $\mathbb{R}^n - \mathrm{C}(\boldsymbol{X}, \boldsymbol{V}_1, \ldots, \boldsymbol{V}_m)$. Denote their corresponding ranks by $r_0, \ldots, r_{m+1}$. If $r_m > 0$, that is $C(X, \boldsymbol{V}_1, \ldots, \boldsymbol{V}_{m-1}) \not\subseteq C(X, \boldsymbol{V}_1, \ldots, \boldsymbol{V}_{m-1}, \boldsymbol{V}_m)$, then $\boldsymbol{Q}'_m\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{0}, \sigma_e^2\boldsymbol{I}_{r_m} + \sigma_m^2\boldsymbol{Q}'_m\boldsymbol{V}_m\boldsymbol{Q}_m)$ and `eLRT`, `eRLRT` and `eScore` can be applied to $\boldsymbol{Q}'_m\boldsymbol{Y}$.

The order of $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_m$ does not matter. If $r_m = 0$, that is $\mathrm{C}(\boldsymbol{X}, \boldsymbol{V}_1, \ldots, \boldsymbol{V}_{m-1}) = \mathrm{C}(\boldsymbol{X}, \boldsymbol{V}_1, \ldots, \boldsymbol{V}_m)$, we construct a test based on the transformed data $\boldsymbol{Q}'_{m-1}\boldsymbol{Y} + \boldsymbol{C}\boldsymbol{Q}'_{m+1}\boldsymbol{Y}$. Without loss of generality we assume $\boldsymbol{Q}_{m-1}$ is nontrivial. If $r_{m-1} = 0$, we use $\boldsymbol{Q}_{m-2}$ and so on. We consider the following cases:

1. If $\boldsymbol{Q}'_{m-1}\boldsymbol{V}_m = \boldsymbol{0}$, e.g., when $C(\boldsymbol{V}_m) \subset C(\boldsymbol{X}, \boldsymbol{V}_1, \ldots, \boldsymbol{V}_{m-2})$, then this test cannot be performed. Shifting the order of $\boldsymbol{X}, \boldsymbol{V}_1, \ldots, \boldsymbol{V}_{m-1}$ might solve the issue.
2. If $\boldsymbol{Q}'_{m-1}\boldsymbol{V}_{m-1}\boldsymbol{Q}_{m-1} = \gamma\boldsymbol{I}_{rm-1}$ and $\gamma \neq 0$, then

$$\boldsymbol{Q}'_{m-1}\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{0}, \sigma_e^2\boldsymbol{I}_{rm-1} + \sigma_{m-1}^2\boldsymbol{Q}'_{m-1}\boldsymbol{V}_{m-1}\boldsymbol{Q}_{m-1} + \sigma_m^2\boldsymbol{Q}'_{m-1}\boldsymbol{V}_m\boldsymbol{Q}_{m-1})$$

$$= \mathcal{N}(0, \sigma_e^2 + \gamma\sigma_{m-1}^2)\boldsymbol{I}_{r_{m-1}} + \sigma_m^2\boldsymbol{Q}'_{m-1}\boldsymbol{V}_m\boldsymbol{Q}_{m-1}$$

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

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

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

where the scalar ζ is chosen such that $\zeta^{-1}\boldsymbol{Q}'_{m-1}\boldsymbol{V}_{m-1}\boldsymbol{Q}_{m-1} - \boldsymbol{I}_{r_{m-1}}$ positive semidefinite. Let $\boldsymbol{Q}'_{m-1}\boldsymbol{V}_{m-1}\boldsymbol{Q}_{m-1} = \boldsymbol{W}\Lambda\boldsymbol{W}' = \boldsymbol{W}\mathrm{diag}(\delta_i)\boldsymbol{W}'$ be the eigen-decomposition, ζ be the smallest positive eigenvalue, and $\boldsymbol{C} = \boldsymbol{W}\mathrm{diag}(\sqrt{\delta_i/\zeta - 1}$. Then the transformed data $\boldsymbol{Q}'_{m-1}\boldsymbol{Y} + \boldsymbol{C}\boldsymbol{Q}'_{m+1}\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{0}, (\sigma_m^2 - 1 + \sigma_e^2/\zeta)\boldsymbol{Q}'_{m-1}\boldsymbol{V}_{m-1}\boldsymbol{Q}_{m-1} + \sigma_m^2\boldsymbol{Q}'_{m-1}\boldsymbol{V}_m\boldsymbol{Q}_{m-1})$and the test for case (2.1) can be applied. A larger ζ leads to a higher signal-to-noise ratio $\frac{\sigma_m^2}{\sigma_{m-1}^2 + \sigma_e^2/\zeta}$ and thus a more powerful test. Finally we test H_0: $\sigma_m^2 = 0$ using `eLRT`, `eRLRT` or `eScore` test on the transformed data,

