Testing hypotheses about the microbiome using the linear decomposition model

Yi-Juan Hu[1*] **and Glen A. Satten**[2]

[1]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA, 30322, USA;
email: yijuan.hu@emory.edu

[2]Centers for Disease Control and Prevention, Atlanta, GA, 30333, USA;
email: gas0@cdc.gov

* Corresponding author

Keywords: 16S rRNA data, clustered data, differential abundance, microbiome association test, permutation test, multivariate regression, PERMANOVA, redundancy analysis, tests of individual taxa.

Abstract

Background: Distance-based methods for analyzing microbiome data are typically restricted to testing the global hypothesis of microbiome effect, but do not test the contribution of individual operational taxonomic units (OTUs). Conversely, tests for individual OTUs do not typically provide a global test of microbiome effect. Without a unified approach, the findings of a global test may be hard to resolve with the findings at the individual OTU level. In addition, many existing methods cannot be applied to complex studies such as those with confounders and correlated data.

Methods: We introduce the linear decomposition model (LDM), that provides a single analysis path that includes global tests of any effect of the microbiome, tests of the effects of individual OTUs while accounting for multiple testing by controlling the false discovery rate (FDR), and a connection to distance-based ordination. The LDM accommodates both continuous and discrete variables (e.g., clinical outcomes, environmental factors) as well as interaction terms to be tested either singly or in combination, allows for adjustment of confounding covariates, and uses permutation-based p-values that can control for correlation (e.g., repeated measurements on the same individual). The LDM can also be applied to transformed data, and an "omnibus" test can easily combine results from analyses conducted on different transformation scales. We also provide a new implementation of PERMANOVA based on our approach.

Results: For global testing, our simulations indicate the LDM provided correct type I error, even with substantial confounding and/or correlations, and can has comparable power to existing distance-based methods. For testing individual OTUs, our simulations indicate the LDM controlled the FDR well. In contrast, DESeq2 often had inflated FDR; MetagenomeSeq generally had the lowest sensitivity. The flexibility of the LDM for a variety of microbiome studies is illustrated by the analysis of data from two microbiome studies. We also show that our implementation of PERMANOVA can outperform existing implementations.

Conclusions: The LDM is a powerful method for global and OTU-specific testing with a natural connection between the two. The LDM is also capable of handling the confounders and correlated data that frequently occur in modern microbiome studies.

Background

Data from studies of the microbiome is accumulating at a rapid rate. The relative ease of conducting a census of bacteria by sequencing the 16S rRNA gene (or, for fungi, the 18S rRNA gene) has led to many studies that examine the association between microbiome and health states or outcomes. Unfortunately, the development of statistical methods to analyze these data has not kept pace. Many microbiome studies have complex design features (e.g., paired, clustered, or longitudinal data) or complexities that frequently arise in medical studies (e.g., the presence of confounding covariates), while existing methods for analyzing microbiome data are often restricted to testing only simple hypotheses.

Statistical methods for analyzing microbiome data seem to fall into one of two camps. One camp comprises methods that test the *global* effect of the microbiome, such as PERMANOVA [1, 2] and MiRKAT [3], which can be used to test the hypothesis that variables of interest (e.g., case-control status) are significantly associated with overall microbial compositions. However, these methods do not provide convenient tests of the effects or contributions of individual operational taxonomic units (OTUs), should a global microbiome effect be found (here we refer to OTUs, although all our results apply equally to data on amplicon sequence variants (ASVs) or count data from shotgun sequencing). The other camp is comprised of OTU-by-OTU tests, often directly using a method developed for RNA-Seq data such as DESeq2 [4], or a modification thereof such as metagenomeSeq [5], or based on a compositional data approach such as ANCOM [6, 7]. While these approaches have been widely applied, they generally do not give a single test of the global null hypothesis. Although test statistics or p-values from OTU-specific tests can of course be combined to give a global test (e.g., aMiSPU [8]), the performance of this kind of global test is often poor [9] since many of the OTU-specific tests only contribute noise.

We introduce here the Linear Decomposition Model (LDM) for analyzing microbial count data such as that obtained in a 16S rRNA study or a metagenomics sequencing study. The

LDM gives a unified approach that allows both global testing of the overall effect of the microbiome on arbitrary traits of interest, while also providing OTU-specific tests that correspond to the contribution of individual OTUs to the global test results. It allows for complex fixed-effects models such as models that include multiple variables of interest (both continuous and categorical), their interactions, as well as confounding covariates. It is permutation based, and so can accommodate clustered data and maintain validity for small sample sizes and when data are subject to overdispersion. Because the permutations are based on the Freedman-Lane approach [10], we can construct powerful type III or "last variable added" tests like those used in most linear regression packages [11, 12]. We also provide a new version of the PERMANOVA test based on our approach that we show outperforms the functions `adonis` and `adonis2` in the *R* package `vegan`, the most commonly used implementations of PERMANOVA for microbiome studies. Recent simulation studies suggest that many microbiome analysis methods fail to control the false discovery rate (FDR) when applied to overdispersed data [13]. We show that the LDM controls FDR in exactly the kind of situations where other methods fail.

We describe the LDM in detail in the methods section. In the results section, we describe the simulation studies and the two real datasets that we use to assess the performance of the LDM, and compare it to results obtained by PERMANOVA, MiRKAT, DESeq2, MetagenomeSeq [5], and the Wilcoxon rank-sum test. We conclude with a discussion section. Some technical details are relegated to an Appendix and Supplementary Materials.

Methods

Microbial composition data are usually summarized in an OTU table of read counts, here denoted by X, which is the $n \times J$ data matrix whose $(i, j)^{\text{th}}$ element is the number of times OTU j is observed in sample i. The total counts in each sample (the library size) can vary widely between samples and this variability must be accounted for. Here we accomplish this by converting counts to frequencies (i.e., relative abundances) by dividing by the library sizes, although other normalizations can be used with the LDM if desired. We show that the LDM

performs well when counts are normalized to frequencies, even with highly overdispersed data. Although the LDM is not explicitly a compositional data analysis method, compositional analyses can also be conducted by applying the LDM to appropriately transformed (e.g., centered log-ratio) data.

Many reasonable models of the relationship between data in an OTU table and covariates that describe traits or characteristics of individual samples can be expressed as a linear model. Because the large number of OTUs is a problem for models in which OTU frequencies predict traits or covariates when the goal is inference, we consider models in which traits and covariates are used as predictors of OTU frequencies. As an example of the kind of model we use, consider a study with n samples and a single binary trait corresponding to presence of a disease we are studying. Let $Y_i = 1$ if the ith participant has been diagnosed with the disease and $Y_i = 0$ if not. Define B_i to be the centered and scaled Y_i, i.e., $B_i = \nu^{-1}(Y_i - n_1/n)$, where $\nu = \sqrt{n_1 n_0/n}$ and n_1 and n_0 are the numbers of participants with and without disease. Suppose that the rows of the OTU table X have been scaled to give the OTU frequencies for each sample. Assume that the frequency of the jth OTU among disease-free participants is (on average) π_{0j} while that among participants with disease is π_{1j}. Then a reasonable model relating the observed OTU frequencies X_{ij} and the trait B_i can be written as

$$\mathrm{E}(X_{ij}|B_i) = \left(\frac{n_1}{n}\pi_{1j} + \frac{n_0}{n}\pi_{0j}\right) + B_i\nu\left(\pi_{1j} - \pi_{0j}\right),$$

which is a linear model for the data in the jth OTU, for each OTU. In matrix form we have

$$\mathrm{E}(X|B) = 1\left(\frac{n_1}{n}\pi_1 + \frac{n_0}{n}\pi_0\right)^{\mathrm{T}} + B\nu(\pi_1 - \pi_0)^{\mathrm{T}},$$

where 1 is the n-dimensional column vector with all entries equal to 1, and B, π_1, and π_0 are column vectors with elements B_i, π_{1j}, and π_{0j}, respectively. The goal of fitting this model might be to estimate π_0 and π_1 or, at least, test hypotheses about $\pi_1 - \pi_0$. These considerations motivate the general model

$$\mathrm{E}(X|B) = BW^{\mathrm{T}}, \tag{1}$$

where B is an $n \times r$ design matrix that we fix and that contains (potentially multiple) covariates including confounders, and W is a $J \times r$ matrix that we must estimate. In the example just considered, B has information on the disease status and an intercept that are known, while W contains the information on OTU frequencies π_0 and π_1 that we wish to estimate (as well as the normalization information in ν). Considered as J models for the columns of X, the jth column of W^{T} has the regression coefficients for the jth regression model. In order to provide a clean decomposition of the sum of squares of X, we will require that the columns of B are orthonormal, as they are in the example above. This also aids in the interpretation of some hypothesis tests, particularly terms that represent interactions with main effect terms that are also being tested.

The least-squares estimators for the matrix W can be obtained by minimizing

$$||X - BW^{\mathrm{T}}||_F^2,$$

where $||A||_F^2 \equiv \mathrm{Tr}(AA^{\mathrm{T}}) \equiv \sum_{i,j} A_{ij}^2$ is the Frobenius (matrix) norm of matrix A and $\mathrm{Tr}(.)$ is the trace operator. Satten et al. [14] showed that the resulting estimators are $W = X^{\mathrm{T}}B$, which are also the estimators obtained by fitting the regression model column by column. In some situations, we may wish to partition covariates into K groups, which we call "submodels". For example, we may wish to group several measures of smoking history into a single smoking "submodel". Thus, we partition B as $(B_1, \ldots, B_K)$, and re-write (1) as

$$\mathrm{E}(X|B) = \sum_{k=1}^{K} B_k W_k^{\mathrm{T}}; \tag{2}$$

then, the least-squares estimators of W_k are given by $W_k = X^{\mathrm{T}}B_k$ for each k.

The LDM as a decomposition

A linear model like equation (1) can also lead to a decomposition of the matrix X, if we choose B so that the columns of B span the column space of X. Then, we can always find W so that

$$X = BW^{\mathrm{T}}$$

holds exactly, corresponding to use of a saturated regression model. If we further write $W = VD$, where D is a diagonal matrix with entries given by the norms of the columns of W, so that the columns of V are normalized, then

$$X = BDV^{\mathrm{T}}. \tag{3}$$

Thus, the LDM mimics the singular value decomposition (SVD) of X, while differing in two important ways: first, in (3) we are free to choose the columns of B in any convenient way; and second, the columns of V are *not* orthogonal in general.

