A Mixed Effect Similarity Matrix Regression Model (SMRmix) for Integrating Multiple Microbiome Datasets at Community Level and its Application in HIV

MENGYU HE
Department of Biostatistics and Bioinformatics, Emory University
NI ZHAO *
Department of Biostatistics, Johns Hopkins University

Abstract

Recent studies have highlighted the importance of human microbiota in our health and diseases. However, in many areas of research, individual microbiome studies often offer inconsistent results due to the limited sample sizes and the heterogeneity in study populations and experimental procedures. Integrative analysis of multiple microbiome datasets is necessary. However, statistical methods that incorporate multiple microbiome datasets and account for the study heterogeneity are not available in the literature. In this paper, we develop a mixed effect similarity matrix regression (SMRmix) approach for identifying community level microbiome shifts between outcomes. SMRmix has a close connection with the microbiome kernel association test, one of the most popular approaches for such a task but is only applicable when we have a single study. Via extensive simulations, we show that SMRmix has well-controlled type I error and higher power than some potential competitors. We also applied SMRmix to data from the HIV-reanalysis consortium, a collective effort that obtained all publicly available data on gut microbiome and HIV at December 2017, and obtained consistent associations of gut microbiome with HIV infection, and with MSM status (i.e. men who have sex with men).

Keywords: microbiome community-level association studies; integrative analysis; distance-based association analysis; similarity matrix regression

1 Introduction

With over 35 million people currently infected and close to one million deaths each year, the HIV/AIDS epidemic remains among the most significant global health challenges today according to World Health Organization. The advent of antiretroviral therapy has transformed HIV from a progressive, typically fatal infection into a chronic condition that may last for years or even decades

*Corresponding author: Ni Zhao (nzhao10@jhu.edu)

(Deeks *et al.*, 2013). Yet, people living with HIV are still at an increased risk of many diseases, including renal failure, cardiovascular diseases and cancers, partially attributed to their elevated levels of systemic inflammation (Lewden, 2004; Triant, 2013; Phair and Palella, 2011; Kirk *et al.*, 2007). The human gut harbors an extensive microbial ecosystem (the gut microbiome) that plays a fundamental role in shaping our immunity and maintaining gut mucosal integrity. Increasing evidence has linked HIV-related inflammation with gut dysbiosis (i.e., microbial disruption) (Dubourg *et al.*, 2016; Mutlu *et al.*, 2014; Noguera-Julian *et al.*, 2016; Nowak *et al.*, 2015; Pinto-Cardoso *et al.*, 2017; Villanueva-Millán *et al.*, 2017).

Despite the volume of literature linking gut microbiome with HIV infection, individual studies detected inconsistent signals. While most studies found a decreased microbial diversity in HIV-infected individuals compared to those not infected (Dubourg *et al.*, 2016; Mutlu *et al.*, 2014; Noguera-Julian *et al.*, 2016; Nowak *et al.*, 2015; Pinto-Cardoso *et al.*, 2017; Villanueva-Millán *et al.*, 2017; Yu *et al.*, 2014; Sun *et al.*, 2016; Vázquez-Castellanos *et al.*, 2015; Monaco *et al.*, 2016; Vesterbacka *et al.*, 2017), some studies detected no significant difference (Volpe *et al.*, 2014; Ling *et al.*, 2016; Dinh *et al.*, 2015; Nowak *et al.*, 2017) and even an increase of microbial diversity (Lozupone *et al.*, 2013). Studies also offer inconsistent results on the microbial community-level structural differences in HIV, which can be captured by the pair-wise beta diversity measurements. Some studies found significant differences in microbiome communities comparing HIV infected individuals and non-infected individuals (Mutlu *et al.*, 2014; Dinh *et al.*, 2015). However, other studies found that the changes of microbiome composition were due to the differences in sexual behavior, especially MSM (men who have sex with men) status, instead of HIV infection (Noguera-Julian *et al.*, 2016). Several reasons contribute to these inconsistency. First of all, most individual studies are small in sample size that their conclusions may differ just due to randomness in data. Secondly, studies are heterogeneous with respect to study populations, study designs and the technical/analytical procedures that are utilized: participants from different countries or regions may have underlying differences in gut microbiome composition regardless of HIV status; different DNA extraction methods, sequencing primer and bioinformatic pipelines for preprocessing sequencing data may introduce differential bias, all of which lead to different conclusions.

Integrative analysis of multiple microbiome datasets is necessary that aggregates information across multiple studies to achieve reproducible discoveries toward microbiome-based treatment and prevention. A recent effort on this is the HIV-reanalysis consortium (Tuddenham *et al.*, 2020), which obtained and re-processed all data that are publicly available on the topic of gut microbiome and HIV since December 2017. However, due to the aformentioned study heterogeneity, it is invalid to directly combine all data without accounting for the between-study heterogeneities. For an intuitive example, please refer to Vaginal Microbiome Consortium *et al.* (2015).

In this paper, we focus on the association test between the overall microbiome community profile and outcomes of interest while incorporating multiple microbiome datasets. For such a task when data are collected from a single study, two approaches are the most popular: distance-based analysis (McArdle and Anderson, 2001; MetaHIT Consortium (additional members) *et al.*, 2011) and the microbiome regression kernel association test (MiRKAT) (Zhao *et al.*, 2015). In the distance-based approach, a phylogenetic distance between individuals' microbiome profiles, usually captured by microbial beta-diversities, is computed and tested for association with outcome of interest via permutation. MiRKAT utilizes a semi-parametric kernel machine regression in which a kernel similarity matrix is calculated and assessed for association with the outcome, in which the kernel matrix is a transformed matrix of the beta-diversity matrix. In both cases, an

assessment of similarity/dissimilarity between all pairs of samples is needed.

Extending to situations when we have multiple studies, neither distance-based methods nor MiRKAT is directly applicable due to the lack of all pairwise distance matrix across all subjects and studies. In situations when data are obtained from public resources without access to the original sequencing files, the different preprocessing steps across studies make it impossible to construct beta-diversity matrices between studies – we may have totally distinct OTU (operational taxonomic unit) assignments and phylogenetic trees. Even in situations when re-processing of the data across studies is possible, differential biases introduced through the various sequencing protocols can invalidate the beta-diversity measurements between samples across studies. In other words, in most cases, only within-study distances are available but between-study distances are missing.

Herein, we propose a novel statistical method to test the association between microbiome communities and outcome of interests aggregating information across multiple studies. Our method is built upon the similarity matrix regression (SMR) and only requires the within-study distances. Instead of modeling the relationship between trait and microbiome similarity directly, SMR regresses pairwise trait similarity on pairwise microbiome similarity for association testing, in which the trait similarity is calculated by outer-product of residuals from a null regression model and the microbiome similarity is calculated from within study beta diversity matrix. Association is then detected by a score test. We extend SMR to incorporate mixed effects by vectorizing similarity matrices and deriving the corresponding score test statistic through maximum likelihood. The new method, SMRmix, makes it possible to take into account study-level random effect and avoid usage of inaccessible distances. Our method also enables development of the "optimal" test in which multiple distances can be utilized for a single study. This method will have high power in the omnibus.

Through extensive simulations, we demonstrate the the proposed SMRmix has well-controlled type I error, and comparable or higher power than potential competitors such as CSKAT (Zhan *et al.*, 2018) and GLMM-MiRKAT (Koh *et al.*, 2019), both of which require the usually inaccessible distance matrices across all studies. We apply SMRmix to data from HIV-reanalysis consortium (Tuddenham *et al.*, 2020) to evaluate the association between HIV status and microbiome composition adjusting for MSM status and age, and between MSM status and microbiome composition adjusting for HIV status and age. The analysis lead to a consistent result for the association studies.

The rest of this paper is organized as follows. In section 2 we present the proposed SMRmix methods for integrative analysis of multiple microbiome datasets. Simulation and real data analysis results are presented in section 4. And we conclude the paper with a brief discussion.

2 Statistical Methods

Our goal is to investigate the relationship between the microbial community and outcomes of interest. In section 2.1, we present a microbiome association test for a single study and its link to Similarity Matrix Regression (SMR). In section 2.2, we propose a novel association test method, SMRmix, for integrating multiple microbiome datasets. In section 2.3, we introduce a perturbation approach for obtaining p-values in the proposed test. Finally, in section 2.4, we extend our method to an omnibus test that can combine information from multiple kernels.

2.1 Microbiome regression-based kernel association test for a single study and its connection with similarity matrix regression

When analyzing data from a single study, one popular approach is the microbiome regression-based kernel association test (MiRKAT) (Zhao *et al.*, 2015). In this method, the data are summarized as independent observations of the form (Y_i, X_i, M_i) for $i = 1, \cdots, n$. Here, Y_i represents the outcome of interest, which could be either continuous or binary. The covariates we want to adjust for, such as age, gender, and MSM status, are included in $X_i = (1, x_{i1}, \cdots, x_{ip})^T$, where p is the number of covariates. The abundances of all taxa/OTUs in the microbial community are represented by $M_i = (m_{i1}, \cdots, m_{iq})$, where q is the total number of taxa/OTUs.

MiRKAT compares pairwise similarity in the outcome variable to pairwise similarity in the microbiome profiles, and high correspondence is suggestive of association. It exploits a kernel machine regression framework (Kwee *et al.*, 2008; Wu *et al.*, 2010, 2011) to relate the covariates and the microbiome profiles to the outcomes. Specifically, for a continuous outcome, we assume a linear kernel machine regression model,

$$Y_i = X_i\beta + f(M_i) + \varepsilon_i \ , \tag{1}$$

and for a binary outcome, we use the logistic kernel machine regression model,

$$\text{logit}(P(Y_i = 1)) = X_i\beta + f(M_i) \ , \tag{2}$$

where ε_i is the random error with mean 0 and a common variance, $f(\cdot)$ captures the association between microbial community and the outcome, and additional covariates X_i are adjusted for through a linear function. The null hypothesis that there is no microbiome effect entails that $f(M) = 0$.

In MiRKAT, $f(M_i)$ is assumed to be from a reproducing kernel Hilbert space generated by a positive definite kernel function $K(\cdot,\cdot)$ that $f(M_i) = \sum_{i'=1}^{n} \alpha_{i'} K(M_i, M_{i'})$. The (i, i')-th element of the kernel matrix K, defined as $K(M_i, M_{i'})$, measures the pairwise similarity between the microbiome profiles of the i-th and the i'-th samples. To utilize the phylogenetic/taxonomic information contained in the microbiome sequences, we can construct the kernel similarity matrix by exploiting the correspondence with the well-defined beta-diversities, which measure dissimilarities between subjects through: $K = -\frac{1}{2}(I - \frac{11'}{n})D^2(I - \frac{11'}{n})$. Here, D depicts pairwise dissimilarities/distances of all subjects calculated with beta-diversity metrics, such as the Bray-Curtis or UniFrac distance. A variance component score test is calculated through the connection between the kernel machine regression model and the linear mixed models (Zhao *et al.*, 2015) that $Q = \frac{1}{2\phi}(Y - \hat{\mu}_0)' K (Y - \hat{\mu}_0)$, where $\hat{\mu}_0$ are the fitted values for Y under the null, and ϕ is the dispersion parameter.

An alternative approach to investigate the association between the outcome variable and a set of "-omics" markers is by the similarity matrix regression (Tzeng *et al.*, 2009). In microbiome association studies, SMR model can be written as

$$Z_{ij} = b \times K_{ij} + e_{ij}, \ i < j, \tag{3}$$

where Z_{ij}, the trait similarity between subject i and j, is defined as the outer product of trait residuals, $Z_{ij} = (Y_i - \hat{\mu}_i^0)(Y_j - \hat{\mu}_j^0)$, and $\hat{\mu}_i^0$ is the fitted values of Y_i under the null hypothesis. K_{ij} is the i, j-th element of microbial kernel similarity matrix. Intuitively, if the outcome is associated with microbial profiles, then subjects with higher level of outcome similarity should have more higher microbiome similarity.

Under the null, assuming that all Y_i's are independent, the score statistic for testing in SMR is $U_b = \sum_{i<j} K_{ij}(v_i^0)^{-1}(v_j^0)^{-1}Z_{ij} = \sum_{i<j} K_{ij}(v_i^0)^{-1}(v_j^0)^{-1}(Y_i - \hat{\mu}_i^0)(Y_j - \hat{\mu}_j^0)$, where $v_i^0 = Var(Y_i|X_i, M_i)$ under the null. The statistic can be written in the quadratic form $U_b = \frac{1}{2}(Y - \hat{\mu}_0)'V^{-1}K_0V^{-1}(Y - \hat{\mu}_0)$, where K_0 is a matrix with diagonal elements 0 and the (i, j)-th off-diagonal element as K_{ij}. Under the null, U_b approximately follows a weighted chi-squared distribution, which can be used to assess the statistical significance.

Formulate the SMR into a vector format, we have

$$Z^{vec} = b \times K^{vec} + e^{vec}, \tag{4}$$

where Z^{vec} and K^{vec} are stacked by those Z_{ij}, K_{ij} with pairwise distance of (i, j) available. There is a close connection between MiRKAT and the SMR (Zhan, 2019). Specifically, the square of the empirical sample correlation between outcome similarity and microbiome similarity in SMR, denoted by R^2, is proportional to the score statistic statistic in MiRKAT. Note that R^2 statistic depicts the proportion of variation explained in SMR.

Even though SMR is statistically equivalent to MiRKAT, the vector format of SMR enables integration of multiple studies when the between-study similarity matrix is not available, because now we can remove those pairs with missing distances from similarity vectors Z^{vec} and K^{vec}.

2.2 SMRmix for integrating multiple microbiome datasets

Now suppose that we have a total of K microbiome studies. The data for the i-th subject in the k-th study is denoted as (Y_{ik}, X_{ik}, M_{ik}), in which $i = 1, \cdots, n_k$, $k = 1, \cdots, K$, n_k is the number of subjects in $k-$th study. Let $N = \sum_{k=1}^{K} n_k$ denote the total number of subjects from K studies. We assume all studies are independent and share no subjects. We also assume that all samples are independent within a study. Common violation of this second assumption includes when multiple samples are longitudinal measurements from the same participant or are clustered within families in a single study. In other words, we assume that samples are clustered based on studies, and no other clustering structure exists even though our method can be extended to allow for correlation within studies.

When microbiome data come from different studies, as samples are processed through different protocols in different studies, subjects from different studies are not directly comparable, and the corresponding beta diversities can not be calculated. As a result, it is possible that only intra-study microbial dissimilarities are available for association testing. Suppose the similarity matrix for all subjects across all studies (now assume it exists) is

$$\mathbb{K}_{\text{full}} = \begin{bmatrix} K_1 & K_{12} & \cdots & K_{1K} \\ K_{21} & K_2 & \cdots & K_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ K_{K1} & K_{K2} & \cdots & K_K \end{bmatrix}$$

Here, the diagonal blocks K_k represent the kernel similarity matrix for samples within the same study k, and the off-diagonals represent the microbiome similarity for pairs of samples from different studies. For existing kernel-based/distance-based methods, all elements in $\mathbb{K}_{\text{full}}$ should be calculated using the same distance metric. Otherwise, the absolute values of two pairwise kernels using different metrics are not on the same scale and not comparable.

In reality, we often only have access to the diagonal blocks as represented as follows:

$$\mathbb{K} = \begin{bmatrix} K_1 & & & \\ & K_2 & & \\ & & \ddots & \\ & & & K_K \end{bmatrix}$$

where only diagonal blocks of the large matrix $\mathbb{K}$ are obtained, cells outside the diagonal blocks are all missing. By having only the block diagonal elements, we have the flexibility to choose each kernel for each study to be different, for example, K_1 is of UniFrac kernel while K_2 is of Bray-Curtis kernel. While it is advantageous to interpret the results when all kernels involved are of the same type, having different kernels for each study would still be valid. This is helpful in situations when the kernels of the same type are not available for some studies. For example, we are able to construct UniFrac kernel for study 1, while UniFrac kernel is not available for study 2 because we don't have the phylogenetic tree information if the data is obtained from public. In this paper, we will focus on the former situations when all individual K_k are of the same type for simplicity of description. Vectorized SMR model enables us to test the association based on these blocks without the need of the off diagonal-blocks, by defining the similarity regression model only when K_{ij} exists.

