Title
Variance Component Selection With Applications to Microbiome Taxonomic Data

Journal
Frontiers in Microbiology, 9(MAR)

ISSN
1664-302X

Authors
Zhai, Jing
Kim, Juhyun
Knox, Kenneth S
et al.

Publication Date
2018

DOI
10.3389/fmicb.2018.00509

Peer reviewed

frontiers in Microbiology

METHODS
published: 28 March 2018
doi: 10.3389/fmicb.2018.00509

Variance Component Selection With Applications to Microbiome Taxonomic Data

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

[1] Department of Epidemiology and Biostatistics, University of Arizona, Tucson, AZ, United States, [2] Department of Biostatistics, University of California, Los Angeles, Los Angeles, CA, United States, [3] Division of Pulmonary, Allergy, Critical Care, and Sleep Medicine, Department of Medicine, University of Arizona, Tucson, AZ, United States, [4] Division of Pulmonary, Critical Care, Sleep, and Occupational Medicine, Indiana University Medical Center, Indianapolis, IN, United States

High-throughput sequencing technology has enabled population-based studies of the role of the human microbiome in disease etiology and exposure response. Microbiome data are summarized as counts or composition of the bacterial taxa at different taxonomic levels. An important problem is to identify the bacterial taxa that are associated with a response. One method is to test the association of specific taxon with phenotypes in a linear mixed effect model, which incorporates phylogenetic information among bacterial communities. Another type of approaches consider all taxa in a joint model and achieves selection via penalization method, which ignores phylogenetic information. In this paper, we consider regression analysis by treating bacterial taxa at different level as multiple random effects. For each taxon, a kernel matrix is calculated based on distance measures in the phylogenetic tree and acts as one variance component in the joint model. Then taxonomic selection is achieved by the lasso (least absolute shrinkage and selection operator) penalty on variance components. Our method integrates biological information into the variable selection problem and greatly improves selection accuracies. Simulation studies demonstrate the superiority of our methods versus existing methods, for example, group-lasso. Finally, we apply our method to a longitudinal microbiome study of Human Immunodeficiency Virus (HIV) infected patients. We implement our method using the high performance computing language `Julia`. Software and detailed documentation are freely available at https://github.com/JingZhai63/VCselection.

Keywords: Human Immunodeficiency Virus (HIV), lasso, longitudinal study, lung microbiome, MM-algorithm, variance component models, variable selection

OPEN ACCESS

Edited by:
Michele Guindani, University of California, Irvine, United States

Reviewed by:
Gwenael Piganeau, FR3724 Observatoire Océanologique de Banyuls sur Mer (OOB), France
Jun Chen, Mayo Clinic, United States

***Correspondence:**
Jin J. Zhou
jzhou@email.arizona.edu

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: *15 November 2017*
Accepted: *06 March 2018*
Published: *28 March 2018*

Citation:
Zhai J, Kim J, Knox KS, Twigg HL III, Zhou H and Zhou JJ (2018) Variance Component Selection With Applications to Microbiome Taxonomic Data. Front. Microbiol. 9:509. doi: 10.3389/fmicb.2018.00509

1. INTRODUCTION

The advent of high-throughput sequencing technologies has produced extensive microbial community data, which reveals the impact of human microbes on health and various diseases (Mardis, 2008; Haas et al., 2011; Hodkinson and Grice, 2015; Kuleshov et al., 2016; Wang and Jia, 2016). Microbial community data collected from oral, skin, and gastrointestinal tract samples have received early attention (Eckburg et al., 2005; Gill et al., 2006; Turnbaugh et al., 2009; Dewhirst et al., 2010; Grice and Segre, 2011). Studies of the respiratory tract microbiome did not start until the discovery of microbiome in the lungs of both healthy (Erb-Downward et al., 2011; Morris et al., 2013; Twigg III et al., 2013) and diseased populations (Zemanick et al., 2011; Lozupone et al., 2013) using culture-independent techniques. A pulmonary microbiome dataset was sampled

longitudinally from 30 HIV-infected individuals after starting highly active antiretroviral therapy (HAART). The objective is to study how the pulmonary microbiome impacts lung function of advanced HIV patients after HAART (Garcia et al., 2013; Lozupone et al., 2013; Twigg III et al., 2016).

After microbiome sequences have been acquired, they are usually clustered into Operational Taxonomic Units (OTUs): groups of sequences that correspond to taxonomic clusters or monophyletic groups (Caporaso et al., 2010). The abundance of an OTU is defined as the number of sequences in that OTU. The microbial community is then described by a list of OTUs, their abundances, and a phylogenetic tree. Regression methods have been a powerful tool to identify clusters of OTUs that are associated with or predictive of host phenotypes (Zhao et al., 2015; Wang and Zhao, 2016; Wang et al., 2017). Microbiome data presents several challenges. First microbiome abundances are sparse and the number of OTUs is usually much bigger than sample size. In our longitudinal data set, there are 2,964 OTUs and only two of them have abundance greater than 5%. When OTUs are included as predictors for clinical phenotypes in a regression model, regularizations are often used to overcome ill-conditioning. For example, Lin et al. (2014) proposed a linear log-contrast model with ℓ_1 regularization. Another possible strategy to overcome the sparsity of microbial data is to cluster multiple OTUs into their higher phylogenetic levels, e.g., genus, order, and phylum. Shi et al. (2016) extended Lin et al.'s (2014) model to allow selecting taxa at different higher taxonomic ranks. However, both methods overlook the distance information in the phylogenetic tree. A network-constrained sparse regression is proposed to achieve better prediction performance through a Laplacian regularization (Chen et al., 2012b, 2015b). Another popular approach for sparse linear regression is the group-wise selection scheme, group-lasso, which selects an entire group for inclusion or exclusion (Yuan and Lin, 2006; Garcia et al., 2013; Simon et al., 2013; Yang and Zou, 2015). Therefore, group-lasso is a natural tool for incorporating group information defined by the phylogenetic tree, but still misses fine level information. To encourage hierarchically close species to have similar effects on the phenotype, Wang and Zhao (2016) and Wang et al. (2017) both used tree topology information and fused variables that stay closer in a tree. However, this assumption may be violated. For example, the bacteria *Clostridia*, some species in this class convert dietary fiber into anti-inflammatory short-chain fatty acids, while others cause severe colitis. We, therefore, need a method that can incorporate biologically meaningful cluster information, phylogenetic distance, or tree information, can encourage sparse feature selection, and can handle possible adverse effect within clusters.

By modeling microbiome cluster effects as random effects, Zhai et al. (2017b) proposed a variance component model

$$\begin{aligned} &\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zb} + \sum_{l}^{L} h_l + \boldsymbol{\varepsilon} \\ &\boldsymbol{b} \sim \mathcal{N}(0, \sigma_d^2 \boldsymbol{I}_n),\ h_l \sim \mathcal{N}(\boldsymbol{0}, \sigma_{gl}^2 \boldsymbol{K}_l),\ \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma_e^2 \boldsymbol{I}_n), \end{aligned} \quad (1)$$

where $\boldsymbol{y}$, $\boldsymbol{X}$, and $\boldsymbol{\varepsilon}$ are the vertically stacked vectors/matrices of $\boldsymbol{y}_i$, $\boldsymbol{X}_i$, and $\boldsymbol{\varepsilon}_i$. The $\boldsymbol{y}_i$ is an $n_i \times 1$ vector of n_i repeated measures of the quantitative phenotype for an individual i. $\boldsymbol{X}_i$ is the $n_i \times p$ covariates. The $\boldsymbol{\varepsilon}_i$ is an $n_i \times 1$ vector of the random error. $\boldsymbol{Z}_i = (1, \ldots, 1)'$ is an $n_i \times 1$ design matrix linking the vector of random effects b_i to $\boldsymbol{y}_i$. $\boldsymbol{Z}$ is a block diagonal matrix with $\boldsymbol{Z}_i$ on its diagonal. $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects. The $\boldsymbol{b} = (b_i)$ is the subject-specific random effects. L is the total number of microbiome taxonomic clusters, N is the total number of individuals and $\sum_{i=1}^{N} n_i$ is the total number of observations. In model (Equation 1), h_l is the random effects generated by microbiome taxa l with covariance $\sigma_{gl}^2 \boldsymbol{K}_l$. $\boldsymbol{K}_l$ is a positive-definite kernel matrix derived from a distance matrix that is calculated based on the OTU abundances of taxa in the phylogenetic tree. Two common distance matrices are UniFrac Distance (Lozupone and Knight, 2005) and Bray-Curtis dissimilarity (Bray and Curtis, 1957). Therefore,

$$\mathrm{Var}(\boldsymbol{y}) = \sigma_d^2 \boldsymbol{Z}'\boldsymbol{Z} + \sum_{l=1}^{L} \sigma_{gl}^2 \boldsymbol{K}_l + \sigma_e^2 \boldsymbol{I}_n, \quad (2)$$

where σ_{gl}^2 and σ_d^2 are the phenotypic variance from microbiome clusters and between subject variance from repeated measurements. σ_e^2 is the within-subject variance that cannot be explained by either microbiome or repeated measurements. To identify associated microbiome taxa at different phylogeny levels is to select non-zero variance components at different phylogeny levels.

In this article, we adopt a penalized likelihood approach by regularizing variance components based on linear mixed effect models: variance component lasso selection (VC-lasso). We incorporate the phylogenetic tree information by using kernel matrices. We reduce the dimensionality of large and very sparse OTU abundances within a cluster by translating them into a random effect. Furthermore, our method can be applied to a longitudinal design, where an unpenalized variance component that captures the correlation of repeated measurements is included. Our Majorization-Minimization (MM) algorithm for variance component selection guarantees estimation and selection computational efficiency (Hunter and Lange, 2004; Hunter and Li, 2005; Zhou et al., 2011, 2015; Lange, 2016). Many statistical methods have been proposed related to the selection of random effects. Ibrahim et al. (2011) considered jointly selecting fixed and random effect in mixed effect model using the maximum likelihood with the smoothly clipped absolute deviation (SCAD) and adaptive lasso penalization. Fan and Li (2012) proposed a group variable selection strategy to select and estimate important random effects. Hui et al. (2017) extended this strategy to generalized linear mixed model by combining the penalized quasi-likelihood (PQL) estimation with sparsity-inducing penalties on the fixed and random coefficients. However, none of these methods can be easily extended to microbiome data and none of them use variance component regularization.

The rest of this paper is organized as follows. We introduce the variance component lasso selection method in section 2.

Section 3 conducts comparative simulation studies. Section 4 presents simulation and real data analysis results. We conclude with a discussion in section 5.