To connect the decomposition view of the LDM to the regression view and to ordination-based analyses, we propose a way to construct the matrix B using regression variables (that are not necessarily orthogonal) and a distance matrix Δ (which we assume has the same rank as X). Given the design matrix M_k for each submodel $k = 1, \ldots, K$, we first construct the hat matrix H_k corresponding to the *cumulative* design matrix $(M_1, \ldots, M_k)$; then define $\overline{H}_k = I - H_k$. We then choose B_1 to be the matrix whose columns are the eigenvectors of $H_1 \Delta H_1$ having non-zero eigenvalues and define the residual distance $\Delta_1 = \overline{H}_1 \Delta \overline{H}_1$. For $2 \leq k \leq K$, we take B_k to be the eigenvectors of $H_k \Delta_{k-1} H_k$ having non-zero eigenvalues and then set $\Delta_k = \overline{H}_k \Delta_{k-1} \overline{H}_k$. Finally, the remaining columns B_{K+1} are chosen to be the eigenvectors of Δ_K having non-zero eigenvalues. Use of the distance matrix here is primarily to allow direct comparison of the variability explained by covariates to the variability explained by the principal components of the distance matrix, described at the end of this section.

Using this partition of the columns of B, we can rewrite (3) as

$$X = \sum_{k=1}^{K+1} B_k W_k^{\mathrm{T}} = \sum_{k=1}^{K+1} B_k D_k V_k^{\mathrm{T}}, \tag{4}$$

which is the decomposition version of the regression model in equation (2); the $(K+1)$th term corresponds to the decomposition of the residual error in (2). As before, $W_k = X^{\mathrm{T}} B_k$ for $k = 1, \ldots, K+1$. To help ensure that the rank of X and Δ agree, if X has been column-centered, then we also center Δ as recommended by Gower [15]. Then, the columns of B are orthogonal to the vector 1, as are the columns of X.

The LDM (4) can be used to decompose the total sum of squares $S_{\text{total}} = ||X||_F^2$ into parts explained by each submodel, and the sum of squares that can be assigned to the residual directions corresponding to the columns of B_{K+1}:

$$S_{\text{total}} = \sum_{k=1}^{K+1} S_k,$$

where $S_k = ||B_k D_k V_k^{\mathrm{T}}||_F^2$. In analogy with Satten et al. [14] we can express S_k in one of two ways:

$$S_k = \text{Tr}\left(D_k^2\right) \tag{5}$$

or

$$S_k = \sum_{j=1}^{J} |W_{k;j\cdot}|^2, \tag{6}$$

where $W_{k;j\cdot}$ is the jth row of W_k and where $|w|$ is the Euclidean norm of vector w. The first representation indicates that, as with a standard SVD, the sum of squares for the kth submodel is given by the sum of the squares of the corresponding singular values (diagonal elements of D_k). The second representation partitions S_k into contributions from each OTU. These results only require B has orthonormal columns, not V. If X is centered, the sums-of-squares in (5) and (6) are proportional to the variance explained by a submodel or an individual OTU in a submodel; for this reason we sometimes refer to these tests as tests of the variance explained (VE).

Because the LDM is a decomposition, we can use S_k in (5) and the diagonal elements of D_{K+1}^2 to construct a scree plot to compare the sum of squares explained by each submodel with directions related to the eigenvectors of the residual distance matrix Δ_K. To accomplish this, we can plot either the absolute sum of squares (S_k) or average per-component sum of squares ($S_k/\text{Dim}(D_k)$) for each submodel $k = 1, \ldots, K$, along with the variability explained by each element of D_{K+1}^2. Because Δ_K may be used for ordination if we wish visualize the observations after removing the (linear) effects of covariates from the distance Δ, this scree plot is a quick and easy way to see which submodels explain a reasonable fraction of the variability we would expect to see in an ordination using distance Δ.

Testing hypotheses using the LDM

We use decomposition of the sum of squares implied by the LDM to test hypotheses about the effect of individual covariates or sets of covariates grouped into submodels as described in the previous section. Here, and in the rest of the paper, we let B, W, D and V consist only of those columns corresponding to the model terms, i.e., we exclude the residual terms B_{K+1}, W_{K+1} etc. unless explicitly stated otherwise.

To test hypotheses about the contribution of the jth OTU to the sum of squares for the kth submodel, we use its contribution given in (6), normalized as an F statistic, to give

$$F_{kj} = \frac{|W_{k;j\cdot}|^2}{|X_{\cdot j}|^2 - \sum_{k=1}^{K} |W_{k;j\cdot}|^2}, \tag{7}$$

where $X_{\cdot j}$ is the jth column of X and where we have dropped the constant of proportionality $\{\text{Rank}(X) - \text{Rank}(B)\}\,/\text{Rank}(B_k)$ found in a typical F test as we intend to use permutation to assess significance. The presence of all submodels in the denominator indicates that this is a type III or "last variable added" test statistic.

To test the global hypothesis we consider the total sum of squares for the kth submodel given in (5), again normalized as an F statistic. Using (6), this statistic can be constructed by summing the numerator and denominator of the OTU-specific statistics separately, i.e.,

$$F_{k,\text{global}} \propto \frac{\sum_{j=1}^{J} |W_{k;j\cdot}|^2}{\sum_{j=1}^{J} \left(|X_{\cdot j}|^2 - \sum_{k=1}^{K} |W_{k;j\cdot}|^2 \right)} = \frac{\text{Tr}(D_k^2)}{||X||_F^2 - \text{Tr}(D^2)}. \tag{8}$$

Assessing significance by permutation

We assess the significance of our test statistics, F_{kj} and $F_{k,\text{global}}$, using a variant of a permutation scheme described by Freedman and Lane [10]. Although the Freedman-Lane procedure formally permutes residuals, we show in Appendix that it is equivalent to a permutation procedure in which the residuals are held fixed but the covariates are permuted.

To describe our permutation approach, define X_k, the residual matrix obtained after fitting a reduced model to X that excludes the kth submodel term B_k, to be

$$X_k = \left(I - \sum_{\substack{k'=1 \\ k' \neq k}}^{K} B_{k'} B_{k'}^{\mathrm{T}} \right) X$$

and note that, because of the orthogonality of the columns of B, we can write $W_k = X^{\mathrm{T}} B_k$ as

$$W_k = X_k^{\mathrm{T}} B_k.$$

As a result, the test statistics F_{kj} in (7) can be rewritten as

$$F_{kj} = \frac{X_{k;\cdot j}^{\mathrm{T}} B_k B_k^{\mathrm{T}} X_{k;\cdot j}}{X_{k;\cdot j}^{\mathrm{T}} \left(I - \sum_{k'=1}^{K} B_{k'} B_{k'}^{\mathrm{T}} \right) X_{k;\cdot j}}, \tag{9}$$

where $X_{k;\cdot j}$ is the jth column of X_k. The analogous result for the global test statistic $F_{k,\mathrm{global}}$ is obtained by summing numerator and denominator over j. In Appendix, we show that if P_π is a permutation matrix corresponding to π, a permutation of the integers $1, \ldots, n$, then the Freedman-Lane permutation procedure is equivalent to forming the test statistics

$$F_{kj}^{(\pi)} = \frac{X_{k;\cdot j}^{\mathrm{T}} B_k^{(\pi)} B_k^{(\pi)^{\mathrm{T}}} X_{k;\cdot j}}{X_{k;\cdot j}^{\mathrm{T}} \left(I - \sum_{k'=1}^{K} B_{k'}^{(\pi)} B_{k'}^{(\pi)^{\mathrm{T}}} \right) X_{k;\cdot j}}, \tag{10}$$

where $B_k^{(\pi)} = P_\pi B_k$ is a row-permuted version of B_k. The test statistic for the global test $F_{k,\mathrm{global}}^{(\pi)}$ is obtained by (separately) summing the numerator and denominator of (10) over OTUs and can be written as

$$F_{k,\mathrm{global}}^{(\pi)} = \frac{\mathrm{Tr}\left[X_k^{\mathrm{T}} B_k^{(\pi)} B_k^{(\pi)^{\mathrm{T}}} X_k \right]}{\mathrm{Tr}\left[X_k^{\mathrm{T}} \left(I - \sum_{k'=1}^{K} B_{k'}^{(\pi)} B_{k'}^{(\pi)^{\mathrm{T}}} \right) X_k \right]}. \tag{11}$$

Although Freedman and Lane [10] only considered independent residuals, some simple but important cases involving correlated data can be tested using the Freedman-Lane approach. Here, we only consider the case of clustered data in which residuals within each cluster can be

considered as exchangeable. The main requirement for a valid permutation replicate dataset is that the dataset preserve the correlation found in the original data. Thus, variables that vary within clusters (sometimes called "plots" in the Ecology literature) can be permuted within each cluster. For example, if each cluster consists of a "before treatment" observation and an "after treatment" observation from the same individual, the effect of treatment can be tested by randomly permuting the "before" and "after" assignment within each cluster (individual). Note that in this situation, the cluster sizes need not be balanced (i.e., have equal size). For variables that are constant for all cluster members (i.e., are assigned at or "above" the cluster level), only permutation replicates that assign the same value to each cluster member are allowed. For example, in a rodent study of the effect of diet on the gut microbiome, rodents housed in the same cage should be treated as a cluster, as rodents are coprophagic. Thus, when permuting diet, rodents in the same cage should always be assigned the same diet. Note that for datasets with variables assigned at or above the cluster level, the cluster sizes must all be equal or the data must be stratified by cluster size with all permutations taking place within strata. Our implementation of the LDM uses the same permutation options available in the R package `vegan`, through the R package `permute`.

Our software implementation of the LDM uses sequential stopping rules to increase computational efficiency. When only the global test is of interest, we adopt the sequential stopping rule of Besag et al. [16] for calculating the p-value of the global test. This algorithm terminates when either a pre-determined number $L_{\min}$ of rejections (i.e., the permutation statistic exceeded the observed test statistic) has been reached or a pre-determined maximum number $K_{\max}$ of permutations have been generated. When the OTU-specific results are desired, we use the algorithm proposed by Sandve et al. [17], which adds a FDR-based sequential stopping criterion to the Besag et al. algorithm. Note that the Sandve et al. algorithm limits the total number of needed permutations to $J \times L_{\min} \times \alpha^{-1}$, which is 200 times the number of OTUs when the nominal FDR $\alpha = 10\%$ and $L_{\min} = 20$. When testing multiple hypotheses (e.g., both the global and OTU-specific hypotheses, or hypotheses corresponding to multiple sets of

variables), we generate permutations until all hypotheses reach their stopping point.