$$\Lambda^{-1/2}\boldsymbol{W}'(\boldsymbol{Q}'_{m-1} + \boldsymbol{C}\boldsymbol{Q}'_{m+1})\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{0}, (\sigma_{m-1}^2 + \sigma_e^2/\zeta)\boldsymbol{I}_{r_{m-1}} + \sigma_m^2\Lambda^{-1/2}\boldsymbol{W}'\boldsymbol{Q}'_{m-1}\boldsymbol{V}_m\boldsymbol{Q}_{m-1}\boldsymbol{W}\Lambda^{-1/2})$$

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

3 Simulation

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

Longitudinal microbiome count data with 2 repeated measurements are simulated using the R package ZIBR (Zero-Inflated Beta Random Effect model) (Chen and 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 dataset contains 30 samples, each with 2 to 4 repeated measurements: baseline, 4 weeks, 1 year and 3 years (Twigg III et al., 2016). OTU alignment at species level was produced by software Mothur (https://www.mothur.org/) (Schloss et al., 2009) and Basic Local Alignment Search Tool (BLAST) (https://blast.ncbi.nlm/.nih.gov/Blast.cg) (Altschul et al., 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 III et al., 2016). We construct the higher taxon level, e.g., phylum, using the phylogenetic tree generator phyloT (http://phylot.biobyte.de/) (Letunic and Bork, 2007, 2011) and NCBI database taxonomy (https://www.ncbi.nlm.nih.gov/taxonomy) (Federhen, 2012). There are 2964 operational taxonomic units (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 Appendix 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 $X_{2i} = h(G_i)_{baseline} + N(0,1)$. Their effects are $\beta_1 = \beta_2 = 0.1$. We set within-individual variance to $\sigma_e^2 = 1$. For longitudinal data simulation, between individual variance σ_d^2 is set to 0.6. This corresponds to 60% of overall baseline phenotypic variance (Twigg III et al., 2016).

Scenario 1: Testing overall microbiome effect.

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

Scenario 2: Localizing fine microbiome cluster effects.

We cluster OTUs into 6 phyla, *Actinobacteria*, *Bacteroidetes*, *Fusobacteria*, *Proteobacteria*, *Firmicutes*, and *other*. We assume that only cluster *other*, $h(\boldsymbol{G}_{1i})$, has effects. That is $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2, \sigma_d^2\boldsymbol{Z}\boldsymbol{Z}' + \Sigma_{l=1}^{6}\sigma_{g_l}^2\boldsymbol{K}_l + \sigma_e^2\boldsymbol{I})$, where $\sigma_{g_1}^2 = 0$, 0.5, 1.5 and $\sigma_{g_1}^2 = 0$ for $l = 2$, …,6. Due to the correlation between phyla, marginal tests of 5 individual phylum may show false signal if we do not adjust for the effects of $h(\boldsymbol{G}_{1i})$. We present testing of variance components in a joint model has correct type I error.

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 three methods when $\sigma_d^2 = 0$. Responses are generated according to simulation scenario 1, where $\sigma_g^2 = 0, \ldots, 1.5$.

In scenarios 1 and 2, the sample size is fixed at $n = 100$. In scenario 3, we compare the performance of 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 higher power than other kernels in scenario 1. 1000 Monte Carlo replicates are generated for all simulations and we use the nominal significance level 0.05 to evaluate type I error and power.

4 Results

Simulation Results

Scenario 1: Testing 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 Figures 1a and 1c, five different kernels are constructed using OTU count data directly. In Figures 1b and 1d, 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 power (Figures 1a and 1c). The pattern of the power increase with effect size differs according to which taxon level count data are used to calculate kernel. The power of five kernels became similar to each other in Figures 1b and 1d. Further, the power of unweighted UniFrac kernel K_{UW}, which is the least powerful kernel in Figures 1a and 1c, greatly improves in Figures 1b and 1d. 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 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.