2. METHODS

2.1. Lasso Penalized Log-Likelihood

We consider model (Equation 2) with model parameters $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_m^2)$. The log-likelihood of our model is:

$$L(\boldsymbol{\beta}, \boldsymbol{\sigma}^2; \boldsymbol{y}, \boldsymbol{X}) = -\frac{1}{2}\ln\det(\boldsymbol{V}) - \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}), \tag{3}$$

where

$$\boldsymbol{V} = \sum_{i=1}^{m} \sigma_i^2 \boldsymbol{V}_i.$$

For the selection of non-zero variance components among a large number of variance components, we estimate the regression parameter $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$ by minimizing the lasso penalized log-likelihood function

$$pl(\boldsymbol{\beta}, \boldsymbol{\sigma}^2; \boldsymbol{y}, \boldsymbol{X}, \lambda) = -L(\boldsymbol{\beta}, \boldsymbol{\sigma}^2) + \lambda \sum_{i=1}^{m} c_i \sigma_i, \tag{4}$$

subject to nonnegativity constraint $\sigma_i \geq 0$. The first part $-L(\boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ of the penalized function (Equation 4) is the negative log-likelihood defined in Equation (3). The second part is the lasso penalty to enforce shrinkage of high-dimensional components. We do not penalize fixed effects $\boldsymbol{\beta}$. λ is the tuning parameter controlling model complexity; $c_i \in \{0, 1\}$ allows differential shrinkage of specific variance components. For example, when modeling longitudinal phenotypes with random intercept model, the corresponding variance component is unpenalized and always stays in the model. c_i can be chosen using different weighting schemes based on prior knowledge such as functional annotations.

2.2. Minimization of Penalized Likelihood via MM Algorithm

Minimizing the penalized negative log-likelihood is challenging due to non-convexity. Based on the Majorization-Minimization (MM) algorithm (Lange et al., 2000; Hunter and Lange, 2004), Zhou et al. (2015) proposed a strategy for maximizing the log-likelihood Equation (3) by alternate updating $\boldsymbol{\beta}$ and variance components $\boldsymbol{\sigma}^2$. We follow the same strategy to solve the lasso penalized likelihood estimation problem (Algorithm 1).

Given $\boldsymbol{\sigma}^{2(t)}$, updating $\boldsymbol{\beta}$ is a general least squares problem with solution

$$\boldsymbol{\beta}^{(t+1)} = \left(\boldsymbol{X}'\boldsymbol{V}^{-(t)}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-(t)}\boldsymbol{y}, \tag{5}$$

where $\boldsymbol{V}^{-(t)}$ represents the tth-step update of $\boldsymbol{V}^{-1}$. Given $\boldsymbol{\beta}^{(t)}$, updating the variance components $\boldsymbol{\sigma}^2$ invokes the MM principle. To minimize the objective function $pl(\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\sigma}^2)$,

Algorithm 1: MM algorithm for minimizing lasso penalized likelihood (Equation 4).

Data: $\boldsymbol{y}$, $\boldsymbol{X}$, $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_m$, λ

Result: $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\sigma}}^2$ such that $pl(\boldsymbol{\beta}, \boldsymbol{\sigma}^2) = -L(\boldsymbol{\beta}, \boldsymbol{\sigma}^2) + \lambda \sum_{i=1}^{m} c_i\sigma_i$ is minimized.

Initialize $\sigma_i^{(0)} > 0$. $i = 1, \ldots, m$ **repeat**

$\boldsymbol{V}^{(t)} \leftarrow \sum_{i=1}^{m} \sigma_i^{2(t)} \boldsymbol{V}_i$;

$\boldsymbol{\beta}^{(t)} \leftarrow \text{argmin}_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{V}^{-(t)}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$;

$\sigma_i^{(t+1)} \leftarrow \sigma_i^{(t)}$ by finding polynomial roots of $\boldsymbol{P}(\cdot) = 0$, $i = 1, \ldots, m$

$$\boldsymbol{P}\left(\sigma_i^{(t+1)}\right) = \sigma_i^{4(t+1)}\text{tr}\left(\boldsymbol{V}^{-(t)}\boldsymbol{V}_i\right) + \lambda\sigma_i^{3(t+1)} - \sigma_i^{4(t)}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)}\right)'\boldsymbol{V}^{-(t)}\boldsymbol{V}_i\boldsymbol{V}^{-(t)}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)}\right)$$

until *objective function pl converges*;

the majorization step operates by creating a surrogate function $g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ that satisfies two conditions

$$\text{dominance condition}: pl(\boldsymbol{\theta}) \leq g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \text{ for all } \boldsymbol{\theta}$$
$$\text{tangent condition}: pl(\boldsymbol{\theta}^{(t)}) = g(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}).$$

The second M of the MM principle minimizes the surrogate function to produce the next iterate $\boldsymbol{\theta}^{(t+1)}$. Then we have

$$pl(\boldsymbol{\theta}^{(t+1)}) \leq g(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) \leq g(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) = pl(\boldsymbol{\theta}^{(t)}).$$

Therefore, when the surrogate function is minimized, the objective function $f(\boldsymbol{\theta})$ is driven downhill. We combine two following majorizations to construct the surrogate function. First, with all $\boldsymbol{V}_i$ being positive semidefinite, Zhou et al. (2015) show that

$$\begin{aligned}
\boldsymbol{V}^{(t)}\boldsymbol{V}^{-1}\boldsymbol{V}^{(t)} &= \left(\sum_{i=1}^{m}\sigma_i^{2(t)}\boldsymbol{V}_i\right)\left(\sum_{i=1}^{m}\sigma_i^{2}\boldsymbol{V}_i\right)^{-1}\left(\sum_{i=1}^{m}\sigma_i^{2(t)}\boldsymbol{V}_i\right) \\
&\preceq \sum_{i=1}^{m}\frac{\sigma_i^{2(t)}}{\sum_j\sigma_j^{2(t)}}\left(\frac{\sum_j\sigma_j^{2(t)}}{\sigma_i^{2(t)}}\sigma_i^{2(t)}\boldsymbol{V}_i\right) \\
&\quad \left(\frac{\sum_j\sigma_j^{2(t)}}{\sigma_i^{2}}\sigma_i^{2}\boldsymbol{V}_i\right)^{-1}\left(\frac{\sum_j\sigma_j^{2(t)}}{\sigma_i^{2(t)}}\sigma_i^{2(t)}\boldsymbol{V}_i\right) \\
&= \sum_{i=1}^{m}\frac{\sigma_i^{4(t)}}{\sigma_i^2}\boldsymbol{V}_i\boldsymbol{V}_i^{-1}\boldsymbol{V}_i = \sum_{i=1}^{m}\frac{\sigma_i^{4(t)}}{\sigma_i^2}\boldsymbol{V}_i,
\end{aligned}$$

leading to the first majorization

$$\begin{aligned}
&(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\
&\preceq (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{V}^{-(t)}\left(\sum_{i=1}^{m}\frac{\sigma_i^{4(t)}}{\sigma_i^2}\boldsymbol{V}_i\right)\boldsymbol{V}^{-(t)}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).
\end{aligned} \tag{6}$$

It separates the variance components $\sigma_1^2, \ldots, \sigma_m^2$ in the quadratic term of the log-likelihood function (Equation 4). By the supporting hyperplane inequality, the second majorization is

$$\ln \det \boldsymbol{V} \leq \ln \det \boldsymbol{V}^{(t)} + \mathrm{tr}\big[\boldsymbol{V}^{-(t)}\big(\boldsymbol{V} - \boldsymbol{V}^{(t)}\big)\big], \tag{7}$$

which separates $\sigma_1^2, \ldots, \sigma_m^2$ in the log-determinant term of Equation (4). The overall majorization $g(\boldsymbol{\sigma}^2|\boldsymbol{\sigma}^{2(t)})$ of $pl(\boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ is obtained by combining Equations (6) and (7)

$$\begin{aligned} g\big(\boldsymbol{\sigma}^2|\boldsymbol{\sigma}^{2(t)}\big) &= \frac{1}{2}\mathrm{tr}\big(\boldsymbol{V}^{-(t)}\boldsymbol{V}\big) + \frac{1}{2}\big(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)}\big)'\boldsymbol{V}^{-(t)} \\ &\quad \left(\sum_{i=1}^{m}\frac{\sigma_i^{4(t)}}{\sigma_i^2}\boldsymbol{V}_i\right)\boldsymbol{V}^{-(t)}\big(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)}\big) + \lambda\sum_{i=1}^{m}\sigma_i + s^{(t)} \\ &= \sum_{i=1}^{m}\left[\frac{\sigma_i^2}{2}\mathrm{tr}\big(\boldsymbol{V}^{-(t)}\boldsymbol{V}_i\big) + \frac{\sigma_i^{4(t)}}{2\sigma_i^2}\big(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)}\big)'\right. \\ &\quad \left.\boldsymbol{V}^{-(t)}\boldsymbol{V}_i\boldsymbol{V}^{-(t)}\big(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)}\big) + \lambda\sigma_i\right] + s^{(t)}, \end{aligned} \tag{8}$$

where $s^{(t)}$ is an irrelevant constant term.

We minimize the surrogate function (Equation 8) by setting the derivative of $g(\boldsymbol{\sigma}^2|\boldsymbol{\sigma}^{2(t)})$ to zero. The update $\sigma_i^{(t+1)}$ for variance component $\sigma_i^{(t)}$ is chosen among the positive roots of the polynomial

$$\begin{aligned} \boldsymbol{P}\left(\sigma_i^{(t+1)}\right) &= \sigma_i^{4(t+1)}\mathrm{tr}\big(\boldsymbol{V}^{-(t)}\boldsymbol{V}_i\big) + \lambda\sigma_i^{3(t+1)} \\ &\quad - \sigma_i^{4(t)}\big(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)}\big)'\boldsymbol{V}^{-(t)}\boldsymbol{V}_i\boldsymbol{V}^{-(t)}\big(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)}\big) \end{aligned}$$

or 0, whichever yields the largest objective value. The alternating updates repeat until

$$|\, pl\big(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\sigma}^{2(t+1)}\big) - pl\big(\boldsymbol{\beta}^{(t)}, \boldsymbol{\sigma}^{2(t)}\big)\, | < tol * (|\, pl\big(\boldsymbol{\beta}^{(t)}, \boldsymbol{\sigma}^{2(t)}\big)\, | + 1),$$

where *tol* is the pre-specified tolerance. The default tolerance is 10^{-4}.

2.3. Tuning Parameter Selection

The tuning parameter λ in the penalized likelihood estimation is chosen by a 5-fold cross-validation procedure based on g-Measure $= \sqrt{\text{sensitivity} * \text{specificity}}$. g-Measure is an indicator of the model selection accuracy. g-Measure $= 1$ indicates the best accuracy and g-Measure $= 0$ the worst (Zhai et al., 2017a). It can counteract the imbalance between the number of of irrelevant and relevant clusters. Therefore, we present g-Measure instead of sensitivity (true positive rate) and specificity (true negative rate) alone (Supplementary Material section 3). Akaike Information Criterion (AIC) (Akaike, 1998) and Schwarzs Bayesian Information Criterion (BIC) (Schwarz et al., 1978) are used in the real data analysis. Performance comparisons between cross-validation and AIC/BIC are provided in the Supplementary Material section 4.

2.4. Software Implementation

We implement our method using the high performance computing language `Julia`. UniFrac distance matrices are computed using our `Julia` package `PhylogeneticDistance`.

TABLE 1 | Simulation parameter configurations.