To motivate the SMR model when integrating multiple microbiome datasets, we first evaluate its counterpart model within the kernel machine regression framework when the overall microbiome kernel similarity matrix $\mathbb{K}_{\text{full}}$ is available. For continuous outcomes, correlated sequence kernel association test (CSKAT) (Zhan *et al.*, 2018), an extension of MiRKAT, could be used to test the association between microbiome profiles and the outcomes when the data from the same study are clustered. CSKAT assumes a linear mixed-effects model with study-specific random effect to link the outcome and microbial composition:

$$Y_{ik} = {X_{ik}}^T \beta + f(M_{ik}) + h_k + \varepsilon_{ik} \ . \tag{5}$$

For dichotomous outcomes, another extension of MiRKAT, GLMM-MiRKAT (Koh *et al.*, 2019), incorporates a study-specific random effect as

$$logit(P(Y_{ik} = 1)) = {X_{ik}}^T \beta + f(M_{ik}) + h_k \ . \tag{6}$$

In these models, h_k is the study-specific random effect that captures the within study correlation and $h_k \sim N(0, \sigma^2)$, ε_{ik} is the subject-level random effect that $\varepsilon_{ik} \sim N(0, \delta_k^2)$. Putting samples from all studies together in a vector format in these models, we have that $h \sim N(0, \sigma^2 \Sigma_h)$, $\Sigma_h = diag\{1_{n_1} 1_{n_1}^T, \cdots, 1_{n_K} 1_{n_K}^T\}$, $\varepsilon \sim N(0, \Sigma_\varepsilon)$, $\Sigma_\varepsilon = diag\{\delta_1^2 I_{n_1}, \cdots, \delta_S^2 I_{n_K}\}$.

If a full kernel matrix across all studies is available, the kernel machine regression model in CSKAT and GLMM-MiRKAT is equivalent to the (generalized) linear mixed model with $f \sim N(0, \tau \mathbb{K}_{\text{full}})$. Hypothesis testing for $H_0 : \tau = 0$ can be conducted as in the literature. However, in reality, because only the block diagonal elements K_k of the overall kernel matrix are available, CSKAT and GLMM-MiRKAT can not be applied. However, we are still able to calculate the trait similarity Z_k and the variance-covariance matrix $V_k = \text{Var}(Z_k | X_k, M_k)$(Please refer to section A of Supplementary Materials).

When only $\mathbb{K}$ is available, we propose the following mixed effect similarity matrix regression model (SMRmix) that

$$Z_{ij,k} = a + bK_{ij,k} + u_k + f_i + e_{ij,k},\ i < j. \tag{7}$$

For subjects i, j in study k, we link the outcome similarity $Z_{ij,k}$ and the microbiome similarity $K_{ij,k}$. Here, $e_{ij,k}$ is a random error, which are allowed to differ in different studies, u_k is a study-specific random effect, f_i is a random effect that captures the additional correlation between two pairs of trait similarities if the two pairs share one sample in common. For example, it captures the additional correlation between $Z_{ij,k}$ and $Z_{ij',k}$ compared to $Z_{ij,k}$ and $Z_{i'j',k}$ with $i \neq i'$ and $j \neq j'$. Compared to the original SMR model that is equivalent to a kernel machine regression model, our model is based to a mixed effect kernel machine regression model, therefore, with the name of SMRmix. Here, all random effects all have mean 0 and constant variance. There is a mathematical relationship between the random effects and the random effects/errors in CSKAT and GLMM-MiRKAT, which we specify in section B of Supplementary Materials.

To test whether there is a microbiome effect, we test $H_0 : b = 0$. We propose the following score statistic:

$$U = \sum_{k=1}^{K} (K_k^{vec})^T (V_k)^{-1} (Z_k^{vec} - \mathbb{E}(Z_k^{vec} | X_k, M_k)), \tag{8}$$

where $K_k^{vec} = (K_{(1,2),k}, \cdots, K_{(1,n_k),k}, K_{(2,3),k}, \cdots, K_{(2,n_k),k}, \cdots, K_{(n_k-1,n_k),k})^T$ is a vector in which we stack all the available within-study pairwise kernels of k-th study. And $Z_k^{vec} = (Z_{(1,2),k}, \cdots, Z_{(1,n_k),k}, Z_{(2,3),k}, \cdots, Z_{(2,n_k),k}, \cdots, Z_{(n_k-1,n_k),k})^T$, $V_k = Var(Z_k^{vec})$ are the corresponding outcome similarities and variances calculated under the null.

This test aggregates different studies by testing summation of the score statistic in each study. In computation, it only requires microbial similarities within studies, while CSKAT requires microbial pair-wise similarities of all samples.

2.3 Perturbation approach for obtaining p-value

In the original SMR approach, the test statistic can be written into a quadratic form and approximated by a weighted chi-square distribution. However, for SMRmix, because of the complexity of variance V_k^{vec} and the vectorization, the asymptotic distribution of U can not be easily approximated. Instead, we propose a perturbation approach for obtaining the null distribution of U and hence the p-value. We note that, each element in $Z_k = (Y_k - \mu_k^0)(Y_k - \mu_k^0)^T$ is obtained from the outer product of $\mathbf{Y}_k - \mu_k^0$, which, after some transformation, asymptotically follows a multivariate normal distribution with mean 0 and covariance matrix that can be estimated under the null hypothesis. The idea behind perturbation is to re-generate a large number of $\tilde{Z}_k^{vec}$ values that asymptotically follows the same distribution as of Z_k^{vec}. By doing so, we can generate a large number of U^* which is of the same distribution of U under the null. Therefore, we will be able to compare our observed test statistics with U^* for p-value calculation.

Perturbation procedures, sometimes referred to as just resampling or Monte-Carlo approaches, have been previously used by many other researchers (Lin, 2005; Conneely and Boehnke, 2007; Chapman and Whittaker, 2008; Pan *et al.*, 2010) in multiple contexts. Our work is related to these approaches, but differs substantially that we need to take into consideration the additional correlations that arises from the clustered nature of our data. To be more specific, denote the

estimated covariance matrix of $Y_k - \mu_k^0$ from the mixed-effect model as Σ_k. For continuous outcomes, Σ_k is a symmetric matrix with diagonal elements equal to $\hat{\sigma}^2 + \hat{\delta}_k^{\,2}$ and off-diagonal elements equal to $\hat{\sigma}^2$. For binary outcomes, $i-$th diagonal elements is $\hat{\mu}_{ik}^0(1-\hat{\mu}_{ik}^0)$, and element $(i,j),\ i \neq j$, in Σ_k is $\hat{\mu}_{ik}^0(1-\hat{\mu}_{ik}^0)\hat{\mu}_{jk}^0(1-\hat{\mu}_{ik}^0)\hat{\sigma}^2$. Let $P_k = \Sigma_k - X_k({X_k}^T\Sigma^{-1}X_k)^{-1}{X_k}^T$, then we get ${P_k}^{\frac{1}{2}} = {\Sigma_k}^{\frac{1}{2}} - X_k({X_k}^T{\Sigma_k}^{-1}X_k)^{-1}{X_k}^T{\Sigma_k}^{-\frac{1}{2}}$. It can be proved that, under the null hypothesis,

$$ {P_k}^{-\frac{1}{2}}(Y_k - \mu_k^0) \sim MVN(0, I_N), \tag{9} $$

where I_N is the identity matrix of size N.

To generate the null distribution of the test statistic, we first generate random sample $\tilde{R}_k$ from k-dimensional normal distributions with mean 0 and identity covariance matrix (i.e., the distribution of ${P_k}^{-\frac{1}{2}}(Y_k - \mu_k^0)$). As $P_k^{\frac{1}{2}}\tilde{R}_k$ has the same distribution as $Y_k - \mu_k^0$, we can replace Z_k^{vec} in equation (8) by a vectorized version of the outer products of $P_k^{\frac{1}{2}}\tilde{R}_k$, which we denote as $\tilde{Z}_k^{vec}$. The process is conducted for a large number of times (e.g., B) to construct the null distribution of our test statistic, which we use to obtain p-values and hypothesis testing.

2.4 Omnibus test

As previous research pointed out, power of statistical methods depends on whether the choice of kernel captures the true association signal (Zhao *et al.*, 2015). For example, the weighted UniFrac and Bray-Curtis kernels utilize the relative abundance information in constructing the pairwise similarity and give higher weight to common taxa, therefore, they are more powerful when the true association happens in common taxa. On the other hand, the unweighted UniFrac and Jaccard kernels only utilize the information of whether a taxon is present in the sample and give higher weight to rare taxa, therefore, they are more powerful when the true association happens in rare taxa. In realty, we never know before analysis the proper choice of kernel. Therefore, in this session we develop an omnibus test that extends SMRmix to simultaneously consider multiple possible kernels for each study. We expect that the omnibus-SMRmix is robust to different patterns of association that it loses little power compared to when the best kernel is used but is substantially more powerful than when a bad kernel is used.

Let $\mathbb{K}_l$, $l = (1, \cdots, L)$ denote one of the concatenated block diagonal kernels (without the off-diagonal blocks) in which the kernel for each study is of the same type, such as Weighted UniFrac. Suppose that there are a total of L of such groups of kernels. For each $\mathbb{K}_l$, we can conduct the SMRmix and the perturbation approach as previously described to obtain a p-value. Then all the p-values are combined through a harmonic mean combination (HMP) (Wilson, 2019). In simulations, the HMP combination shows superior type I error control compared to other competitors, such as the Cauchy combination (Liu and Xie, 2020), Simes combination (Simes, 1986) and the Tippett combination (Tippett, 1931). Note that even though we have L sets of SMRmix models, we only need to generate one set of B multivariate normal distributions for the perturbation across all sets of kernels. This reduces the computational time drastically.

3 Simulation studies

We conduct extensive simulations to evaluate the type I error and power performance of SM-Rmix compared to CSKAT and GLMM-MiRKAT. We conduct two sets of simulations. In the first set of simulation, we simulate microbiome data using a Dirichlet-Multinomial distribution for which the hyper-parameters are estimated from a real data with a common tree structure (Charlson *et al.*, 2010). Because studies differ in their way of data collections, we introduce a multiplicative and differential bias to all studies to capture this heterogeneity. In this way, all microbiome data from all studies share the same phylogenetic tree and it is possible to calculate a common kernel matrix that approaches like CSKAT or GLMM-MiRKAT can be applied. However, with differential bias across studies, the kernel components corresponding to samples across different studies are not directly comparable, which may cause problem in its conclusions. This simulation is similar to situations when we have access to the raw sequencing files and are able to reprocess the data through the same bioinformatic pipeline. We name this set of simulation as Simulation **A**.

In the second set of simulations (**B**), we reprocess the raw sequencing (fastq) files from a real gut microbiome study (Monaco *et al.*, 2016) through different bioinformatic pipelines. This way, we generate multiple OTU tables that have distinct tree structures. For each dataset processed by a distinct bioinformatic pipelines, we estimate its hyper-parameters via the Dirichlet-Multinomial distribution and subsequently simulate additional microbiome data with its corresponding tree structure. This simulation mimics the situation that we collect multiple microbiome datasets that are pre-processed using different bioinformatic pipelines, but we have no access to the raw sequencing data, which is often the case when people deal with public data. In this scenario, generating a single kernel matrix is not possible and approaches that require a common kernel matrix such as CSKAT and GLMM-MiRKAT can not be used.

3.1 Simulation A: Microbiome data processed via the same pipeline with a common tree

We used the Dirichilet-multinomial simulation framework as in previous studies (Chen and Li, 2013; Zhao *et al.*, 2015; Zhan *et al.*, 2018) to generate OTU count tables. Panel A in Figure 1 demonstrate the general workflow of this simulation. To be more specific, we first generate a large OTU table via a Dirichilet Multinomial distribution with hyper-parameters estimated from a real throat microbiome dataset (Charlson *et al.*, 2010), which contains 60 samples with 856 OTUs. We consider these microbiome data as the "true" microbiome abundance, based on which the outcome variable is simulated, with the association effect can be the same or different across studies depending on the different simulation scenarios. The total number of counts for all samples are simulated to be 10^6, which represent the total number of 16s rRNA molecules in the community. We denote this "true" OTU tables as M^* which are unobserved. From this community, we subsample to generate the observed microbiome OTU table (denoted as M) and use M for association testing.

After simulating the true OTU table, we partition the 856 taxa into 20 clusters according to cophenetic distances calculated from the phylogenetic tree. The abundance of these OTU clusters vary substantially. With the simulated microbial community, for a continuous trait, we generate

the outcomes via the following:

$$Y_{ik} = 0.5 \cdot (1 + X_{ik,1} + X_{ik,2}) + f_k \cdot scale(\sum_{l \in \mathscr{A}} M^*_{ik,l}) + h_k + \varepsilon_{ik} \ , \tag{10}$$

where, $i = 1, \cdots, n_k$, $k = 1, \cdots, K$, $X_{\cdot,1}$ and $X_{\cdot,2}$ are other covariates with $X_{\cdot,1} \sim Bernoulli(0.5)$ and $X_{\cdot,2} \sim N(0,1)$. Here, $M^*_{ik,l}$ is the unobserved "true" counts of lth taxon for individual i in kth study, $l \in \mathscr{A}$ indicates whether taxon l belongs to a specific taxon cluster $\mathscr{A}$. $scale(\cdot)$ is the function standardizing vector $\sum_{l \in \mathscr{A}} \{M^*_{ik,l}\}_i$ to a vector with zero mean and standard deviation of 1. Because the total number of counts is 10^6 for all samples, this is not different from using the true relative abundances. f_k is the effect size for k-th study. Between-study variation and within-study variation are described by $h_k \sim N(0, \sigma^2)$, $\varepsilon_{ik} \sim N(0, \delta_k^2)$ separately. We set $\sigma^2 = 1$ and the same $\delta_k^2 = 1$ for all k, or different combinations of δ_k^2 using $\delta_k^2 = 1$, for $k = 1,2,3$, $\delta_k^2 = 3$, for $k = 4,5,6,7$, $\delta_k^2 = 5$, for $k = 8,9,10$ to evaluate type I error. We set $n_k = 50, K = 10$ so that we have a total of 10 studies with each study having 50 study subjects.

For binary outcome, we generated outcomes by

$$logit[P(Y_{ik} = 1)] = 0.5 \cdot (1 + X_{ik,1} + X_{ik,2}) + f_k \cdot scale(\sum_{l \in \mathscr{A}} M^*_{ik,l}) + h_k, \tag{11}$$

in which $X_{\cdot,1}$, $X_{\cdot,2}$, $M^*_{ik,l}$ and h_k are defined the same way as before.

Note that the above outcome generating procedure links the outcome with the "true" microbiome community, even though the observed data can be biased, as we will describe later.

To evaluate type I error, we simulated data with $f_k = 0$ for all k, i.e. no taxon is associated with the outcome in any of the studies. We assess power under two scenarios that differ in the effects of microbial profiles within multiple studies:

Scenario A1 the microbiome effect is the same across all studies.

Scenario A2 the microbiome effects are different across studies.

For scenario A1, the underlying microbiome effect on the outcomes is the same across all studies; we set $f_k = c$ for all k and $\mathscr{A}$ is selected to be the 16-th cluster, which is a very common cluster constituting 19.6% of the total true OTU reads. We vary the constant c to assess the statistical power. In scenario A2 , we set $f_k = c$ for $k = 1, \cdots, 5$ and $f_k = -c$ for $k = 6, \cdots, 10$. Other clusters were not selected to be related to the outcome Y.