A Freedman-Lane PERMANOVA test (PERMANOVA-FL)

The operations that lead to the F-statistics for the LDM can also be used to develop an improved PERMANOVA test statistic. Following [2] we write Euclidean distance $\Delta = ZZ^{\mathrm{T}}$, then write a linear model of the form (2) in which Z replaces X. Here the only tests of interest are the global tests; the analogues of the OTU-specific tests are tests of the effect of covariates on the j^{th} component (column) of Z and are only used as intermediate steps. After replacing X with Z in (8) and using the invariance of the trace to cyclic permutations, the statistic $F_{k,\text{global}}$ can be rewritten as

$$F_{k,\text{PERMANOVA}} \propto \frac{\text{Tr}\left[B_k B_k^{\mathrm{T}} \Delta B_k B_k^{\mathrm{T}}\right]}{\text{Tr}\left[\left(I - \sum_{k'=1}^{K} B_{k'} B_{k'}^{\mathrm{T}}\right) \Delta \left(I - \sum_{k'=1}^{K} B_{k'} B_{k'}^{\mathrm{T}}\right)\right]},$$

which is the usual form of the PERMANOVA F statistic. The same argument leading to (9) yields

$$F_{k,\text{PERMANOVA}} \propto= \frac{\text{Tr}\left[B_k B_k^{\mathrm{T}} \widetilde{\Delta}_k B_k B_k^{\mathrm{T}}\right]}{\text{Tr}\left[\left(I - \sum_{k'=1}^{K} B_{k'} B_{k'}^{\mathrm{T}}\right) \widetilde{\Delta}_k \left(I - \sum_{k'=1}^{K} B_{k'} B_{k'}^{\mathrm{T}}\right)\right]},$$

where

$$\widetilde{\Delta}_k = \left(I - \sum_{\substack{k'=1 \\ k' \neq k}}^{K} B_{k'} B_{k'}^{\mathrm{T}}\right) \Delta \left(I - \sum_{\substack{k'=1 \\ k' \neq k}}^{K} B_{k'} B_{k'}^{\mathrm{T}}\right).$$

Thus, for a replicate dataset having covariates $B_k^{(\pi)}$, the Friedman-Lane PERMANOVA test statistic can be obtained by replacing X by Z in (11):

$$F_{k,\text{PERMANOVA}}^{(\pi)} \propto \frac{\text{Tr}\left[B_k^{(\pi)} B_k^{(\pi)^{\mathrm{T}}} \widetilde{\Delta}_k B_k^{(\pi)} B_k^{(\pi)^{\mathrm{T}}}\right]}{\text{Tr}\left[\left(I - \sum_{k'=1}^{K} B_{k'}^{(\pi)} B_{k'}^{(\pi)^{\mathrm{T}}}\right) \widetilde{\Delta}_k \left(I - \sum_{k'=1}^{K} B_{k'}^{(\pi)} B_{k'}^{(\pi)^{\mathrm{T}}}\right)\right]}. \tag{12}$$

We refer to this test as PERMANOVA-FL. The same kinds of restricted permutations as in our implementation of the LDM are available in PERMANOVA-FL.

The permutation scheme implemented in the `adonis` function in the *R* package `vegan` is similar to (12) except that the $\widetilde{\Delta}_k$s are all replaced by Δ. We further note that our proposed permutation replicates in (12) have the same advantages as the PERMANOVA replications implemented `adonis`, in that they only require functions of the distance matrix Δ (which, in our approach, are the projected distance matrices $\widetilde{\Delta}_k$). As a result, our approach, like other implementations of PERMANOVA, can be computed even if the distance matrix is non-Euclidean. Further, the distance matrices $\widetilde{\Delta}_k$ do not need to be recalculated for each replicate.

The arcsin-root transformation

The LDM can also be applied to transformed data. Because we consider frequency data, we show we achieve good results using the arcsin-root transformation, which is variance-stabilizing for Multinomial and Dirichlet-Multinomial (DM) counts. Thus we write $\Theta_{ij} = \sin^{-1}\sqrt{X_{ij}/N_i}$ where X_{ij} are the raw counts and N_i are the library sizes. We can additionally center Θ, replacing it by $\left(I - n^{-1}11^{\mathrm{T}}\right)\Theta$ if we also plan to center Δ. We can now replace X by Θ in (1) or (2) and proceed as before. This approach is related to an approach of Berkson [18, 19] for fitting logistic models to bioassay data. We also had considered a logit-based model using Haldane's [20] unbiased logit by forming $\Theta_{ij} = \ln\{(X_{ij} + 0.5)/(N_i - X_{ij} + 0.5)\}$ but found that the arcsin-root transform performed better in all cases we examined. We expect the LDM applied to (untransformed) frequency data will work best when the associated OTUs are abundant, while we expect the LDM applied to arcsin-root-transformed frequencies to work best when the associated OTUs are less abundant. Since we do not know the association mechanism *a priori*, we also consider an omnibus strategy that simultaneously applies LDM on both data scales. For the omnibus tests, we use the minimum of the p-value obtained from the frequency and arcsin-root-transformed data as the final test statistic and use the corresponding minima from the permuted data to simulate the null distribution [21].

The LDM and Redundancy Analysis

The LDM bears some resemblance to Redundancy Analysis (RA), but also differs in notable respects. RA seeks to describe how much of a matrix X can be explained by a single set of variables B_1, also concluding that the variability explained is $||B_1 B_1^{\mathrm{T}} X||_F^2$. RA also calculates a matrix like W_1; however, RA requires that W_1 have orthogonal columns, which is unnecessary for calculating either $\mathrm{Tr}(D_1^2)$ or $|w_{1;j\cdot}|^2$. Further, RA only allows analysis of one set of variables at a time, so only a single matrix W_1 is produced; this is presumably because the non-orthogonality of multiple W_ks implies that it is impossible to find W_1 and W_2 that satisfy $W_1^T W_2 = 0$ for arbitrary submodels B_1 and B_2. Thus in RA, the effect of each submodel B_k must be tested sequentially using a separate linear model like

$$\widetilde{X}_k = B_k W_k^{\mathrm{T}} + \epsilon,$$

where

$$\widetilde{X}_k = \left(I - \sum_{k'<k} B_{k'} B_{k'}^{\mathrm{T}} \right) X.$$

As a result, the F tests available in the LDM are expected to be more powerful than the type I or "order of variables added" tests available in RA when there is more than one submodel [11]. This is because the residual sums of squares in the denominator of the type III tests used in the LDM include *all* submodels tested, rather than only submodels with $k' < k$ used in sequential RA. Use of the restricted model in RA can thus result in an incorrect estimate of the residual sum of squares, which may affect power even in a permutation setting as the test is then not (asymptotically) pivotal. A second advantage of the LDM is that it is that we can assign significance to all submodels with a single permutation experiment, while RA requires a separate set of permutations for each submodel B_k tested.

Results

Simulation studies

We conducted several simulation studies to evaluate the performance of the LDM and compare it to competing methods. To evaluate the global test, we compared our results

to those obtained using our own implementation of PERMANOVA and the PERMANOVA implemented in the `adonis2` function. We also calculated OTU-specific tests using the LDM, which we compared to results from DESeq2. We only performed limited comparisons to results from MetagenomeSeq and Wilcoxon rank-sum test (applied to OTU frequencies), as they do not allow for confounding covariates.

To generate our simulation data, we used the same motivating dataset as Zhao et al. [3], specifically data on the upper-respiratory-tract (URT) microbiome first described by Charlson et al. [22]. To simulate read count data for the 856 OTUs reported in this study, we adopted a DM model using the empirical frequencies calculated from the study data; we set the overdispersion parameter to the estimate 0.02 obtained from these data, which is also the median value we observed in an admittedly brief survey of the literature [23–25]. While the original microbiome dataset was generated from 454 pyrosequencing with mean library size $\sim$1500, we increased the mean library size to 10000 to reflect Illumina MiSeq sequencing which is currently in common usage. For each simulation, we generated data for 100 samples unless otherwise noted. We also conducted sensitivity analysis with a wide range of library sizes, overdispersion parameters, and sample sizes, and by replacing the DM model with a Poisson log-normal model (PLNM) to generate the read count data (the PLNM is described in Supplementary Text S1).

We focused on two complementary scenarios. The first scenario (S1) assumed that a large number of moderately abundant and rare OTUs were differentially abundant between cases and controls, and the second scenario (S2) assumed the top 10 most abundant OTUs were differentially abundant. Both scenarios have a one-way, case-control design with a confounder and independent samples. Later we varied these scenarios to simulate a continuous trait, a two-way design, or clustered data.

In both scenarios S1 and S2, we let Y denote case-control status and assumed an equal number of cases ($Y = 1$) and controls ($Y = 0$). We simulated a confounder, $C = 0.5Y + \epsilon$, where ϵ was drawn from a uniform distribution $U[0, 1]$. In S1, we uniformly and independently

sampled two (overlapping) sets of 428 OTUs (half of all OTUs), the first set associated with Y and the second set associated with C; the set for Y was sampled after excluding the top three most abundant OTUs to focus on less abundant OTUs. In S2, we assumed the ten most abundant OTUs were associated with Y and the next forty most abundant OTUs were associated with C. These OTU sets were held fixed across replicates of data. We denoted the OTU frequencies estimated from the real data by the vector π_1 and formed vectors π_2 and π_3 by first setting π_2 and π_3 equal to π_1 and then randomly permuting those frequencies in π_2 and π_3 that belong to the selected set of OTUs associated with Y and C, respectively. Note that the frequencies for OTUs not selected to be associated with Y (or C) remain the same in π_1 and π_2 (or π_3). We then defined a sample-specific frequency vector as $\widetilde{\pi}(Y, C) = p_1(Y, C)\pi_1 + p_2(Y)\pi_2 + p_3(C)\pi_3$, where $p_2(Y) = \beta Y$, $p_3(C) = \beta_C C$, $p_1(Y, C) = 1 - p_2(Y) - p_3(C)$. In this model, β and β_C are the effect sizes of Y and C on the sample-specific OTU frequencies, respectively; here we set β_C to 0.3 except for simulations with no confounding, for which we set β_C to zero. We then generated the OTU count data for each sample using the DM model with $\widetilde{\pi}(Y, C)$, overdispersion parameter of 0.02, and library size sampled from $N(10000, 10000/3)$ and left-truncated at 500. By mixing π_1, π_2, and π_3 in a way that depends on the values of Y and C, we induced associations between the selected OTUs and Y and C. Note that π_1 serves as the "reference" OTU frequencies that characterizes samples for which Y and C are both zero. In addition, the correlations among Y, C, and OTU frequencies establish C as a confounder of the association between Y and the OTUs. Finally, note that when $\beta = 0$, $\widetilde{\pi}(Y, C)$ does not depend on Y, so that the null hypothesis of no association between Y and OTU frequencies holds.

To generate clustered data, we assume that we had samples from 50 distinct individuals, each of whom contributed 2 samples. We modified scenarios S1 and S2 by assuming that half of the individuals were cases ($Y = 1$) and the remaining individuals were controls ($Y = 0$). We generated the confounder C at the individual level in the same way as for unclustered data; to induce within-cluster correlations, we generated individual-specific OTU frequencies from

the Dirichlet distribution with mean frequencies $\widetilde{\pi}(Y, C)$ (defined earlier) and overdispersion parameter 0.02, and then generated counts for each samples from the same individual using the Multinomial distribution with mean being the individual-specific OTU frequencies, using library sizes that were generated independently for each sample. Note that if each individual had a single sample, the combination of Dirichlet and Multinomial sampling would reproduce the DM mixture model used for unclustered data.