Scenario 2. Localizing fine microbiome cluster effects.—Table 3 shows the type I error rates for testing microbiome effect at the phylum level, with and without adjusting for the effect contributed by cluster, *other*. Most type I error rates are inflated when not adjusting for cluster *other* effects. In 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-value = 0.70, 0.24, respectively). After adjustment, type I error rates stay correct even when confounding effects are large (Table 3).

In practice, symbiosis of bacteria causes correlation between them (Xu et al., 2007; Dickson et al., 2013; 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 (Hicks et al., 2013; Blaser, 2016). Simulation scenario 2 demonstrates that our method is capable of localizing fine microbiome cluster effects.

Scenario 3: Comparing to existing methods MiRKAT and LinScore.—Table 4 presents type I 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 comparing to LinScore when association strength is weak. Microbiome studies usually have limited sample size due to the high cost. Higher power of the exact tests at small sample sizes will be particularly valuable for biologists and physicians to identify the associated microbiome clusters.

Analysis of Longitudinal Pulmonary Microbiome Data

It is well-known that HIV infection is associated with alterations in the respiratory microbiome (Twigg III et al., 2016). However due to the limited investigation, the clinical implications of lung microbial dysbiosis are currently unknown. As an initial step to reveal the connection of 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/mm^3). 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. Details of microbiome composition have been discussed in Section 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 (FVC), forced expiratory volume in 1 second (FEV1), and average forced expiratory flow (FEF). Diffusion capacity of the lungs for carbon monoxide (DLCO) measures how much oxygen travels from lung alveoli to 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). Missing covariate is imputed by its mean. For overall microbiome association test, no tests find significant associations. However at the phylum level, *Bacteroidetes* shows significant association with spirometry while *Firmicutes* shows significant association with diffusing capacity measures. Similar results have been reported by Molyneaux et al. (2012) 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 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 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 previous studies that *Parvimonas* abundance changed in subjects with pulmonary disease (e.g., asthma or COPD) comparing to the control group. (Pragman et al., 2012; Kim et al., 2018) 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 shed lights for future clinical causality research (Twigg III et al., 2016; Weiden et al., 2017; Segal 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 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.

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 extend 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 fine-grained level. The latter has been emerging as the focus of many current microbiome studies (Nayfach et al., 2016; Lloyd-Price et al., 2017; Truong et al., 2017). Unlike Qu et al. (2013) and Zhao et al. (2015)'s score test that uses moment-matching to approximate null distribution, our tests are exact in finite samples, therefore beneficial to the studies with limited sample size. Compared to score test, our `eLRT` and `eRLRT` tests can further boost power when microbiome effects are weak. Simulation studies verify that our exact tests have 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 demonstrate the ability of our exact tests to differentiate associated genus by two real data examples. Although the derivation of eLRT and eRLRT require normality assumption, a sensitivity simulation shows that even with a misspecified phenotypic distribution, like t-distribution, our tests still preserve 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, variance adjusted weighted 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 include both random intercepts and random slopes (Drikvandi et al., 2013). Our methods extend to these cases naturally and we defer them to future research. Second direction is to incorporate multiple types of kernels into exact tests. Last 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 (PQL) (Lin, 1997; Chen et al., 2016).

6 Software

A Julia package is freely available at https://github.com/JingZhai63/VCmicrobiome. In the real longitudinal data analysis with sample size 30 and 2964 OTUs, the elapsed CPU time are 0.1 seconds and 0.04 seconds for `eRLRT` and `eScore`, respectively. The analysis was performed by a MacBook Pro with 2.3GHz Intel Core i7 processor and 8GB 1600MHz DDR3 memory.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

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

References

Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ (1990). Basic local alignment search tool. Journal of Molecular Biology 215, 403–410. [PubMed: 2231712]

Blaser MJ (2016). Antibiotic use and its consequences for the normal microbiome. Science 352, 544–545. [PubMed: 27126037]

Bray JR and Curtis JT (1957). An ordination of the upland forest communities of southern Wisconsin. Ecological Monographs 27, 325–349.