Note that we generate the outcomes according to the covariates and the "true" microbiome abundance, which are unobserved. We further generate the observed OTU table (M) by a sub-sampling of the "true" OTU table and use M for association testing. Here, we consider the sequencing procedure as a random sampling (with or without bias) from the 16s rRNA molecules (a total of 10^6 per sample) in the true community. Two methods of the random sub-sampling are considered. In the first method, each molecule has the same probability being sampled and there is no bias in the sequencing process. In specific, we fix the library size (i.e., total number of sequencing reads in a sample) of all samples as 1,000, and for sample i in study k with the l-th taxon true counts as $M^*_{ik,l}$, we generate observed OTU counts as a multinomial random sample from with probability $M^*_{ik,l} \times 10^{-6}$ and total read counts of 1,000. In this case, the observed OTU counts are a faithful representation of the underlying community with only random errors. In the second method, there is bias in the sequencing procedure and some taxa are observed with a higher (or lower) relative abundances than in the true community. In specific, we first generate an observed OTU table the

same way as in method one, and then for study 1-5, we inflate the read counts of common taxa by 1.1 times. Here, the "common" taxa are defined as the top 10% taxa with the highest relative abundances in the "true" unobserved community. This way, the common taxa are observed with a higher (and biased) relative abundances in these studies than in the true community and the rest are observed with a lower relative abundances. Due to such a bias, the beta-diversities from samples from the different studies are not trustworthy.

All hypothesis testings and model evaluations are conducted using these "observed" datasets because these are the data that we observe in reality.

We apply the proposed SMRmix (including the version with a single kernel and the ominus version), CKSAT (continuous) or GLMM-MiRKAT(binary) to the simulated data. In each evaluation, we make 10,000 and 5,000 replicates to calculate type I error and and power, respectively. We test the null hypothesis that there is no association between microbiome profiles and outcomes. For the test with a single kernel, we use three kernels, i.e., the weighted UniFrac(K_w), the unweighted UniFrac(K_u) and the Bray-Curtis (K_{bc}) kernels. The UniFrac kernels utilize phylogenetic information while the Bray-Curtis does not. For SMRmix-Omnibus test, we synthesize test results of five kernels including the weighted UniFrac, the unweighted UniFrac, the Bray-Curtis and the generalized UniFrac with weight parameter of 0 and 0.5. We also include in our comparison a meta-analysis procedure, i.e., we apply the classic MiRKAT to each individual study, obtain their p-values and combine the p-values using the HMP combination (Wilson, 2019).

3.2 Simulation B: Microbiome data processed through different pipelines with different tree structures

In this simulation, we generate microbiome read table and multiple phylogenetic trees based on the raw sequencing files from a real microbiome study (Monaco *et al.*, 2016), which contains stool microbiome data from 82 HIV-positive and 40 HIV-negative individuals. In this study, the samples underwent PCR amplification of the V4 region and were sequenced on the Illumina MiSeq platform. Panel B in Figure 1 summarizes the general workflow of this simulation. In specific, we download the pair-end sequencing data in FASTQ format from Synapse (ID: syn22273201), and reprocess it through different pipelines to generate a total of 9 dataset, which each has distinct phylogenetic tree structure and OTU definition.

Among them, six datasets are generated through the QIIME2 pipeline (Bolyen *et al.*, 2019), in which the pair-end sequencing data are joined and filtered before subsequent steps. Two data denoising techniques (DADA2 (Callahan *et al.*, 2016) and deblur (Amir *et al.*, 2017)) and three phylogeny construction approaches (FastTree (Price *et al.*, 2009), RAxML (Stamatakis, 2006) and IQ-TREE (Nguyen *et al.*, 2015)) are used in Qiime2 to generate a total of 6 (2×3) datasets. The other 3 datasets are generated through the original QIIME pipeline (Caporaso *et al.*, 2010) with three different strategies: closed-reference OTU picking based on references SILVA (Release 132) (Quast *et al.*, 2012; Yilmaz *et al.*, 2013), Greengenes (Version 13_8) (McDonald *et al.*, 2012), and De novo OTU picking. Because the differences in the phylogenetic trees and the taxa definitions across the processing steps, a common kernel matrix is not available in this simulation. This represents the situation in which we use public dataset without access to the raw sequencing data when we integrate multiple datasets.

After generating the nine datasets, we partition the taxa in each dataset into 20 clusters. We

further simulate new count tables based on these datasets using Dirichlet-Multinomial distribution and the hyper-parameters from the nine count tables. The outcomes are generated using equation (10) for the continuous case and equation (11) for the binary case. As in scenario A, we evaluate type I error and power using $\sigma^2 = 1$ and either the same $\delta_k^2 = 1$ for all k, or different combinations of δ_k^2 where $\delta_k^2 = 1$ for $k = 1,2,3$, $\delta_k^2 = 3$ for $k = 4,5,6,7$, $\delta_k^2 = 5$ for $k = 8,9,10$. To assess power, we assume that the microbiome effect is the same across the nine tables because they originate from the same microbiome community. We specify $f_k = c$ for $k \in \{1, \cdots, 9\}$ and select the 16-th cluster in each dataset as $\mathscr{A}$.

For this simulation, we do not attempt to simulate a bias-process because there is no "true" community available. We apply the proposed SMRmix and the meta-analysis procedures to the simulated data, while noting that CSKAT/GLMM-MiRKAT could not be applied.

4 Results

4.1 Simulation results

Table 1 summarizes the type I error rates under different scenarios for continuous and dichotomous outcomes under the multiple scenarios of simulation A. SMRmix, CSKAT and GLMM-MiRKAT all have valid type I error rates across all simulations, with or without bias in the sequencing procedure. SMRmix tends to be slightly conservative with binary outcomes.

In simulation A1 with both continuous and binary outcomes, (left panels in Figure 2) when all studies share the same microbiome effect and there is no bias in the sequencing procedure, comparing across methods with the same types of kernel (such as the Bray-Curtis kernel), CSKAT/GLMM-MiRKAT has higher power than SMRmix. This is expected because this is a scenario when the microbiome data across studies are comparable with the same microbiome effect. CSKAT/GLMM-MiRKAT utilize the similarities for pairs of samples that come from different studies, offering more efficient use of the data. The meta-analysis approach that combines the p-values from multiple MiRKAT models has substantial power loss. Comparing within the same method across different kernels, for both CSKAT and SMRmix, the weighted UniFrac is the most powerful, followed by Bray-Curtis kernels and the unweighted UniFrac is the least powerful, due to that the microbiome effect focuses on a phylogenetically-informed cluster of common taxa. Similar trend can be observed even when the sequencing procedure are biased (left panels in Figure S1 in Supplementary Materials), possibly due to the simple/small bias generation procedure utilized in this simulation.

For scenarios A2 when the studies have very different microbiome effect (middle panels in Figure 2) with no sequencing bias, CSKAT/GLMM-MiRKAT with any kernel matrices has very little power: CSKAT/GLMM-MiRKAT implicitly assumes that the microbiome-outcome relationship is the same via the same kernel function, which is clearly violated in this scenario. On the other hand, SMRmix model reports much higher power. The power of meta-analysis is substantially lower than SMRmix, but higher than CSKAT/GLMM-MiRKAT in this context. Comparing different kernels, the conclusion is similar to in scenario A1. Interestingly, SMRmix almost lose no power comparing simulations A1 and A2, indicating that its ability to maintain power when different studies have different effect sizes. The conclusion is the same with the bias that we introduce (right panels in Figure S1 of Supplementary Materials).

For simulation scenario B, CKSAT/GLMM-MiRKAT can not be applied. SMRmix shows

much higher power than the meta-analysis procedure (right panels in Figure 2).

4.2 Application to HIV studies

We utilized SMRmix to analyze data from the HIV Microbiome Re-analysis Consortium (Tuddenham *et al.*, 2020), which re-analyzed $\alpha-$diversity using raw 16S rRNA sequencing data from 17 studies. These studies aimed to characterize the gut dysbiosis in HIV+ individuals. The microbiome amplicon sequencing data and metadata of these studies are available on Synapse (ID: syn18406854) and NCBI. The 17 studies differ in their study design, institution of participants recruitment, study population, type of samples and various factors in the sequencing protocol including the 16s rRNA variable regions that were sequenced and the sequencing platform. The consortium reprocessed the raw sequencing data using standardized protocols and conducted taxonomic assignments in species level through Resphera Insight.

Our goal was to evaluate the association between microbiome composition and HIV status, MSM status by integrating all these studies and draw a consistent conclusion. Out of the 17 datasets, 14 studies(763 samples and 666 taxa) have subject-level information on MSM status, age, and HIV status, which we used for further analysis, which are the data that we use for this analysis. Because a common phylogenetic tree is not available from Synapse webpage, we created a phylogentic tree using the "tax_tree" function in R package "LTN" based on the taxonmy relationship Wang *et al.* (2021). The Principal Coordinates Analysis (PCoA) plots, which visualize the distance matrix between study subjects, are presented in Supplementary Figure S3.

We conducted the association tests utilizing single-kernel SMRmix based on kernels including weighted UniFrac(K_w), unweighted UniFrac(K_u), Jaccard(K_{jac}), Bray-Curtis(K_{bc}) and omnibus SMRmix using the above kernels. Suggested by previous studies (Noguera-Julian *et al.*, 2016), we adjusted HIV status when the outcome variable is MSM, and adjusted MSM when the outcome variable is HIV status. The statistical model was established as

$$logit(P(HIV_{ik}=1)) = \beta_1 age_{ik} + \beta_2 MSM_{ik} + f(M_{ik}) + h_k,$$
$$logit(P(MSM_{ik}=1)) = \beta_1 age_{ik} + \beta_2 HIV_{ik} + f(M_{ik}) + h_k.$$

The p-values of association test were summarized in Table 2.

For the outcome HIV status, SMRmix with weighted UniFrac, UniFrac, Jaccard, and Bray-Curtis kernels all discovered significant signals (p-values < 0.0001). Synthesizing the results based on above kernels, omnibus SMRmix found HIV status and microbiome composition associated (p-value < 0.0001). For the outcome MSM, similar conclusions were obtained that SMRmix with all kernels show significant connection between MSM and microbiome composition.

5 Discussion

In this paper, we present a novel approach, SMRmix, for testing the association between microbiome composition and outcomes of interest, by combining the concepts of similarity matrix regression and kernel machine regression. This method is particularly beneficial in the context of integrative analysis of microbiome data from multiple studie, cohorts or protocols, where inter-study pairwise distances are either absent or unreliable. SMRmix effectively addresses the limitations of existing kernel regression tests by leveraging only within-study pairwise distances for

testing. In simulations, SMRmix showed well-controlled Type 1 error rates. We compared SMRmix to two other methods, CSKAT and GLMM-MiRKAT (which require the usually non-available between-study pairwise distance), for microbiome compositional analysis. Our results showed that if distances of all pairs of samples are available and when the microbiome effect is homogeneous , SMRmix has relatively lower power. However, SMRmix exhibits superior performance in scenarios where the effects of taxa are a mixture of positive and negative effects, and when different effect sizes result in a significant reduction in power for the CSKAT and GLMM-MiRKAT. This highlights the robustness and effectiveness of SMRmix, particularly when inter-study pairwise distances may not be reliable or available. We also compared SMRmix with a meta-analysis procedure (which, like SMRmix, doesn't require the between-study pairwise distance), and showed higher power of SMRmix across all simulation setups. In summary, our proposed SMRmix method represents a novel and effective approach for microbiome compositional analysis, with the potential to contribute to ongoing efforts to better understand the complex interactions between microbial communities and host health.

Kernel selection is a critical challenge for any kernel-based approaches, as choosing an unsuitable kernel can severely reduce the statistical power despite controlling the type I error. In microbiome studies, each kernel function assumes a distinct microbiome effect in the community, and improper choice of kernel can compromise the analysis (Plantinga *et al.*, 2017). To address this issue, SMRmix provides an omnibus test that integrates multiple kernel matrices to enhance the robustness of the analysis. SMRmix also allows for the use of different kernels for each study, such as Bray-Curtis distances in one study and UniFrac distances in another, thereby offering greater flexibility depending on the data availability. Our simulations demonstrate that single-kernel tests can be sensitive to kernel selection. In contrast, our omnibus SMRmix approach synthesizes results from multiple kernel-based tests by combining p-values and only incurs a small loss of power relative to using the best kernel. By contrast, the omnibus test improves the power substantially compared to a poor kernel choice. Given that the best kernel is often unknown in practice, the omnibus test is a practical and useful alternative.

We applied SMRmix to investigate the relationship between the human gut microbiome composition and HIV infection using data from 14 studies. Although many studies have reported gut microbiome dysbiosis associated with HIV infection and the possible role of "leaky gut" in HIV pathologenesis (Dubourg *et al.*, 2016; Mutlu *et al.*, 2014; Noguera-Julian *et al.*, 2016; Nowak *et al.*, 2015; Pinto-Cardoso *et al.*, 2017; Villanueva-Millán *et al.*, 2017; Yu *et al.*, 2014; Sun *et al.*, 2016; Vázquez-Castellanos *et al.*, 2015; Monaco *et al.*, 2016; Vesterbacka *et al.*, 2017), some have also suggested that MSM may be a confounding factor in the HIV-microbiome relationship (Noguera-Julian *et al.*, 2016). Our analysis adjusted for age and MSM status to assess the association between microbiome composition and HIV status, as well as for age and HIV infection to examine the association between microbiome composition and MSM status. By integrating information from multiple studies on this topic, we discovered significant associations between microbiome composition and HIV (adjusting for age and MSM) and between microbiome and MSM (adjusting for age and HIV), which provides compelling evidence of HIV dysbiosis that is independent of MSM status, contradicting the notion that MSM is a confounding factor in this relationship.

Software

The implementation of SMRmix in R codes are available in the GitHub repository `https://github.com/mengyu-he/SMRmix`.

Funding

This work was funded by by NIH grants R21AI154236, R01GM147162 and U24OD023382.

References

AMIR, AMNON, MCDONALD, DANIEL, NAVAS-MOLINA, JOSE A., KOPYLOVA, EVGUENIA, MORTON, JAMES T., ZECH XU, ZHENJIANG, KIGHTLEY, ERIC P., THOMPSON, LUKE R., HYDE, EMBRIETTE R., GONZALEZ, ANTONIO *et al.* (2017, April). Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems* **2**(2).

BOLYEN, EVAN, RIDEOUT, JAI RAM, DILLON, MATTHEW R., BOKULICH, NICHOLAS A., ABNET, CHRISTIAN C., AL-GHALITH, GABRIEL A., ALEXANDER, HARRIET, ALM, ERIC J., ARUMUGAM, MANIMOZHIYAN, ASNICAR, FRANCESCO, BAI, YANG, BISANZ, JORDAN E., BITTINGER, KYLE, BREJNROD, ASKER, BRISLAWN, COLIN J., BROWN, C. TITUS, CALLAHAN, BENJAMIN J., CARABALLO-RODRÍGUEZ, ANDRÉS MAURICIO, CHASE, JOHN, COPE, EMILY K., DA SILVA, RICARDO, DIENER, CHRISTIAN, DORRESTEIN, PIETER C., DOUGLAS, GAVIN M., DURALL, DANIEL M., DUVALLET, CLAIRE, EDWARDSON, CHRISTIAN F., ERNST, MADELEINE, ESTAKI, MEHRBOD, FOUQUIER, JENNIFER, GAUGLITZ, JULIA M., GIBBONS, SEAN M., GIBSON, DEANNA L., GONZALEZ, ANTONIO, GORLICK, KESTREL, GUO, JIARONG, HILLMANN, BENJAMIN, HOLMES, SUSAN, HOLSTE, HANNES, HUTTENHOWER, CURTIS, HUTTLEY, GAVIN A., JANSSEN, STEFAN, JARMUSCH, ALAN K., JIANG, LINGJING, KAEHLER, BENJAMIN D., KANG, KYO BIN, KEEFE, CHRISTOPHER R., KEIM, PAUL, KELLEY, SCOTT T., KNIGHTS, DAN, KOESTER, IRINA, KOSCIOLEK, TOMASZ, KREPS, JORDEN, LANGILLE, MORGAN G. I., LEE, JOSLYNN, LEY, RUTH, LIU, YONG-XIN, LOFTFIELD, ERIKKA, LOZUPONE, CATHERINE, MAHER, MASSOUD, MAROTZ, CLARISSE, MARTIN, BRYAN D., MCDONALD, DANIEL, MCIVER, LAUREN J., MELNIK, ALEXEY V., METCALF, JESSICA L., MORGAN, SYDNEY C., MORTON, JAMIE T., NAIMEY, AHMAD TURAN, NAVAS-MOLINA, JOSE A., NOTHIAS, LOUIS FELIX, ORCHANIAN, STEPHANIE B., PEARSON, TALIMA, PEOPLES, SAMUEL L., PETRAS, DANIEL, PREUSS, MARY LAI, PRUESSE, ELMAR, RASMUSSEN, LASSE BUUR, RIVERS, ADAM, ROBESON, MICHAEL S., ROSENTHAL, PATRICK, SEGATA, NICOLA, SHAFFER, MICHAEL, SHIFFER, ARRON, SINHA, RASHMI, SONG, SE JIN, SPEAR, JOHN R., SWAFFORD, AUSTIN D., THOMPSON, LUKE R., TORRES, PEDRO J., TRINH, PAULINE, TRIPATHI, ANUPRIYA, TURNBAUGH, PETER J., UL-HASAN, SABAH, VAN DER HOOFT, JUSTIN J. J., VARGAS, FERNANDO, VÁZQUEZ-BAEZA, YOSHIKI, VOGTMANN, EMILY, VON HIPPEL, MAX, WALTERS, WILLIAM, WAN, YUNHU, WANG, MINGXUN, WARREN, JONATHAN, WEBER, KYLE C., WILLIAMSON, CHARLES H. D., WILLIS, AMY D., XU, ZHENJIANG ZECH, ZANEVELD, JESSE R., ZHANG, YILONG, ZHU, QIYUN, KNIGHT, ROB *et al.* (2019, August).

Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology* **37**(8), 852–857.

CALLAHAN, BENJAMIN J., MCMURDIE, PAUL J., ROSEN, MICHAEL J., HAN, ANDREW W., JOHNSON, AMY JO A. AND HOLMES, SUSAN P. (2016). Dada2: High-resolution sample inference from illumina amplicon data. *Nature Methods* **13**.

CAPORASO, J GREGORY, KUCZYNSKI, JUSTIN, STOMBAUGH, JESSE, BITTINGER, KYLE, BUSHMAN, FREDERIC D, COSTELLO, ELIZABETH K, FIERER, NOAH, PEÑA, ANTONIO GONZALEZ, GOODRICH, JULIA K, GORDON, JEFFREY I, HUTTLEY, GAVIN A, KELLEY, SCOTT T, KNIGHTS, DAN, KOENIG, JEREMY E, LEY, RUTH E, LOZUPONE, CATHERINE A, MCDONALD, DANIEL, MUEGGE, BRIAN D, PIRRUNG, MEG, REEDER, JENS, SEVINSKY, JOEL R, TURNBAUGH, PETER J, WALTERS, WILLIAM A, WIDMANN, JEREMY, YATSUNENKO, TANYA, ZANEVELD, JESSE *et al.* (2010, May). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**(5), 335–336.

CHAPMAN, JULIET AND WHITTAKER, JOHN. (2008, September). Analysis of multiple SNPs in a candidate gene or region. *Genetic Epidemiology* **32**(6), 560–566.

CHARLSON, EMILY S., CHEN, JUN, CUSTERS-ALLEN, REBECCA, BITTINGER, KYLE, LI, HONGZHE, SINHA, ROHINI, HWANG, JENNIFER, BUSHMAN, FREDERIC D. AND COLLMAN, RONALD G. (2010, December). Disordered Microbial Communities in the Upper Respiratory Tract of Cigarette Smokers. *PLoS ONE* **5**(12), e15216.

CHEN, JUN AND LI, HONGZHE. (2013). Kernel methods for regression analysis of microbiome compositional data. Volume 55.

CONNEELY, KAREN N. AND BOEHNKE, MICHAEL. (2007, December). So Many Correlated Tests, So Little Time! Rapid Adjustment of P Values for Multiple Correlated Tests. *The American Journal of Human Genetics* **81**(6), 1158–1168.

DEEKS, STEVEN G, LEWIN, SHARON R AND HAVLIR, DIANE V. (2013, November). The end of AIDS: HIV infection as a chronic disease. *The Lancet* **382**(9903), 1525–1533.

DINH, DUY M., VOLPE, GRETCHEN E., DUFFALO, CHAD, BHALCHANDRA, SEEMA, TAI, ALBERT K., KANE, ANNE V., WANKE, CHRISTINE A. AND WARD, HONORINE D. (2015, January). Intestinal Microbiota, Microbial Translocation, and Systemic Inflammation in Chronic HIV Infection. *Journal of Infectious Diseases* **211**(1), 19–27.

DUBOURG, GRÉGORY, LAGIER, JEAN-CHRISTOPHE, HÜE, SOPHIE, SURENAUD, MATHIEU, BACHAR, DIPANKAR, ROBERT, CATHERINE, MICHELLE, CAROLINE, RAVAUX, ISABELLE, MOKHTARI, SAADIA, MILLION, MATTHIEU, STEIN, ANDREAS, BROUQUI, PHILIPPE, LEVY, YVES *et al.* (2016, July). Gut microbiota associated with HIV infection is significantly enriched in bacteria tolerant to oxygen. *BMJ Open Gastroenterology* **3**(1), e000080.

KIRK, G. D., MERLO, C., O'DRISCOLL, P., MEHTA, S. H., GALAI, N., VLAHOV, D., SAMET, J. AND ENGELS, E. A. (2007, July). HIV Infection Is Associated with an Increased Risk for Lung Cancer, Independent of Smoking. *Clinical Infectious Diseases* **45**(1), 103–110.

KOH, HYUNWOOK, LI, YUTONG, ZHAN, XIANG, CHEN, JUN AND ZHAO, NI. (2019, May). A Distance-Based Kernel Association Test Based on the Generalized Linear Mixed Model for Correlated Microbiome Studies. *Frontiers in Genetics* **10**, 458.

KWEE, LYDIA COULTER, LIU, DAWEI, LIN, XIHONG, GHOSH, DEBASHIS AND EPSTEIN, MICHAEL P. (2008, February). A Powerful and Flexible Multilocus Association Test for Quantitative Traits. *The American Journal of Human Genetics* **82**(2), 386–397.

LEWDEN, C. (2004, August). Causes of death among human immunodeficiency virus (HIV)-infected adults in the era of potent antiretroviral therapy: emerging role of hepatitis and cancers, persistent role of AIDS. *International Journal of Epidemiology* **34**(1), 121–130.

LIN, D. Y. (2005, March). An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics* **21**(6), 781–787.

LING, ZONGXIN, JIN, CHANGZHONG, XIE, TIANSHENG, CHENG, YIWEN, LI, LANJUAN AND WU, NANPING. (2016, August). Alterations in the Fecal Microbiota of Patients with HIV-1 Infection: An Observational Study in A Chinese Population. *Scientific Reports* **6**(1), 30673.

LIU, YAOWU AND XIE, JUN. (2020). Cauchy combination test: A powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association* **115**.

LOZUPONE, CATHERINE A., LI, MARCELLA, CAMPBELL, THOMAS B., FLORES, SONIA C., LINDERMAN, DEREK, GEBERT, MATTHEW J., KNIGHT, ROB, FONTENOT, ANDREW P. AND PALMER, BRENT E. (2013, September). Alterations in the Gut Microbiota Associated with HIV-1 Infection. *Cell Host & Microbe* **14**(3), 329–339.

MCARDLE, BRIAN H. AND ANDERSON, MARTI J. (2001, January). FITTING MULTIVARIATE MODELS TO COMMUNITY DATA: A COMMENT ON DISTANCE-BASED REDUNDANCY ANALYSIS. *Ecology* **82**(1), 290–297.

MCDONALD, DANIEL, PRICE, MORGAN N, GOODRICH, JULIA, NAWROCKI, ERIC P, DESANTIS, TODD Z, PROBST, ALEXANDER, ANDERSEN, GARY L, KNIGHT, ROB AND HUGENHOLTZ, PHILIP. (2012, March). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal* **6**(3), 610–618.

METAHIT CONSORTIUM (ADDITIONAL MEMBERS), ARUMUGAM, MANIMOZHIYAN, RAES, JEROEN, PELLETIER, ERIC, LE PASLIER, DENIS, YAMADA, TAKUJI, MENDE, DANIEL R., FERNANDES, GABRIEL R., TAP, JULIEN, BRULS, THOMAS, BATTO, JEAN-MICHEL, BERTALAN, MARCELO, BORRUEL, NATALIA, CASELLAS, FRANCESC, FERNANDEZ, LEYDEN, GAUTIER, LAURENT, HANSEN, TORBEN, HATTORI, MASAHIRA, HAYASHI, TETSUYA, KLEEREBEZEM, MICHIEL, KUROKAWA, KEN, LECLERC, MARION, LEVENEZ, FLORENCE, MANICHANH, CHAYSAVANH, NIELSEN, H. BJØRN, NIELSEN, TRINE, PONS, NICOLAS, POULAIN, JULIE, QIN, JUNJIE, SICHERITZ-PONTEN, THOMAS, TIMS, SEBASTIAN, TORRENTS, DAVID, UGARTE, EDGARDO, ZOETENDAL, ERWIN G., WANG, JUN, GUARNER, FRANCISCO, PEDERSEN, OLUF, DE VOS, WILLEM M., BRUNAK, SØREN, DORÉ, JOEL,

WEISSENBACH, JEAN, EHRLICH, S. DUSKO *et al.* (2011, May). Enterotypes of the human gut microbiome. *Nature* **473**(7346), 174–180.

MONACO, CYNTHIA L., GOOTENBERG, DAVID B., ZHAO, GUOYAN, HANDLEY, SCOTT A., GHEBREMICHAEL, MUSIE S., LIM, EFREM S., LANKOWSKI, ALEX, BALDRIDGE, MEGAN T., WILEN, CRAIG B., FLAGG, MEAGHAN, NORMAN, JASON M., KELLER, BRIAN C., LUÉVANO, JESÚS MARIO, WANG, DAVID, BOUM, YAP, MARTIN, JEFFREY N., HUNT, PETER W., BANGSBERG, DAVID R., SIEDNER, MARK J., KWON, DOUGLAS S. *et al.* (2016, March). Altered Virome and Bacterial Microbiome in Human Immunodeficiency Virus-Associated Acquired Immunodeficiency Syndrome. *Cell Host & Microbe* **19**(3), 311–322.

MUTLU, ECE A., KESHAVARZIAN, ALI, LOSURDO, JOHN, SWANSON, GARTH, SIEWE, BASILE, FORSYTH, CHRISTOPHER, FRENCH, AUDREY, DEMARAIS, PATRICIA, SUN, YAN, KOENIG, LARS, COX, STEPHEN, ENGEN, PHILLIP, CHAKRADEO, PRACHI, ABBASI, RAWAN, GORENZ, ANNIKA, BURNS, CHARLES *et al.* (2014, February). A Compositional Look at the Human Gastrointestinal Microbiome and Immune Activation Parameters in HIV Infected Subjects. *PLoS Pathogens* **10**(2), e1003829.

NGUYEN, LAM-TUNG, SCHMIDT, HEIKO A., VON HAESELER, ARNDT AND MINH, BUI QUANG. (2015, January). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* **32**(1), 268–274.

NOGUERA-JULIAN, MARC, ROCAFORT, MUNTSA, GUILLÉN, YOLANDA, RIVERA, JAVIER, CASADELLÀ, MARIA, NOWAK, PIOTR, HILDEBRAND, FALK, ZELLER, GEORG, PARERA, MARIONA, BELLIDO, ROCÍO, RODRÍGUEZ, CRISTINA, CARRILLO, JORGE, MOTHE, BEATRIZ, COLL, JOSEP, BRAVO, ISABEL, ESTANY, CARLA, HERRERO, CRISTINA, SAZ, JORGE, SIRERA, GUILLEM, TORRELA, ARIADNA, NAVARRO, JORDI, CRESPO, MANEL, BRANDER, CHRISTIAN, NEGREDO, EUGÈNIA, BLANCO, JULIÀ, GUARNER, FRANCISCO, CALLE, MARIA LUZ, BORK, PEER, SÖNNERBORG, ANDERS, CLOTET, BONAVENTURA *et al.* (2016, March). Gut Microbiota Linked to Sexual Preference and HIV Infection. *EBioMedicine* **5**, 135–146.

NOWAK, PIOTR, TROSEID, MARIUS, AVERSHINA, EKATARINA, BARQASHO, BABILONIA, NEOGI, UJJWAL, HOLM, KRISTIAN, HOV, JOHANNES R., NOYAN, KAJSA, VESTERBACKA, JAN, SVÄRD, JENNY, RUDI, KNUT *et al.* (2015, November). Gut microbiota diversity predicts immune status in HIV-1 infection. *AIDS* **29**(18), 2409–2418.

NOWAK, REBECCA G., BENTZEN, SØREN M., RAVEL, JACQUES, CROWELL, TREVOR A., DAUDA, WUESE, MA, BING, LIU, HONGJIE, BLATTNER, WILLIAM A., BARAL, STEFAN D. AND CHARURAT, MANHATTAN E. (2017, March). Rectal microbiota among HIV-uninfected, untreated HIV, and treated HIV-infected in Nigeria. *AIDS* **31**(6), 857–862.

PAN, WEI, HAN, FANG AND SHEN, XIAOTONG. (2010). Test Selection with Application to Detecting Disease Association with Multiple SNPs. *Human Heredity* **69**(2), 120–130.

PHAIR, JOHN AND PALELLA, FRANK. (2011, July). Renal disease in HIV-infected individuals:. *Current Opinion in HIV and AIDS* **6**(4), 285–289.

PINTO-CARDOSO, SANDRA, LOZUPONE, CATHERINE, BRICEÑO, OLIVIA, ALVA-HERNÁNDEZ, SELMA, TÉLLEZ, NORMA, ADRIANA, AGUILAR, MURAKAMI-OGASAWARA, AKIO AND REYES-TERÁN, GUSTAVO. (2017, March). Fecal Bacterial Communities in treated HIV infected individuals on two antiretroviral regimens. *Scientific Reports* **7**(1), 43741.

PLANTINGA, ALEXANDER, ZHAN, XIANG, ZHAO, NI, CHEN, JUN, JENQ, ROBERT R AND WU, MICHAEL C. (2017, Feb). Mirkat-s: a community-level test of association between the microbiota and survival times. *Microbiome* **5**(1), 17.

PRICE, M. N., DEHAL, P. S. AND ARKIN, A. P. (2009, July). FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Molecular Biology and Evolution* **26**(7), 1641–1650.

QUAST, CHRISTIAN, PRUESSE, ELMAR, YILMAZ, PELIN, GERKEN, JAN, SCHWEER, TIMMY, YARZA, PABLO, PEPLIES, JÖRG AND GLÖCKNER, FRANK OLIVER. (2012, November). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* **41**(D1), D590–D596.

SIMES, R. J. (1986). An improved bonferroni procedure for multiple tests of significance. *Biometrika* **73**.

STAMATAKIS, ALEXANDROS. (2006, November). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**(21), 2688–2690.

SUN, YANG, MA, YINGFEI, LIN, PING, TANG, YI-WEI, YANG, LIYING, SHEN, YINZHONG, ZHANG, RENFAN, LIU, LI, CHENG, JUN, SHAO, JIASHEN, QI, TANGKAI, TANG, YAN, CAI, RENTIAN, GUAN, LIQIAN, LUO, BIN, SUN, MEIYAN, LI, BEN, PEI, ZHIHENG *et al.* (2016, January). Fecal bacterial microbiome diversity in chronic HIV-infected patients in China. *Emerging Microbes & Infections* **5**(1), 1–7.

TIPPETT, L H C. (1931). *The methods of statistics*, Volume 2d ed., re. London: Williams Norgate ltd.

TRIANT, VIRGINIA A. (2013, September). Cardiovascular Disease and HIV Infection. *Current HIV/AIDS Reports* **10**(3), 199–206.

