Testing hypotheses about the microbiome using an ordination-based 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

Abstract

Background: Data from a microbiome study is often analyzed by calculating a matrix of pairwise distances between observations; ordination using principal components or multidimensional scaling of the distance matrix is frequently successful in separating meaningful groups of observations (e.g., cases and controls). Although distance-based analyses like PERMANOVA can be used to test association using only the distance matrix, the connection between data on individual species (or operational taxonomic units, OTUs) and the information in the distance matrix is lost. This makes it hard to know which species contribute to the patterns seen in ordination for the high-dimensional data we gather in a microbiome study.

Methods: We introduce a novel approach called the linear decomposition model (LDM) that provides a single analysis path that includes distance-based ordination, global tests of any effect of the microbiome, and tests of the effects of individual OTUs with false discovery rate (FDR)-based correction for multiple testing. 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.

Results: For global testing, our simulations indicate the LDM provides correct type I error, even with substantial confounding and/or correlations, and can have substantially higher power than existing distance-based methods. For testing individual OTUs, our simulations indicate the LDM generally controlled the FDR well. In contrast, DESeq2 had inflated FDR in the presence of confounding and inferior sensitivity with sparse OTU data; 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.

Conclusions: We present the LDM, a method that integrates global testing of any microbiome

effect and detection of differentially abundant OTUs. The LDM is generally more powerful than existing methods, and is capable of handling the confounders and correlated data that frequently occur in modern microbiome studies.

Keywords: Clustered data, differentially abundant OTUs, distance measure, linear model, microbiome association test, multifactorial, multivariate outcomes, resampling methods, tests of interactions

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: methods that originate in Ecology, which are based on distance measures and ordination (i.e., representation of pairwise distances between samples in a 2- or 3-dimensional plot, for example using principal components (PCs)), and methods that originate in analysis of RNA-Seq data, which typically test the effects of each individual OTUs. Distance-based methods are popular because ordination frequently shows good separation between meaningful groups (e.g., cases and controls) in a study. Distance-based methods such as PERMANOVA [1, 2] and MiRKAT [3] can be used to test the global hypothesis that variables of interest (e.g., case-control status) are significantly associated with overall microbial compositions. However, unless distance is measured by the Euclidean distance between (possibly transformed) OTU frequencies, it is not possible to determine which OTUs contribute to the features seen in the ordination plots [4]. While the significance of individual OTUs can subsequently be ascertained using an OTU-by-OTU approach, these results cannot be be connected back to the features seen in ordination. Finally, while test statistics from OTU-specific tests can be combined to give a global test (e.g., aMiSPU [5]), the performance of this kind of global test is often poor [6] since many of the OTU-specific tests only contribute noise.

We introduce here the Linear Decomposition Model (LDM) for analyzing microbial com-

position data such as that obtained in a 16S rRNA study. The LDM is distance-based, yet connects the PCs of a pre-specified distance matrix with linear combinations of OTUs. Thus, it can simultaneously perform both the global test of any association between microbial composition and arbitrary variables (i.e., categorical or continuous, univariate or multivariate) and OTU-specific tests in a coherent way, so that the OTU-specific effects add up to the global effect. It allows for complex fixed-effects models such as models that include multiple confounding covariates and/or multiple variables of interest as well as interaction terms. It is permutation based, and so can accommodate block-structured (i.e., correlated) data and maintains validity for small sample sizes and when data are subject to overdispersion.

Microbial composition data are usually summarized in an OTU table of read counts. The total counts in each sample (i.e., library size) can vary widely between samples and this variability must be accounted for. Here we accomplish this by converting counts to frequencies by dividing by the library sizes. We show that the LDM performs well with this normalization, even with highly overdispersed data. We eschew the normalizations that have been proposed for RNA-Seq data as inappropriate for microbial data. Because the expectation in RNA-Seq is that only a few genes are differentially expressed even when comparing case to control expression profiles, the goal of RNA-Seq normalization is to ensure that the 'typical' gene expression in every sample is the same [7]. The situation in the microbiome could be quite different; since microbes form a community, we may expect many or even most OTUs may be found in different abundances in samples corresponding to different biological or environmental conditions. Genes probably do not compete for limited resources to express, while microbes definitely compete for resources. Further, microbes can change their environment (e.g., by changing the pH) so that the success of one species may suppress growth of a different species. In fact, the default normalization for RNA-Seq packages like DESeq2 [8] often fail for microbiome data because, unlike RNA-seq data, most cells in an OTU table are empty.

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, and MetagenomeSeq [9]. We conclude with a discussion section. Some technical details are relegated to two appendices and Supplementary Materials.

Methods

The starting point for our analysis is the OTU table, 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. We will typically scale the rows of X by the library size to convert counts to frequencies, and then center the columns; this is the standard processing for Principal Coordinates Analysis (PCoA). Let Δ be the matrix with the $(i, i')^{\text{th}}$ element denoting the dissimilarity between the ith and i'^{th} observations. Although many of the dissimilarities we consider are not formally distances, we follow common practice and refer to Δ as the distance matrix. Following Gower [10], we typically assume that the elements of Δ have been squared and that Δ has been centered if X has. Concordant centering (or not) of X and Δ ensures that the column vector having each element equal to 1 (i.e., unit column) is absent from (or present in) the column space of both X and Δ.

To motivate our approach, we first consider an ordination analysis in which we process the counts as for PCoA, and then measure similarity using the Euclidean distance [11]; then Gower's [10] squared, centered distance is $\Delta = XX^T$, the variance-covariance matrix between observations. With this choice of distance matrix, the eigenvectors and eigenvalues of $\Delta = LEL^T$ can equivalently be obtained by forming the singular value decomposition (SVD) of $X = L\Sigma R^T$ with $E = \Sigma^2$. Let $A_{\cdot j}$ denote the jth column of matrix A. Suppose that after ordination (e.g., representation of samples using eigenvectors in L corresponding to the two or three largest eigenvalues in E) we find that the first eigenvector (i.e., first PC) $L_{\cdot 1}$ separates groups that we are exactly interested in comparing (e.g., cases and controls). We can then use the SVD of X to write $L_{\cdot 1} = \Sigma_{11}^{-1} X \cdot R_{\cdot 1}$, where Σ_{11} is the singular value corresponding to

$L_{\cdot 1}$. Thus, $R_{\cdot 1}$ gives the factor loadings for each OTU corresponding to $L_{\cdot 1}$, and can be used to determine the contribution of each OTU to the separation seen in the ordination.