Chang Q, Luan Y, and Sun F (2011). Variance adjusted weighted UniFrac: a powerful beta diversity measure for comparing communities based on phylogeny. BMC Bioinformatics 12, 118. [PubMed: 21518444]

Chen EZ and Li H (2016). A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. Bioinformatics 32, 2611–2617. [PubMed: 27187200]

Chen H, Wang C, Conomos MP, Stilp AM, Li Z, Sofer T, Szpiro AA, Chen W, Brehm JM, Celedón JC, et al. (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. [PubMed: 27018471]

Chen J, Bittinger K, Charlson ES, Hoffmann C, Lewis J, Wu GD, Collman RG, Bushman FD, and Li H (2012). Associating microbiome composition with environmental covariates using generalized UniFrac distances. Bioinformatics 28, 2106–2113. [PubMed: 22711789]

Chen J, Chen W, Zhao N, Wu MC, and Schaid DJ (2016). Small sample kernel association tests for human genetic and microbiome association studies. Genetic Epidemiology 40, 5–19. [PubMed: 26643881]

Chen Z and Dunson DB (2003). Random effects selection in linear mixed models. Biometrics 59, 762–769. [PubMed: 14969453]

Christensen R (1996). Exact tests for variance components. Biometrics 52, 309–314.

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

Cui L, Morris A, Huang L, Beck JM, Twigg HL, III, Von Mutius E, and Ghedin E (2014). The microbiome and the lung. Annals of the American Thoracic Society 11, S227–S232. [PubMed: 25148429]

Dickson RP, Erb-Downward JR, and Huffnagle GB (2013). The role of the bacterial microbiome in lung disease. Expert Review of Respiratory Medicine 7, 245–257. [PubMed: 23734647]

Drikvandi R, Verbeke G, Khodadadi A, and Partovi Nia V (2013). Testing multiple variance components in linear mixed-effects models. Biostatistics 14, 144–59. [PubMed: 22930674]

Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, and Relman DA (2005). Diversity of the human intestinal microbial flora. Science 308, 1635–1638. [PubMed: 15831718]

Erb-Downward JR, Thompson DL, Han MK, Freeman CM, McCloskey L, Schmidt LA, Young VB, Toews GB, Curtis JL, Sundaram B, et al. (2011). Analysis of the lung microbiome in the “healthy” smoker and in COPD. PloS One 6, e16384. [PubMed: 21364979]

Federhen S (2012). The NCBI Taxonomy database. Nucleic Acids Research 40, D136–D143. [PubMed: 22139910]

Gilbert JA, Quinn RA, Debelius J, Xu ZZ, Morton J, Garg N, Jansson JK, Dorrestein PC, and Knight R (2016). Microbiome-wide association studies link dynamic microbial consortia to disease. Nature 535, 94–103. [PubMed: 27383984]

Greven S, Crainiceanu CM, Küchenhoff H, and Peters A (2008). Restricted likelihood ratio testing for zero variance components in linear mixed models. Journal of Computational and Graphical Statistics 17, 870–891.

Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G, Ciulla D, Tabbaa D, Highlander SK, Sodergren E, et al. (2011). Chimeric 16s rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Research 21, 494–504. [PubMed: 21212162]

Hicks LA, Taylor TH, Jr, and Hunkler RJ (2013). US outpatient antibiotic prescribing, 2010. New England Journal of Medicine 368, 1461–1462. [PubMed: 23574140]

Hodkinson BP and Grice EA (2015). Next-generation sequencing: a review of technologies and tools for wound microbiome research. Advances in Wound Care 4, 50–58. [PubMed: 25566414]

Jangi S, Gandhi R, Cox LM, Li N, Von Glehn F, Yan R, Patel B, Mazzola MA, Liu S, Glanz BL, et al. (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, Park J, Won S, Lee S-Y, and Hong S-J (2018). Different functional genes of upper airway microbiome associated with natural course of childhood asthma. Allergy 73, 644–652. [PubMed: 29052232]

Kinney SK and Dunson DB (2007). Fixed and random effects selection in linear and logistic models. Biometrics 63, 690–698. [PubMed: 17403104]

Kuleshov V, Jiang C, Zhou W, Jahanbani F, Batzoglou S, and Snyder M (2016). Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. Nature Biotechnology 34, 64–69.