TUDDENHAM, SUSAN A., KOAY, WEI LI A., ZHAO, NI, WHITE, JAMES R., GHANEM, KHALIL G., SEARS, CYNTHIA L., DILLON, STEPHANIE, WILSON, CARA, LOZUPONE, CATHERINE, WARD, HONORINE, WANKE, CHRISTINE, DUBOURG, GREGORY, RAOULT, DIDIER, PALMER, BRENT, MONACO, CYNTHIA, KWON, DOUGLAS, MUTLU, ECE, LANDAY, ALAN, PAREDES, ROGER, NOGUERA-JULIAN, MARC, SÖNNERBORG, ANDERS, NEOGI, UJJWAL, NOWAK, REBECCA, RAVEL, JACQUES, PÉREZ-SANTIAGO, JOSUÉ, SMITH, DAVEY M., PINTO-CARDOSO, SANDRA, REYES-TERÁN, GUSTAVO, SERRANO-VILLAR, SERGIO, SOLER, MARIA JOSE GOSALBES, VESTERBACKA, JAN, NOWAK, PIOTR, PÉREZ-MATUTE, P., OTEO, JOSÉ A., D'AURIA, GIUSEPPE, VILLAR-GARCÍA, JUDIT, YU, GUOQIN *et al.* (2020). The impact of human immunodeficiency virus infection on gut microbiota α-diversity: An individual-level meta-analysis. *Clinical Infectious Diseases* **70**.

TZENG, JUNG-YING, ZHANG, DAOWEN, CHANG, SHENG-MAO, THOMAS, DUNCAN C. AND DAVIDIAN, MARIE. (2009, September). Gene-Trait Similarity Regression for Multimarker-Based Association Analysis. *Biometrics* **65**(3), 822–832.

VAGINAL MICROBIOME CONSORTIUM, BROOKS, J PAUL, EDWARDS, DAVID J, HARWICH, MICHAEL D, RIVERA, MARIA C, FETTWEIS, JENNIFER M, SERRANO, MYRNA G, RERIS, ROBERT A, SHETH, NIHAR U, HUANG, BERNICE, GIRERD, PHILIPPE, STRAUSS, JEROME F, JEFFERSON, KIMBERLY K *et al.* (2015, December). The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiology* **15**(1), 66.

VESTERBACKA, JAN, RIVERA, JAVIER, NOYAN, KAJSA, PARERA, MARIONA, NEOGI, UJJWAL, CALLE, MALU, PAREDES, ROGER, SÖNNERBORG, ANDERS, NOGUERA-JULIAN, MARC AND NOWAK, PIOTR. (2017, December). Richer gut microbiota with distinct metabolic profile in HIV infected Elite Controllers. *Scientific Reports* **7**(1), 6269.

VILLANUEVA-MILLÁN, MARÍA J., PÉREZ-MATUTE, PATRICIA, RECIO-FERNÁNDEZ, EMMA, LEZANA ROSALES, JOSÉ M. AND OTEO, JOSÉ A. (2017). Differential effects of antiretrovirals on microbial translocation and gut microbiota composition of HIV-infected patients: Villanueva-Millán MJ et al. *Journal of the International AIDS Society* **20**(1), 21526.

VOLPE, GRETCHEN E, WARD, HONORINE, MWAMBURI, MKAYA, DINH, DUY, BHALCHANDRA, SEEMA, WANKE, CHRISTINE AND KANE, ANNE V. (2014, March). Associations of Cocaine Use and HIV Infection With the Intestinal Microbiota, Microbial Translocation, and Inflammation. *Journal of Studies on Alcohol and Drugs* **75**(2), 347–357.

VÁZQUEZ-CASTELLANOS, J F, SERRANO-VILLAR, S, LATORRE, A, ARTACHO, A, FERRÚS, M L, MADRID, N, VALLEJO, A, SAINZ, T, MARTÍNEZ-BOTAS, J, FERRANDO-MARTÍNEZ, S, VERA, M, DRONDA, F, LEAL, M, DEL ROMERO, J, MORENO, S, ESTRADA, V, GOSALBES, M J *et al.* (2015, July). Altered metabolism of gut microbiota contributes to chronic immune activation in HIV-infected individuals. *Mucosal Immunology* **8**(4), 760–772.

WANG, ZHENWEI, MAO, JIANSHI AND MA, LI. (2021). Microbiome compositional analysis with logistic-tree normal models. *arXiv preprint arXiv:2106.15051*.

WILSON, DANIEL J. (2019, January). The harmonic mean p -value for combining dependent tests. *Proceedings of the National Academy of Sciences* **116**(4), 1195–1200.

WU, MICHAEL C., LEE, SEUNGGEUN, CAI, TIANXI, LI, YUN, BOEHNKE, MICHAEL AND LIN, XIHONG. (2011, July). Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *The American Journal of Human Genetics* **89**(1), 82–93.

WU, MICHAEL C., KRAFT, PETER, EPSTEIN, MICHAEL P., TAYLOR, DEANNE M., CHANOCK, STEPHEN J., HUNTER, DAVID J. AND LIN, XIHONG. (2010, June). Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies. *The American Journal of Human Genetics* **86**(6), 929–942.

YILMAZ, PELIN, PARFREY, LAURA WEGENER, YARZA, PABLO, GERKEN, JAN, PRUESSE, ELMAR, QUAST, CHRISTIAN, SCHWEER, TIMMY, PEPLIES, JÖRG, LUDWIG, WOLFGANG

AND GLÖCKNER, FRANK OLIVER. (2013, November). The SILVA and "all-species living tree project (LTP)" taxonomic frameworks. *Nucleic Acids Research* **42**(D1), D643–D648.

YU, GUOQIN, FADROSH, DOUG, MA, BING, RAVEL, JACQUES AND GOEDERT, JAMES J. (2014, March). Anal microbiota profiles in HIV-positive and HIV-negative MSM. *AIDS* **28**(5), 753–760.

ZHAN, XIANG. (2019, February). Relationship Between MiRKAT and Coefficient of Determination in Similarity Matrix Regression. *Processes* **7**(2), 79.

ZHAN, XIANG, XUE, LINGZHOU, ZHENG, HAOTIAN, PLANTINGA, ANNA, WU, MICHAEL C., SCHAID, DANIEL J., ZHAO, NI AND CHEN, JUN. (2018, December). A small-sample kernel association test for correlated data with application to microbiome association studies. *Genetic Epidemiology* **42**(8), 772–782.

ZHAO, NI, CHEN, JUN, CARROLL, IAN M., RINGEL-KULKA, TAMAR, EPSTEIN, MICHAEL P., ZHOU, HUA, ZHOU, JIN J., RINGEL, YEHUDA, LI, HONGZHE AND WU, MICHAEL C. (2015, May). Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. *The American Journal of Human Genetics* **96**(5), 797–807.

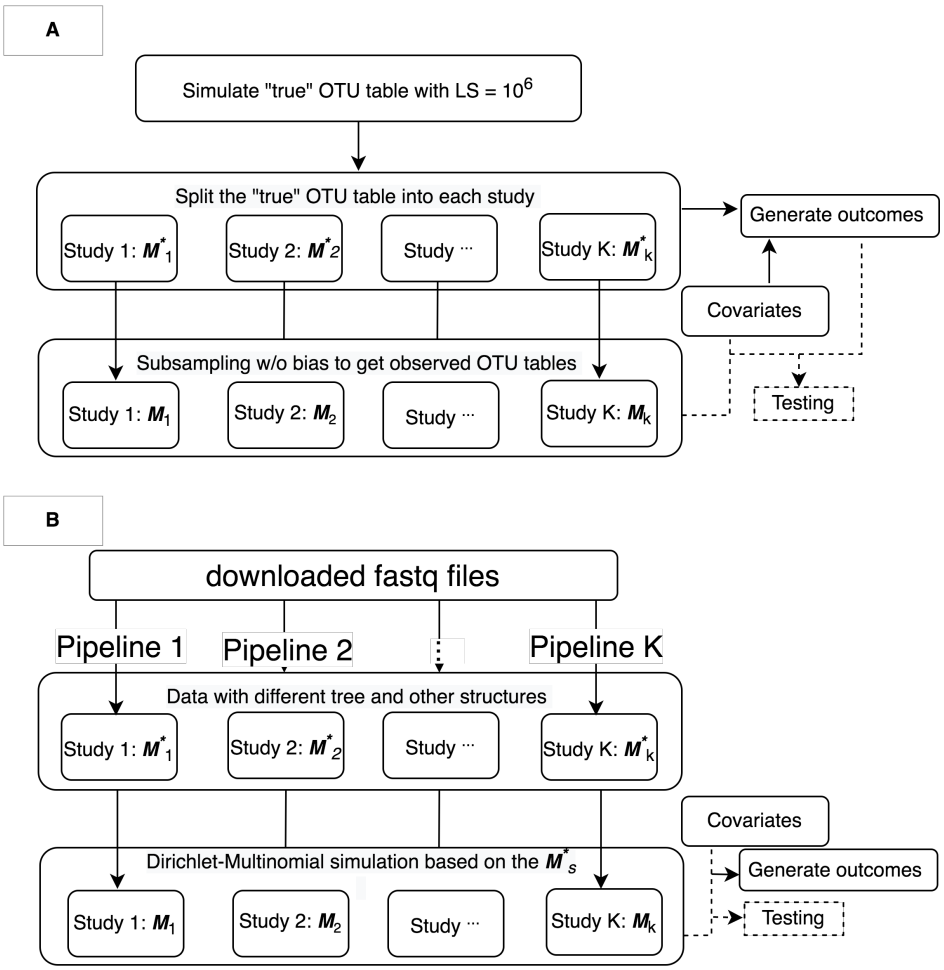

Figure 1: Workflow for simulation studies. Panel A: simulation scenario A, where microbiome data are processed from the same pipeline with a common tree structure. Panel B: simulation scenario B, where microbiome data are processed from different pipelines with different tree structures.

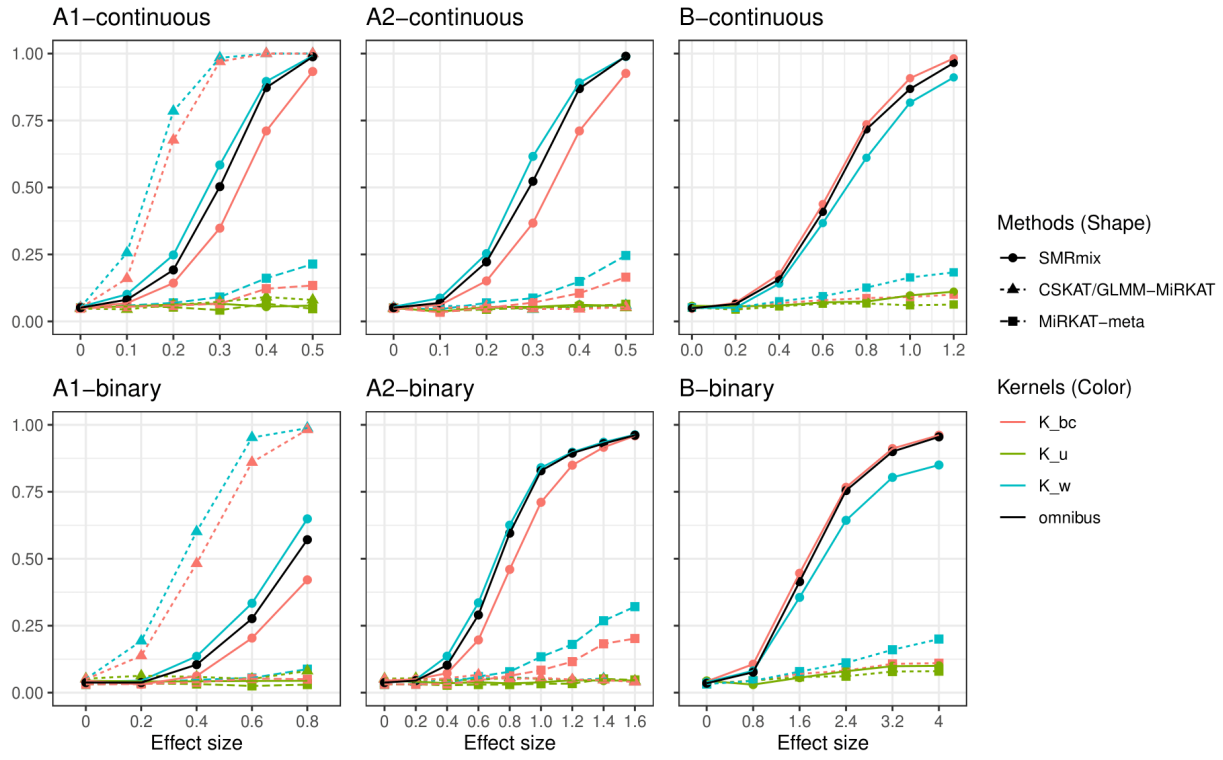

Figure 2: Statistical power for simulations A and B when there is no bias in sequencing sampling. For continuous scenarios, $\delta_k^2 = 1$, $\forall k$ and $\sigma^2 = 1$.

Table 1: Empirical type I error rates comparing CSKAT, GLMM-MiRKAT and SMRmix when there is no bias. For continuous outcome with the same δ_k^2, $\delta_k^2 = 1, \forall k$; for continuous outcome with different δ_k^2, $\delta_k^2 = 1$ for $k = 1,2,3$, $\delta_k^2 = 3$ for $k = 4,5,6,7$, $\delta_k^2 = 5$ for $k = 8,9,10$. For binary scenarios, $\sigma^2 = 1$.

		$\alpha = \mathbf{0.05}$				$\alpha = \mathbf{0.01}$			
Continuous		K_w	K_u	K_{bc}	Omnibus	K_w	K_u	K_{bc}	Omnibus
A : same δ_k^2	CSKAT	0.0528	0.0488	0.0502	/	0.0106	0.0081	0.0087	/
	SMRmix	0.0508	0.0522	0.0509	0.0494	0.0110	0.0107	0.0110	0.0111
A : different δ_k^2	CSKAT	0.0475	0.0481	0.0482	/	0.0100	0.0098	0.0092	/
	SMRmix	0.0459	0.0525	0.0492	0.0481	0.0101	0.0094	0.0098	0.0105
B	SMRmix	0.0469	0.0583	0.0504	0.0493	0.0083	0.0106	0.0101	0.0107
Binary		K_w	K_u	K_{bc}	Omnibus	K_w	K_u	K_{bc}	Omnibus
A	GLMM-MiRKAT	0.0513	0.0522	0.0500	/	0.0097	0.0105	0.0096	/
	SMRmix	0.0403	0.0444	0.0443	0.0377	0.0083	0.0077	0.0071	0.0057
B	SMRmix	0.0471	0.0438	0.0393	0.0452	0.0084	0.0064	0.0060	0.0072

Table 2: P-values of using SMRmix to test associations between microbiome composition and outcome variables HIV Status and MSM adjusting for age.

	Kernels				
	K_w	K_u	K_{jac}	K_{bc}	Omnibus
HIV	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
MSM	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001

Supplementary Materials for

"A Mixed Effect Similarity Matrix Regression Model (SMRmix) for Integrating Multiple Microbiome Datasets at Community Level and its Application in HIV"

MENGYU HE
Department of Biostatistics and Bioinformatics, Emory University
NI ZHAO *
Department of Biostatistics, Johns Hopkins University

A Connection Between SMRmix and Mixed effect Kernel Machine Regression Models

When data are obtained from a single study, the corresponding kernel model regression model, as specified by equations (2.1) and (2.2) in the main manuscript, are $Y_i = \mathbf{X_i}\beta + f(\mathbf{M_i}) + \varepsilon_i$, and $\text{logit}(P(Y_i = 1)) = \mathbf{X_i}\beta + f(\mathbf{M_i})$, respectively. Let $Z_{ij} = (Y_i - \hat{\mu}_i^0)(Y_j - \hat{\mu}_j^0)$, where μ_i^0 is the expectatio of Y_i under the null hypothesis, the original SMR model is $Z_{ij} = b \times K_{ij,k} + e_{ij},\ i < j$. The correspondence between SMRmix and kernel machine regression model has been established via Zhan (2019). Here, no intercept is needed because $Z_{ij} = (Y_i - \hat{\mu}_i^0)(Y_j - \hat{\mu}_j^0)$ has already adjusted for the effect of covariates including the intercept (Tzeng *et al.*, 2009).