To simulate data with a two-way design (without confounding), we considered two factors, Y_1 and Y_2, that each have two levels and that are orthogonal. The samples were randomly split into two groups with equal size, one group being assigned $Y_1 = 0$ and the other $Y_1 = 1$. Samples in each group were then further split randomly into two subgroups with equal size, with one group assigned $Y_2 = 0$ and the other $Y_2 = 1$. Then we induced association of Y_1 and Y_2 with the OTUs in the same way as Y and C by assigning Y_1 and Y_2 the same sets of OTUs as assigned to Y and C in scenarios S1 and S2 and using the sample sample-specific frequency vector $\widetilde{\pi}(Y_1, Y_2) = (1 - \beta_1 Y_1 - \beta_2 Y_2)\pi_1 + \beta_1 Y_1 \pi_2 + \beta_2 Y_2 \pi_3$. Note that β_1 and β_2 are the effect sizes of Y_1 and Y_2, respectively.

To generate data with a continuous trait, we used a model considered by Zhao et al. [3]. We first generated OTU counts for each sample using the DM model with frequency vector π_1, overdispersion parameter 0.02, and library size sampled from $N(10000, 10000/3)$. Let $S = \sum_{j \in \mathcal{A}} X_{ij}/\overline{X}_j$, where $\mathcal{A}$ is the set of the ten most abundant OTUs, X_{ij} is the frequency of the jth OTU in the ith sample, and $\overline{X}_j$ is the average frequency for the jth OTU across samples. We generated a confounder $C = \text{scale}(S) + \widetilde{\epsilon}$, where scale($v$) centers and normalizes vector v to have unit variance and $\widetilde{\epsilon} \sim N(0, 1)$. Finally, we simulated the continuous trait as $Y = \beta_C \text{scale}(C) + \beta \text{scale}(S) + \epsilon$, where $\beta_C = 0.3$ and $\epsilon \sim N(0, 1)$. Note that when $\beta = 0$ there is no association between Y and the OTU frequencies.

We evaluated the type I error and power for testing the global hypothesis at nominal significance level 0.05, and we assessed empirical sensitivity (proportion of truly associated OTUs that are detected) and empirical FDR for testing individual OTUs at nominal FDR of

10%. Results for type I error were based on 10000 replicates; all other results are based on 1000 replicates. In all simulations with confounders we treated C and Y as separate submodels M_1 and M_2, respectively, when fitting the LDM. For the two-way simulations, Y_1 and Y_2 were considered as separate submodels M_1 and M_2.

Results for testing global hypotheses with independent samples in the one-way case-control design

For testing the global hypothesis $H_0 : \beta = 0$ of no association between microbiome composition and Y, we applied the LDM on the frequency and arcsin-root scales and also calculated the omnibus test; these results are presented as VE-freq, VE-arcsin, and VE-omni, respectively, where VE denotes variance explained. We also applied our own implementation of PERMANOVA as well as the `adonis2` implementation; we refer to them as PERMANOVA-FL and adonis2, respectively.

Table 1 (top panel) shows our results for type I error. All methods, after adjusting for confounders, had correct type I error; with the small sample size 20, the type I error rates of PERMANOVA-FL and LDM methods were slightly conservative, which is consistent with the findings of Anderson and Legendre (1999) [26]. There was substantial inflation of type I error for S1 and modest inflation for S2 when the confounder was not accounted for, demonstrating that our methods are effective in accounting for confounders, with either modest or substantial confounding. The type I error rates were also close to 0.05 when the PLNM was adopted for count data simulation (Table S1).

Figure 1 (top panel) displays our results for power. We can see that VE-arcsin is more powerful than VE-freq under S1 and vice versa under S2; this is presumably because the variance stabilization of the arcsin-root transformation gives greater power to detect association with the rare OTUs that carry association in S1, while the untransformed data gives increased power to detect the common-OTU associations that characterize S2. In both cases, VE-omni achieved almost the same power as the most powerful test, without having to know whether common or rare OTUs were most important. PERMANOVA-FL has varying power depending

on the choice of distance measure: the power is lowest with the weighted-UniFrac distance, since the association was induced without reference to any phylogenetic tree, and the power is highest with the Hellinger distance in S1 and Bray-Curtis in S2. In both S1 and S2, our best-performing method has comparable power as the best-performing PERMANOVA-FL. Our sensitivity analysis showed that the relative performance of these methods persist for a wide range of library sizes, overdispersion parameters, and sample sizes (Figure S1), as well as with the PLNM (Figure S2). For this set of studies, PERMANOVA-FL and adonis2 yielded very similar power.

Results for testing individual OTUs with independent samples in the one-way case-control design

Because PERMANOVA-FL and adonis2 does not provide OTU-specific results, we compared our results on testing individual OTUs to DESeq2. When applying DESeq2, we replaced the default normalization by GMPR normalization [27], which was specifically developed for zero-inflated microbiome data.

Figure 1 middle and bottom panels display results on empirical sensitivity and empirical FDR, respectively. The LDM-based methods controlled FDR at 10% in all cases; their empirical FDRs are conservative (and the sensitivity values are low) in S1 because this scenario permuted frequencies among 428 OTUs selected for Y, majority of which are rare, and thus generated many weakly associated OTUs that are essentially null OTUs. The sensitivity of VE-omni tracks the method between VE-freq and VE-arcsin that performs better. Note that VE-arcsin has a higher sensitivity than VE-freq in both S1 and S2, but the order can be reversed in the scenario with a continuous trait (Figure 3). In contrast, the empirical FDRs for DESeq2 are modestly inflated under S1 and highly inflated under S2.

Because MetagenomeSeq and the Wilcoxon rank-sum test do not allow for adjustment of confounding covariates, we have not included results from these methods in Figure 1. To compare with MetagenomeSeq and the Wilcoxon rank-sum test, we set $\beta_C = 0$ for both scenarios S1 and S2 to remove confounders. MetagenomeSeq always controlled FDR but was

extremely conservative (FDR$<$ 2% for nominal FDR of 10% and sample size 100) in detecting associated OTUs for the simulations we conducted (Figure S3). The Wilcoxon test controlled FDR and achieved good sensitivity when the DM model was used for generating the read count data (Figure S3); when the PLNM was used (with no confounders), the data appeared less overdispersed and Wilcoxon had consistently lower sensitivity than VE-omni (Figure S2). Finally, DESeq2 failed to control FDR even in absence of any confounders.

Results for the two-way design

We set $\beta_1 = 0$ and $\beta_2 = 0.5$ to ascertain the type I error of the test of Y_1, and $\beta_2 = 0$ and $\beta_1 = 0.5$ to ascertain the type I error of the test of Y_2. Both the LDM and PERMANOVA-FL yielded correct type I error for testing each factor, whereas adonis2 had conservative type I error in scenario S1. The type I error using adonis2 was about a factor of 3 smaller for testing Y_1 than for testing Y_2 because the sampled OTUs for association with Y_2 (or C) included the top two most abundant OTUs and, as a result, Y_2 had a stronger global effect on the OTUs than Y_1. Consistent with the conservative type I error, adonis2 had lower power than PERMANOVA-FL (Figure 2). LDM (VE-omni) continued to maintain good power relative to PERMANOVA-FL for either factor (Figure 2). Further, LDM controlled FDR for OTUs that were detected to be associated with either factor (Figure 2).

Results for clustered data

In Table 1, we can see that permuting the case-control status over clusters rather than observations yields the correct type I error for all methods. We also calculated the type I error we would have obtained if we had incorrectly ignored the clustering structure when performing the permutations. Note that failure to account for the clustering structure result in a type I error of 100%. In Figure 3, we see the LDM controled FDR for these data, although the power and sensitivity is lower than was observed with the same number of samples which were unclustered (Figure 1). This is reasonable, as data with within-individual correlation is typically not as informative as data from an equivalent number of independent samples.

Results for continuous trait

From Table 1, we again see that all methods (adjusting for the confounder) have the correct type I error for data with a continuous trait; there was inflation of type I error when the confounder was not accounted for. In Figure 4, we see that the power of most methods is about the same. Although the sensitivity remains low as the effect size β increases, this appears to be related to the sample size, as we also show that the sensitivity increases rapidly as the sample size increases (at fixed $\beta = 3$). The LDM continues to control FDR as the sample size and sensitivity increase, while the empirical FDR for DESeq2 is never less than 40% for the range of sample sizes we considered.

Analysis of two microbiome datasets

To show the performance of the LDM in real microbiome data, we reanalyzed two datasets that were previously analyzed using MiRKAT and MMiRKAT [28] (a variant of MiRKAT for testing association between multiple continuous covariates and microbiome composition). The first is from a study of the association between the upper-respiratory-tract (URT) microbiome and smoking, and the second is from a study of the association between the prepouch-ileum (PPI) microbiome and host gene expression in patients with inflammatory bowel disease (IBD). We compared the performance of our global test (VE-omni) with PERMANOVA-FL, MiRKAT, and MMiRKAT; we also compared our OTU-specific results with results from DESeq2.

URT microbiome and smoking association study

The data for our first example were generated as part of a study to examine the effect of cigarette smoking on the orpharyngeal and nospharyngeal microbiome [22]. The 16S sequence data are summarized in an OTU table consisting of data from 60 samples and 856 OTUs, with mean library size 1500; metadata on smoking status (28 smokers and 32 nonsmokers) and two additional covariates (gender and antibiotic use within the last 3 months) was also available.

An imbalance in the proportion of male subjects by smoking status (75% in smokers, 56% in non-smokers) indicates potential for confounding. Zhao et al. (2015) [3] analyzed these data using MiRKAT, finding a significant global association between microbiome composition and smoking status after adjusting for potential confounders gender and antibiotic use. We used the Bray-Curtis distance for our analysis because it led to the smallest p-values compared to other distances in Zhao et al. (2015). We combined gender and antibiotic use into a single submodel M_1 and treated smoking status as M_2 when fitting the LDM.

We first constructed the ordination plots in Figure 5 using the Bray-Curtis distance after removing the effects of gender and antibiotic use (i.e., using Δ_1 as the distance matrix); these plots demonstrate a clear shift in smokers compared with nonsmokers even after removing the effect of potential confounders. The accompanying scree plots (Figure 5) on both frequency and arcsin-root scales further suggest that smoking explains an important faction of the variability in the OTU table. The residual (non-model) components are plotted in decreasing order of the size of the eigenvalue of the component in the spectral decomposition of Δ_2 (after removing the effect of confounders and smoking); the high correlation between the order of values D_k from the LDM and the order of eigenvalues of Δ_2 is noteworthy. We filtered out OTUs with presence in less than 5 samples, retaining 233 OTUs for analysis. The results of the LDM global tests, along with results from PERMANOVA-FL and MiRKAT, are presented in top-left panel of Table 2. VE-omni gave a smaller p-value than MiRKAT or PERMANOVA-FL based on the Bray-Curtis distance. In the top-right panel of Table 2, we show the results of our OTU-specific tests. VE-omni detected 5 OTUs (which include the 4 OTUs detected by VE-freq and constitute 5 of the 14 OTUs detected by VE-arcsin) whereas DESeq2 detected none. The inefficiency of DESeq2 is consistent with our simulation studies when the mean library size was 1500 (results not shown).