	Non-zero variance components	Cluster/kernel	Design	$\sigma_g^{2\dagger}$	Method
Scenario 1: *Selection under different sample sizes*					
$n = 20, 50, 100$; simulated count data	$l = 1, 2, 3, 4, 5$	genus; $\boldsymbol{K}_W$	longitudinal; cross-sectional	1, 5, 25, 100	VC-lasso group-lasso
Scenario 2: *Selection under different number of non-zero variance components*					
$n = 50$; simulated count data	(i) $l = 20, 30$; (ii) $l = 1, 2, 3, 4, 5$; (iii) $l = 1, 2, 3, \ldots, 15$;	genus; $\boldsymbol{K}_W$	longitudinal; cross-sectional	1, 5, 25, 100	VC-lasso group-lasso
Scenario 3: *Selection under different UniFrac distance kernels*					
$n = 50$; simulated count data	$l = 1, 2, 3, 4, 5$	genus; $\boldsymbol{K}_W$, $\boldsymbol{K}_{UW}$, $\boldsymbol{K}_{VAW}$, $\boldsymbol{K}_0$, $\boldsymbol{K}_{0.5}$	longitudinal; cross-sectional	1, 5, 25, 100	VC-lasso group-lasso
Scenario 4: *Selection under fixed effect model*					
$n = 50$; simulated count data	$l = 20, 30$; $l = 1, 2, 3, 4, 5$; $l = 1, 2, 3, \ldots, 15$;	genus; $\boldsymbol{K}_W$	cross-sectional	1, 5, 25, 100	VC-lasso group-lasso

Throughout simulations, $\sigma_e^2 = 1$, $\beta_1 = \beta_2 = 0.1$. We use $\sigma_d^2 = 0.6$ and 3 repeated measurements in the longitudinal design. We use $\sigma_d^2 = 0$ for the cross-sectional design. Group-lasso is performed only in the cross-sectional design.

†*The non-zero variance components are assumed to have equal effect strength in each simulation setting.*

3. SIMULATION

In this section, we conduct simulation studies to evaluate the variable selection and prediction performance of VC-lasso and compare the results with the conventional method group-lasso as implemented in the `gglasso` package (Yang and Zou, 2015). Phenotypes are simulated based on one real pulmonary microbiome dataset and one simulated longitudinal microbiome dataset. We first describe real and simulated microbiome abundance data, phylogenetic tree, and then detail our four phenotype simulation schemes (**Table 1**).

The real pulmonary microbiome data has been discussed in Twigg III et al. (2016). Thirty individuals were recruited. During up to three-years follow-up, lung functions and microbiome composition were measured 2–4 times for each individual. The longitudinal microbiome taxonomic data is summarized as 2,964 OTUs with a phylogenetic tree (Twigg III et al., 2016). Longitudinal microbiome abundance data is generated by a Zero-Inflated Beta Random Effect model using `R` package `ZIBR` in Supplementary Material section 2 (Chen and Li, 2016). For cross-sectional design, we generate taxonomic data using a Dirichlet-Multinomial (DM) model (Chen et al., 2012a). Simulation parameters, such as proportion of each OTU and the overall dispersion, are estimated from our real pulmonary microbiome abundance data.

Given simulated microbiome count data and taxonomic information, we classify 2,353 of 2,964 OTUs to 30 genera (taxa clusters) and the remaining 611 of 2,964 OTUs are grouped into the 31st cluster named *other* (**Table 2**). As described in Supplementary Material section 1, UniFrac distance matrices ($\boldsymbol{D}$) of the 31 clusters are computed and converted to kernel matrices as

$$\boldsymbol{K} = -\frac{1}{2}(\boldsymbol{I} - \frac{\boldsymbol{1}\boldsymbol{1}'}{n})\boldsymbol{D}^2(\boldsymbol{I} - \frac{\boldsymbol{1}\boldsymbol{1}'}{n}) \quad (9)$$

followed by a positive definiteness correction (Chen and Li, 2013; Zhao et al., 2015). All of the microbiome kernel matrices $\boldsymbol{K}$ are scaled to have unit Frobenius norm.

Phenotypes are simulated based on the following scenarios.

3.1. Scenario 1: Selection Under Different Sample Size

Longitudinal and cross-sectional responses are generated by

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2\beta_2,\ \sigma_d^2\boldsymbol{Z}\boldsymbol{Z}' + \sum_{l=1}^{L}\sigma_{gl}^2\boldsymbol{K}_l + \sigma_e^2\boldsymbol{I}), \quad (10)$$

where $\sigma_{gl}^2 > 0$ for $l = 1, \ldots, 5$ and $\sigma_{gl}^2 = 0$ otherwise. The total number of variance components for microbiome clusters is $L = 31$. The true model has five non-zero variance components including *Anaerococcus*, *Atopobium*, *Actinomyces*, *Campylobacter*, and *Capnocytophaga*. We compare the selection performance at three sample sizes: $n = 20,\ 50,\ 100$. For cross-sectional design, responses are simulated by setting $\sigma_d^2 = 0$.

TABLE 2 | List of 31 Genera.

	Genus	Phylum	No of OTU	Mean Reads
1	*Actinomyces*	*Actinobacteria*	150	230.59
2	*Anaerococcus*	*Firmicutes*	17	2.90
3	*Atopobium*	*Actinobacteria*	22	40.83
4	*Campylobacter*	*Proteobacteria*	31	51.05
5	*Capnocytophaga*	*Bacteroidetes*	31	70.81
6	*Catonella*	*Firmicutes*	22	40.09
7	*Corynebacterium*	*Actinobacteria*	47	12.22
8	*Flavobacterium*	*Bacteroidetes*	25	5.08
9	*Fusobacterium*	*Fusobacteria*	55	174.29
10	*Gemella*	*Firmicutes*	17	72.11
11	*Lactobacillus*	*Firmicutes*	33	141.10
12	*Leptotrichia*	*Fusobacteria*	15	12.40
13	*Megasphaera*	*Firmicutes*	14	36.99
14	*Methylobacterium*	*Proteobacteria*	11	2.88
15	*Neisseria*	*Proteobacteria*	18	109.61
16	*OD1_genera_incertae_sedis*	*OD1*	75	0.92
17	*Parvimonas*	*Firmicutes*	20	76.46
18	*Peptoniphilus*	*Firmicutes*	11	1.16
19	*Porphyromonas*	*Bacteroidetes*	42	134.41
20	*Prevotella*	*Bacteroidetes*	304	833.35
21	*Rothia*	*Actinobacteria*	16	49.83
22	*Selenomonas*	*Firmicutes*	50	16.16
23	*Sneathia*	*Fusobacteria*	12	37.09
24	*Sphingomonas*	*Proteobacteria*	14	0.61
25	*SR1_genera_incertae_sedis*	*SR1*	17	5.95
26	*Streptococcus*	*Firmicutes*	66	1,107.81
27	*TM7_genera_incertae_sedis*	*TM7*	61	40.54
28	*Treponema*	*Spirochaetes*	60	51.62
29	*Unclassified*	*Unclassified*†	1,068	258.65
30	*Veillonella*	*Firmicutes*	29	370.85
31	*Others*	*Others*	611	1,009.88

Summary of phylum information, the number of OTUs, and the average abundance (across sample and time points) within each genus from the pulmonary microbiome dataset are shown.

† *The genus unclassified may belong to phylum unclassified or other 12 phyla.*

3.2. Scenario 2: Selection Under Different Numbers of Non-zero Variance Components

The sample size is fixed at $n = 50$ in this scenario. Responses are generated by model (Equation 10) with different numbers of non-zero variance components. In Supplementary Material section 5, VC-lasso is evaluated when the number of variance components in the model is large.

(1) 2 non-zero variance components: $\sigma_{g20}^2 > 0$, $\sigma_{g30}^2 > 0$, and $\sigma_{gl}^2 = 0$ otherwise. Two associated genera are *prevolleta* and *veillonella*.

(2) 5 non-zero variance components: $\sigma_{gl}^2 > 0$ for $l = 1, 2, \ldots, 5$ and $\sigma_{gl}^2 = 0$ otherwise. Associated clusters are *Anaerococcus*, *Atopobium*, *Actinomyces*, *Campylobacter*, and *Capnocytophaga*.

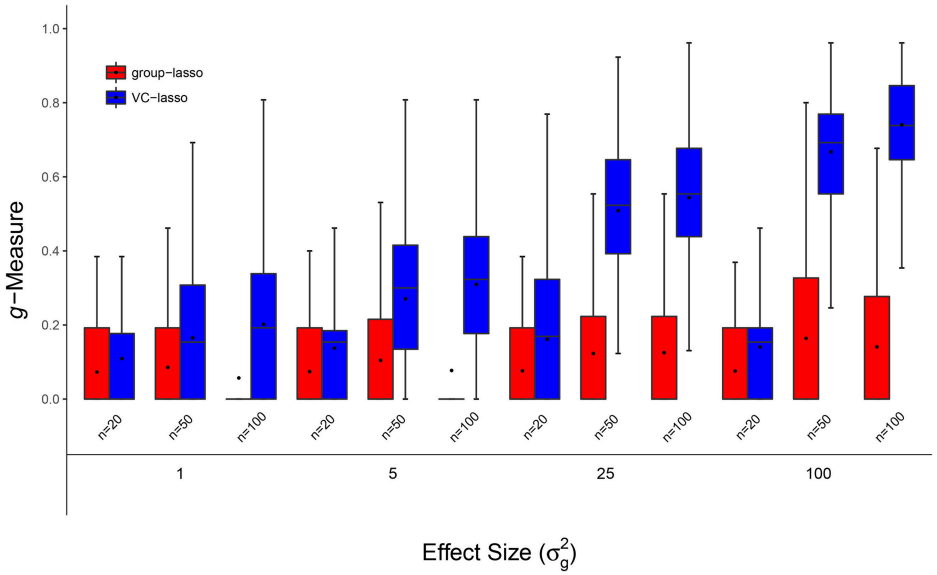

FIGURE 1 | Scenario 1: Estimated ***g***-Measure of both VC-lasso and group-lasso under different sample sizes for models with 5 non-zero variance components in a cross-sectional design. Three sample sizes, $n = 20,\ 50,\ 100$, are compared and $\sigma_d^2 = 0$.