When data are obtained from multiple studies, extra terms should be considered in the similarity matrix model to account for the possible extra correlations among similarities. Here, the counterpart mixed-effect kernel machine regression models (as in CSKAT and GLMM-MiRKAT, Equations (2.5) and (2.6) in the main manuscript) are,

$$
\begin{aligned}
Y_{ik} &= \mathbf{X}_{ik}^T\beta + f(\mathbf{M}_{ik}) + h_k + \varepsilon_{ik} \\
logit(P(Y_{ik} = 1)) &= \mathbf{X}_{ik}^T\beta + f(\mathbf{M}_{ik}) + h_k \ .
\end{aligned}
$$

where $h_k \sim N(0, \sigma^2),\ \varepsilon_{ik} \sim N(0, \delta_k^2)$, and $f(\mathbf{M}_{ik})$ depicts the microbiome effect on the outcome. Let $\gamma_{ik} = f(\mathbf{M}_{ik})$, $\gamma_k = (\gamma_{1k}, \cdots, \gamma_{n_k k}) \sim \text{MVN}(\mathbf{0}, \tau\mathbf{K}_k)$, where $\mathbf{K}_k$ is the kernel similarity matrix in k-th study. Under the null hypothesis, $\tau = 0$, i.e. $\gamma_{ik} = 0$ for all i and k.

To understand the trait similarity between pairs of $Z_{ij,k} = (Y_{ik} - \hat{\mu}_{ik}^0)(Y_{jk} - \hat{\mu}_{jk}^0)$, and relate it to the corresponding microbiome similarity $K_{ij,k}$, the conditional expectation and the following four quantities of conditional variance-covariance matrix are required,

1. $\text{Var}(Z_{ij,k}|\mathbf{X}, \mathbf{M})$: the variance of trait similarity of subject pair (i, j) within the k-th study.

*Corresponding author: Ni Zhao (nzhao10@jhu.edu)

2. $\mathrm{Cov}(Z_{ij,k}, Z_{it,k}|\mathbf{X},\mathbf{M})$: the covariance of trait similarities between subject pairs (i,j) and (i,t) in the k-th study, where the pairs share the same subject i ($i \neq j \neq t$).

3. $\mathrm{Cov}(Z_{ij,k}, Z_{st,k}|\mathbf{X},\mathbf{M})$: the covariance of trait similarities between subject pairs (i,j) and (s,t) in the k-th study, where the pairs involve four different subjects ($i \neq j \neq s \neq t$).

4. $\mathrm{Cov}(Z_{ij,k}, Z_{st,k'}|\mathbf{X},\mathbf{M})$: the covariance of trait similarities between subject pairs (i,j) in k-th study and (s,t) in k'-th study ($i \neq j, s \neq t$).

As detailed in section B, we show that, for both the continuous and binary outcomes,

$$\mathbb{E}(Z_{ij,k}|\mathbf{X},\mathbf{M}) = a + bK_{ij,k}, \tag{A.1}$$

where a and b are two constants, $K_{ij,k}$ is the i,j-th element of microbiome similarity matrix $\mathbf{K}_k$.

Building on the counterpart model, the conditional variance and covariances for continuous outcomes (See details in section B.1),

$$\begin{aligned}
\mathrm{Var}(Z_{ij,k}|\mathbf{X},\mathbf{M}) =& \tau\sigma^2(K_{ii,k} + 2K_{ij,k} + K_{jj,k}) + \tau^2(K_{ii,k}K_{jj,k} + K_{ij,k}^2) \\
&+ 2\sigma^4 + \tau\delta_k^2(K_{ii,k} + K_{jj,k}) + 2\delta_k^2\sigma^2 + \delta_k^4 && \text{(A.2)} \\
\mathrm{Cov}(Z_{ij,k}, Z_{it,k}|\mathbf{X},\mathbf{M}) =& \tau\sigma^2(K_{it,k} + K_{ii,k} + K_{ji,k} + K_{jt,k}) + \tau^2(K_{ii,k}K_{jt,k} + K_{it,k}K_{ij,k}) \\
&+ 2\sigma^4 + \tau\delta_k^2 K_{jt,k} + \delta_k^2\sigma^2 && \text{(A.3)} \\
\mathrm{Cov}(Z_{ij,k}, Z_{st,k}|\mathbf{X},\mathbf{M}) =& \tau\sigma^2(K_{it,k} + K_{is,k} + K_{js,k} + K_{jt,k}) + \tau^2(K_{is,k}K_{jt,k} + K_{it,k}K_{js,k}) \\
&+ 2\sigma^4 && \text{(A.4)} \\
\mathrm{Cov}(Z_{ij,k}, Z_{st,k'}|\mathbf{X},\mathbf{M}) =& \tau^2(K_{is,kk'}K_{jt,kk'} + K_{it,kk'}K_{js,kk'}) && \text{(A.5)}
\end{aligned}$$

where δ_k^2 denotes the variance of ε_{ik}, σ^2 denotes the variance of h_k, $K_{ij,k}$ represents that microbiome similarity between subject i and subject j within the study k, and $K_{is,kk'}$ represents the microbiome similarity between the sample i and sample s when sample i is from study k and the sample s is from the study k'. Note that similarities $K_{is,kk'}, K_{jt,kk'}, K_{it,kk'}, K_{js,kk'}$ are the underlying similarities between subjects from different studies, which may not be available.

Note that, under the null hypothesis, $\tau = 0$, and the corresponding variance-covariance reduces. For example, $\mathrm{Cov}(Z_{ij,k}, Z_{st,k}|\mathbf{X},\mathbf{M}) = 0$ under the null, indicating that the trait similarity between pairs of samples in study k is uncorrelated to the trait similarity between pairs of samples in study k'.

We can combine the mean relationship (A.1) and the variance-covariance (A.2), (A.3), (A.4), (A.5) of the trait similarities to construct SMRmix. Clearly, from (A.1), an intercept is needed.

Comparing (A.4) to (A.5), if pairs of samples come from the same study, an additional covariance term arises under the null. Thus, we would need an extra random effect in the SMRmix to capture the study effect.

Comparing (A.3) to (A.4), $(Z_{ij,k},\ Z_{it,k})$ has higher covariance than $(Z_{ij,k}, Z_{st,k})$ because they share the i-th sample. Thus, we need an extra random effect to capture this additional correlation.

Based on (A.2), a random error is needed. Thus, the proposed SMRmix model can be formulated as follows,

$$Z_{ij,k} = a + bK_{ij,k} + u_k + f_i + e_{ij,k},\ i < j,$$

where a is an intercept, b is the regression coefficient, u_k is the study-specific random effect, f_i is the individual-level random effect, and $e_{ij,k}$ is the pairwise similarity random effect.

Similarly, for binary outcomes, the same SMRmix model can be justified by comparing the following conditional variance and covariances (See details in section B.2),

$$\begin{aligned}
\mathrm{Var}(Z_{ij,k}|\mathbf{X},\mathbf{M}) =& [\mu_{ik}^0(1-\mu_{ik}^0)\mu_{jk}^0(1-\mu_{jk}^0)]^2 \\
&\cdot [2\sigma^4 + \tau\sigma^2(K_{ii,k} + 2K_{ij,k} + K_{jj,k}) + \tau^2(K_{ii,k}K_{jj,k} + K_{ij,k}^2)] \\
&+ [\mu_{ik}^0(1-\mu_{ik}^0)\mu_{jk}^0(1-\mu_{jk}^0)][\mu_{ik}^0(1-\mu_{ik}^0)(\sigma^2+\tau K_{ii,k}) + \mu_{jk}^0(1-\mu_{jk}^0)(\sigma^2+\tau K_{jj,k})] \\
&+ \mu_{ik}^0(1-\mu_{ik}^0)\mu_{jk}^0(1-\mu_{jk}^0)[(1-2\mu_{ik}^0)(1-2\mu_{jk}^0)(\sigma^2+\tau K_{ij,k}) + 1]
\end{aligned} \tag{A.6}$$

$$\begin{aligned}
\mathrm{Cov}(Z_{ij,k}, Z_{it,k}|\mathbf{X},\mathbf{M}) =& [\mu_{ik}^0(1-\mu_{ik}^0)]^2\mu_{jk}^0(1-\mu_{jk}^0)\mu_{tk}^0(1-\mu_{tk}^0) \\
&\cdot [2\sigma^4 + \tau\sigma^2(K_{it,k} + K_{ii,k} + K_{ji,k} + K_{jt,k}) + \tau^2(K_{ii,k}K_{jt,k} + K_{it,k}K_{ij,k})] \\
&+ \mu_{ik}^0(1-\mu_{ik}^0)\mu_{jk}^0(1-\mu_{jk}^0)\mu_{tk}^0(1-\mu_{tk}^0)(\sigma^2+\tau K_{jt,k})
\end{aligned} \tag{A.7}$$

$$\begin{aligned}
\mathrm{Cov}(Z_{ij,k}, Z_{st,k}|\mathbf{X},\mathbf{M}) =& \mu_{ik}^0(1-\mu_{ik}^0)\mu_{jk}^0(1-\mu_{jk}^0)\mu_{sk}^0(1-\mu_{sk}^0)\mu_{tk}^0(1-\mu_{tk}^0) \\
&\cdot [2\sigma^4 + \tau\sigma^2(K_{it,k} + K_{is,k} + K_{js,k} + K_{jt,k}) + \tau^2(K_{is,k}K_{jt,k} + K_{it,k}K_{js,k})]
\end{aligned} \tag{A.8}$$

$$\begin{aligned}
\mathrm{Cov}(Z_{ij,k}, Z_{st,k'}|\mathbf{X},\mathbf{M}) =& \mu_{ik}^0(1-\mu_{ik}^0)\mu_{jk}^0(1-\mu_{jk}^0)\mu_{sk'}^0(1-\mu_{sk'}^0)\mu_{tk'}^0(1-\mu_{tk'}^0) \\
&\cdot [\tau^2(K_{is,kk'}K_{jt,kk'} + K_{it,kk'}K_{js,kk'})]
\end{aligned} \tag{A.9}$$

B Conditional Variance-covariance Matrix of Z

In this section, we present the expressions of variance-covariance matrix $\mathbf{V} = \mathrm{Var}(\mathbf{Z}|\mathbf{X},\mathbf{M})$ for both continuous case and binary case.

B.1 Continuous outcomes

For continuous outcomes, we assume a linear mixed-effect model,

$$Y_{ik} = \mathbf{X}_{ik}^T\beta + f(\mathbf{M}_{ik}) + h_k + \varepsilon_{ik},$$

where $h_k \sim N(0,\sigma^2)$, $\varepsilon_{ik} \sim N(0,\delta_k^2)$, and $f(\mathbf{M}_{ik})$ depicts the microbiome effect on the outcome. Let $\gamma_{ik} = f(\mathbf{M}_{ik})$, $\gamma_k = (\gamma_{1k},\cdots,\gamma_{n_k k}) \sim \mathrm{MVN}(\mathbf{0},\tau\mathbf{K}_k)$.

In similarity matrix regression mixed effect model (SMRmix), the trait similarity between i-th subject and j-th subject in k-th study is

$$Z_{ij,k} = (Y_{ik} - \hat{\mu}_{ik}^0)(Y_{jk} - \hat{\mu}_{jk}^0),$$

where $\hat{\mu}_{ik}^0$ is the fitted values of Y_{ik} under the null hypothesis. Let $\gamma_{ik} = f(\mathbf{M}_{ik})$. The conditional expectation of $Z_{ij,k}$ $\mathbb{E}(Z_{ij,k}|\mathbf{X},\mathbf{M})$ is

$$\begin{aligned}
&\mathbb{E}(Z_{ij,k}|\mathbf{X},\mathbf{M}) = \mathbb{E}[(Y_{ik} - \mu_{ik}^0)(Y_{jk} - \mu_{jk}^0)|\mathbf{X},\mathbf{M}]\\
=&\mathbb{E}[(\gamma_{ik} + h_k + \varepsilon_{ik})(\gamma_{jk} + h_k + \varepsilon_{jk})|\mathbf{X},\mathbf{M}]\\
=&\sigma^2 + \tau K_{ij,k}.
\end{aligned}$$

By the law of total variance, the variance $\mathrm{Var}(Z_{ij,k}|\mathbf{X},\mathbf{M})$ is

$$\mathrm{Var}(Z_{ij,k}|\mathbf{X},\mathbf{M}) = \mathrm{Var}[\mathbb{E}(Z_{ij,k}|\mathbf{X},\mathbf{M},\gamma)|\mathbf{X},\mathbf{M}] + \mathbb{E}[\mathrm{Var}(Z_{ij,k}|\mathbf{X},\mathbf{M},\gamma)|\mathbf{X},\mathbf{M}],$$

where

$$\begin{aligned}
&\mathrm{Var}(Z_{ij,k}|\mathbf{X},\mathbf{M},\gamma) = \mathrm{Var}((\gamma_{ik} + h_k + \varepsilon_{ik})(\gamma_{jk} + h_k + \varepsilon_{jk})|\mathbf{X},\mathbf{M},\gamma)\\
=&\mathrm{Var}((\gamma_{ik} + \gamma_{jk} + \varepsilon_{ik} + \varepsilon_{jk})h_k + \gamma_{ik}\varepsilon_{jk} + \gamma_{jk}\varepsilon_{ik} + h_k^2 + \varepsilon_{ik}\varepsilon_{jk}|\mathbf{X},\mathbf{M},\gamma)\\
=&(\gamma_{ik} + \gamma_{jk})^2\sigma^2 + (\gamma_{ik}^2 + \gamma_{jk}^2)\delta_k^2 + 2\sigma^4 + 2\delta_k^2\sigma^2 + \delta_k^4,
\end{aligned}$$

and

$$\mathbb{E}(Z_{ij,k}|\mathbf{X},\mathbf{M},\gamma) = \gamma_{ik}\gamma_{jk} + \sigma^2.$$

Therefore,

$$\begin{aligned}
\mathrm{Var}(Z_{ij,k}|\mathbf{X},\mathbf{M}) =&\tau\sigma^2(K_{ii,k} + 2K_{ij,k} + K_{jj,k}) + \tau^2(K_{ii,k}K_{jj,k} + K_{ij,k}^2)\\
&+ 2\sigma^4 + \tau\delta_k^2(K_{ii,k} + K_{jj,k}) + 2\delta_k^2\sigma^2 + \delta_k^4
\end{aligned}$$

By the law of total covariance, the conditional covariance $\mathrm{Cov}(Z_{ij,k}, Z_{it,k}|\mathbf{X},\mathbf{M})$ is

$$\begin{aligned}
&\mathrm{Cov}(Z_{ij,k}, Z_{it,k}|\mathbf{X},\mathbf{M})\\
=&\mathbb{E}[\mathrm{Cov}(Z_{ij,k}, Z_{it,k}|\mathbf{X},\mathbf{M},\gamma)|\mathbf{X},\mathbf{M}] + \mathrm{Cov}[\mathbb{E}(Z_{ij,k}|\mathbf{X},\mathbf{M},\gamma), \mathbb{E}(Z_{it,k}|\mathbf{X},\mathbf{M},\gamma)|\mathbf{X},\mathbf{M}]
\end{aligned}$$

where

$$\begin{aligned}&\mathbb{E}[\mathrm{Cov}(Z_{ij,k},Z_{it,k}|\mathbf{X},\mathbf{M},\gamma)|\mathbf{X},\mathbf{M}]=\mathbb{E}[(\gamma_{ik}+\gamma_{jk})(\gamma_{ik}+\gamma_{tk})\sigma^2+\gamma_{jk}\gamma_{tk}+\delta_k^2\sigma^2+2\sigma^4|\mathbf{X},\mathbf{M},\gamma]\\&=\tau\sigma^2(K_{it,k}+K_{ii,k}+K_{ji,k}+K_{jt,k})+\tau\delta_k^2K_{jt,k}+\delta_k^2\sigma^2+2\sigma^4,\end{aligned}$$

and

$$\begin{aligned}&\mathrm{Cov}[\mathbb{E}(Z_{ij,k}|\mathbf{X},\mathbf{M},\gamma),\mathbb{E}(Z_{it,k}|\mathbf{X},\mathbf{M},\gamma)|\mathbf{X},\mathbf{M}]=\mathrm{Cov}(\gamma_{ik}\gamma_{jk}+\sigma^2,\gamma_{ik}\gamma_{tk}+\sigma^2)\\&=\tau^2(K_{ii,k}K_{jt,k}+K_{it,k}K_{ij,k})\end{aligned}$$