Unfortunately, the analysis just described fails in the absence of two elements: use of the Euclidean distance measure, and a clear separation of meaningful groups in the data by one of the important (i.e., first two or three) PCs. When an arbitrary distance matrix Δ is used, its PCs do not necessarily coincide with the left singular vectors of X and so the factor loadings of OTUs corresponding to the PCs are not directly available, unless the data can be transformed so that the desired distance is Euclidean [12]. We can solve the first problem by using the orthogonal decomposition approach of Satten et al. [4], which provides an 'approximate' SVD of X that allows the 'left singular vectors', denoted by B, to be any fixed matrix with orthogonal columns; in particular, we can choose B to be the PCs of a distance matrix Δ that we wish to use. Then,

$$X = BDV^T + \epsilon \tag{1}$$

where D is a diagonal matrix with non-negative entries, V is a matrix with orthogonal columns, ϵ is a matrix of residuals, and D and V minimize $||X - BDV^T||_F^2$ subject to $V^TV = I$, where $||A||_F^2 \equiv \mathrm{Tr}(AA^T) \equiv \sum_{i,j} A_{ij}^2$ is the Frobenius norm of matrix A and $\mathrm{Tr}(.)$ is the trace operator. An algorithm for finding D and V given B, called the tandem algorithm by Everson [13], is also found in Satten et al. [4] and is implemented in the R package `ordinationDuality`. When PCs of any arbitrary distance matrix are used for B in fitting (1), the jth column of V gives the 'factor loadings' corresponding to the jth PC.

The LDM for ordination-based testing

The ordination-based analysis also fails if we do not see good separation between groups when plotting the first few PCs. Even if the first few PCs do separate groups, simply testing the magnitude of factor loadings for each PC is not sufficient since it may be some linear combination of PCs that separate groups. For example, suppose cases fall in the upper right quadrant of the ordination plot constructed using the first two PCs (say, PC1 and PC2), while

controls fall in the lower left quadrant. Then, the best variable to separate the two groups would be a linear combination of PC1 and PC2. To make sure the best linear combination is always available, we force the existence of 'PCs' corresponding to the hypotheses we wish to test (e.g., difference between case and control participants), and then pick the remaining PCs using the distance matrix.

Specifically, we first construct an $n \times s$ design matrix M for variables we wish to test jointly; like a standard linear model, each univariate continuous variable contributes one column and categorical variables with c levels can be recoded into c columns corresponding to c indicators. Inspired by PERMANOVA, we form the projection operator (i.e., hat matrix) $H = M(M^T M)^- M^T$, where A^- denotes the Moore-Penrose inverse of matrix A, and then calculate the eigendecomposition of $H\Delta H = B_m E_m B_m^T$. We adopt the convention that E_m only includes non-zero eigenvalues, so that the number of columns in B_m corresponds to the rank m of $H\Delta H$. To generate a full-rank decomposition of X, we also include in B the eigenvectors B_r of $\widetilde{H}\Delta\widetilde{H} = B_r E_r B_r^T$ where we again take E_r to have only non-zero eigenvalues and where $\widetilde{H} = I - H$. Then, we use $B = (B_m, B_r)$ when fitting (1). With this choice, we are assured that the first m columns of B will separate groups of observations by the variables included in M, although we must still decide if this separation is significant. Note that while we use eigenvectors B_m and B_r for our model, PERMANOVA uses eigenvalues E_m and E_r to form the test statistic $F \propto \mathrm{Tr}(E_M)/\mathrm{Tr}(E_r)$.

In some situations we may wish to test for two or more sets of variables individually and sequentially. For example, in a study with two active treatments and a placebo, we may wish to first test for any treatment effect, and then test for a difference between the two active treatments. This would give two one-degree-of-freedom tests, rather than the two-degree-of-freedom test we would obtain by putting all treatment indicators into a single design matrix M. For K sets of variables, we form design matrices M_1, M_2, $\cdots$, M_K and define H_k ($k = 1, \cdots, K$) to be the projection matrix corresponding to the cumulative design matrix (M_1, $\cdots$, M_k) with orthogonal complement $\widetilde{H}_k = I - H_k$. We then choose B_1 to be the

matrix whose columns are the eigenvectors of $H_1 \Delta H_1$ corresponding to non-zero eigenvalues and define $\widetilde{\Delta}_1 = \widetilde{H}_1 \Delta \widetilde{H}_1$. For $2 \leq k \leq K$, we take B_k to be the eigenvectors of $H_k \widetilde{\Delta}_{k-1} H_k$ and then set $\widetilde{\Delta}_k = \widetilde{H}_k \widetilde{\Delta}_{k-1} \widetilde{H}_k$. Finally, we calculate B_r, the eigenvectors of $\widetilde{\Delta}_K$. Then we use $B = (B_1, B_2, \cdots, B_K, B_r)$ when fitting (1).

It is worth noting that an equivalent procedure would be to form design matrices $\overline{M}_1$, $\overline{M}_2$, $\overline{M}_3$, $\cdots$, $\overline{M}_K$, where $\overline{M}_1 = M_1$ and the columns of each successive $\overline{M}_k$ were chosen to be orthogonal to the columns of all previous matrices $\overline{M}_{k'} (k' < k)$. In this approach, we could then calculate the hat matrix $\overline{H}_k$ corresponding to each $\overline{M}_k$ and choose B_k to be the eigenvectors of $\overline{H}_k \Delta \overline{H}_k$, which is now calculated using the original distance matrix. If we took $\overline{H}_r = I - \prod_{k=1}^{K} \overline{H}_k$ then B_r would be given by the eigenvectors of $\overline{H}_r \Delta \overline{H}_r$. Thus, when using the LDM, it should be understood that the effect of variables in M_1 are tested first, then the effect of the variables in M_2 that remain after accounting for M_1 are tested, and so on. This interpretation also shows how interactions between variables are handled in the LDM: if the main effects enter the model before the interaction term, then the interaction term is automatically taken to be the part of the interaction that is orthogonal to the main effects.