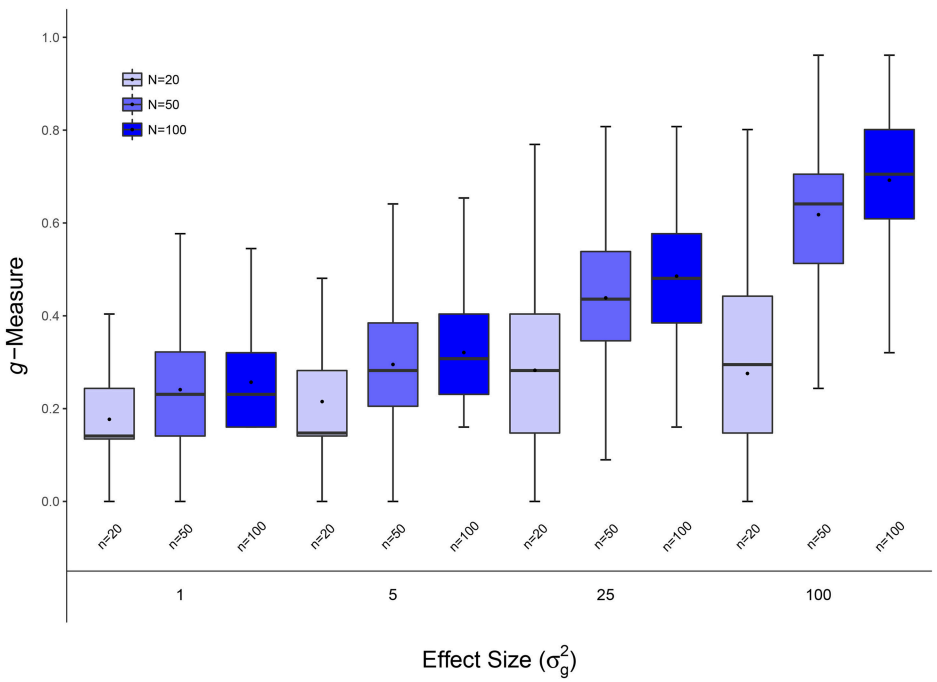

FIGURE 2 | Scenario 1: Estimated ***g***-Measure of VC-lasso under different sample sizes for models with 5 non-zero variance components in a longitudinal design. Three sample sizes, $n = 20,\ 50,\ 100$, are compared and $\sigma_d^2 = 0.6$.

(3) 15 non-zero variance components: $\sigma_{g_l}^2 > 0$ for $l = 1, 2, \ldots,$ 15 and $\sigma_{g_l}^2 = 0$ otherwise. Associated clusters, including *Actinomyces*, *Anaerococcus*, ..., and *Neisseria* are listed in **Table 2**.

3.3. Scenario 3: Selection Under Different UniFrac Distance Kernels

The sample size is fixed at $n = 50$ with 5 non-zero variance components. We compare the selection performance

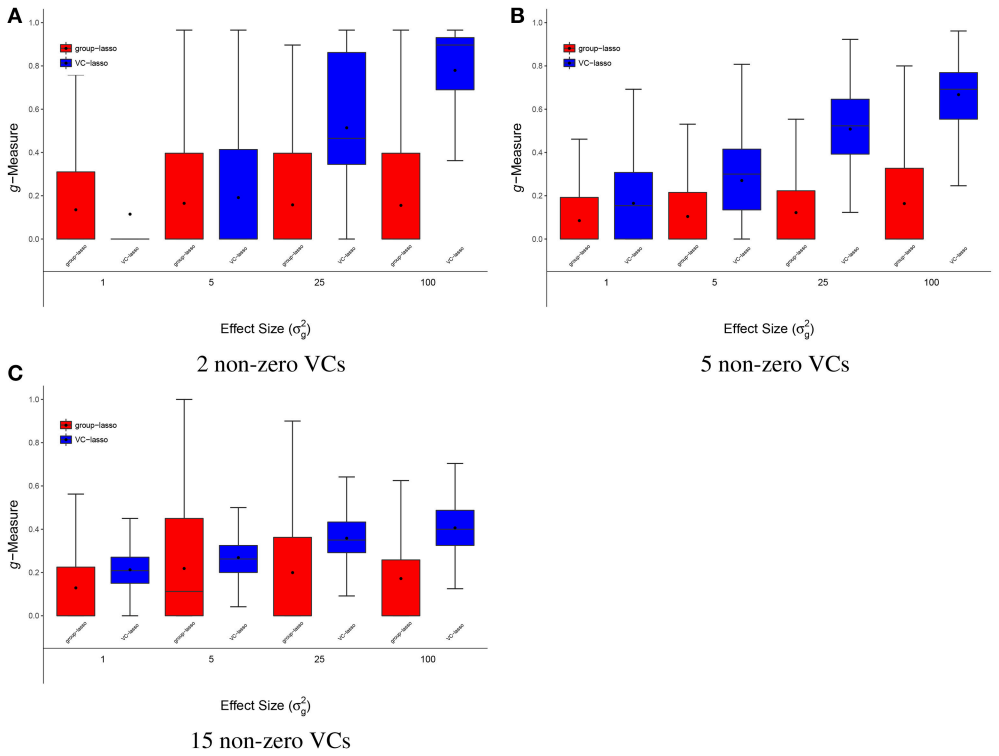

FIGURE 3 | Scenario 2: Estimated ***g***-Measure of both VC-lasso and group-lasso under different number of non-zero variance components in a cross-sectional design. The number of non-zero variance components (VCs) are set to 2 **(A)**, 5 **(B)**, 15 **(C)**, sample size is $n = 50$, and $\sigma_d^2 = 0$.

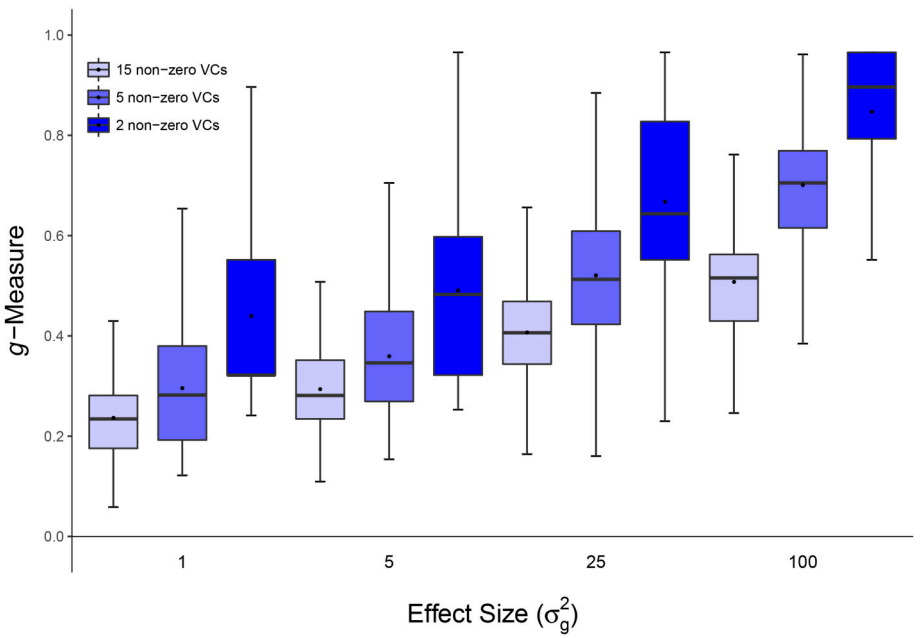

FIGURE 4 | Scenario 2: Estimated ***g***-Measure of VC-lasso under different number of non-zero variance components in a longitudinal design. Three different numbers of non-zero variance components (VCs), 2, 5, 15, are shown, sample size is set to $n = 50$ and $\sigma_d^2 = 0.6$.

using kernels defined by 5 different distance measures: variance adjusted weighted UniFrac distance ($\boldsymbol{K}_{VAW}$) Chang et al., 2011), generalized UniFrac distance ($\boldsymbol{K}_0$, $\boldsymbol{K}_{0.5}$) (Chen et al., 2012a), unweighted UniFrac distance ($\boldsymbol{K}_{UW}$) (Lozupone and Knight, 2005), and weighted UniFrac distance ($\boldsymbol{K}_W$) (Lozupone et al., 2007).

3.4. Scenario 4: Selection Under Fixed Effect Model

We again use the sample size $n = 50$ and vary the number of clusters containing signal. Responses are simulated by a fixed effect model

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2\beta_2 + \boldsymbol{G}_1^*\gamma_1 + \boldsymbol{G}_2^*\gamma_2 + \ldots + \boldsymbol{G}_u^*\gamma_u,\ \sigma_e^2\boldsymbol{I}), \quad (11)$$

where $\boldsymbol{G}_1^*, \boldsymbol{G}_2^*, \ldots, \boldsymbol{G}_u^*$ are OTU count matrices of different clusters scaled by their sample maximum. u is the total number of clusters with effects that ranges from 2 to 15. Fixed effect vector $\boldsymbol{\gamma}_l$ for cluster l are generated from $\boldsymbol{\gamma}_l \sim \mathcal{N}(0, \sigma_{gl}^2\boldsymbol{I})$ and are fixed for each simulation replicate.

We applied VC-lasso to scenarios 1-3 using both longitudinal and cross-sectional designs. Scenario 4 is performed using a cross-sectional design only. We compare our approach with group-lasso (R package `gglasso`) in all 4 scenarios for cross-sectional design because the `gglasso` package cannot handle longitudinal data.

We set the within-individual variance $\sigma_e^2 = 1$ throughout simulations. The between individual variance of random intercept is set to $\sigma_d^2 = 0.6$ for longitudinal design and $\sigma_d^2 = 0$ otherwise (Twigg III et al., 2016). The effect strength is set to $\sigma_g^2 = 1, 5, 25, 100$ (Chen et al., 2015a). We set the non-zero variance components to have the same effect strength under each setting, therefore omit subscript l. Two covariates $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are generated from the standard normal distribution and effect sizes are set to $\beta_1 = \beta_2 = 0.1$. 1000 Monte Carlo simulation replicates