Letunic I and Bork P (2007). Interactive Tree of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23, 127–128. [PubMed: 17050570]

Letunic I and Bork P (2011). Interactive Tree of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Research 39, W475–W478. [PubMed: 21470960]

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 AB, Brady A, Creasy HH, McCracken C, Giglio MG, et al. (2017). Strains, functions and dynamics in the expanded human microbiome project. Nature 550, 61. [PubMed: 28953883]

Lozupone C, Cota-Gomez A, Palmer BE, Linderman DJ, Charlson ES, Sodergren E, Mitreva M, Abubucker S, Martin J, Yao G, et al. (2013). Widespread colonization of the lung by Tropheryma whipplei in HIV infection. American Journal of Respiratory and Critical Care Medicine 187, 1110–1117. [PubMed: 23392441]

Lozupone C and Knight R (2005). UniFrac: a new phylogenetic method for comparing microbial communities. Applied and Environmental Microbiology 71, 8228–8235. [PubMed: 16332807]

Lozupone CA, Hamady M, Kelley ST, and 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. [PubMed: 17220268]

Maidak BL, Olsen GJ, Larsen N, Overbeek R, McCaughey MJ, and Woese CR (1996). The Ribosomal Database Project (RDP). Nucleic Acids Research 24, 82–85. [PubMed: 8594608]

Molyneaux PL, Russell AM, Cox MJ, Moffatt MF, Cookson WO, and Maher TM (2012). The respiratory microbiome in idiopathic pulmonary fibrosis. In C103. pathogenesis, biomarkers, and risk factors for interstitial lung disease: from bench to bedside, pages A5174–A5174. American Thoracic Society.

Morris A, Beck JM, Schloss PD, Campbell TB, Crothers K, Curtis JL, Flores SC, Fontenot AP, Ghedin E, Huang L, et al. (2013). Comparison of the respiratory microbiome in healthy nonsmokers and smokers. American Journal of Respiratory and Critical Care Medicine 187, 1067–1075. [PubMed: 23491408]

Nayfach S, Rodriguez-Mueller B, Garud N, and Pollard KS (2016). An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. Genome research 26, 1612–1625. [PubMed: 27803195]

Ofversten J (1993). Exact tests for variance components in unbalanced mixed linear models. Biometrics 49, 45–57.

Pragman AA, Kim HB, Reilly CS, Wendt C, and Isaacson RE (2012). The lung microbiome in moderate and severe chronic obstructive pulmonary disease. PloS one 7, e47305. [PubMed: 23071781]

Qu L, Guennel T, and Marshall SL (2013). Linear score tests for variance components in linear mixed models and applications to genetic association studies. Biometrics 69, 883–892. [PubMed: 24328714]

Rogers GB, Carroll M, Hoffman L, Walker A, Fine D, and Bruce K (2010). Comparing the microbiota of the cystic fibrosis lung and human gut. Gut Microbes 1, 85–93. [PubMed: 21326915]

Saville BR and Herring AH (2009). Testing random effects in the linear mixed model using approximate bayes factors. Biometrics 65, 369–76. [PubMed: 18759835]

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and Environmental Microbiology 75, 7537–7541. [PubMed: 19801464]

Segal LN, Clemente JC, Li Y, Ruan C, Cao J, Danckers M, Morris A, Tapyrik S, Wu BG, Diaz P, et al. (2017). Anaerobic bacterial fermentation products increase tuberculosis risk in antiretroviral-drug-treated HIV patients. Cell Host & Microbe 21, 530–537. [PubMed: 28366509]

Self SG and 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 MJ and Sen PK (2011). Constrained statistical inference: order, inequality, and shape constraints, volume 912 John Wiley & Sons.

Stram DO and Lee JW (1994). Variance components testing in the longitudinal mixed effects model. Biometrics pages 1171–1177. [PubMed: 7786999]

Truong DT, Tett A, Pasolli E, Huttenhower C, and Segata N (2017). Microbial strain-level population structure and genetic diversity from metagenomes. Genome research 27, 626–638. [PubMed: 28167665]

Tunney MM, Einarsson GG, Wei L, Drain M, Klem ER, Cardwell C, Ennis M, Boucher RC, Wolfgang MC, and Elborn JS (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. [PubMed: 23348972]

Twigg HL, Morris A, Ghedin E, Curtis JL, Huffnagle GB, Crothers K, Campbell TB, Flores SC, Fontenot AP, Beck JM, et al. (2013). Use of bronchoalveolar lavage to assess the respiratory microbiome: signal in the noise. The Lancet Respiratory Medicine 1, 354–356. [PubMed: 24429191]