Therefore,

$$\begin{aligned}\mathrm{Cov}(Z_{ij,k},Z_{it,k}|\mathbf{X},\mathbf{M})=&\tau\sigma^2(K_{it,k}+K_{ii,k}+K_{ji,k}+K_{jt,k})+\tau^2(K_{ii,k}K_{jt,k}+K_{it,k}K_{ij,k})\\&+2\sigma^4+\tau\delta_k^2K_{jt,k}+\delta_k^2\sigma^2\end{aligned}$$

By the law of total covariance, the conditional covariance $\mathrm{Cov}(Z_{ij,k},Z_{st,k}|\mathbf{X},\mathbf{M})$ is

$$\begin{aligned}&\mathrm{Cov}(Z_{ij,k},Z_{it,k}|\mathbf{X},\mathbf{M})\\&=\mathbb{E}[\mathrm{Cov}(Z_{ij,k},Z_{st,k}|\mathbf{X},\mathbf{M},\gamma)|\mathbf{X},\mathbf{M}]+\mathrm{Cov}[\mathbb{E}(Z_{ij,k}|\mathbf{X},\mathbf{M},\gamma),\mathbb{E}(Z_{st,k}|\mathbf{X},\mathbf{M},\gamma)|\mathbf{X},\mathbf{M}]\end{aligned}$$

where

$$\begin{aligned}&\mathbb{E}[\mathrm{Cov}(Z_{ij,k},Z_{st,k}|\mathbf{X},\mathbf{M},\gamma)|\mathbf{X},\mathbf{M}]=\mathbb{E}[(\gamma_{ik}+\gamma_{jk})(\gamma_{sk}+\gamma_{tk})\sigma^2+2\sigma^4|\mathbf{X},\mathbf{M}]\\&=\tau\sigma^2(K_{it,k}+K_{is,k}+K_{js,k}+K_{jt,k})+2\sigma^4,\end{aligned}$$

and

$$\begin{aligned}&\mathrm{Cov}[\mathbb{E}(Z_{ij,k}|\mathbf{X},\mathbf{M},\gamma),\mathbb{E}(Z_{st,k}|\mathbf{X},\mathbf{M},\gamma)|\mathbf{X},\mathbf{M}]=\mathrm{Cov}(\gamma_{ik}\gamma_{jk}+\sigma^2,\gamma_{sk}\gamma_{tk}+\sigma^2|\mathbf{X},\mathbf{M})\\&=\tau^2(K_{is,k}K_{jt,k}+K_{it,k}K_{js,k})\end{aligned}$$

Therefore,

$$\mathrm{Cov}(Z_{ij,k},Z_{st,k}|\mathbf{X},\mathbf{M})=\tau\sigma^2(K_{it,k}+K_{is,k}+K_{js,k}+K_{jt,k})+\tau^2(K_{is,k}K_{jt,k}+K_{it,k}K_{js,k})+2\sigma^4$$

By the law of total covariance, the conditional covariance $\mathrm{Cov}(Z_{ij,k},Z_{st,k'}|\mathbf{X},\mathbf{M})$ is

$$\begin{aligned}&\mathrm{Cov}(Z_{ij,k},Z_{st,k'}|\mathbf{X},\mathbf{M})\\&=\mathbb{E}[\mathrm{Cov}(Z_{ij,k},Z_{st,k'}|\mathbf{X},\mathbf{M},\gamma)|\mathbf{X},\mathbf{M}]+\mathrm{Cov}[\mathbb{E}(Z_{ij,k}|\mathbf{X},\mathbf{M},\gamma),\mathbb{E}(Z_{st,k'}|\mathbf{X},\mathbf{M},\gamma)|\mathbf{X},\mathbf{M}]\\&=\tau^2(K_{is,kk'}K_{jt,kk'}+K_{it,kk'}K_{js,kk'})\end{aligned}$$

where we assume $K_{is,kk'}$ is a known value of similarity between i-th subject from k-th study and s-th subject from k'-th study.

B.2 Binary outcomes

For binary outcomes, the mixed-effect logistic model is

$$g(\mathbb{P}(Y_{ik}=1)) = \mathbf{X}_{ik}^T\beta + f(\mathbf{M}_{ik}) + h_k,$$

where $g(\cdot)$ is the logit function, $h_k \sim N(0,\sigma^2)$, and $f(\mathbf{M}_{ik})$ depicts the microbiome effect on the outcome. Let $\gamma_{ik} = f(\mathbf{M}_{ik})$, $\gamma_k = (\gamma_{1k},\cdots,\gamma_{n_k k}) \sim \mathrm{MVN}(\mathbf{0},\tau\mathbf{K}_k)$.

Let $\mu_{ik}(\gamma_{ik},h_k) = \mathbb{E}(Y_{ik}|\mathbf{X},\mathbf{M},\gamma,\mathbf{h}) = g^{-1}(\mathbf{X}_{ik}^T\beta+\gamma_{ik}+h_k)$, and $\mu_{ik}^0 = \mu_{ik}(0,0) = g^{-1}(\mathbf{X}_{ik}^T\beta)$. By Taylor expansion of $\mu_{ik}(\gamma_{ik},h_k)$ around $(0,0)$, the conditional expectation of $Z_{ij,k}$ is,

$$\begin{aligned}
&\mathbb{E}(Z_{ij,k}|\mathbf{X},\mathbf{M}) = \mathbb{E}((Y_{ik}-\mu_{ik}^0)(Y_{jk}-\mu_{jk}^0)|\mathbf{X},\mathbf{M})\\
=&\mathbb{E}\{\mathbb{E}[(Y_{ik}-\mu_{ik}(\gamma_{ik},h_k)+\mu_{ik}(\gamma_{ik},h_k)-\mu_{ik}^0)(Y_{jk}-\mu_{jk}(\gamma_{jk},h_k)+\mu_{jk}(\gamma_{jk},h_k)-\mu_{jk}^0)|\mathbf{X},\mathbf{M},\gamma,\mathbf{h}]|\mathbf{X},\mathbf{M}\}\\
=&\mathbb{E}[(\mu_{ik}(\gamma_{ik},h_k)-\mu_{ik}^0)(\mu_{jk}(\gamma_{jk},h_k)-\mu_{jk}^0)|\mathbf{X},\mathbf{M}]\\
\approx&\mathbb{E}[(g^{-1})'(\mathbf{X}_{ik}^T\beta)(\gamma_{ik}+h_k)(g^{-1})'(\mathbf{X}_{jk}^T\beta)(\gamma_{jk}+h_k)|\mathbf{X},\mathbf{M}]\\
=&\mu_{ik}^0(1-\mu_{ik}^0)\mu_{jk}^0(1-\mu_{jk}^0)(\sigma^2+\tau K_{ij,k})
\end{aligned}$$

By the law of total variance, the variance $\mathrm{Var}(Z_{ij,k}|\mathbf{X},\mathbf{M})$ is

$$\mathrm{Var}(Z_{ij,k}|\mathbf{X},\mathbf{M}) = \mathrm{Var}[\mathbb{E}(Z_{ij,k}|\mathbf{X},\mathbf{M},\gamma,\mathbf{h})|\mathbf{X},\mathbf{M}] + \mathbb{E}[\mathrm{Var}(Z_{ij,k}|\mathbf{X},\mathbf{M},\gamma,\mathbf{h})|\mathbf{X},\mathbf{M}],$$

where

$$\begin{aligned}
&\mathrm{Var}[\mathbb{E}(Z_{ij,k}|\mathbf{X},\mathbf{M},\gamma,\mathbf{h})|\mathbf{X},\mathbf{M}] \approx \mathrm{Var}[(g^{-1})'(\mathbf{X}_{ik}^T\beta)(\gamma_{ik}+h_k)(g^{-1})'(\mathbf{X}_{jk}^T\beta)(\gamma_{jk}+h_k)|\mathbf{X},\mathbf{M}]\\
&= [\mu_{ik}^0(1-\mu_{ik}^0)\mu_{jk}^0(1-\mu_{jk}^0)]^2[2\sigma^4+\tau\sigma^2(K_{ii,k}+2K_{ij,k}+K_{jj,k})+\tau^2(K_{ii,k}K_{jj,k}+K_{ij,k}^2)], 2qw
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}[\mathrm{Var}(Z_{ij,k}|\mathbf{X},\mathbf{M},\gamma,\mathbf{h})|\mathbf{X},\mathbf{M}]\\
=&\mathbb{E}\{\mathrm{Var}[(Y_{ik}-\mu_{ik}(\gamma_{ik},h_k)+\mu_{ik}(\gamma_{ik},h_k)-\mu_{ik}^0)(Y_{jk}-\mu_{jk}(\gamma_{jk},h_k)+\mu_{jk}(\gamma_{jk},h_k)-\mu_{jk}^0)|\mathbf{X},\mathbf{M},\gamma,\mathbf{h}]|\mathbf{X},\mathbf{M}\}\\
\approx&\mathbb{E}\{\mathrm{Var}[(Y_{ik}-\mu_{ik}(\gamma_{ik},h_k))(Y_{jk}-\mu_{jk}(\gamma_{jk},h_k))|\mathbf{X},\mathbf{M},\gamma,\mathbf{h}]\\
&+(\mu_{jk}(\gamma_{jk},h_k)-\mu_{jk}^0)(Y_{ik}-\mu_{ik}(\gamma_{ik},h_k))+(\mu_{ik}(\gamma_{ik},h_k)-\mu_{ik}^0)(Y_{jk}-\mu_{jk}(\gamma_{jk},h_k))|\mathbf{X},\mathbf{M}\}\\
=&\mathbb{E}[\mu_{ik}(\gamma_{ik},h_k)(1-\mu_{ik}(\gamma_{ik},h_k))\mu_{jk}(\gamma_{jk},h_k)(1-\mu_{jk}(\gamma_{jk},h_k))\\
&+(\mu_{ik}(\gamma_{ik},h_k)-\mu_{ik}^0)^2\mu_{jk}(\gamma_{jk},h_k)(1-\mu_{jk}(\gamma_{jk},h_k))\\
&+(\mu_{jk}(\gamma_{jk},h_k)-\mu_{jk}^0)^2\mu_{ik}(\gamma_{ik},h_k)(1-\mu_{ik}(\gamma_{ik},h_k))|\mathbf{X},\mathbf{M}]
\end{aligned}$$

$$
\begin{aligned}
\approx&\mathbb{E}\{(1-2\mu_{ik}^0)(g^{-1})'(\mathbf{X}_{ik}^T\beta)(\gamma_{ik}+h_k)(1-2\mu_{jk}^0)(g^{-1})'(\mathbf{X}_{jk}^T\beta)(\gamma_{jk}+h_k)|\mathbf{X},\mathbf{M}\}+\mu_{ik}^0(1-\mu_{ik}^0)\mu_{jk}^0(1-\mu_{jk}^0)\\
&+\mu_{jk}^0(1-\mu_{jk}^0)[(g^{-1})'(\mathbf{X}_{ik}^T\beta)]^2(\sigma^2+\tau K_{ii,k})+\mu_{ik}^0(1-\mu_{ik}^0)[(g^{-1})'(\mathbf{X}_{jk}^T\beta)]^2(\sigma^2+\tau K_{jj,k})\\
=&\mu_{ik}^0(1-\mu_{ik}^0)\mu_{jk}^0(1-\mu_{jk}^0)[(1-2\mu_{ik}^0)(1-2\mu_{jk}^0)(\sigma^2+\tau K_{ij,k})+1]\\
&+[\mu_{ik}^0(1-\mu_{ik}^0)\mu_{jk}^0(1-\mu_{jk}^0)][\mu_{ik}^0(1-\mu_{ik}^0)(\sigma^2+\tau K_{ii,k})+\mu_{jk}^0(1-\mu_{jk}^0)(\sigma^2+\tau K_{jj,k})].
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\mathrm{Var}(Z_{ij,k}|\mathbf{X},\mathbf{M})=&[\mu_{ik}^0(1-\mu_{ik}^0)\mu_{jk}^0(1-\mu_{jk}^0)]^2[2\sigma^4+\tau\sigma^2(K_{ii,k}+2K_{ij,k}+K_{jj,k})+\tau^2(K_{ii,k}K_{jj,k}+K_{ij,k}^2)]\\
&+\mu_{ik}^0(1-\mu_{ik}^0)\mu_{jk}^0(1-\mu_{jk}^0)[(1-2\mu_{ik}^0)(1-2\mu_{jk}^0)(\sigma^2+\tau K_{ij,k})+1]\\
&+[\mu_{ik}^0(1-\mu_{ik}^0)\mu_{jk}^0(1-\mu_{jk}^0)][\mu_{ik}^0(1-\mu_{ik}^0)(\sigma^2+\tau K_{ii,k})+\mu_{jk}^0(1-\mu_{jk}^0)(\sigma^2+\tau K_{jj,k})].
\end{aligned}
$$

By the law of total covariance, the conditional covariance $\mathrm{Cov}(Z_{ij,k},Z_{it,k}|\mathbf{X},\mathbf{M})$ is

$$
\begin{aligned}
&\mathrm{Cov}(Z_{ij,k},Z_{it,k}|\mathbf{X},\mathbf{M})\\
=&\mathbb{E}[\mathrm{Cov}(Z_{ij,k},Z_{it,k}|\mathbf{X},\mathbf{M},\gamma,\mathbf{h})|\mathbf{X},\mathbf{M}]+\mathrm{Cov}[\mathbb{E}(Z_{ij,k}|\mathbf{X},\mathbf{M},\gamma,\mathbf{h}),\mathbb{E}(Z_{it,k}|\mathbf{X},\mathbf{M},\gamma,\mathbf{h})|\mathbf{X},\mathbf{M}]
\end{aligned}
$$

where

$$
\begin{aligned}
&\mathbb{E}[\mathrm{Cov}(Z_{ij,k},Z_{it,k}|\mathbf{X},\mathbf{M},\gamma,\mathbf{h})|\mathbf{X},\mathbf{M}]\\
=&\mathbb{E}[\mathrm{Cov}((Y_{ik}-\mu_{ik}(\gamma_{ik},h_k)+\mu_{ik}(\gamma_{ik},h_k)-\mu_{ik}^0)(Y_{jk}-\mu_{jk}(\gamma_{jk},h_k)+\mu_{jk}(\gamma_{jk},h_k)-\mu_{jk}^0),\\
&\qquad(Y_{ik}-\mu_{ik}(\gamma_{ik},h_k)+\mu_{ik}(\gamma_{ik},h_k)-\mu_{ik}^0)(Y_{tk}-\mu_{tk}(\gamma_{tk},h_k)+\mu_{tk}(\gamma_{tk},h_k)-\mu_{tk}^0)|\mathbf{X},\mathbf{M},\gamma,\mathbf{h})|\mathbf{X},\mathbf{M}]\\
=&\mathbb{E}[(\mu_{jk}-\mu_{jk}^0)(\mu_{tk}-\mu_{tk}^0)\mu_{ik}(\gamma_{ik},h_k)(1-\mu_{ik}(\gamma_{ik},h_k))|\mathbf{X},\mathbf{M}]\\
\approx&(g^{-1})'(\mu_{jk}^0)(g^{-1})'(\mu_{tk}^0)\mathbb{E}[(\gamma_{jk}+h_k)(\gamma_{tk}+h_k)(\mu_{ik}^0(1-\mu_{ik}^0)+(1-2\mu_{ik}^0)(g^{-1})'(\mu_{ik}^0)(\gamma_{ik}+h_k))]\\
=&\mu_{ik}^0(1-\mu_{ik}^0)\mu_{jk}^0(1-\mu_{jk}^0)\mu_{tk}^0(1-\mu_{tk}^0)(\sigma^2+\tau K_{jt,k}),
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathrm{Cov}[\mathbb{E}(Z_{ij,k}|\mathbf{X},\mathbf{M},\gamma,\mathbf{h}),\mathbb{E}(Z_{it,k}|\mathbf{X},\mathbf{M},\gamma,\mathbf{h})|\mathbf{X},\mathbf{M}]\\
\approx&\mathrm{Cov}((g^{-1})'(\mathbf{X}_{ik}^T\beta)(g^{-1})'(\mathbf{X}_{jk}^T\beta)(\gamma_{ik}+h_k)(\gamma_{jk}+h_k),(g^{-1})'(\mathbf{X}_{ik}^T\beta)(g^{-1})'(\mathbf{X}_{tk}^T\beta)(\gamma_{ik}+h_k)(\gamma_{tk}+h_k)|\mathbf{X},\mathbf{M})\\
=&[\mu_{ik}^0(1-\mu_{ik}^0)]^2\mu_{jk}^0(1-\mu_{jk}^0)\mu_{tk}^0(1-\mu_{tk}^0)[\tau^2(K_{ii,k}K_{jt,k}+K_{it,k}K_{ij,k})+\tau\sigma^2(K_{it,k}+K_{ii,k}+K_{ji,k}+K_{jt,k})+2\sigma^4]
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\mathrm{Cov}(Z_{ij,k},Z_{it,k}|\mathbf{X},\mathbf{M})=\mu_{ik}^0(1-\mu_{ik}^0)\mu_{jk}^0(1-\mu_{jk}^0)\mu_{tk}^0(1-\mu_{tk}^0)(\sigma^2+\tau K_{jt,k})\\
&+[\mu_{ik}^0(1-\mu_{ik}^0)]^2\mu_{jk}^0(1-\mu_{jk}^0)\mu_{tk}^0(1-\mu_{tk}^0)[\tau^2(K_{ii,k}K_{jt,k}+K_{it,k}K_{ij,k})+\tau\sigma^2(K_{it,k}+K_{ii,k}+K_{ji,k}+K_{jt,k})+2\sigma^4]
\end{aligned}
$$