Note that if B is partitioned as $B = (B_1, B_2, \cdots, B_K, B_r)$, we can rewrite (1) in terms of the K models (i.e., sets of variables) as

$$X = \sum_{k=1}^{K} B_k D_k V_k^T + B_r D_r V_r^T + \epsilon \tag{2}$$

It is important that B gives a full-rank approximate decomposition of X; in this way our approach differs crucially from Redundancy Analysis (RA) [11], since this feature allows us to incorporate both information on variables of interest and a distance matrix Δ when explaining variability in X. We now restate some results from Satten et al. [4] that suggest test statistics to use for testing hypotheses implied by (1) or (2). Satten et al. [4] showed that the total sum of squares, $S_{\text{total}} = ||X||_F^2$, can be partitioned into sums of squares for each model, plus a residual sum of squares:

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

where $S_k = ||B_k D_k V_k^T||_F^2$ and S_{resid} includes both contributions from residual directions B_r as well as ϵ. Let $W_k = D_k V_k^T$ so that

$$(W_k)_{\cdot j} = (D_k V_k^T)_{\cdot j} \tag{3}$$

where we recall $A_{\cdot j}$ is the jth column of A. Satten et al. [4] give two representations: $S_k = \text{Tr}(D_k^2)$ and $S_k = \sum_{j=1}^{J} w_{kj}^2$ where $w_{kj}^2 = ||(W_k)_{\cdot j}||^2$ and where $||v||$ is the Euclidean norm of vector v. The first representation indicates that, as with a standard SVD, the sum of squares for the kth model is given by the sum of the squares of the corresponding singular values. The second representation partitions S_k into contributions from each OTU.

To test the global hypothesis whether there is any relationship between the variables in M_k and the microbiome after accounting for all variables in $M_{k'}$ ($k' < k$), we propose to use S_k as our test statistic. To test whether the jth OTU has an effect on this hypothesis, we propose to use w_{kj}^2. We use permutation to establish the significance of these test statistics and defer the details to a later section. Note that if the columns of X are centered, $S_{\text{total}} = ||X||_F^2$ can be interpreted as a variance. Since Δ should also be centered in this case, the columns of B are also centered so that $S_k = ||B_k D_k V_k^T||_F^2$ can also be interpreted as a variance. Thus, we refer to S_k as the variance explained by the variables in model k and w_{kj}^2 as the contribution of the jth OTU to the variance explained by model k. Comparing the LDM to RA, we note that RA does not partition variability by separating variables into separate testable models, but rather uses 'partial RA' to test a succession of models that have some variables tested and others controlled as confounders, then assigns sums of squares for individual variables by subtraction. For this reason, some combinations of variables may be assigned a negative 'sum of squares' [14]. Even if the the overall sum of squares assigned in this way is positive, the contribution of individual OTUs can also be negative. Finally, RA is not distance-based (except in those special situations where XX^T can be used as a distance [12]).

The LDM as a Linear Model

We refer to our procedure as a linear decomposition model because it shares features of both a matrix decomposition and a linear model. The sums of squares results just presented are features that are common to matrix decompositions like the SVD. To see the linear-model-like features, we rewrite (2) as

$$X_{\cdot j} = \sum_{k=1}^{K} B_k (W_k)_{\cdot j} + \widetilde{\epsilon}_{\cdot j} \tag{4}$$

where $(W_k)_{\cdot j}$ is given in (3) and we include the effects of residual columns B_r in $\widetilde{\epsilon}$. Here, $B = (B_1, \cdots, B_K)$ plays the role of the design matrix and $(W_k)_{\cdot j}$ is a vector of coefficients to be estimated that relates the variables of interest B_k to the data for the jth OTU. The Wald test for significance of $(W_k)_{\cdot j}$ would be $\{(\widehat{W}_k)_{\cdot j}\}^T \widehat{\mathrm{Var}}^{-1} \{(\widehat{W}_k)_{\cdot j}\} (\widehat{W}_k)_{\cdot j}$ but the ordinary least squares estimator for $\widehat{\mathrm{Var}}\{(\widehat{W}_k)_{\cdot j}\} \propto \left({B_k}^T B_k\right)^{-1} = I$ since the columns of B are orthogonal. Thus, the Wald test would reduce to $||(\widehat{W}_k)_{\cdot j}||^2 \equiv w_{kj}^2$, the contribution of OTU j to S_k that we have proposed as a test statistic. The form of (4) overstates the similarity between the LDM and RA, since the largest group of terms in (2) is presumably B_r, which was incorporated into $\widetilde{\epsilon}$ in (4).

Controlling for confounding effects in the LDM

In some cases, we may wish to control for a set of potentially confounding covariates, rather than test all variables in the model. Without loss of generality, we assume that all confounding covariates (and only these variables) are included in M_1. After controlling for confounders in M_1, we then wish to test variables in $M_2, \cdots, M_K$.

Assume B_1 defined earlier is a matrix having orthogonal columns that span the columns of M_1; if M_1 contains columns not spanned by the columns of B_1 because these directions are in the null space of Δ, then we re-define B_1 to be any matrix having columns that form an orthonormal basis for M_1. Using the regression analogy, to control confounding represented by B_1, we should estimate the corresponding regression coefficients W_1 from the data under

the composite null model without variables in $M_2, \cdots, M_K$, and then keep these values fixed when we test the other variables. In the usual linear model, this is equivalent to replacing each column of X by the appropriate column of $\widetilde{X} \equiv X - B_1 W_1$ and then fitting the linear model of $M_2, \cdots, M_K$ for each column of the residual $\widetilde{X}$. This in turn is equivalent to using RA to estimate X using the confounding variables in M_1 by defining $\widehat{X}_1 = B_1 B_1^T X$, where $B_1 B_1^T$ is the projection operator into the space spanned by M_1, and then replacing X by $\widetilde{X} = X - \widehat{X}_1$.