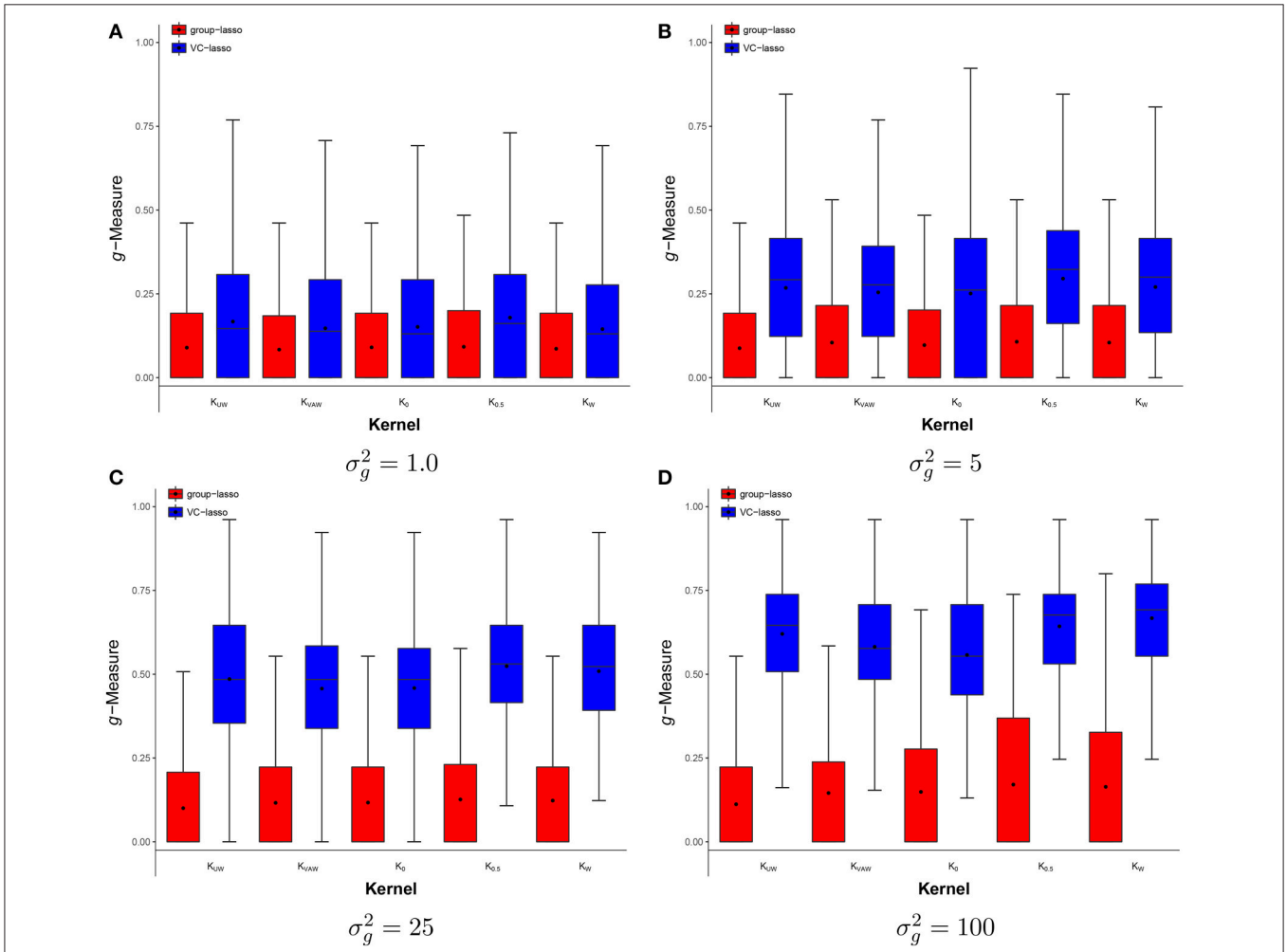

FIGURE 5 | Scenario 3: Estimated ***g***-Measure of both VC-lasso and group-lasso under different UniFrac distance kernels in a cross-sectional design. Five different kernels, K_{UW}, K_{VAW}, K_0, $K_{0.5}$ and K_W, and two methods, VC-lasso and group-lasso, are displayed in a cross-sectional design. Four effect strengths, 1 **(A)**, 5 **(B)**, 25 **(C)**, and 100 **(D)** are shown. There are 5 non-zero variance components, sample size is $n = 50$, and $\sigma_d^2 = 0$.

are generated. We split each dataset to training (80%) and testing (20%). Five-fold cross-validation is performed in training set to estimate the optimal λ^*. Selection performance is evaluated and reported by applying λ^* to the testing set.

4. RESULTS

4.1. Analysis of Simulated Data

The simulation results are summarized in **Figures 1–9** including variable selection performance under different sample sizes (**Figures 1**, **2**), different numbers of non-zero variance components (**Figures 3**, **4**), and different UniFrac distance measures (**Figures 5**, **6**) for both cross-sectional and longitudinal designs. Comparisons between VC-lasso and group-lasso are shown in all cross-sectional simulation studies.

The trajectories of g-Measure versus tuning parameter λ from cross-validation is presented in **Figure 9**. g-Measures remain stable or slightly decrease as λ getting larger under moderate effect size when $\sigma_g^2 = 1$ and 5. It starts to decrease when λ is greater than 0.6. **Figure 9** suggests that the trajectories of tuning criteria is generally consistent across sample sizes, effect sizes, and study designs.

4.1.1. Scenario 1: Selection Under Different Sample Sizes

Figures 1, **2**, **8A** display performance of selection (g-Measure) and prediction (area under the receiver operating characteristic curve, AUROC). In **Figure 1**, we compare VC-lasso (blue bar) and group-lasso (red bar) using cross-sectional design. In **Figure 2**, we compare the g-Measure of VC-lasso under different sample sizes using a longitudinal design. For both cross-sectional and longitudinal designs, g-Measure of VC-lasso boosts with increased sample size and effect sizes. Except for the third quartile of g-Measure over 1,000 replicates for sample size, $n = 20$, VC-lasso always outperforms group-lasso in this scenario.

Area under receiver operating characteristic (AUROC) is used to evaluate the prediction performance (**Figure 8A**) when effect size is fixed at $\sigma_g^2 = 25$. Larger AUROC represents better prediction ability. For VC-lasso, AUROC increases with sample size under cross-sectional design. For longitudinal study, $n = 50$ has similar AUROC with $n = 100$, which indicates the optimal prediction we can receive under this simulation setting. The AUROC of group-lasso (red bar) is similar under different sample sizes and shows no advantages compared to the VC-lasso.

4.1.2. Scenario 2: Selection Under Different Number of Non-zero Variance Components

Figures 3, **4**, **8B** show simulation results for the selection under different number of non-zero variance components. Specifically, **Figure 3** shows g-Measure for both VC-lasso and group-lasso in a cross-sectional design, while **Figure 4** presents g-Measure for VC-lasso in a longitudinal design.

In **Figures 3**, **4**, the performance of VC-lasso selection improves when effect size increases. For a model with 2 non-zero variance components, the true discovery rate (TDR, or sensitivity) is either 0, 0.5 or 1.0, which lead to a large variation of the g-Measure (**Figure 3A**). As more non-zero variance

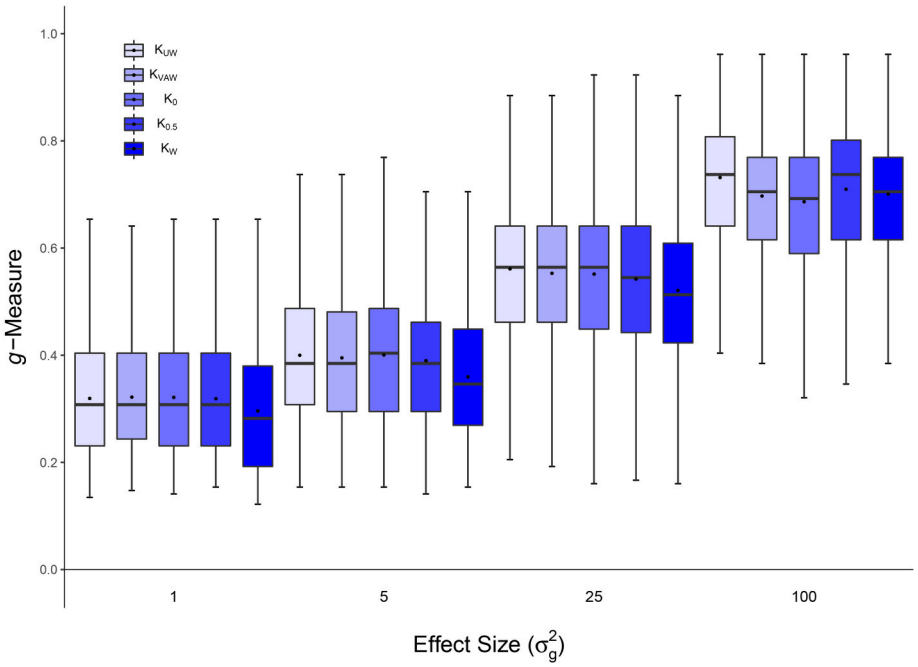

FIGURE 6 | Scenario 3: Estimated ***g***-Measure of VC-lasso under different UniFrac distance kernels in a longitudinal design.Five different kernels, K_{UW}, K_{VAW}, K_0, $K_{0.5}$ and K_W, are compared. There are 5 non-zero variance components. Sample size is $n = 50$ and $\sigma_d^2 = 0.6$.

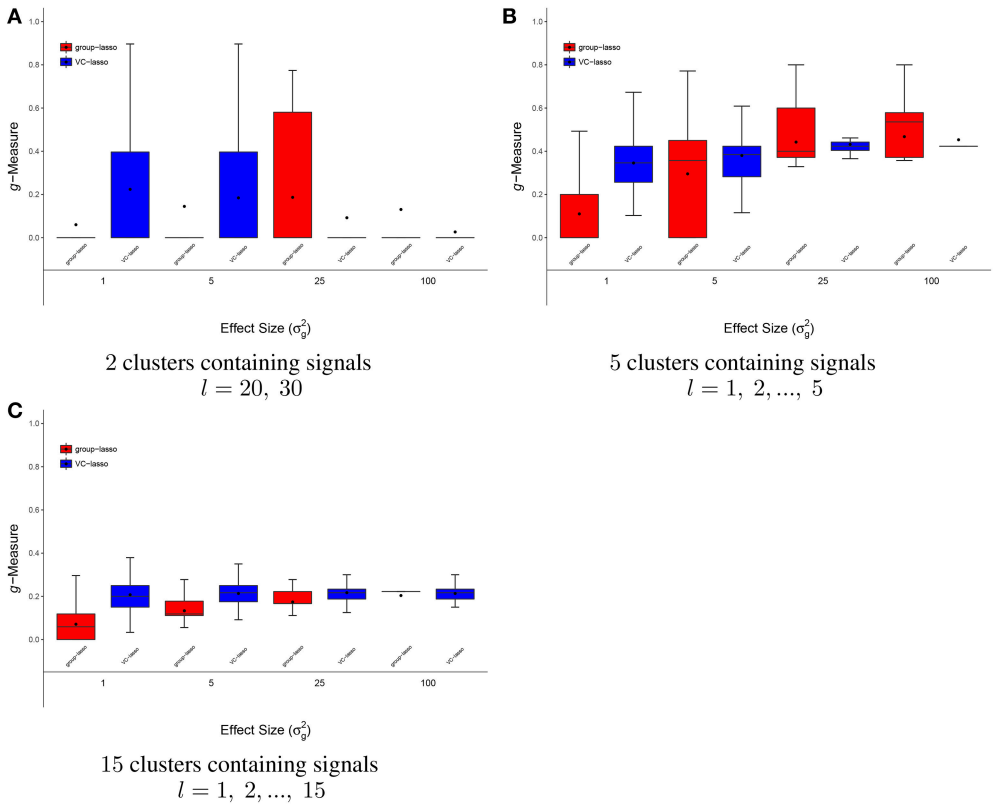

FIGURE 7 | Scenario 4: Estimated ***g***-Measure of VC-lasso and group-lasso under fixed effect model in a cross-sectional design. There are 2 **(A)**, 5 **(B)**, 15 **(C)** clusters with signals. Sample size is $n = 50$ and $\sigma_d^2 = 0$.