Twigg HL, III, Knox KS, Zhou J, Crothers KA, Nelson DE, Toh E, Day RB, Lin H, Gao X, Dong Q, et al. (2016). Effect of advanced HIV infection on the respiratory microbiome. American Journal of Respiratory and Critical Care Medicine 194, 226–235. [PubMed: 26835554]

Wang J and Jia H (2016). Metagenome-wide association studies: fine-mining the microbiome. Nature Reviews Microbiology 14, 508–522. [PubMed: 27396567]

Weiden MD, Segal LN, Clemente J, Li Y, Danckers-Degregory M, Morris AM, Tapyrik S, Diaz P, Dawson R, Van Zyl-Smit R, et al. (2017). Lung microbiome dysbiosis is a risk factor for pulmonary diffusion abnormalities in antiretroviral treated HIV-infection. In A13. Role of dysbiosis in lung disease, pages A1002–A1002. American Thoracic Society.

Xu J, Mahowald MA, Ley RE, Lozupone CA, Hamady M, Martens EC, Henrissat B, Coutinho PM, Minx P, Latreille P, et al. (2007). Evolution of symbiotic bacteria in the distal human intestine. PLoS Biology 5, 1–13.

Zemanick ET, Sagel SD, and Harris JK (2011). The airway microbiome in cystic fibrosis and implications for treatment. Current Opinion in Pediatrics 23, 319–324. [PubMed: 21494150]

Zeng MY, Cisalpino D, Varadarajan S, Hellman J, Warren HS, Cascalho M, Inohara N, and Núñez G (2016). Gut microbiota-induced immunoglobulin G controls systemic infection by symbiotic bacteria and pathogens. Immunity 44, 647–658. [PubMed: 26944199]

Zhan X, Tong X, Zhao N, Maity A, Wu MC, and Chen J (2017). A small-sample multivariate kernel machine test for microbiome association studies. Genetic Epidemiology 41, 210–220. [PubMed: 28019040]

Zhao N, Chen J, Carroll IM, Ringel-Kulka T, Epstein MP, Zhou H, Zhou JJ, Ringel Y, Li H, and Wu MC (2015). Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. American Journal of Human Genetics 96, 797–807. [PubMed: 25957468]

Zhou JJ, Hu T, Qiao D, Cho MH, and Zhou H (2016). Boosting gene mapping power and efficiency with efficient exact variance component tests of single nucleotide polymorphism sets. Genetics 204, 921–931. [PubMed: 27646141]