Note that since $B_1^T \widetilde{X} = 0$, the column rank of $\widetilde{X}$ has been reduced (compared to the column rank of X) by the rank of B_1; therefore a similar reduction in the row rank of $\widetilde{X}$ must occur. In Appendix 1 we show that if $Z_1 = X^- B_1$, then $\widetilde{X} Z_1 = 0$. Thus, the directions that are removed when replacing X by $\widetilde{X}$ are the least-squares predictors of B_1, since Z_1 satisfies $X Z_1 = B_1$.

Since the directions in B_1 have been removed from the column space of the data matrix $\widetilde{X}$, these directions should also be removed from the distance matrix by replacing Δ with $\widetilde{\Delta}_1 = \widetilde{H}_1 \Delta \widetilde{H}_1$, where we recall $\widetilde{H}_1$ is the projection matrix into the orthogonal complement of the space spanned by M_1; this ensures that the column spaces of the adjusted data and distance matrices coincide. For reasons we discuss in the next section, we also replace M_k by $\widetilde{M}_k = (I - H_1) M_k$ for $k \geq 2$. These adjusted data can then be analyzed using the LDM that does not adjust for confounding. When we fit the LDM to $\widetilde{X}$ to estimate D_k and V_k $(k = 2, \cdots, K, r)$ by minimizing

$$||\widetilde{X} - \sum_{k=2}^{K} B_k D_k V_k^T - B_r D_r V_r^T||_F^2, \tag{5}$$

where B_2, $\cdots$, B_K, and B_r are calculated using $\widetilde{\Delta}_1$ and $\widetilde{M}_k$ $(k \geq 2)$, it would appear in addition to the usual constraints

$$V_k^T V_{k'} = \delta_{kk'} I \ , \ k, k' = 2, \cdots, K, r, \tag{6}$$

we need to apply the additional constraints

$$Z_1^T V_k = 0 \ , \ k = 2, \cdots, K, r. \tag{7}$$

However, we show in Appendix 2 that the estimators of D_k and V_k ($k = 2, \cdots, K, r$) obtained by minimizing (5) subject only to (6) also satisfy the additional constraints (7). This result holds precisely because $\widetilde{X}Z_1 = 0$, which ensures that the direction Z_1 is never generated in the minimization of (5).

Assessing significance by permutation

We assess the significance of our test statistics, both global and OTU-specific, using a permutation scheme described by Winkler et al. [15]. Specifically, we permute the variables we are testing rather than the data in the OTU table X. This choice corresponds to permuting the quantities that are under experimental control, and is similar to the permutation structure used in PERMANOVA. An advantage to this approach is that the distance matrix Δ only needs to be calculated once. In this work we consider only models with fixed effects; we plan to consider models with random effects in a later publication.

In the simplest case when there is no confounder and no block design (i.e., correlation), we permute the rows of M_1, M_2, $\ldots, M_K$, using the same permutation for each M_k, while preserving X and Δ. Then we recalculate B_1, $\cdots$, B_K, B_r for each permutation, refit the LDM, and obtain test statistics that are compared with the observed statistics for calculating p-values.

In the presence of confounders in M_1, each permutation dataset should have the same amount of confounding as the original dataset; for this reason, variables in M_1 should not be permuted. This can be accomplished by using the adjusted data that replace X by $\widetilde{X}$ and Δ by $\widetilde{\Delta}_1$ as discussed previously. We also replace M_k by $\widetilde{M_k}$ ($k \geq 2$), which ensures logical consistency in the situation when M_k has interactions with variables we have kept fixed in M_1; when we permute the rows of $\widetilde{M_k}$ instead of the rows of M_k, we only permute the part of the interaction that is orthogonal to the main effect. (Note that if we had permuted M_k in place of $\widetilde{M_k}$, using $\widetilde{\Delta}_1$ in place of Δ would have been sufficient to ensure that the B_k, $k = 2, \cdots, K, r$, were orthogonal to the original B_1 for each permutation since B_1 is not in the column space

of $\widetilde{\Delta}_1$.)

The same types of block-structured (i.e., clustered) data can be analyzed using the LDM as can be analyzed using PERMANOVA. Here we explore only some of the simplest cases. In all cases, it is required that each permutation of the rows of (or the residuals of) the design matrices respect the experimental design. First, for variables that are constant for all block members (i.e., are assigned 'above' the block level), only permutations that assign each block member the same but possibly new value for these variables 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 block, as rodents are coprophagic. Thus, when permuting diet, rodents in the same cage should always be assigned the same diet. (Assignments for blocks with a single observation, i.e., singleton blocks, can be made from an appropriate distribution, such as the distribution of values among singleton blocks - possibly sampled without replacement - or by randomly selecting first a block then randomly selecting the value from one of the block members.) Second, variables that vary within blocks can be permuted within each block. For example, if each block consists of a 'before treatment' and 'after treatment' observation from the same individual, the effect of treatment can be tested by randomly permuting the 'before' and 'after' assignment within each block. (The singleton blocks can be treated as independent observations and their 'before' or 'after' labels can be permuted among these singleton blocks, or can be drawn from another appropriate distribution such as the cluster-weighted marginal distribution.)

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 100 times the number of OTUs when $\alpha = 10\%$ and $L_{\min} = 10$. 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.

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 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 $\left(I - n^{-1}11^T\right)\Theta$ 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]. These transformations also have the salutary effect of removing the compositional constraint on the data, since it is generally not the case that the row sums of Θ are zero, even if the column sums have been centered.

Results

Simulation studies

We conducted several simulation studies to evaluate the performance of the LDM and to compare it to competing methods. To evaluate the global test, we compared our results to results obtained using our own implementation of PERMANOVA that allows for confounding covariates (i.e., replacing the distance matrix Δ by $\widetilde{\Delta}_1$, replacing the design matrices M_k by $\widetilde{M}_k$ ($k \geq 2$), and permuting the rows of $\widetilde{W}_k$). We also calculated OTU-specific results, which we compared to results from DESeq2. We only performed limited comparisons to results from MetagenomeSeq, as MetagenomeSeq does not allow for confounding covariates.