PPI microbiome and host gene expression association study

The data for our second example were generated in a study of the association between the

mucosal microbiome in the prepouch-ileum (PPI) and host gene expression among patients with IBD [25]. The PPI microbiome data are summarized in an OTU table with data from 196 IBD patients and 7,000 OTUs; gene expression data at 33,297 host transcripts, as well as clinical metadata such as antibiotic use (yes/no), inflammation score (0–13), and disease type (familial adenomatous polyposis/FAP and non-FAP) were also available. The data also included nine gene principal components (gPCs) that together explain 50% of the total variance in host gene expression. Zhan et al. [28] gave a joint test of all nine gPCs for association with microbiome composition, using MMiRKAT based on the Bray-Curtis distance measure and adjusting for antibiotic use, inflammation score, and disease type (FAP/non-FAP). Here we performed the same joint test using the LDM by putting the confounders in submodel matrix M_1 and then including all nine gPCs in a single submodel matrix M_2; however, we followed Morgan et al. (2015) [25] in only analyzing the original 196 PPI samples, not an additional 59 pouch samples from some of the same individuals included in the analysis of Zhan et al. [28]. We filtered out OTUs found in fewer than 5% of samples, retaining 2096 OTUs for analysis. VE-freq, VE-arcsin, and VE-omni yielded p-values 0.023, 0.0084, and 0.015, respectively, and detected 0, 4, and 3 OTUs (the 3 OTUs detected by VE-omni are included in the 4 OTUs detected by VE-arcsin) that significantly accounted for the global association at a nominal FDR rate of 10%. PERMANOVA-FL had p-value 0.0076 and MMiRKAT 0.0049, both based on the Bray-Curtis distance.

We also followed Zhan et al. (2016) to conduct individual tests of each of the nine gPCs. We treated each gPCs as a separate submodel (i.e., gPC1–gPC9 in M_2–M_{10}) in a single LDM, with M_1 accounting for the same confounders as in the joint test. Note that the gPCs are orthogonal. Scree plots for frequency and arcsin-root transformed data are shown in Figure 6, and indicate that gPC4 and gPC5 are most likely to be associated with microbial composition although any association is likely to be marginal. In fact, Table 2 confirms that gPC4 and gPC5 showed significant associations (at the 0.05 significance level) with the overall microbiome composition by the global tests; no other gPCs were found to be associated. Both VE-omni

and VE-arcsin detected (the same) 3 associated OTUs for gPC5, while VE-arcsin additionally found one OTU associated with gPC4. Both VE-omni and VE-arcsin also detected (the same) 1 OTU for gPC6, which was not significantly associated with the microbiome in a global test by any method. In contrast to the results obtained by the LDM, DESeq2 detected between 4 and 59 OTUs for each of the nine gPCs, which seems implausible given the results of the global tests. These findings may be related to the failure of DESeq2 to control FDR in the presence of confounders in our simulation studies.

Discussion

We have presented the LDM, a linear model for testing association hypotheses for microbiome data that can account for the complex designs found in microbiome studies. We have shown that the LDM has good power for global tests of association between variables of interest and the microbiome when compared to existing methods such as PERMANOVA and MiRKAT (the simulation results of MiRKAT were similar to those of PERMANOVA and thus not shown), but also provides OTU-specific tests. This is true even when confounding covariates are present, or when the study design results in correlated data. We have additionally shown that the OTUs identified by the LDM preserve FDR, while those identified by RNA-Seq-based approaches such as DESeq2 typically do not; further, since global and OTU-specific tests are unified, our analysis of the PPI microbiome data show that the LDM is less likely to identify "significant" OTUs for variables that are not globally significant. In the analyses we show here, there was only one instance where the LDM discovered OTUs that were significantly associated with a variable but the LDM global test for that variable was non-significant (gPC6 in the PPI data); in our simulations there were no such cases, although that may be because we only evaluated sensitivity for effect sizes that were large enough that the global test was always positive. While some analysts may choose to only calculate OTU-specific tests in the presence of a significant global test, this restriction is not required to control FDR at the OTU level. We have evaluated our approach using simulated data with realistic amounts

of overdispersion, confounding covariates and clustered data, and have shown how it can be applied to two real datasets.

We implemented the LDM in the R package `LDM` for use on any operating system, available at `http://web1.sph.emory.edu/users/yhu30/software.html` or on GitHub at `https://github.com/yhu/LDM`. The program is scalable to large sample sizes. Using a single thread of a MacBook Pro laptop (2.9 GHz Intel Core i7, 16 GB memory) and the default value $L_{\min} = 20$, it took 8, 19, 833 seconds to perform integrated global and OTU tests with a simulated dataset that consists of 20, 100, and 1000 samples. In our applications to real data, we used $L_{\min} = 100$ to ensure stability of results which are based on Monte Carlo sampling. It took 23 seconds (and stopped after 29400 permutations) to perform global and OTU tests with the URT data, and 5 hours (and stopped after 1273000 permutations) to perform global and OTU tests of nine gPCs separately with the PPI data.

The LDM easily handles complex designs. One important area the LDM is well-suited to is accounting for experimental artifacts that may be introduced by the way the samples are processed. For example, if the samples in a study are run on two different plates, and if some samples are included on both plates, then a variable that represent the difference between the microbiome profiles of the same sample on different plates can be included for each replicate pair. In this situation, we could either test for a plate effect or simply decide *a priori* to control for possible plate effects as confounders. Note that each variable we add (representing a pair of samples) is balanced by the presence of an extra row in the data matrix, so that the number of variables we have after adjusting for confounding remains equal to the number of distinct samples. Note that in clustered data, the confounders (e.g., plate effect) could vary within clusters. We plan to consider application of the LDM to design issues, such as how many samples to replicate, in a separate publication.

The LDM has features in common with RA (Redundancy Analysis), a multivariate technique to describe how much variability in one matrix (say, an OTU table X) can be described by variables in a second matrix (say, a model or design matrix M). The LDM differs from RA

most importantly in its ability to simultaneously obtain results for several submodels (groupings of one or more variables). To fit more than one submodel using RA, it is necessary to fit RA to each submodel, using data for which the previous submodels have been projected off. This precludes use of type III (last variable added) tests, which are known to be the most powerful [11]. Our use of the Freedman-Lane approach also gives superior performance; in simulations, our PERMANOVA-FL had higher power than the `adonis` and `adonis2` functions in the *R* package `vegan`, even though `adonis2` is based on some form of permutation of residuals (according to `adonis2` output).

Although the LDM is primarily based on the Euclidean distance in its focus on sums of squares, variability explained and F-like tests, we have shown how information on arbitrary distances can be incorporated in exploratory analyses, and use a distance matrix to choose analysis directions when submodels contain multiple terms. Although the Euclidean distance has been criticized when used for ecological analysis, Chao and Chiu [29] have recently suggested the problems associated with use of the Euclidean distance as a measure of beta diversity are related to normalization, rather than any intrinsic failure of the Euclidean distance. Finally, while *distance-based* Redundancy Analysis [30] does incorporate distance information, like PERMANOVA, it removes information on the effects of individual OTUs, and so was not included in our discussion.

In our examples here, we have put all confounders into the first submodel M_1. This conforms with practice in epidemiology in which confounders are not tested for inclusion into a model, but rather are included based on subject-area knowledge [31]. With this in mind, our implementation of the LDM does not provide p-values for the set of variables that are designated as confounders, which makes the code run faster. However, for those who want to estimate and test the individual effects of confounders, each confounder can be treated as separate submodel, and the LDM will calculate a p-value for each confounder. The results obtained in this way for the remaining variables are identical.

In this report we have concentrated on count data rather than presence-absence data.

Although we plan to consider presence-absence data in a separate publication, preliminary results indicate that if the data matrix is rarefied to a common library size and then converted into presence-absence information, then the LDM works well with presence-absence data and presence-absence-based distance measures. We plan to incorporate averages over rarefactions as part of the Monte-Carlo procedure for hypothesis testing to eliminate the possible loss of information that rarefaction may imply.

Among OTU-specific tests in absence of confounders, we found that metagenomeSeq controlled FDR in the simulations we conducted, while Hawinkel et al. (2016) [13] claimed that metagenomeSeq failed to control FDR. We noticed that Hawinkel et al. (2016) [13] adopted the zero-inflated Gaussian mixture distribution (i.e., the `fitZig` function), whereas we adopted the zero-inflated log-normal mixture model (i.e., the `fitFeatureModel` function) as recommended by the metagenomeSeq R package. We also found that the Wilcoxon rank-sum test is a robust and powerful choice for detecting differentially abundant OTUs when testing a single binary covariate. A recently-developed version of the rank-sum test [32] that uses inverse-probability-of-treatment weights could provide an interesting extension for categorical testing when adjustment for confounding covariates is required. However, OTU-specific tests based on the rank-sum test do not provide coherent results with any global test.

Conclusions

We propose the LDM, a method for testing association hypotheses for microbiome data. It integrates distance-based analysis, global testing of any microbiome association, and detection of individual OTU association to give coherent results. It is capable of handling complex design features such as confounders, interactions, and correlated data, and is thus widely applicable in modern microbiome studies. The LDM is generally as powerful as existing methods for testing global associations, and controls FDR better than existing methods for finding individual OTU effects. As such, it can accelerate the search for associations between the microbiome and variables of interest.

Disclaimer

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Funding

This research was supported by the National Institutes of Health awards R01GM116065 (Hu).

Availability of data and materials

The URT microbiome data used in this article are available in the `MiRKAT` R package `http://research.fhcrc.org/wu/en.html`. The PPI microbiome data used in this article are publicly available as Bioproject PRJNA269954 (16S sequence data) and as GEO GSE65270 (microarray data). The R package `LDM` is available at `http://web1.sph.emory.edu/users/yhu30/software.html` or on GitHub at `https://github.com/yhu/LDM` in formats appropriate for Macintosh or Windows systems. In addition, each of these sites includes a vignette demonstrating use of the package.

Authors' contributions

YJH and GAS conceived the study, developed the method, analyzed the data, and wrote the manuscript. Both authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publications

No consent for publication was required for this study; all microbiome datasets used here are publicly available.

Ethics approval and consent to participate

This study only involved secondary analyses of existing, de-identified datasets; as such it does not require IRB approval.