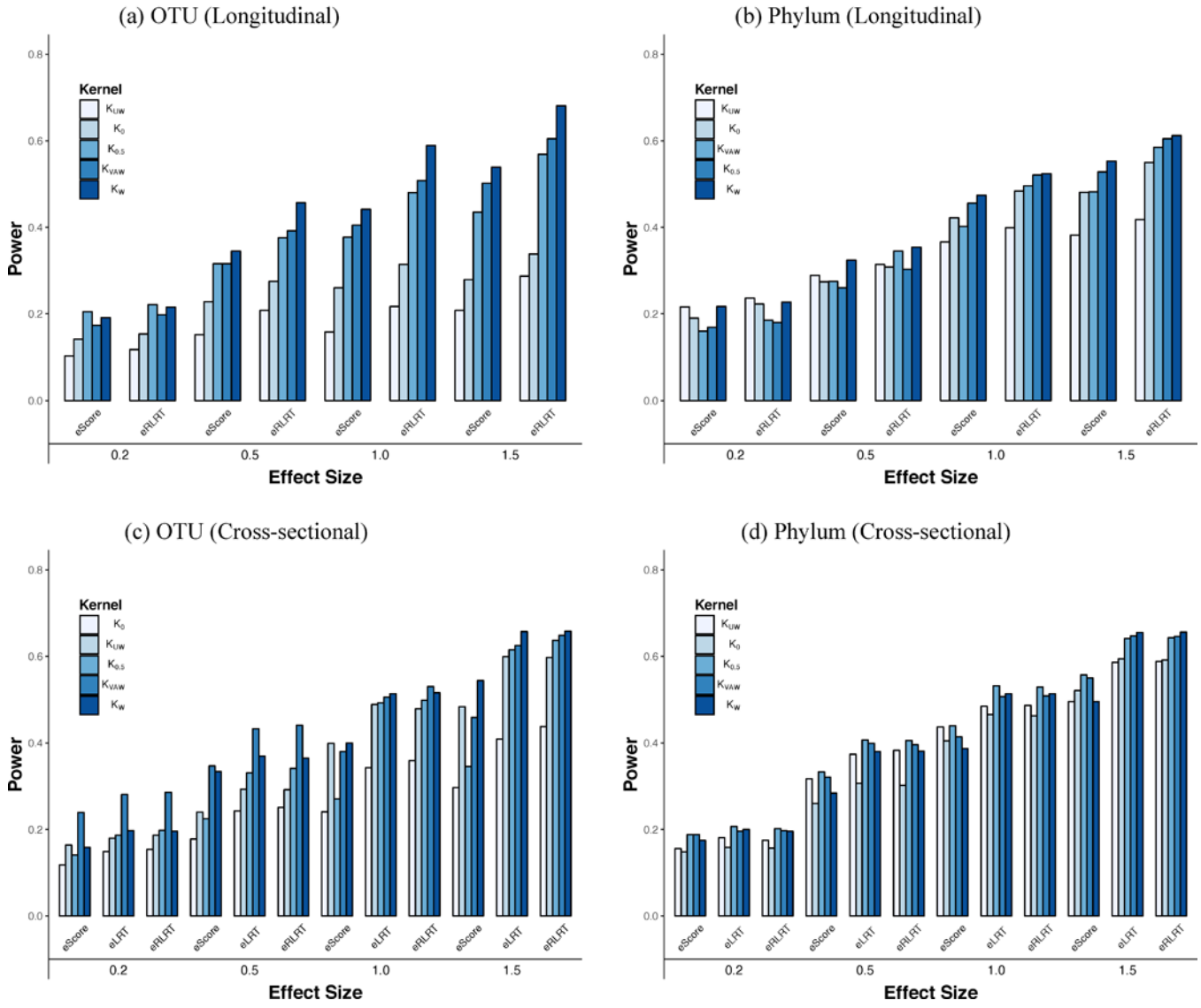

FIGURE 1:

Scenario 1: Power of `eRLRT`, `eLRT` and `eScore` using different distance measures. Figure to the left shows results where the OTU counts are used to calculate distances, figures to the right shows that OTU counts are summarized at phylum level to construct the distances. K_0, $K_{0.5}$, K_W, K_U and K_{VAW} represent generalized UniFrac distance with α = 0, 0.5, weighted UniFrac distance, unweighted UniFrac distance and variance adjusted weighted UniFrac distance, respectively.

TABLE 1

Simulation configurations. For all simulations, $\sigma_e^1 = 2$ and $\sigma_d^2 = 0$ when number of repeats (# Repeat) = 1 or $\sigma_e^2 = 1$ and $\sigma_d^2 = 0.6$ when number of repeats >1. There are 2964 OTUs presented in simulated count data. A phylogenetic tree generated using the real pulmonary microbiome data is used for kernel calculation and phenotype simulations. K_W: weighted UniFrac kernel; K_U: unweighted UniFrac kernel; $K_{V\ AW}$: variance adjusted weighted UniFrac kernel; K_a: generalized UniFrac kernels with = 0 and 0:5.

Scenario 1: Testing overall microbiome effect					
Sample size	Kernel type	Clustering	# Repeat	σ_g^2	Method
100	K_W, K_U, K_{VAW}, K_a	None	2	0 – 1.5	`eRLRT` `eScore`
100	K_W, K_U, K_{VAW}, K_a	None	1	0 – 1.5	`eRLRT` `eLRT` `eScore`
Scenario 2: Localizing fine microbiome cluster effects					
Sample size	Kernel type	Clustering	# Repeat	σ_g^2	Method
100	K_W	Yes	2	0 – 1.5	`eRLRT` `eScore`
100	K_W	Yes	1	0 – 1.5	`eRLRT` `eScore`
Scenario 3: Comparing to existing methods					
Sample size	Kernel type	Clustering	# Repeat	σ_g^2	Method
20; 30; 50; 100	K_W	None	2	0 – 1.5	`eRLRT` `eScore` `LinScore`
20; 30; 50; 100	K_W	None	1	0 – 1.5	`eRLRT` `eLRT` `eScore` `LinScore` `MiRKAT`

TABLE 2

Scenario 1: Type I error of `eLRT`, `eRLRT` and `eScore` for detecting overall microbiome effects. Five distance measures, weighted UniFrac kernel (K_W), unweighted UniFrac kernel (K_U), variance adjusted weighted UniFrac kernel ($K_{V\,AW}$), and generalized UniFrac kernels with = 0 (K_0) and 0:5 ($K_{0:5}$) are compared.