To generate our simulation data, we used the same motivating dataset as Zhao et al. [3]. Specifically, we simulated read count data for 856 OTUs assuming a Dirichlet-Multinomial (DM) model using frequencies and overdispersion parameter that were estimated from a real upper-respiratory-tract (URT) microbiome dataset [22]. 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 of ∼1500, we increased the mean library size here to 10000 to reflect Illumina MiSeq sequencing which is currently in common usage. We set the total sample size to 100 unless otherwise noted. We also conducted sensitivity analysis with a wide range of library sizes, overdispersion parameters, and sample sizes.

We considered two complementary scenarios; the first scenario (S1) assumed that a large number of moderately abundant and rare OTUs are associated with the outcome, and the second scenario (S2) assumed the top 10 most abundant OTUs were associated with the outcome. Further, S1 has a binary outcome while S2 has a continuous outcome; both scenarios have confounders. S2 is as considered by Zhao et al. [3].

In S1, we assumed an equal number of cases and controls and denoted the case-control

status by Y. We simulated two covariates, $C_1 = 0.2Y + \epsilon_1$ and $C_2 = 0.1Y + 0.1C_1 + \epsilon_2$, where ϵ_1 and ϵ_2 were drawn independently from a uniform distribution $U[0, 1]$. We randomly sampled three sets of 428 OTUs (half of the total number of OTUs) to be associated with Y, C_1, and C_2; these sets of OTUs may overlap and the set for Y was sampled after excluding the top three most abundant OTUs. We denoted the OTU frequencies estimated from the real data by the vector π_1 and formed vectors π_2, π_3, and π_4 by first setting π_2, π_3 and π_4 equal to π_1 and then randomly permuting those frequencies in π_2, π_3, and π_4 that belong to the selected set of OTUs for Y, C_1, and C_2, respectively. Note that the frequencies for OTUs not selected to be associated with Y (or C_1, C_2) remain the same in π_1 and π_2 (or π_3, π_4). We then defined the sample-specific frequency vector as $\widetilde{\pi}(Y, C_1, C_2) = p_1(Y, C_1, C_2)\pi_1 + p_2(Y)\pi_2 + p_3(C_1)\pi_3 + p_4(C_2)\pi_4$, where $p_2(Y) = \beta Y$, $p_3(C_1) = 0.1C_1$, $p_4(C_2) = 0.1C_2$, and $p_1(Y, C_1, C_2) = 1 - p_2(Y) - p_3(C_1) - p_4(C_2)$, where β is a parameter we choose to adjust the strength of association between Y and OTU frequencies. We then generated the OTU count data for each sample using the DM model with $\widetilde{\pi}(Y, C_1, C_2)$ and overdispersion parameter of 0.02 and library size sampled from $N(10000, 10000/3)$. By mixing π_1, π_2, π_3, and π_4 in a way that depends on the values of Y, C_1, and C_2, we induced associations between the selected OTUs and Y, C_1, and C_2. Note that π_1 serves as the 'reference' OTU frequencies that characterizes samples for which Y, C_1, and C_2 are all zero. In addition, the correlations among Y, C_1, C_2, and OTU frequencies establish C_1 and C_2 as confounders of the association between Y and the OTUs. Finally, note that when $\beta = 0$, $\widetilde{\pi}(Y, C_1, C_2)$ does not depend on the case-control status Y, so that the null hypothesis of no association between Y and OTU frequencies holds.

In S2, the ten most abundant OTUs are associated with a continuous outcome. We first generated OTU counts for each sample using the DM model with frequency vector π_1, overdispersion parameter of 0.02, and library size sampled from $N(10000, 10000/3)$. We considered one confounder in this scenario, denoted by C, which has value $C = \text{scale}(S) + \widetilde{\epsilon}$, where S is the sum of the observed OTU frequencies at the ten OTUs, scale(v) centers and normalizes

vector v to have unit variance, and $\widetilde{\epsilon} \sim N(0,1)$. Finally, we simulated the continuous outcome Y by $Y = 0.5\text{scale}(C) + \beta\text{scale}(S) + \epsilon$, where $\epsilon \sim N(0,1)$. As in S1, note that when $\beta = 0$ there is no association between Y and the OTU frequencies.

We also generated clustered data by assuming that we had observations from 40 distinct individuals, 20 of whom contributed 2 observations and 20 contributed 3 observations (for a total of 100 observations). For S1 we assumed that 7 of the individuals contributing 2 observations and 12 of the individuals contributing 3 observations were cases; the remaining individuals were controls, yielding a total of 50 observations from cases and 50 observations from controls. We generated C_1 and C_2 at the individual level; to generate within-cluster association, we generated individual-specific OTU frequencies from the Dirichlet distribution with mean frequencies $\widetilde{\pi}(Y, C_1, C_2)$ and overdispersion parameter 0.02, and then generated counts for each observation 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 observation as described previously. For S2 we similarly generated individual-specific frequencies from the Dirichlet distribution with mean frequencies π_1 and overdispersion parameter 0.02 and then generated Multinomial data for each observation. We then redefined S to be the *averaged* sum of observed OTU frequencies at the top ten most abundant OTUs across repeated samples for an individual, and generated C and Y based on S at the individual level as before. Note that if each individual had a single observation, the combination of Dirichlet and Multinomial sampling would exactly reproduce the DM mixture model used for unclustered data.

Simulation results for the global test

For testing the global hypothesis of no association between microbiome composition and Y, i.e., $H_0 : \beta = 0$, 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. We also calculated the PERMANOVA test statistic using our approach to accounting for

confounders.

We first examined the results for unclustered data. Table 1 (upper panel) shows the type I error results using the Bray-Curtis distance. All methods, after adjusting for confounders, have correct type I error, even with a small sample size of 50. There was modest inflation of size for S1 and substantial inflation of size for S2 when confounders were not accounted for, demonstrating that our methods are effective in accounting for single or multiple confounders, with either modest or substantial confounding.