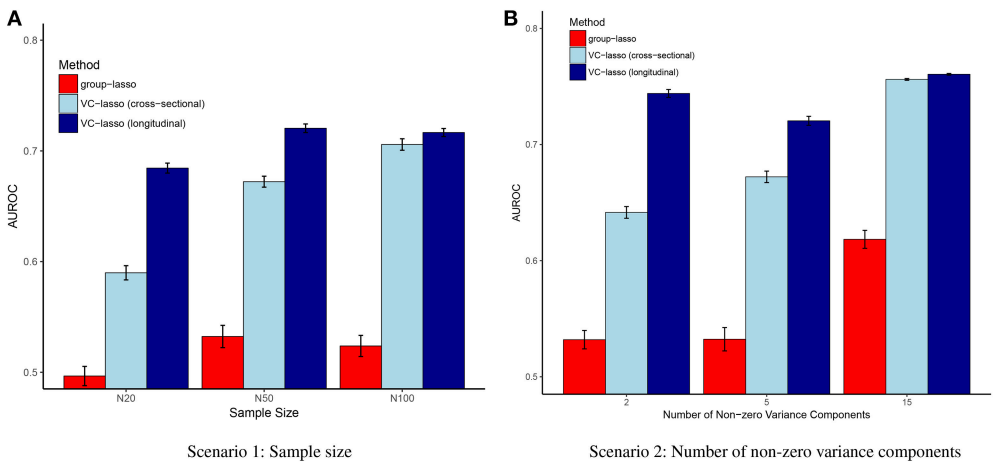

FIGURE 8 | Scenario 1 & 2: AUROC. The AUROC is presented as the mean ± 95% confidence interval based on 1,000 simulation replicates for each simulation scenario when $\sigma_g^2 = 25$. **(A)** Scenario 1; **(B)** Scenario 2.

components are included, in **Figure 3B,C** the trajectory of *g*-Measures becomes smoother. The *g*-Measures of VC-lasso are higher than the group-lasso in most simulation settings except that group-lasso has larger third quartile when $\sigma_g^2 = 1$ in **Figure 3A** and $\sigma_g^2 = 5$ in **Figure 3C**. As shown in **Figure 8B**, VC-lasso has a better prediction ability with an increased number of non-zero variance components. Compared with our method, group-lasso is uncompetitive in predictive ability.

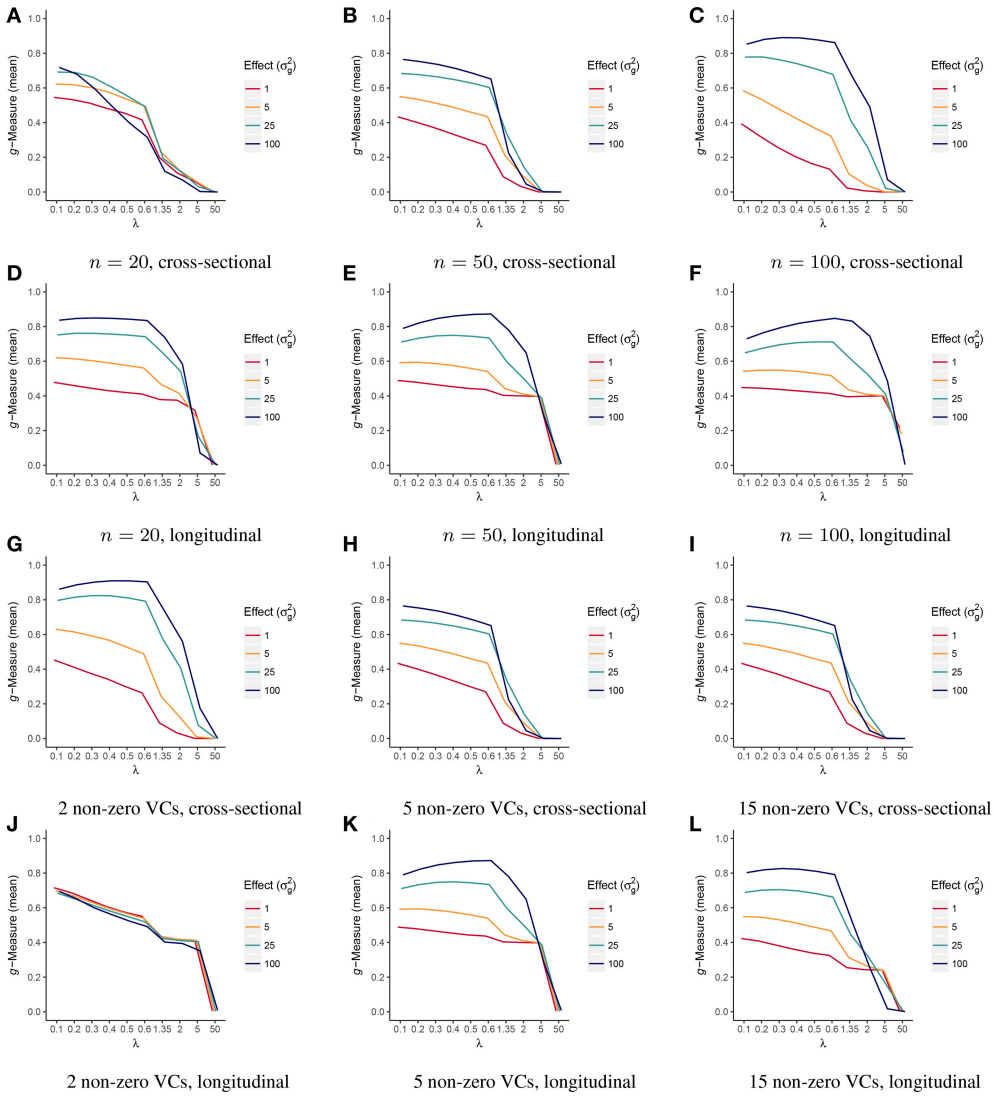

FIGURE 9 | Trajectories of estimated ***g***-Measure as a function of tuning parameter **λ** (scenario 1 & 2). Estimated *g*-Measure is displayed as the mean of 5-fold cross-validation under sample sizes, $n = 20$ **(A,D)**, 50 **(B,E)**, 100 **(C,F)**, or 2 **(G,J)**, 5 **(H,K)**, 15 **(I,L)** non-zero variance components (VCs) in both cross-sectional and longitudinal designs.

4.1.3. Scenario 3: Selection Under Different UniFrac Distance Kernels

We compare the *g*-Measure of five different kernels in **Figures 5**, **6** for the cross-sectional and longitudinal design, respectively. Using longitudinal simulated data, the box-plot of *g*-Measure shows that the five kernels have similar performance except that the K_W has the lowest third quartile and K_0 has the lowest first quartile when σ_g^2 is large. Under the same effect strength (σ_g^2) in the cross-sectional design (**Figure 5**), the *g*-Measure of five kernels are almost identical except that K_0 has slightly smaller *g*-Measure and wider range than other kernels. For example, K_0 has the lowest first quartile in **Figures 5B**. This suggests that the kernels computed from different UniFrac distance play a minor part in the selection performance and our method is superior to group-lasso regardless of kernel types.

4.1.4. Scenario 4: Selection Under Fixed Effect Model

Figure 7 has a distinctive pattern from the above scenarios. For the case that only two microbiome clusters contain signals (*Prevotella* and *Veillonella*), both methods do not perform well (**Figure 7A**). In **Figures 7B,C**, *g*-Measures for both methods improve with increased effect sizes and VC-lasso outplays group-lasso with $\sigma_g^2 = 1$. For $\sigma_g^2 = 5$, 25, average and median *g*-Measure of VC-lasso across simulation replicates outperform group-lasso. Besides, we notice that the range of *g*-Measure for VC-lasso becomes smaller as signal strengths increase, suggesting the prediction performance stabilizes as the association with the

TABLE 3 | Analysis of Forced expiratory volume in one second (FEV1) at genus level in the real pulmonary microbiome cohort using variance component lasso selection (VC-lasso) and exact tests.

	VC-lasso			Exact tests		
	Rank	**Genus**	**Phylum info**	**eRLRT**	**eLRT**	**eScore**
Baseline	1	*Corynebacterium*	*Actinobacteria*	0.28	0.30	0.30
	2	*TM7_genera_incertae_sedis*	*TM7*	1.00	1.00	1.00
	3	*Anaerococcus*	*Firmicutes*	0.06	0.06	0.07
	4	*Neisseria*	*Proteobacteria*	1.00	1.00	1.00
	5	*Treponema*	*Spirochaetes*	0.13	0.14	0.14
Longitudinal	1	*Corynebacterium*	*Actinobacteria*	1.00	–	1.00
	2	*Actinomyces*	*Actinobacteria*	0.00	–	0.01
	3	*Prevotella*	*Bacteroidetes*	0.01	–	0.01
	4	*TM7_genera_incertae_sedis*	*TM7*	1.00	–	1.00
	5	*Porphyromonas*	*Bacteroidetes*	0.00	–	0.00
	6	*Megasphaera*	*Firmicutes*	0.06	–	0.06

The phylum information is provided for selected genera. Tuning parameter $\lambda*$ for baseline and longitudinal data is set to 0.01 and 0.2, respectively. Rank represents the order of genera that appear in the solution path. Results of eLRT are omitted as it is equivalent to eRLRT in a longitudinal design.

TABLE 4 | Analysis of forced expiratory flow (FEF) at genus level in the real pulmonary microbiome cohort using variance component lasso selection (VC-lasso) and exact tests.

	VC-lasso			Exact tests		
	Rank	**Genus**	**Phylum info**	**eRLRT**	**eLRT**	**eScore**
Baseline		–	–	–	–	–
Longitudinal	1	*Methylobacterium*	*Proteobacteria*	1.00	–	1.00
	4	*Prevotella*	*Bacteroidetes*	<0.01	–	<0.01
	2	*Rothia*	*Actinobacteria*	0.01	–	0.03
	3	*Campylobacter*	*Proteobacteria*	0.03	–	0.03
	5	*TM7_genera_incertae_sedis*	*TM7*	0.00	–	0.01
	6	*Corynebacterium*	*Actinobacteria*	0.32	–	0.31

The phylum information is provided for selected genus. Tuning parameter $\lambda^* = 0.035$ for longitudinal data. Rank represents the order of genera that appear in the solution path. No genus is chosen using baseline data only. Results of eLRT are omitted in longitudinal design as it is equivalent to eRLRT.

outcome increases. In general, VC-lasso has a distinctively better selection performance even when model is misspecified.

4.2. Application to Longitudinal Pulmonary Microbiome Data

We apply VC-lasso to a longitudinal dataset of pulmonary microbiome study. Bronchoalveolar lavage (BAL) fluid were collected for microbiome profiling. The inclusion criterion for this cohort were: (1) HIV infection and (2) CD4 count less than 500 $cells/mm^3$ before HAART (Twigg III et al., 2016). Two most common pulmonary function tests were performed repeatedly: spirometry and diffusing capacity for carbon monoxide. In this report we focus on spirometry measures. Spirometry is to measure the lung volume and how well the lung exhales, such as average forced expiratory flow (FEF) and forced expiratory volume in 1s (FEV1). Both spirometry and diffusing capacity were evaluated as percent predicted values as pulmonary function tests are usually interpreted by comparing the patient's value to predicted value of the healthy subject with similar age, height and ethnicity (Twigg III et al., 2016).