		Kernel Type				
Simulation Design	Method	K_W	K_U	$K_{V\,AW}$	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

TABLE 3

Scenario 2: Type I error rate of localizing fine microbiome cluster effects. Only cluster "Other" contains effects, 0, 0:5 and 1:5. Type I 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 bottom panel shows results using cross-sectional data only.

Longitudinal Design						
	No Adjustment for Other			Adjustment for Other		
Phylum	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
Actinobacteria	0.050, 0.038	0.108, 0.075	0.151, 0.100	0.049, 0.038	0.051, 0.048	0.033, 0.040
Bacteroidetes	0.045, 0.040	0.060, 0.055	0.061, 0.055	0.041, 0.040	0.047, 0.042	0.042, 0.037
Firmicutes	0.043, 0.043	0.049, 0.044	0.063, 0.067	0.042, 0.043	0.041, 0.043	0.052, 0.051
Fusobacteria	0.052, 0.048	0.038, 0.041	0.060, 0.048	0.052, 0.048	0.045, 0.044	0.048, 0.037
Proteobacteria	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						
	No Adjustment for Other			Adjustment for Other		
Phylum	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
Actinobacteria	0.041, 0.036	0.117, 0.065	0.111, 0.083	0.050, 0.040	0.052, 0.043	0.048, 0.035
Bacteroidetes	0.051, 0.047	0.048, 0.049	0.051, 0.041	0.051, 0.041	0.048, 0.043	0.048, 0.037
Firmicutes	0.037, 0.038	0.059, 0.052	0.068, 0.062	0.044, 0.038	0.051, 0.045	0.052, 0.048
Fusobacteria	0.053, 0.050	0.070, 0.060	0.078, 0.065	0.052, 0.033	0.051, 0.041	0.048, 0.040
Proteobacteria	0.042, 0.035	0.038, 0.042	0.053, 0.047	0.048, 0.047	0.049, 0.050	0.041, 0.033

TABLE 4

Scenario 3: Comparing to existing methods. Type I 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.

			Effect Size (σ_g^2)						
n	**#Repeat**	**Method**	**0**	**0.10**	**0.2**	**0.5**	**0.8**	**1.0**	**1.5**
20		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
	1	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
		eScore	0.050	0.055	0.040	0.057	0.068	0.077	0.088
	2	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		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
	1	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
		eScore	0.045	0.058	0.067	0.093	0.114	0.127	0.151
	2	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		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
	1	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
		eScore	0.047	0.069	0.084	0.110	0.148	0.177	0.257
	2	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		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
	1	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

Author Manuscript

Author Manuscript

Author Manuscript

			Effect Size $\left(\sigma_g^2\right)$						
n	**#Repeat**	**Method**	**0**	**0.10**	**0.2**	**0.5**	**0.8**	**1.0**	**1.5**
		eScore	0.037	0.140	0.205	0.277	0.378	0.411	0.525
	2	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

TABLE 5

Application to the longitudinal pulmonary microbiome studies. `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 phylum level, while lower panel shows the results at 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 significant genus in the same phylum are included.

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	*Bacteroidetes*						*Firmicutes*					
	Porphyromonas			*Prevotella*			*Veillonella*			*Parvimonas*		
	`eRLRT`	`eScore`	`LinScore`	`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	1.0	1.0	0.46
DsbHb	0.33	0.32	0.87	1.0	1.0	0.07	0.26	0.25	0.62	0.07	0.08	0.11
FEV1	0.05(1.0)	0.05(1.0)	0.75	0.03 (0.35)	0.03 (1.0)	0.16	0.30	0.31	0.25	0.03	0.03	0.88
FVC	0.12	0.10	0.50	0.17	0.16	0.25	1.0	1.0	0.90	0.11	0.11	0.95
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	0.02 (0.05)	0.02 (0.07)	0.60