We evaluated the power of the global test using the Bray-Curtis, weighted-UniFrac, and Hellinger distance measures. Figure 1 shows that unlike PERMANOVA, the power of our methods are relatively insensitive to the choice of distance measure. Among our methods, 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 all 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. With the Bray-Curtis distance, PERMANOVA lost substantial power to VE-omni in both S1 and S2. The power loss is even more dramatic with the weighted-UniFrac distance, since the association was induced without reference to any phylogenetic tree. With the Hellinger distance, PERMANOVA had slightly higher but comparable power to VE-omni under S1 but nevertheless had markedly lower power under S2. Our sensitive analysis showed that the power gain of VE-omni over PERMANOVA persist for a wide range of library sizes, overdispersion parameters, and sample sizes (see Supplementary Text S1 and Supplementary Figure S1).

In the lower panel of Table 1, we can see that even in the presence of correlated observations, our methods and PERMANOVA all have correct type I error. We also calculated the size we would have obtained if we had incorrectly ignored the block structure when performing the permutations. Note that failure to account for the block structure result in a size of

100%. Supplementary Figure S2 shows that the power of the LDM-based methods relative to PERMANOVA was very similar to the simulation results we obtained with independent samples.

Simulation results for testing individual OTUs

Because PERMANOVA 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 [26], which was specifically developed for zero-inflated microbiome data.

The upper panel of Table 2 summarizes results for unclustered data. At the nominal FDR of 10%, the LDM-based methods have the correct empirical FDR (except for VE-arcsin under S2; however, the empirical FDR in this instance was based on only a few replicates that had non-zero detected OTUs and is thus not reliably estimated). Note also that the performance of VE-omni, as measured by the number of detected OTUs, sensitivity, and specificity, tracks VE-freq when it performs best or VE-arcsin when it performs best. In contrast, the empirical FDRs for DESeq2 are modestly inflated under S1 and highly inflated under S2.

Because MetagenomeSeq does not allow for adjustment of confounding covariates, we have not included results from this method in Table 2. To compare with MetagenomeSeq, we modified S1 and S2 to remove all confounders. In Supplementary Table S1 (upper panel), we show MetagenomeSeq was extremely conservative in detecting associated OTUs for the simulations we conducted. Since MetagenomeSeq was primarily developed for more sparse OTU data, we repeated our simulations but with mean library size of 1500 (as observed in the real URT microbiome data [22]). Even in this setting, MetagenomeSeq detected many fewer OTUs than VE-omni (see the lower panel of Supplementary Table S1).

Finally, in the lower panel of Table 2 we consider the effect of correlation on detection of individual OTUs. We find that while our approach continues to control FDR in the presence of correlation, the number of OTUs detected is smaller. This is reasonable, as data with within-

individual correlation is typically not as informative as data from an equivalent number of independent observations, for testing effects at the individual level. We only report results from S1 since the power to detect even the global null hypothesis was very limited for the clustered version of S2. Results from DESeq2 are unreliable as controlling for within-block correlation is not possible in DESeq2.

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 [27] (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 URT (i.e., upper-respiratory-tract) 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 (using our approach to accounting for confounders) and MiRKAT; we also compare 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 the mean library size 1500; metadata on smoking status (28 smokers and 32 nonsmokers) and two additional covariates (i.e., gender and antibiotic use within the last 3 months) was also available. The imbalance in the proportion of male subjects in smokers and nonsmokers, i.e., 75% and 56%, indicates potential for confounding. Ordination plots in Figure 2 using the Bray-Curtis, weighted-UniFrac, and Hellinger distances and after removing the effects of gender and antibiotic use (i.e., using $\widetilde{\Delta}_1$ as the distance matrix) demonstrate a clear shift in

smokers compared with nonsmokers. Zhao et al. (2015) used MiRKAT to find a significant global association between microbiome composition and smoking status after adjusting for gender and antibiotic use ($p = 0.002$ for Bray-Curtis distance, $p = 0.0048$ for weighted-UniFrac distance). The results of the LDM global tests, along with results from PERMANOVA and MiRKAT, are presented in the left panel of Table 3. VE-omni gave a smaller p-value than MiRKAT or PERMANOVA for each choice of distance.

In the right panel of Table 3, we show the results of our OTU-specific tests. VE-omni detected over 100 OTUs whereas DESeq2 detected none. The inefficiency of DESeq2 is consistent with our simulation study when the mean library size was 1500. To interpret the results we found, we grouped OTUs into 7 Phyla (and excluded 3 Phyla with fewer than 10 OTUs) and calculated the proportion of OTUs in each Phylum that we detected as significant. Using Fisher's exact test we found evidence that the proportion of detected OTUs varied significantly by Phylum ($p = 0.025$). Overall the proportion of significant OTUs was 0.149; *Proteobacteria* had a higher-than-expected proportion (0.248), while *Actinobacteria* and bacteria in Phylum *TM7* had lower-than-expected proportions (0.045 and 0.0). We tested the significance of these observations using a logistic regression in which the three Phyla were entered individually and the remaining Phyla were taken as the 'reference', and the Firth correction [28] was used to account for the zero rate among *TM7* bacteria; p-values in the last column of Supplementary Table S2 indicate that these observations are significant.

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 outcome (familial adenomatous polyposis/FAP and non-FAP) were also available. The data

also included nine gene principal components (gPCs) that explain 50% of the total variance in host gene expression. Zhan et al. [27] gave a joint test of all nine gPCs for association with microbiome composition, using the Bray-Curtis distance measure and adjusting for antibiotic use, inflammation score, and outcome (FAP/non-FAP). Here we performed the same joint test using the LDM, which includes all nine gPCs in the same model matrix M_2; however, we followed Morgan et al. (2015) to analyze only 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. VE-omni yielded the smallest p-value (0.0014), followed by PERMANOVA (0.0070) and then MMiRKAT (0.015). In addition, we detected 2 OTUs that significantly accounted for the global association by VE-omni at a nominal FDR rate of 10%.