Twigg III et al. (2016) compared microbiome abundance differences at overall community level between (1) uninfected and baseline; (2) uninfected and 1 year after treatment; and (3) uninfected and 3 year treated subjects. They suggest that the lung microbiome in healthy HIV-infected individuals with preserved CD4 counts is similar to uninfected individuals. Among individuals with more advanced disease, there is an altered alveolar microbiome characterized by a loss of richness and evenness (alpha diversity). This alteration might impact pulmonary complications (often characterized by the measure of lung functions) in HIV-infected patients on antiretroviral therapy (ART). In this application, we therefore aim to identify microbiome genera associated with pulmonary function in both longitudinal and baseline studies. Ethnicity, gender, smoking history, CD4 count, and HIV viral load are included as the covariates. Missing covariates are imputed

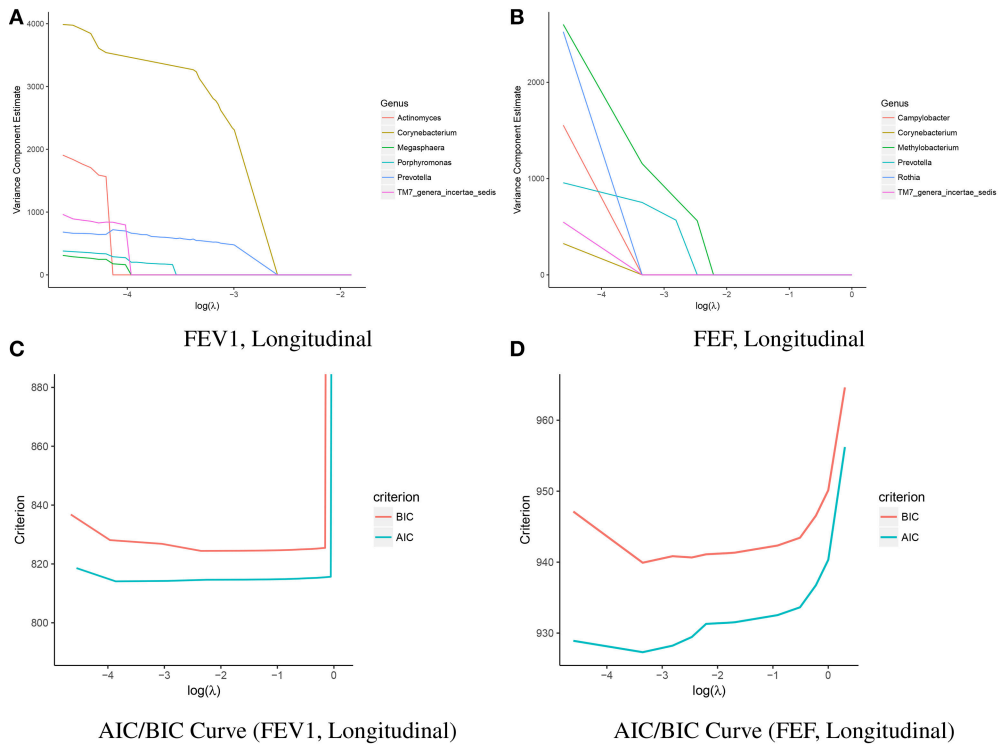

FIGURE 10 | Solution path and AIC/BIC curve of the VC-lasso method in the analysis of 31 genera and the pulmonary function. The solution paths with penalty parameter are presented for FEV1 **(A)** and FEF **(B)** in longitudinal study (upper panel). AIC/BIC curves as a function of tuning parameter for FEV1 **(C)** and FEF **(D)** are shown in the lower panel.

by their mean. Penalized variance component selection is performed among all 31 genera. Due to limited sample sizes, we choose the optimal tuning parameter λ^* by AIC and BIC.

Tables 3, **4** show selected genera with their phylum information and the corresponding *p*-values from exact tests, i.e., score test (eScore), likelihood ratio test (eLRT), and restricted likelihood ratio test (eRLRT) (Zhai et al., 2017b). The genera are ranked in the order they appear in the solution path (**Figures 10A,B**). VC-lasso selects 6 genera associated with FEV1 using longitudinal data and $\lambda^* = 0.2$ (**Table 3** and **Figure 10C**). Three out of six selected genera have eRLRT *p*-values < 0.05 (**Table 3**), including *Actinomyces* ($p < 0.01$), *Prevotella* ($p = 0.01$), and *Porphyromonas* ($p < 0.01$). Using baseline data, we identify five genera associated with FEV1, among which *Corynebacterium* and *TM7 genera incertae sedis* are also selected by using longitudinal data. Several selected genera received insufficient attention in HIV-infected populations previously, for example, *Anaerococcus* and *Megasphaera*. Studies have shown that *Anaerococcus* became more abundant in children with asthma after azithromycin treatment (Slater et al., 2013; Riiser, 2015) and *Megasphaera* has higher relative abundance in smoking population (Segal et al., 2014). However, none of them has been reported in HIV infected pulmonary microbiome (Rogers et al., 2004; Chen et al., 2007; Twigg III et al., 2016).

For variance component selection on FEF (**Table 4**), VC-lasso selects 6 genera in total using longitudinal data with $\lambda^* = 0.035$. Considering the exact test results (eRLRT and eScore), four of them show significant association with FEF (*p*-value < 0.05), i.e., *Prevotella*, *Rothia*, *Campylobacter* and *TM7_genera_incertae_sedis*. Twigg III et al. (2016) reported that HIV-positive BAL samples contained an increased abundance of *Prevotella* after 1-year HAART treatment while significantly decreased abundances during 3 years of treatment. *Campylobacter* is another noteworthy genus that has significant association with inflammation markers of HIV-infected population (Iwai et al., 2014). Additionally, significantly increased abundance of *Rothia* and *TM7_genera_incertae_sedis* in oral wash microbiome has been reported in HAART treatment group (Iwai et al., 2012; Beck et al., 2015). In conclusion, VC-lasso provides innovative association evidence between fine level pulmonary microbiome clusters with lung function phenotypes. Our report is a hypothesis generation procedure. Association results need to be further validated in a separate population or by laboratory experiments.

5. DISCUSSION

In this paper, we propose the variance component selection scheme VC-lasso for sparse and high-dimensional taxonomic data analysis. To reduce the dimensionality, we first aggregate

the dispersed individual OTUs to clusters at higher phylogenetic level, such as genus, family, or phylum. By translating the phylogenetic distance information to kernel matrices, we treat the aggregated taxonomic clusters as multiple random effects in a variance component model. Then, VC-lasso is performed for parsimonious variable selection of variance components. The MM algorithm with lasso penalization derived in Algorithm 1 for parameter estimation extremely simple and computationally efficient for variance component estimation. The group-lasso as a comparison can also be used for the microbiome cluster selection and incorporating higher phylogenetic group information (Yuan and Lin, 2006; Garcia et al., 2013; Yang and Zou, 2015). However, group-lasso suffers from the high-dimensionality and sparsity of OTUs within clusters. And group-lasso is not easy to accommodate phylogenic information. Beyond that, our novel approach VC-lasso can be applied to longitudinal designs. In such cases, we do not penalize the variance component that contains the repeat measurement correlation. Software and detailed documentation are freely available at https://github.com/JingZhai63/VCselection.

The VC-lasso is not limited to random intercept model for longitudinal studies. More complex random effect models, such as random intercept and random slope model, can also be used. More generally, the extension of our method to multivariate responses is expected to have better prediction performances. In the precision medicine era, with the rapid development of sequencing techniques and decreasing costs, the personal microbiome sequencing is already available to the consumer, e.g., American Gut (http://americangut.org/) and uBiome (https://ubiome.com/). Selection for higher-order interactions with random effect, such as microbiome and treatment regime interactions (Gopalakrishnan et al., 2017), will be a straightforward, yet interesting, implementation (Maity and Lin, 2011; Lin et al., 2016).

In practice, knowledge is needed about which taxonomy level should be aimed at to develop strategies for intervention. Considering multiple level taxonomic data, one can extend VC-lasso to include tree topologies (Wang and Zhao, 2016; Wang et al., 2017). For example, overlapping or subgroup VC-lasso can be developed by using both ℓ_1 and ℓ_2 regularizations (Jacob et al., 2009; Bien et al., 2013). Last but not the least, the variance components model requires specification of a kernel function or kernel matrix a priori, but it is often unclear which distance kernel to use in practice. To deal with the uncertainty, we can consider obtaining a composite kernel by utilizing a multiple kernel learning algorithm, such as a multi-kernel boosting algorithm (Xia and Hoi, 2013). In conclusion, with its competitive performance and many potential extensions, our variance components model with regularization, VC-lasso, is a powerful tool for mining the emerging microbiome data.

AUTHOR CONTRIBUTIONS

JZ implemented method and carried out data analysis. JZ wrote the manuscript with support from JK, HZ, and JJZ. HZ helped supervise the project. HT and KK provided pulmonary microbiome data. JJZ supervised the project.

ACKNOWLEDGMENTS

JJZ is supported by NIH grant K01DK106116 and Arizona Biomedical Research Commission (ABRC) grant. HZ is partially supported by NIH grants HG006139, GM105785, GM53275 and NSF grant DMS-1645093. HT and KK is supported by NIH grant UO1 HL121831 and UO1 HL098960. An allocation of computer time from the UA Research Computing High Performance Computing (HPC) and High Throughput Computing (HTC) at the University of Arizona is gratefully acknowledged.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.00509/full#supplementary-material

REFERENCES

Akaike, H. (1998). "Information theory and an extension of the maximum likelihood principle," in *Selected Papers of Hirotugu Akaike*, eds E. Parzen, K. Tanabe, and G. Kitagawa (New York, NY: Springer), 199–213.

Beck, J. M., Schloss, P. D., Venkataraman, A., Twigg III, H., Jablonski, K. A., Bushman, F. D., et al. (2015). Multicenter comparison of lung and oral microbiomes of HIV-infected and HIV-uninfected individuals. *Am. J. Respirat. Crit. Care Med.* 192, 1335–1344. doi: 10.1164/rccm.201501-0128OC

Bien, J., Taylor, J., and Tibshirani, R. (2013). A lasso for hierarchical interactions. *Ann. Statist.* 41:1111. doi: 10.1214/13-AOS1096

Bray, J. R., and Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecol. Monogr.* 27, 325–349. doi: 10.2307/1942268

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336. doi: 10.1038/nmeth.f.303

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. doi: 10.1186/1471-2105-12-118

Chen, E. Z., and Li, H. (2016). A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics* 32, 2611–2617. doi: 10.1093/bioinformatics/btw308

Chen, H. I., Kao, S. J., and Hsu, Y.-H. (2007). Pathophysiological mechanism of lung injury in patients with leptospirosis. *Pathology* 39, 339–344. doi: 10.1080/00313020701329740

Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., Wu, G. D., et al. (2012a). Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics* 28, 2106–2113. doi: 10.1093/bioinformatics/bts342

Chen, J., Bushman, F. D., Lewis, J. D., Wu, G. D., and Li, H. (2012b). Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics* 14, 244–258. doi: 10.1093/biostatistics/kxs038

Chen, J., Just, A. C., Schwartz, J., Hou, L., Jafari, N., Sun, Z., et al. (2015a). CpGFilter: model-based CpG probe filtering with replicates for epigenome-wide association studies. *Bioinformatics* 32, 469–471. doi: 10.1093/bioinformatics/btv577

Chen, J., and Li, H. (2013). *Kernel Methods for Regression Analysis of Microbiome Compositional Data*. New York, NY: Springer.