References

1. Anderson MJ. A new method for non-parametric multivariate analysis of variance. Austral ecology. 2001;26(1):32–46.

2. McArdle BH, Anderson MJ. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. Ecology. 2001;82(1):290–297.

3. Zhao N, Chen J, Carroll IM, Ringel-Kulka T, Epstein MP, Zhou H, et al. Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. The American Journal of Human Genetics. 2015;96(5):797–807.

4. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology. 2014;15(12):550.

5. Paulson JN, Stine OC, Bravo HC, Pop M. Differential abundance analysis for microbial marker-gene surveys. Nature methods. 2013;10(12):1200–1202.

6. Mandal S, Van Treuren W, White RA, Eggesbø M, Knight R, Peddada SD. Analysis of composition of microbiomes: a novel method for studying microbial composition. Microbial ecology in health and disease. 2015;26(1):27663.

7. Kaul A, Mandal S, Davidov O, Peddada SD. Analysis of microbiome data in the presence of excess zeros. Frontiers in microbiology. 2017;8:2114.

8. Wu C, Chen J, Kim J, Pan W. An adaptive association test for microbiome data. Genome medicine. 2016;8(1):56.

9. Koh H, Blaser MJ, Li H. A powerful microbiome-based association test and a microbial taxa discovery framework for comprehensive association mapping. Microbiome. 2017;5(1):45.

10. Freedman D, Lane D. A nonstochastic interpretation of reported significance levels. Journal of Business & Economic Statistics. 1983;1(4):292–298.

11. Muller KE, Fetterman BA. Regression and ANOVA: An Integrated Approach using SAS Software. SAS Institute; 2012.

12. Kleinbaum DG, Kupper LL, Nizam A, Muller KG. Applied Regression Analysis and Other Multivariable Methods. Duxbury Press; 2007.

13. Hawinkel S, Mattiello F, Bijnens L, Thas O. A broken promise: microbiome differential abundance methods do not control the false discovery rate. Briefings in bioinformatics. 2017;.

14. Satten GA, Tyx RE, Rivera AJ, Stanfill S. Restoring the Duality between Principal Components of a Distance Matrix and Linear Combinations of Predictors, with Application to Studies of the Microbiome. PloS one. 2017;12(1):e0168131.

15. Gower JC. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika. 1966;53(3-4):325–338.

16. Besag J, Clifford P. Sequential Monte Carlo p-values. Biometrika. 1991;78(2):301–304.

17. Sandve GK, Ferkingstad E, Nygård S. Sequential Monte Carlo multiple testing. Bioinformatics. 2011;27(23):3235–3241.

18. Berkson J. Application of the logistic function to bio-assay. Journal of the American Statistical Association. 1944;39(227):357–365.

19. Berkson J. A statistically precise and relatively simple method of estimating the bio-assay with quantal response, based on the logistic function. Journal of the American Statistical Association. 1953;48(263):565–599.

20. Haldane J. The estimation and significance of the logarithm of a ratio of frequencies. Annals of human genetics. 1956;20(4):309–311.

21. Westfall PH, Young SS. Resampling-based multiple testing: Examples and methods for p-value adjustment. John Wiley & Sons; 1993.

22. Charlson ES, Chen J, Custers-Allen R, Bittinger K, Li H, Sinha R, et al. Disordered microbial communities in the upper respiratory tract of cigarette smokers. PloS one. 2010;5(12):e15216.

23. La Rosa PS, Brooks JP, Deych E, Boone EL, Edwards DJ, Wang Q, et al. Hypothesis testing and power calculations for taxonomic-based human microbiome data. PloS one. 2012;7(12):e52078.

24. Chen J, Li H. Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. The annals of applied statistics. 2013;7(1).

25. Morgan XC, Kabakchiev B, Waldron L, Tyler AD, Tickle TL, Milgrom R, et al. Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease. Genome biology. 2015;16(1):67.

26. Anderson MJ, Legendre P. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. Journal of statistical computation and simulation. 1999;62(3):271–303.

27. Chen J, Chen L. GMPR: A novel normalization method for microbiome sequencing data. bioRxiv. 2017;p. 112565.

28. Zhan X, Tong X, Zhao N, Maity A, Wu MC, Chen J. A small-sample multivariate kernel machine test for microbiome association studies. Genetic epidemiology. 2017;41(3):210–220.

29. Chao A, Chiu CH. Bridging the variance and diversity decomposition approaches to beta diversity vis similarity and differentiation measures. Methods in Ecology and Evolution. 2016;7(8):919–928.

30. Legendre P, Anderson MJ. Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. Ecological monographs. 1999;69(1):1–24.

31. VanderWeele TJ, Shpitser I. A new criterion for confounder selection. Biometrics. 2011;67(4):1406–1413.

32. Satten GA, Kong M, Datta S. Multisample adjusted U-statistics that account for confounding covariates. Statistics in Medicine. 2018;37(2):3357–3372.

Appendix

In this appendix, we describe the permutation scheme we use to assess the significance of our test statistics for both the LDM and PERMANOVA-FL. We make use of the fact that, since the columns of B are orthogonal, $B_k B_k^{\mathrm{T}}$ is the orthogonal projection operator (hat matrix) corresponding to variables in submodel k. Consider the linear model for the jth column of the matrix X given by

$$X_{\cdot j} = \sum_{k=1}^{K} B_k W_{k;j\cdot}^{\mathrm{T}} + \epsilon_{\cdot j}, \tag{A1}$$

where $\epsilon_{\cdot j} = B_{K+1} W_{K+1;j\cdot}^{\mathrm{T}}$. Suppose we wish to test the kth submodel. The Freedman-Lane approach is to form residuals from the reduced model that excludes the term B_k, writing

$$X_{k;\cdot j} = \left(I - \sum_{\substack{k'=1 \\ k' \neq k}}^{K} B_{k'} B_{k'}^{\mathrm{T}} \right) X_{\cdot j},$$

where we have substituted the least-squares estimator of $W_{k;j\cdot} = X_{\cdot j}^{\mathrm{T}} B_k$, and further note this estimator is the same regardless of other terms in the model, since the columns of B are orthogonal. We then generate a new set of values $X_{\cdot j}^{(\pi)}$ for $X_{\cdot j}$ in which all linear effects except those corresponding to B_k are preserved, but the residuals are permuted, by writing

$$X_{\cdot j}^{(\pi)} = \left(\sum_{\substack{k'=1 \\ k' \neq k}}^{K} B_{k'} B_{k'}^{\mathrm{T}} \right) X_{\cdot j} + P_\pi X_{k;\cdot j},$$

where P_π is a permutation matrix. In order to construct the F tests we have described, we need to calculate the residuals we would obtain by fitting either the full model (A1) to the permuted data $X_{\cdot j}^{(\pi)}$, and the reduced model that excludes the term B_k. These quantities are most easily obtained by left-multiplying by an appropriate projection operator. The residual for fitting the full model is given by

$$\left(I - \sum_{k'=1}^{K} B_{k'} B_{k'}^{\mathrm{T}} \right) X_{\cdot j}^{(\pi)} = \left(I - \sum_{k'=1}^{K} B_{k'} B_{k'}^{\mathrm{T}} \right) P_\pi X_{k;\cdot j},$$

so that the residual sum of squares after fitting the full model is

$$X_{k;\cdot j}^{\mathrm{T}} P_\pi^{\mathrm{T}} \left(I - \sum_{k'=1}^{K} B_{k'} B_{k'}^{\mathrm{T}} \right) P_\pi X_{k;\cdot j}.$$

Because the B_ks are orthogonal, we can see that the difference between the residual sum of squares for the full and restricted models is simply the contribution to the sum of squares for B_k, given by

$$X_{k;\cdot j}^{\mathrm{T}} P_\pi^{\mathrm{T}} B_k B_k^{\mathrm{T}} P_\pi X_{k;\cdot j}.$$

Finally, we note that if P_π is a permutation matrix, then P_π^{T} is also a permutation matrix corresponding to the permutation that reverses the effect of P_π, i.e., $P_\pi P_\pi^{\mathrm{T}} = I$. Thus, we define $B_k^{(\pi)} = P_\pi^{\mathrm{T}} B_k$ to be a row-permuted version of B_k and note that the columns of $B^{(\pi)}$ remain orthogonal, so that $B_k^{(\pi)} B_k^{(\pi)^{\mathrm{T}}}$ is the orthogonal projection (hat) matrix corresponding to fitting a model in which the variables have been permuted according to permutation matrix P_π^{T}. With this observation, we note that the residual sum of squares after fitting the full model is

$$X_{k;\cdot j}^{\mathrm{T}} \left(I - \sum_{k'=1}^{K} B_{k'}^{(\pi)} B_{k'}^{(\pi)^{\mathrm{T}}} \right) X_{k;\cdot j},$$

which is the denominator of F_{kj} given in (10). Similarly, the contribution to the sum of squares due to $B_k^{(\pi)}$ alone is

$$X_{k;\cdot j}^{\mathrm{T}} B_k^{(\pi)} B_k^{(\pi)^{\mathrm{T}}} X_{k;\cdot j},$$

which is the numerator of F_{kj}. Note also that $B_k^{(\pi)^{\mathrm{T}}} B_{k'}^{(\pi)} = \delta_{kk'} I$ since $P_\pi P_\pi^{\mathrm{T}} = I$.

Table 1 Type I error for testing the global hypothesis at nominal significance level 0.05

	Scenario	n	PERMANOVA-FL	adonis2	VE-freq	VE-arcsin	VE-omni
Independent samples, one-way case-control design							
Adjusting for confounder	S1	20	0.040	0.052	0.045	0.031	0.038
		100	0.053	0.052	0.050	0.054	0.053
	S2	20	0.039	0.054	0.043	0.026	0.034
		100	0.050	0.050	0.052	0.047	0.053
Not adjusting for confounder	S1	20	0.299	0.296	0.215	0.408	0.362
		100	0.987	0.987	0.913	0.998	0.997
	S2	20	0.065	0.064	0.056	0.076	0.067
		100	0.151	0.151	0.083	0.245	0.194
Independent samples, two-way design							
Testing for Y_1	S1	100	0.046	0.013	0.048	0.048	0.050
	S2	100	0.049	0.048	0.051	0.051	0.050
Testing for Y_2	S1	100	0.049	0.036	0.049	0.046	0.049
	S2	100	0.053	0.048	0.053	0.054	0.054
Clustered samples, one-way case-control design							
Accounting for clustering	S1	100	0.050	0.050	0.048	0.051	0.053
	S2	100	0.044	0.047	0.047	0.044	0.045
Not accounting for clustering	S1	100	1	1	1	1	1
	S2	100	1	1	0.999	1	1
Independent samples, continuous trait							
Adjusting for confounder		100	0.049	0.051	0.055	0.049	0.051
Not adjusting for confounder		100	0.095	0.095	0.090	0.093	0.091

Results for PERMANOVA-FL and adonis2 are based on the Bray-Curtis distance. n is the number of samples. When testing Y_1, we set $\beta_2 = 0.5$; when testing Y_2, we set $\beta_1 = 0.5$. All analyses of clustered data adjust for the confounder.