We also followed Zhan et al. (2016) to conduct more detailed individual tests of each of the nine gPCs. We treated the nine gPCs in separate design matrices (i.e., gPC1–gPC9 in M_2–M_{10}) in a single LDM, accounting for the same confounders as in the joint test. Note that the gPCs are orthogonal, so that their order in the model is irrelevant. Table 4 shows that only gPC5 showed a significant association by any method ($p = 0.00036$ by VE-omni, $p = 0.017$ by PERMANOVA, $p = 0.031$ by MiRKAT). VE-omni only detected associated OTUs for gPC5, where the global test was highly significant ($p = 0.00036$). In contrast, DESeq2 detected OTUs for each of the nine gPCs, which seems implausible given the results from the global tests, and may be related to the failure of DESeq2 to control FDR in the presence of confounders as shown in our simulation studies.

Discussion

We have presented the LDM, a novel ordination-based 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, but also provides OTU-specific tests. This is true even when

confounding covariates are present, or when the design calls for correlated data. We have additionally shown that the OTUs identified by the LDM preserve FDR better than those identified by RNA-Seq-based approaches such as DESeq2; 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 no instance where the LDM discovered OTUs that were significantly associated with a variable but the LDM global test for that variable was non-significant; while some analysts may choose to only calculate OTU-specific tests in the presence of a significant global test, this restriction is not required. We have evaluated our approach using simulated data with realistic amounts of overdispersion and have shown how it can be applied to two real datasets.

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 blocks. 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 some 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 in important ways: first, the LDM uses a distance matrix to conduct a full-rank approxi-

mate decomposition to the OTU table X, while RA does not use a distance matrix and only decomposes the OTU table X into a part explained by variables and a residual matrix. Second, RA partitions variability across multiple sets of variables by a complex series of 'partial Redundancy Analyses' that alternatively assign sets of variables to be confounders or to be tested [14]; this approach may result in negative sums of squares assigned to certain variables or interactions, making it harder to assign a contribution from each OTU to the variance explained. It is possible to develop a non-standard variety of RA along the same lines as the LDM (see Supplementary Text S2), in which groups of variables are entered in a specific order; even though the resulting combinations of OTU frequencies (the directions that correspond to the matrices V in the LDM) are not typically orthogonal, the decomposition of sum of squares and contribution from each OTU is as described for the LDM. Further, confounders can be adjusted for in the same way as for the LDM. When we applied this approach, however, we found that the global tests of significance were less powerful than the LDM (Supplementary Figure S3 and Supplementary Table S3); further, fewer individual OTUs were found significant (Supplementary Table S4). We believe the increased power enjoyed by the LDM is because a full-rank approximate decomposition is formed; in fact, when we applied the LDM without the addition of the residual terms $B_r D_r V_r^T$, our results were similar (results not shown) to those we obtained from RA. In the absence of a full-rank decomposition, it appears that model variables are allowed to explain more of the variability found in the permuted datasets, resulting in lower power to detect real effects. Finally, distance-based Redundancy Analysis [29] does incorporate distance information, but like PERMANOVA, removes information on the effects of individual OTUs, and so was not included in our discussion.

In this report we have concentrated on distance measures that depend on OTU prevalence, such as the Bray-Curtis and weighted-UniFrac distances. 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. 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.

Conclusions

We propose the LDM, a method for testing association hypotheses for microbiome data. It integrates distance-based ordination, 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 more powerful than existing methods and thus 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 separate IRB consent.

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. 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.

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

6. 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.

7. Li J, Witten DM, Johnstone IM, Tibshirani R. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. Biostatistics. 2012;13(3):523–538.

8. 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.

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

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

11. Legendre P, Legendre LF. Numerical ecology. vol. 24. Elsevier; 2012.

12. Legendre P, Gallagher ED. Ecologically meaningful transformations for ordination of species data. Oecologia. 2001;129(2):271–280.

13. Everson R. Orthogonal, but not orthonormal, procrustes problems. Advances in computational Mathematics. 1998;3(4).

14. Rune Halvorsen Ø. Partitioning the variation in a plot-by-species data matrix that is related to n sets of explanatory variables. Journal of Vegetation Science. 2003;14(5):693–700.

15. Winkler AM, Ridgway GR, Webster MA, Smith SM, Nichols TE. Permutation inference for the general linear model. Neuroimage. 2014;92:381–397.

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. vol. 279. 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. Chen J, Chen L. GMPR: A novel normalization method for microbiome sequencing data. bioRxiv. 2017;p. 112565.

27. 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.

28. Firth D. Bias reduction of maximum likelihood estimates. Biometrika. 1993;80(1):27–38.

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

30. Bezdek JC, Hathaway RJ. Convergence of alternating optimization. Neural, Parallel & Scientific Computations. 2003;11(4):351–368.

Appendix 1. Directions removed from the row space of X when X is replaced by $\widetilde{X}$

We construct the directions Z_1 that are removed from the row space of X when we replace X by $\widetilde{X} = (I - B_1 B_1^T)X$. Following Satten et al. [4], let $Z_1 = X^- B_1$, where X^- is the Moore-Penrose generalized inverse of X. Here we can write $X^- = R\Sigma^{-1}L^T$ (given the SVD of $X = L\Sigma R^T$) where it is understood that the dimension of Σ is determined by the rank of X. We assume that $LL^T B_1 = B_1$ so that the column space of X contains the column space of B_1. With this assumption, $XZ_1 = B_1$ and $\widehat{X}_1 Z_1 = B_1 B_1^T X Z_1 = B_1 B_1^T B_1 = B_1$, so that $\widetilde{X} Z_1 = (X - \widehat{X}_1)Z_1 = 0$.

Appendix 2. Solving for constrained optimization

We first establish a preliminary lemma:

Lemma 1: If D and V minimize $||X - BDV^T||^2$ subject to $D_{ii} > 0$ (all diagonal elements of D are positive) and $V^T V = I$, and if $X = L\Sigma R^T$ by the SVD, then $V = RQ$ where $Q^T Q = I$.