Chen, L., Liu, H., Kocher, J.-P. A., Li, H., and Chen, J. (2015b). glmgraph: an R package for variable selection and predictive modeling of structured genomic data. *Bioinformatics* 31, 3991–3993. doi: 10.1093/bioinformatics/btv497

Dewhirst, F. E., Chen, T., Izard, J., Paster, B. J., Tanner, A. C., Yu, W.-H., et al. (2010). The human oral microbiome. *J. Bacteriol.* 192, 5002–5017. doi: 10.1128/JB.00542-10

Eckburg, P. B., Bik, E. M., Bernstein, C. N., Purdom, E., Dethlefsen, L., Sargent, M., et al. (2005). Diversity of the human intestinal microbial flora. *Science* 308, 1635–1638. doi: 10.1126/science.1110591

Erb-Downward, J. R., Thompson, D. L., Han, M. K., Freeman, C. M., McCloskey, L., Schmidt, L. A., et al. (2011). Analysis of the lung microbiome in the "healthy" smoker and in COPD. *PLoS ONE* 6:e16384. doi: 10.1371/journal.pone.0016384

Fan, Y., and Li, R. (2012). Variable selection in linear mixed effects models. *Ann. Statist.* 40:2043. doi: 10.1214/12-AOS1028

Garcia, T. P., Müller, S., Carroll, R. J., and Walzem, R. L. (2013). Identification of important regressor groups, subgroups and individuals via regularization methods: application to gut microbiome data. *Bioinformatics* 30, 831–837. doi: 10.1093/bioinformatics/btt608

Gill, S. R., Pop, M., DeBoy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., et al. (2006). Metagenomic analysis of the human distal gut microbiome. *Science* 312, 1355–1359. doi: 10.1126/science.1124234

Gopalakrishnan, V., Spencer, C. N., Nezi, L., Reuben, A., Andrews, M. C., Karpinets, T. V., et al. (2017). Gut microbiome modulates response to anti–pd-1 immunotherapy in melanoma patients. *Science* 359, 97–103. doi: 10.1126/science.aan4236

Grice, E. A., and Segre, J. A. (2011). The skin microbiome. *Nat. Rev. Microbiol.* 9, 244–253. doi: 10.1038/nrmicro2537

Haas, B. J., Gevers, D., Earl, A. M., Feldgarden, M., Ward, D. V., Giannoukos, G., et al. (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res.* 21, 494–504. doi: 10.1101/gr.112730.110

Hodkinson, B. P., and Grice, E. A. (2015). Next-generation sequencing: a review of technologies and tools for wound microbiome research. *Adv. Wound Care* 4, 50–58. doi: 10.1089/wound.2014.0542

Hui, F. K., Müller, S., and Welsh, A. (2017). Joint selection in mixed models using regularized PQL. *J. Am. Statist. Assoc.* 112, 1323–1333. doi: 10.1080/01621459.2016.1215989

Hunter, D. R., and Lange, K. (2004). A tutorial on MM algorithms. *Am. Statist.* 58, 30–37. doi: 10.1198/0003130042836

Hunter, D. R., and Li, R. (2005). Variable selection using MM algorithms. *Ann. Statist.* 3:1617. doi: 10.1214/009053605000000200

Ibrahim, J. G., Zhu, H., Garcia, R. I., and Guo, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics* 67, 495–503. doi: 10.1111/j.1541-0420.2010.01463.x

Iwai, S., Fei, M., Huang, D., Fong, S., Subramanian, A., Grieco, K., et al. (2012). Oral and airway microbiota in HIV-infected pneumonia patients. *J. Clin. Microbiol.* 50, 2995–3002. doi: 10.1128/JCM.00278-12

Iwai, S., Huang, D., Fong, S., Jarlsberg, L. G., Worodria, W., Yoo, S., et al. (2014). The lung microbiome of Ugandan HIV-infected pneumonia patients is compositionally and functionally distinct from that of San Franciscan patients. *PLoS ONE* 9:e95726. doi: 10.1371/journal.pone.0095726

Jacob, L., Obozinski, G., and Vert, J.-P. (2009). "Group lasso with overlap and graph lasso," in *Proceedings of the 26th Annual International Conference on Machine Learning* (Montreal, QC: ACM).

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. *Nat. Biotechnol.* 34, 64–69. doi: 10.1038/nbt.3416

Lange, K. (2016). *MM Optimization Algorithms*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Lange, K., Hunter, D. R., and Yang, I. (2000). Optimization transfer using surrogate objective functions. *J. Comput. Graphic. Statist.* 9, 1–20. doi: 10.1080/10618600.2000.10474858

Lin, W., Shi, P., Feng, R., and Li, H. (2014). Variable selection in regression with compositional covariates. *Biometrika* 101, 785–797. doi: 10.1093/biomet/asu031

Lin, X., Lee, S., Wu, M. C., Wang, C., Chen, H., Li, Z., et al. (2016). Test for rare variants by environment interactions in sequencing association studies. *Biometrics* 72, 156–164. doi: 10.1111/biom.12368

Lozupone, C., Cota-Gomez, A., Palmer, B. E., Linderman, D. J., Charlson, E. S., Sodergren, E., et al. (2013). Widespread colonization of the lung by Tropheryma whipplei in HIV infection. *Am. J. Respirat. Crit. Care Med.* 187, 1110–1117. doi: 10.1164/rccm.201211-2145OC

Lozupone, C., and Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* 71, 8228–8235. doi: 10.1128/AEM.71.12.8228-8235.2005

Lozupone, C. A., Hamady, M., Kelley, S. T., and Knight, R. (2007). Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Appl. Environ. Microbiol.* 73, 1576–1585. doi: 10.1128/AEM.01996-06

Maity, A., and Lin, X. (2011). Powerful tests for detecting a gene effect in the presence of possible gene–gene interactions using garrote kernel machines. *Biometrics* 67, 1271–1284. doi: 10.1111/j.1541-0420.2011.01598.x

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends Genet.* 24, 133–141. doi: 10.1016/j.tig.2007.12.007

Morris, A., Beck, J. M., Schloss, P. D., Campbell, T. B., Crothers, K., Curtis, J. L., et al. (2013). Comparison of the respiratory microbiome in healthy nonsmokers and smokers. *A. J. Respirat. Crit. Care Med.* 187, 1067–1075. doi: 10.1164/rccm.201210-1913OC

Riiser, A. (2015). The human microbiome, asthma, and allergy. *Allergy Asthma Clin. Immunol.* 11:35. doi: 10.1186/s13223-015-0102-0

Rogers, G., Carroll, M., Serisier, D., Hockey, P., Jones, G., and Bruce, K. (2004). Characterization of bacterial community diversity in cystic fibrosis lung infections by use of 16S ribosomal DNA terminal restriction fragment length polymorphism profiling. *J. Clin. Microbiol.* 42, 5176–5183. doi: 10.1128/JCM.42.11.5176-5183.2004

Schwarz, G. et al. (1978). Estimating the dimension of a model. *Ann. Statist.* 6, 461–464. doi: 10.1214/aos/1176344136

Segal, L. N., Rom, W. N., and Weiden, M. D. (2014). Lung microbiome for clinicians. new discoveries about bugs in healthy and diseased lungs. *Ann. Am. Thoracic Soc.* 11, 108–116. doi: 10.1513/AnnalsATS.201310-339FR

Shi, P., Zhang, A., and Li, H. (2016). Regression analysis for microbiome compositional data. *Ann. Appl. Statist.* 10, 1019–1040. doi: 10.1214/16-AOAS928

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *J. Comput. Graphic. Statist.* 22, 231–245. doi: 10.1080/10618600.2012.681250

Slater, M., Rivett, D. W., Williams, L., Martin, M., Harrison, T., Sayers, I., et al. (2013). The impact of azithromycin therapy on the airway microbiota in asthma. *Thorax* 69, 673–674. doi: 10.1136/thoraxjnl-2013-204517

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., et al. (2009). A core gut microbiome in obese and lean twins. *Nature* 457, 480–484. doi: 10.1038/nature07540

Twigg III, H. L., Knox, K. S., Zhou, J., Crothers, K. A., Nelson, D. E., Toh, E., et al. (2016). Effect of advanced HIV infection on the respiratory microbiome. *Am. J. Respirat. Crit. Care Med.* 194, 226–235. doi: 10.1164/rccm.201509-1875OC

Twigg III, H. L., Morris, A., Ghedin, E., Curtis, J. L., Huffnagle, G. B., Crothers, K., et al. (2013). Use of bronchoalveolar lavage to assess the respiratory microbiome: signal in the noise. *Lancet Respirat. Med.* 1, 354–356. doi: 10.1016/S2213-2600(13)70117-6

Wang, J., and Jia, H. (2016). Metagenome-wide association studies: fine-mining the microbiome. *Nat. Rev. Microbiol.* 14, 508–522. doi: 10.1038/nrmicro.2016.83

Wang, T., and Zhao, H. (2016). Constructing predictive microbial signatures at multiple taxonomic levels. *J. Am. Statist. Associat.* 112, 1022–1031. doi: 10.1080/01621459.2016.1270213

Wang, T., and Zhao, H. (2017). Structured subcomposition selection in regression and its application to microbiome data analysis. *Ann. Appl. Statist.* 11, 771–791. doi: 10.1214/16-AOAS1017

Xia, H., and Hoi, S. C. (2013). Mkboost: A framework of multiple kernel boosting. *IEEE Trans. Knowledge Data Eng.* 25, 1574–1586. doi: 10.1109/TKDE.2012.89

Yang, Y., and Zou, H. (2015). A fast unified algorithm for solving group-lasso penalize learning problems. *Statist. Comput.* 25, 1129–1141. doi: 10.1007/s11222-014-9498-5

Yuan, M., and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B* 68, 49–67. doi: 10.1111/j.1467-9868.2005.00532.x

Zemanick, E. T., Sagel, S. D., and Harris, J. K. (2011). The airway microbiome in cystic fibrosis and implications for treatment. *Curr. Opin. Pediatr.* 23, 319–324. doi: 10.1097/MOP.0b013e32834604f2

Zhai, J., Hsu, C.-H., and Daye, Z. J. (2017a). Ridle for sparse regression with mandatory covariates with application to the genetic assessment of histologic grades of breast cancer. *BMC Med. Res. Methodol.* 17:12. doi: 10.1186/s12874-017-0291-y

Zhai, J., Knox, K. S., Twigg III, H. L., Zhou, H., and Zhou, J. (2017b). Exact tests of zero variance component in presence of multiple variance components with application to longitudinal microbiome study. *bioRxiv* doi: 10.1101/281246

Zhao, N., Chen, J., Carroll, I. M., Ringel-Kulka, T., Epstein, M. P., Zhou, H., et al. (2015). Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. *Am. J. Hum. Genet.* 96, 797–807. doi: 10.1016/j.ajhg.2015.04.003

Zhou, H., Alexander, D., and Lange, K. (2011). A quasi-Newton acceleration for high-dimensional optimization algorithms. *Statist. Comput.* 21, 261–273. doi: 10.1007/s11222-009-9166-3

Zhou, H., Hu, L., Zhou, J., and Lange, K. (2015). MM algorithms for variance components models. *arXiv preprint arXiv:1509.07426*.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.