Table 2 Results in analysis of the two real datasets

	Testing the global hypothesis					Testing individual OTUs			
trait	MiRKAT	PERMANOVA-FL	VE-freq	VE-arcsin	VE-omni	VE-freq	VE-arcsin	VE-omni	DESeq2
URT microbiome data									
Smoking	0.0019	0.0018	0.0070	0.0006	0.001	4	14	5	0
PPI microbiome data									
gPC1	0.22	0.19	0.13	0.47	0.19	0	0	0	14
gPC2	0.36	0.19	0.19	0.19	0.26	0	0	0	49
gPC3	0.24	0.31	0.30	0.21	0.29	0	0	0	29
gPC4	0.16	0.088	0.013	0.080	0.021	0	1	0	13
gPC5	0.0094	0.015	0.034	0.010	0.015	0	3	3	48
gPC6	0.19	0.41	0.43	0.49	0.53	0	1	1	23
gPC7	0.15	0.21	0.76	0.16	0.22	0	0	0	59
gPC8	0.21	0.33	0.64	0.36	0.47	1	0	0	4
gPC9	0.15	0.12	0.10	0.12	0.15	0	0	0	20

For the global hypotheses, reported results are p-values. For the individual OTU tests, results reported are the number of OTUs detected at FDR = 10%. MiRKAT and PERMANOVA-FL results are based on the Bray-Curtis distance.

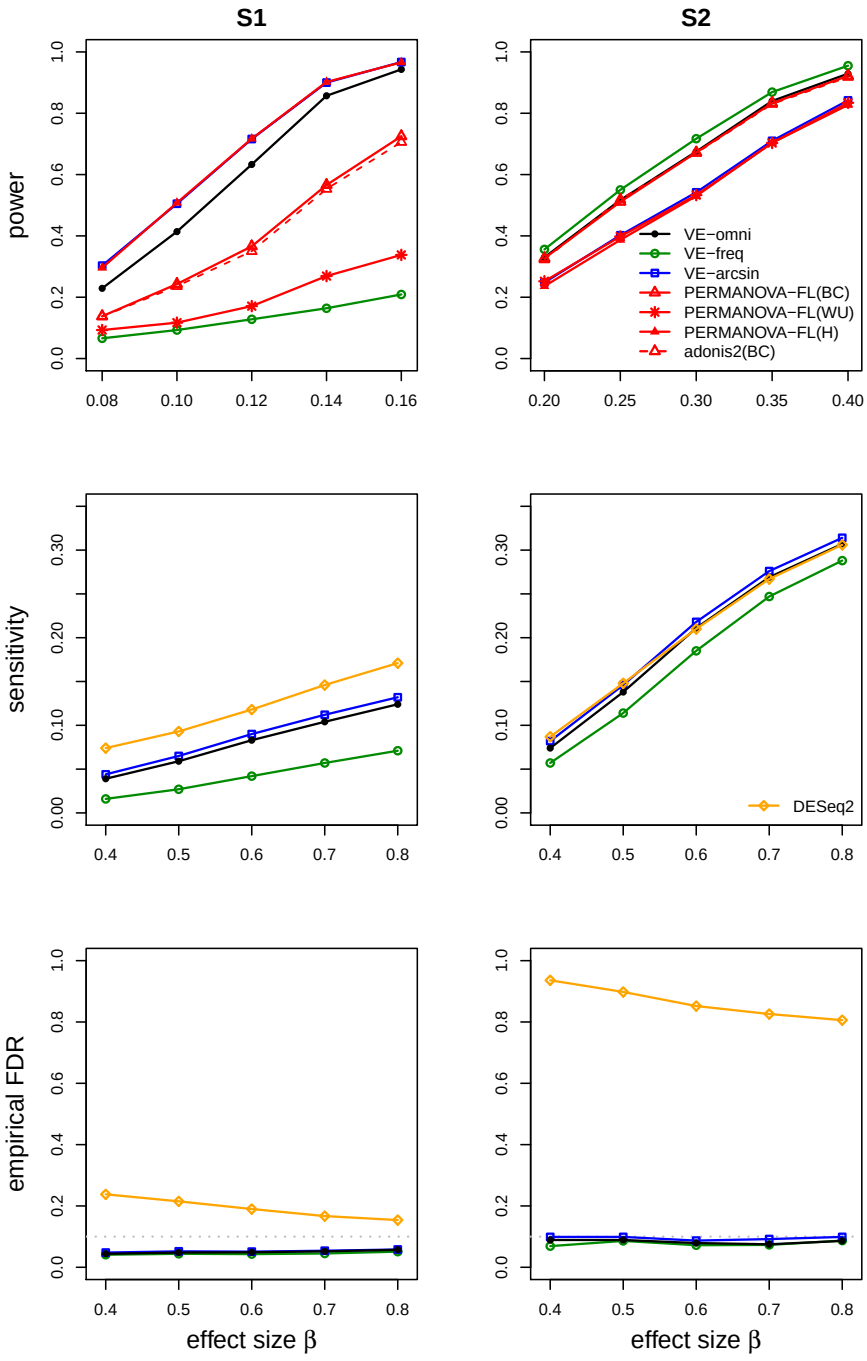

Fig. 1 Simulation results for one-way, case-control studies with independent samples. The gray dotted lines represent the nominal FDR=0.1. BC: Bray-Curtis; WU: weighted UniFrac; H: Hellinger.

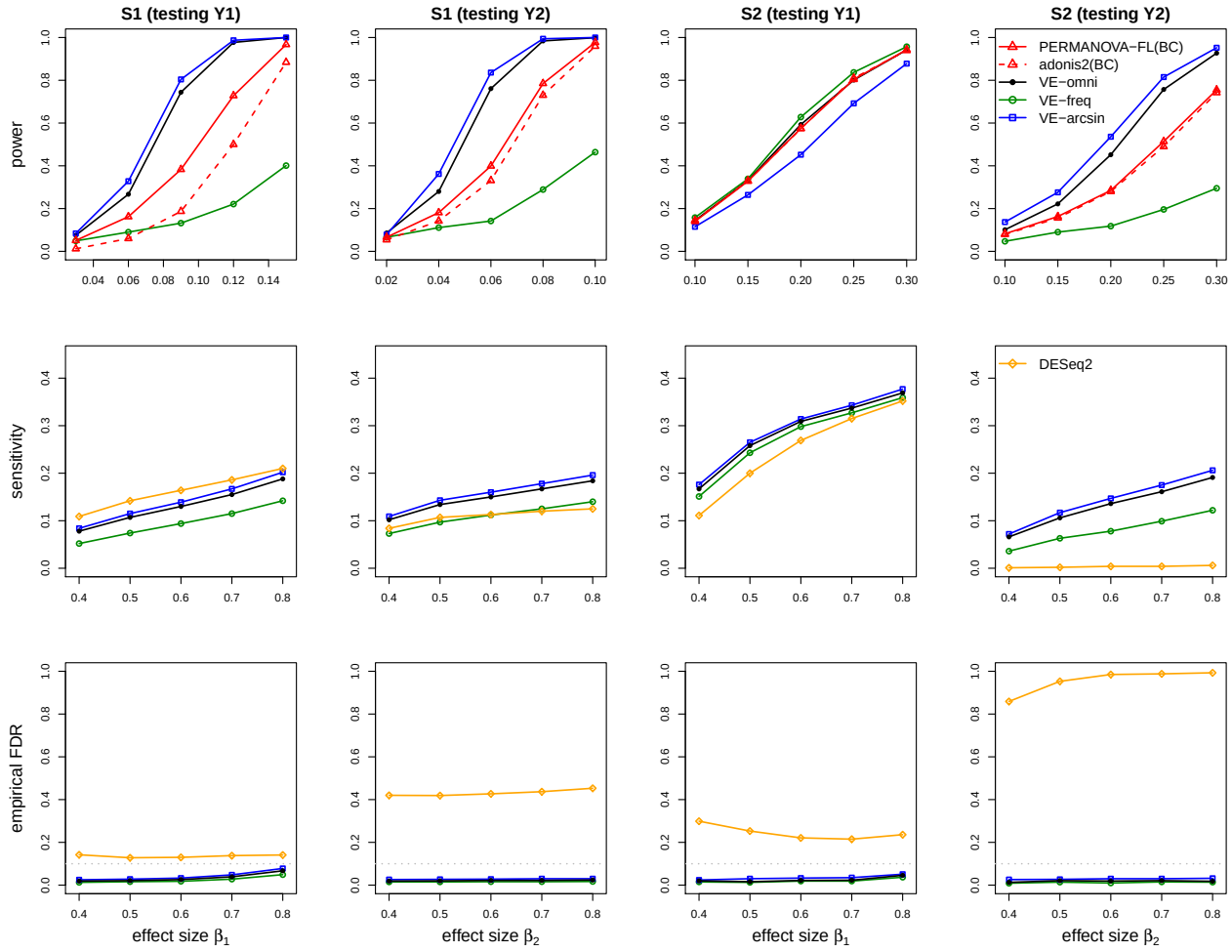

Fig. 2 Simulation results for studies with the two-way design and independent samples. The gray dotted lines represent the nominal FDR=0.1. BC: Bray-Curtis. When testing Y_1, we set $\beta_2 = 0.5$; when testing Y_2, we set $\beta_1 = 0.5$.