By the law of total covariance and conditional independence, $\mathrm{Cov}(Z_{ij,k},Z_{st,k}|\mathbf{X},\mathbf{M})$ is

$$
\begin{aligned}
&\mathrm{Cov}(Z_{ij,k},Z_{st,k}|\mathbf{X},\mathbf{M})=\mathrm{Cov}[\mathbb{E}(Z_{ij,k}|\mathbf{X},\mathbf{M},\gamma,\mathbf{h}),\mathbb{E}(Z_{st,k}|\mathbf{X},\mathbf{M},\gamma,\mathbf{h})|\mathbf{X},\mathbf{M}]\\
\approx&\mathrm{Cov}((g^{-1})'(\mathbf{X}_{ik}^T\beta)(g^{-1})'(\mathbf{X}_{jk}^T\beta)(\gamma_{ik}+h_k)(\gamma_{jk}+h_k),(g^{-1})'(\mathbf{X}_{sk}^T\beta)(g^{-1})'(\mathbf{X}_{tk}^T\beta)(\gamma_{sk}+h_k)(\gamma_{tk}+h_k)|\mathbf{X},\mathbf{M})\\
=&\mu_{ik}^0(1-\mu_{ik}^0)\mu_{jk}^0(1-\mu_{jk}^0)\mu_{sk}^0(1-\mu_{sk}^0)\mu_{tk}^0(1-\mu_{tk}^0)\\
&\cdot[2\sigma^4+\tau\sigma^2(K_{it,k}+K_{is,k}+K_{js,k}+K_{jt,k})+\tau^2(K_{is,k}K_{jt,k}+K_{it,k}K_{js,k})]
\end{aligned}
$$

Similarly, by the law of total covariance and conditional independence, $\mathrm{Cov}(Z_{ij,k}, Z_{st,k'}|\mathbf{X},\mathbf{M})$ is

$$\begin{aligned}
&\mathrm{Cov}(Z_{ij,k}, Z_{st,k'}|\mathbf{X},\mathbf{M}) = \mathrm{Cov}[\mathbb{E}(Z_{ij,k}|\mathbf{X},\mathbf{M},\gamma,\mathbf{h}), \mathbb{E}(Z_{st,k'}|\mathbf{X},\mathbf{M},\gamma,\mathbf{h})|\mathbf{X},\mathbf{M}] \\
\approx &\mathrm{Cov}((g^{-1})'(\mathbf{X}_{ik}^T\beta)(g^{-1})'(\mathbf{X}_{jk}^T\beta)(\gamma_{ik}+h_k)(\gamma_{jk}+h_k), (g^{-1})'(\mu_{sk'}^0)(g^{-1})'(\mu_{tk'}^0)(\gamma_{sk'}+h_{k'})(\gamma_{tk'}+h_{k'})|\mathbf{X},\mathbf{M}) \\
= &\mu_{ik}^0(1-\mu_{ik}^0)\mu_{jk}^0(1-\mu_{jk}^0)\mu_{sk'}^0(1-\mu_{sk'}^0)\mu_{tk'}^0(1-\mu_{tk'}^0)[\tau^2(K_{is,kk'}K_{jt,kk'}+K_{it,kk'}K_{js,kk'})]
\end{aligned}$$

Supplementary Figures and Tables

Table S1: Empirical Type I error Rates Comparing CSKAT, GLMM-MiRKAT and SMRmix in Simulation A when the data are simulated with bias at type I error threshold of 0.05 and 0.01. Note that we do not have a biased version of simulation B. For continuous outcome with the same δ_k^2, $\delta_k^2 = 1, \forall k$; for continuous outcome with different δ_k^2, $\delta_k^2 = 1$ for $k = 1,2,3$, $\delta_k^2 = 3$ for $k = 4,5,6,7$, $\delta_k^2 = 5$ for $k = 8,9,10$. For continuous scenario, $\sigma^2 = 1$.

With Bias		$\alpha = \mathbf{0.05}$				$\alpha = \mathbf{0.01}$			
Continuous		K_w	K_u	K_{bc}	Omnibus	K_w	K_u	K_{bc}	Omnibus
A: same δ_k^2	CSKAT	0.0494	0.0477	0.0498	/	0.0082	0.0094	0.0087	/
	SMRmix	0.0526	0.0512	0.0504	0.0522	0.0106	0.0099	0.0113	0.0111
: different δ_k^2	CSKAT	0.0552	0.0493	0.0528	/	0.0112	0.0094	0.0098	/
	SMRmix	0.0513	0.0506	0.0514	0.0513	0.0104	0.0105	0.0107	0.0112
Binary		K_w	K_u	K_{bc}	Omnibus	K_w	K_u	K_{bc}	Omnibus
A	GLMM-MiRKAT	0.0474	0.0511	0.0485	/	0.0075	0.0107	0.0091	/
	SMRmix	0.0403	0.0387	0.0397	0.0388	0.0047	0.0056	0.0047	0.0051

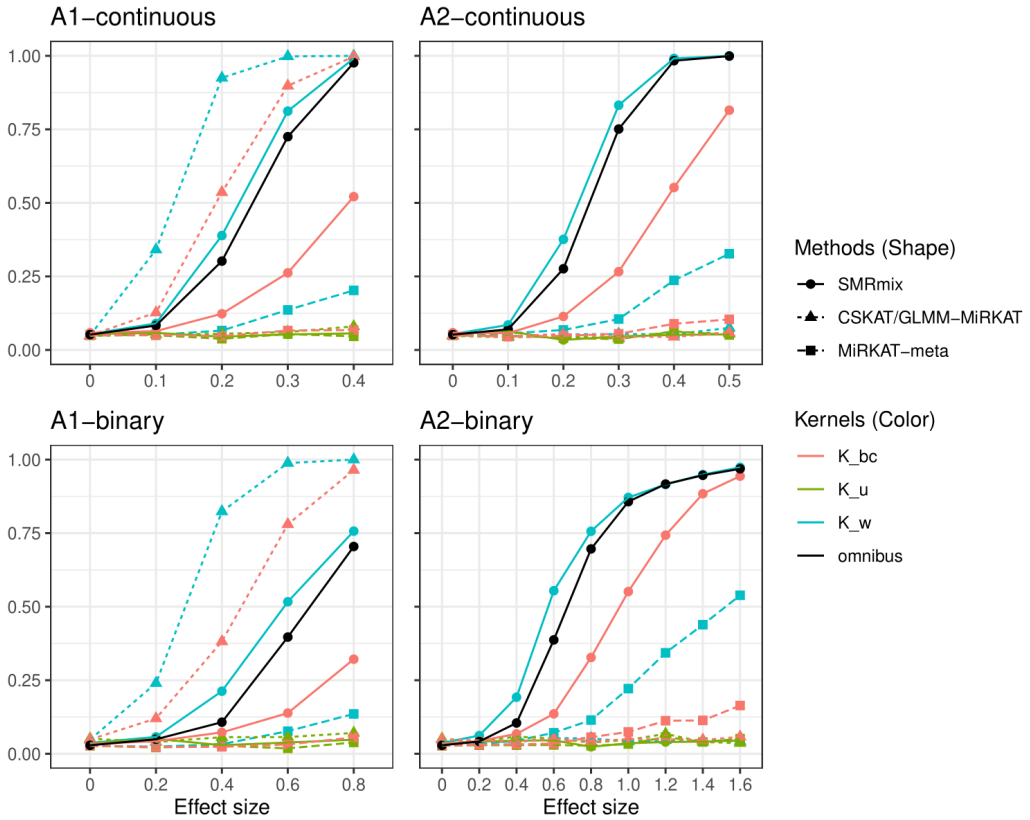

Figure S1: Type I Error and Power comparing CSKAT/GLMM-MiRKAT and SMRmix based on different kernels when differential biases exist. Continuous outcome, $\delta_k^2 = 1$, $\forall k$, and $\sigma^2 = 1$.

References

TZENG, JUNG-YING, ZHANG, DAOWEN, CHANG, SHENG-MAO, THOMAS, DUNCAN C. AND DAVIDIAN, MARIE. (2009, September). Gene-Trait Similarity Regression for Multimarker-Based Association Analysis. *Biometrics* **65**(3), 822–832.

ZHAN, XIANG. (2019, February). Relationship Between MiRKAT and Coefficient of Determination in Similarity Matrix Regression. *Processes* **7**(2), 79.

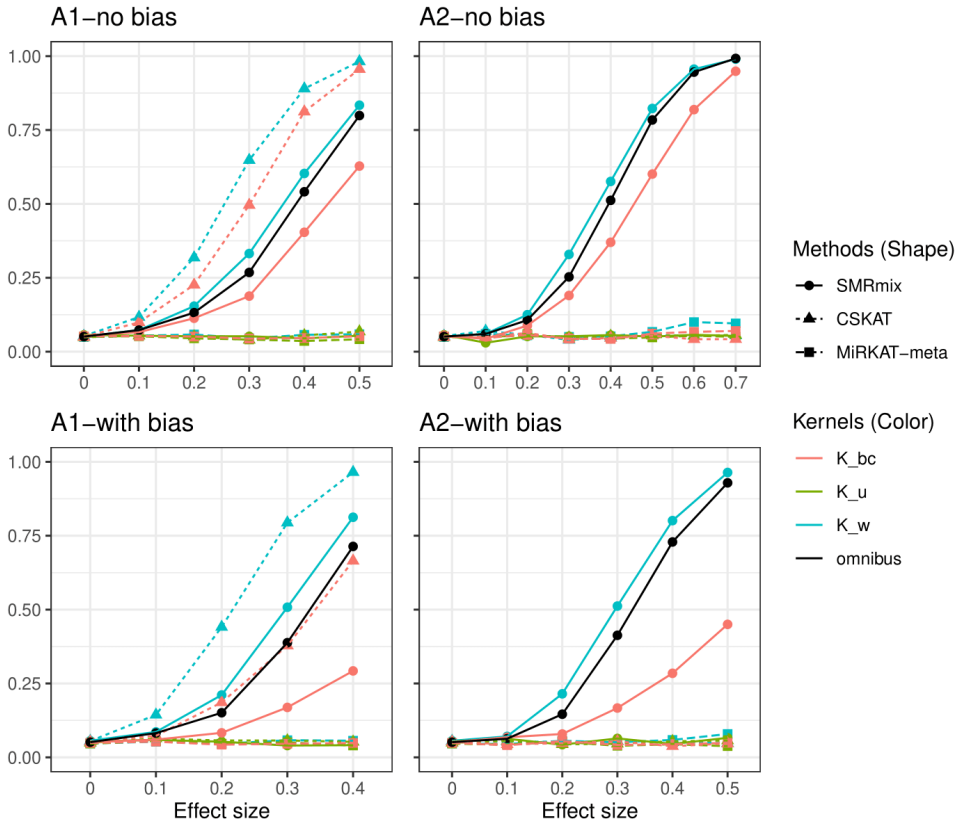

Figure S2: Type I Error and Power comparing CSKAT and SMRmix based on different kernels when the outcome is continuous with $\sigma^2 = 1$ and studies have different δ_k^2. $\delta_k^2 = 1$, for $k = 1,2,3$, $\delta_k^2 = 3$, for $k = 4,5,6,7$, $\delta_k^2 = 5$, for $k = 8,9,10$.

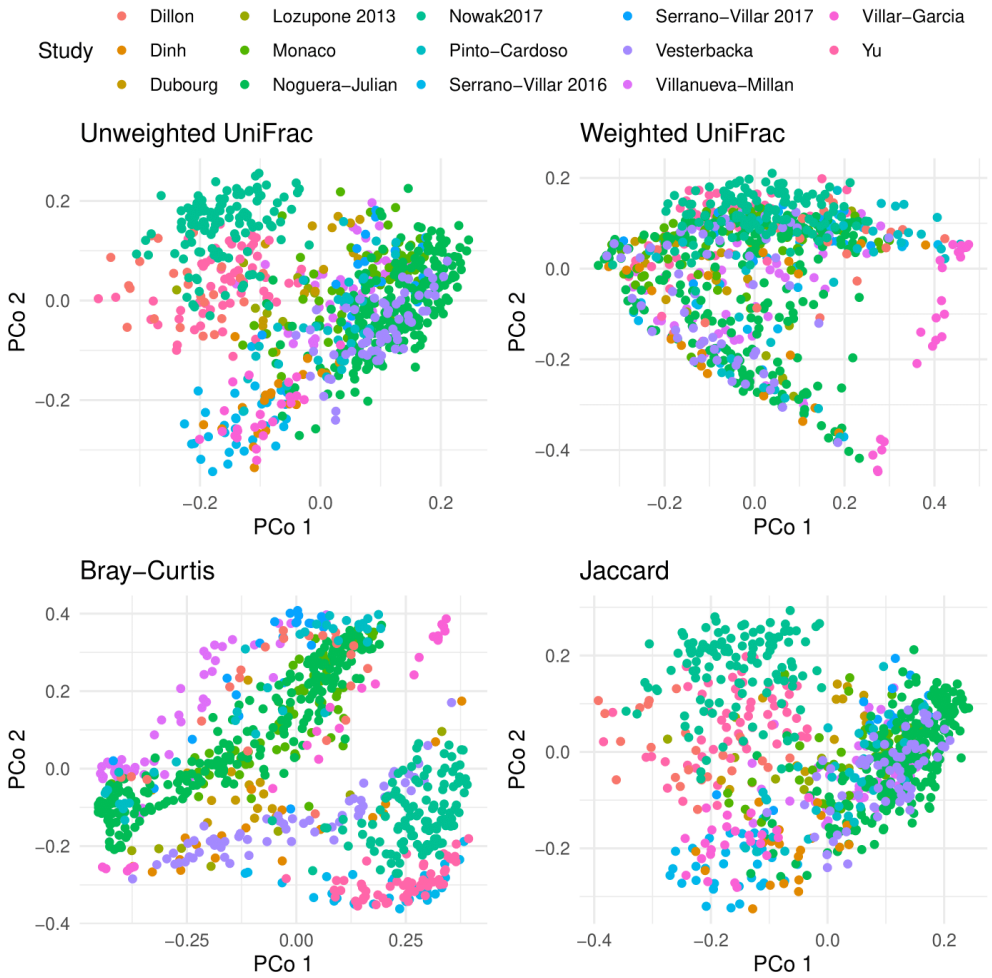

Figure S3: Visualization of Beta-diversities for 14 HIV Datasets