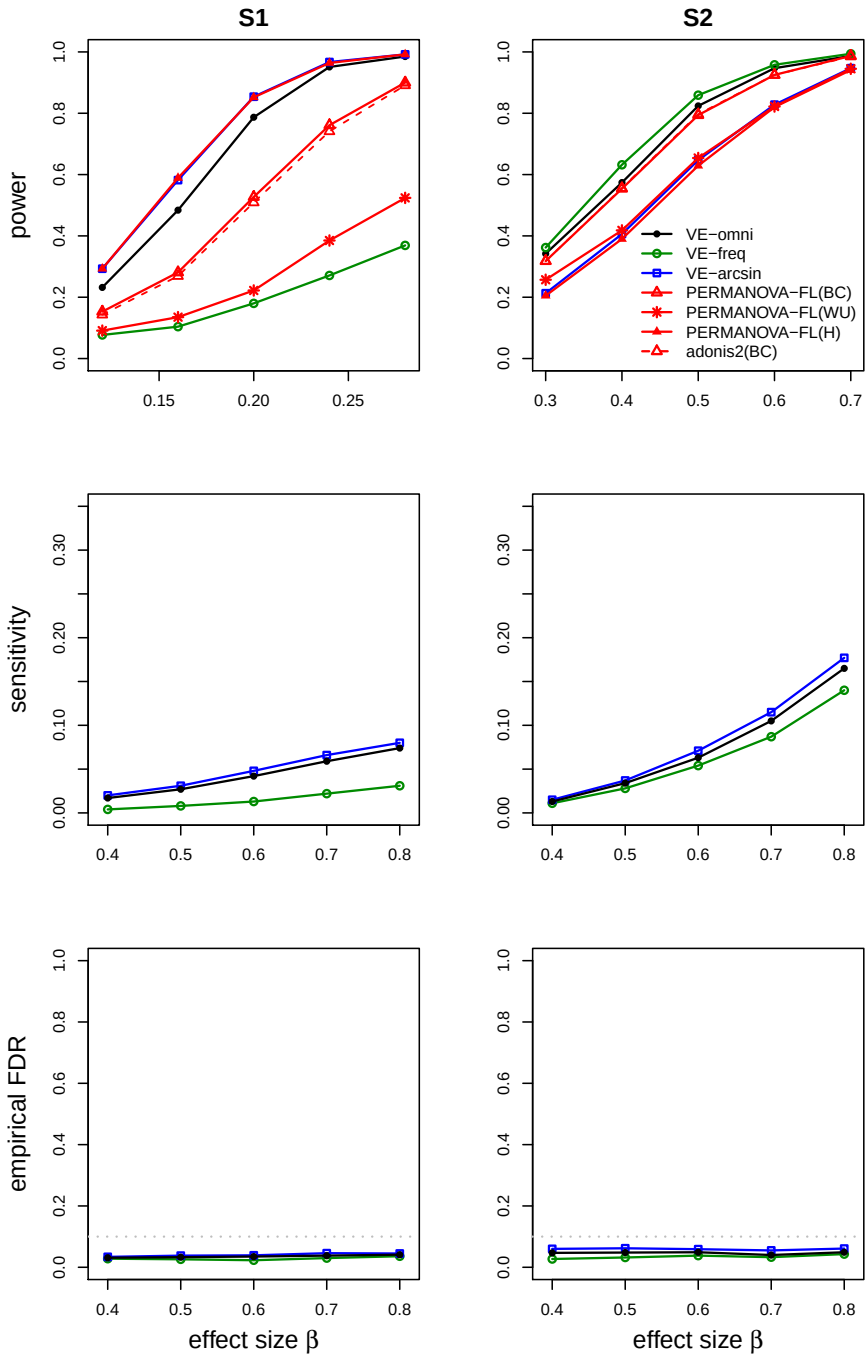

Fig. 3 Simulation results for clustered data. The gray dotted lines represent the nominal FDR=0.1. BC: Bray-Curtis; WU: weighted UniFrac; H: Hellinger. The DESeq2 program is not applicable for this type of clustered data.

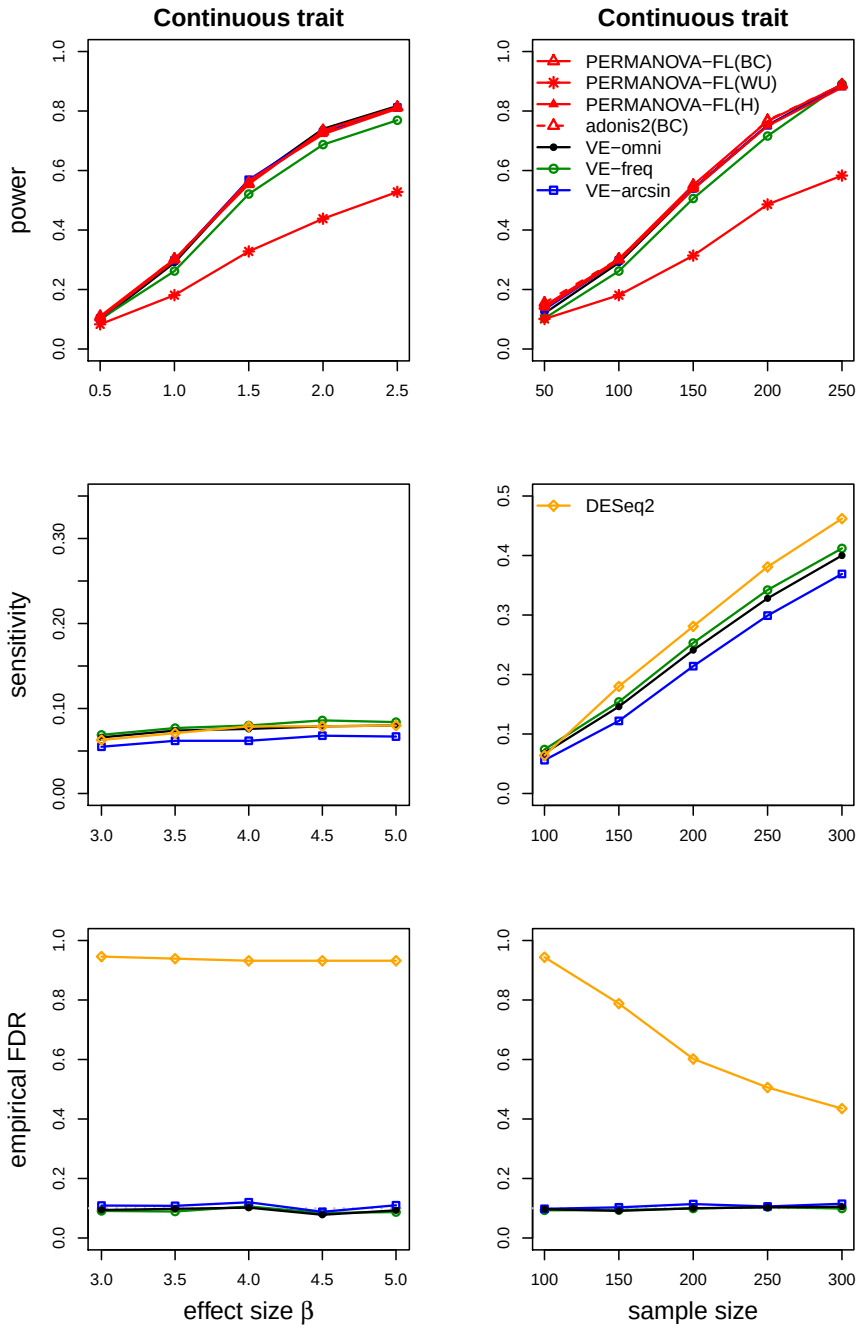

Fig. 4 Simulation results for a continuous trait. The first and second columns correspond to results as the effect size and the sample size, respectively, increase. The gray dotted lines represent the nominal FDR=0.1. When varying the sample size, we set $\beta = 1$ for evaluating power and $\beta = 3$ for sensitivity and empirical FDR.

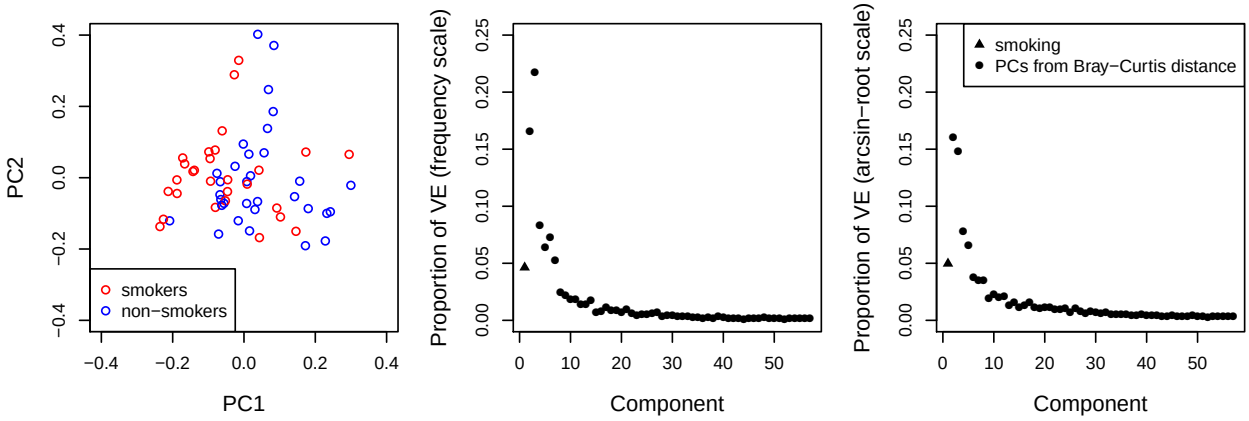

Fig. 5 Exploratory analysis of the URT microbiome data based on the Bray-Curtis distance. Left plot: ordination plot after removing the effects of confounders gender and antibiotic use, colored by smoking status. Center and right plots: proportions of variance explained by smoking and the PCs of the (residual) distance measure after removing the effects of gender and antibiotic use; the PCs are ordered by their Bray-Curtis eigenvalues. The center plot is based on frequency data and the right plot is based on arcsin-root transformed data. The components are ordered by the magnitude of their corresponding eigenvalue in a spectral decomposition of Δ_2 (the distance matrix after removing the effect of the confounders and smoking).

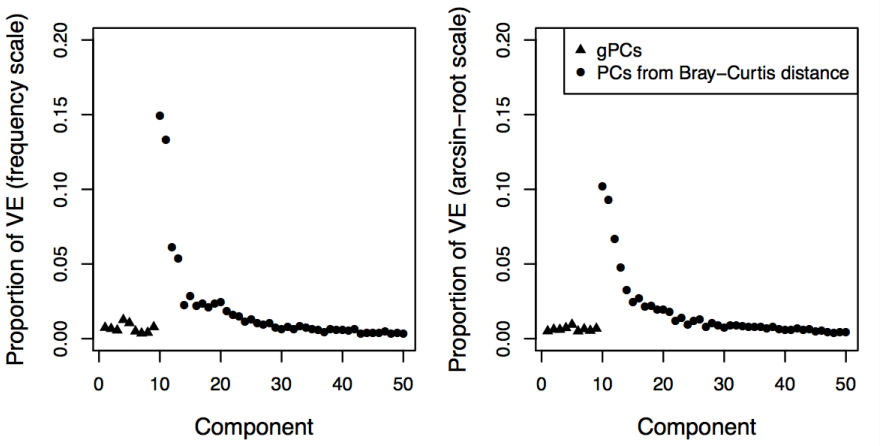

Fig. 6 Exploratory analysis of the PPI microbiome data based on the Bray-Curtis distance. The proportions of variance explained by the 9 gPCs and the PCs of the (residual) distance measure are obtained after removing the effects of confounders antibiotic use, inflammation score, and disease type. The PCs are ordered by their Bray-Curtis eigenvalues. The left plot is based on frequency data and the right plot is based on arcsin-root transformed data. The components are ordered by the magnitude of their corresponding eigenvalue in a spectral decomposition of Δ_{10} (the distance matrix after removing the effect of confounders and the 9 gPCs). Only the first 50 (of 195 total) components are shown.