Proof: We first note that the results in Bezdek and Hathaway [30] imply convergence of the tandem algorithm. We next show that $V = RQ$ for each step in the tandem algorithm. To see this, note that the step to update V given D is to write the SVD of $X^T BD$ as $X^T BD = L'\Sigma' R'^T$ and then choose $V = L'R'^T$. Using the SVD of $X = L\Sigma R^T$ we can also write $X^T BD = R\Sigma L^T BD$. If we write the SVD of $\Sigma L^T BD$ as $\Sigma L^T BD = L''\Sigma'' R''^T$, we have

$$X^T BD = (RL'')\,\Sigma'' R''^T. \tag{A.1}$$

Since $(RL'')^T (RL'') = I$, $(A.1)$ is an SVD of $X^T BD$ and so $V = (RL'')\,R''^T = R\left(L''R''^T\right) \equiv RQ$. Since the $\dim(D) \leq \dim(\Sigma)$, R'' is square and hence orthogonal, so that $Q^T Q = I$. $\square$

We now show that the constrained optimization in (5), (6) and (7) is solved by minimizing (7) subject to (6) without considering (5). Using the Lemma, the minimizer of $||\widetilde{X} - BDV^T||_F^2$, where $BDV^T = \sum_{k=2}^{K} B_k D_k V_k^T + B_r D_r V_r^T$, taken without considering the additional constraint (5) has the form $V = \widetilde{R}Q$, where $\widetilde{R}$ is the matrix of right singular values of $\widetilde{X}$. Since the rows of $\widetilde{X}$ are orthogonal to the columns of Z_1, this implies $Z_1^T V = 0$.

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

	Scenario	n	PERMANOVA	VE-freq	VE-arcsin	VE-omni
Unclustered data						
Adjusting for confounders	S1	50	0.046	0.048	0.048	0.049
		100	0.051	0.050	0.050	0.051
	S2	50	0.050	0.049	0.047	0.051
		100	0.051	0.049	0.051	0.053
Not adjusting for confounders	S1	50	0.056	0.054	0.063	0.060
		100	0.060	0.057	0.076	0.068
	S2	50	0.094	0.100	0.090	0.096
		100	0.171	0.222	0.167	0.208
Clustered data (adjusting for confounders)						
Accounting for block design	S1	100	0.048	0.047	0.049	0.050
	S2	100	0.049	0.049	0.050	0.050
Not accounting for block design	S1	100	1	1	1	1
	S2	100	1	1	1	1

n is the number of observations. Results are based on 10000 replicates, overdispersion parameter of 0.02, mean library size of 10000, and Bray-Curtis distance measure.

Table 2 Simulation results for detecting associated OTUs at the nominal FDR of 10%

		VE-freq			VE-arcsin			VE-omni			DESeq2		
	β	m	s-s	fdr	m	s-s	fdr	m	s-s	fdr	m	s-s	fdr
Unclustered data													
S1	0.5	32.6	(0.067, 0.990)	0.113	47.7	(0.104, 0.992)	0.07	48.1	(0.101, 0.988)	0.099	72.7	(0.127, 0.956)	0.254
	0.8	47.5	(0.103, 0.991)	0.077	77.5	(0.172, 0.989)	0.058	74.0	(0.162, 0.987)	0.071	113.8	(0.216, 0.948)	0.192
S2	3	4.1	(0.358, 0.999)	0.087	2.0	(0.145, 0.999)	0.177	4.1	(0.333, 0.999)	0.128	11.4	(0.065, 0.987)	0.942
	5	5.0	(0.436, 0.999)	0.084	2.5	(0.197, 0.999)	0.158	4.9	(0.410, 0.999)	0.124	11.9	(0.087, 0.987)	0.926
Clustered data													
S1	0.5	5.5	(0.013, 0.999)	0.026	19.4	(0.046, 0.997)	0.049	17.3	(0.041, 0.998)	0.039	187.6	(0.272, 0.801)	0.416
	0.8	21.2	(0.05, 0.998)	0.038	43.4	(0.102, 0.994)	0.047	39.8	(0.094, 0.995)	0.044	212.5	(0.327, 0.796)	0.377

β: effect size. m: number of detected OTUs. s-s: sensitivity and specificity formatted as (sensitivity, specificity). fdr: empirical FDR. Results are based on 1000 replicates, sample size of 100, overdispersion parameter of 0.02, mean library size of 10000, and Bray-Curtis distance measure.

Table 3 Results in analysis of the URT microbiome data

	Testing the global hypothesis					Testing individual OTUs			
Distance	MiRKAT	PERM	VE-freq	VE-arcsin	VE-omni	VE-freq	VE-arcsin	VE-omni	DESeq2
BC	0.0020	0.0020	0.0010	0.00031	0.00058	112	116	126	0
WU	0.0048	0.0045	0.0015	0.00031	0.00056	98	95	115	0
Hellinger	0.00076	0.00092	0.0017	0.00040	0.00078	134	118	150	0

BC: Bray-Curtis. WU: Weighted-UniFrac. PERM: PERMANOVA. The nominal significance level is 0.05. The nominal FDR is 10%.

Table 4 Results for testing individuals gPCs in analysis of the PPI microbiome data

	Testing the global hypothesis					Testing individual OTUs			
gPC	MMiRKAT	PERM	VE-freq	VE-arcsin	VE-omni	VE-freq	VE-arcsin	VE-omni	DESeq2
gPC1	0.18	0.18	0.17	0.36	0.23	0	0	0	13
gPC2	0.28	0.21	0.41	0.35	0.45	0	0	0	43
gPC3	0.28	0.28	0.27	0.23	0.32	0	0	0	19
gPC4	0.13	0.085	0.059	0.086	0.091	0	0	0	10
gPC5	0.031	0.017	0.0087	0.00020	0.00036	79	63	90	42
gPC6	0.39	0.39	0.69	0.34	0.43	0	0	0	19
gPC7	0.17	0.21	0.19	0.11	0.16	0	0	0	39
gPC8	0.28	0.30	0.72	0.67	0.77	0	0	0	1
gPC9	0.17	0.12	0.12	0.044	0.071	0	0	0	16

PERM: PERMANOVA. The nominal significance level is 0.05. The nominal FDR is 10%. The results are based on the Bray-Curtis distance measure.

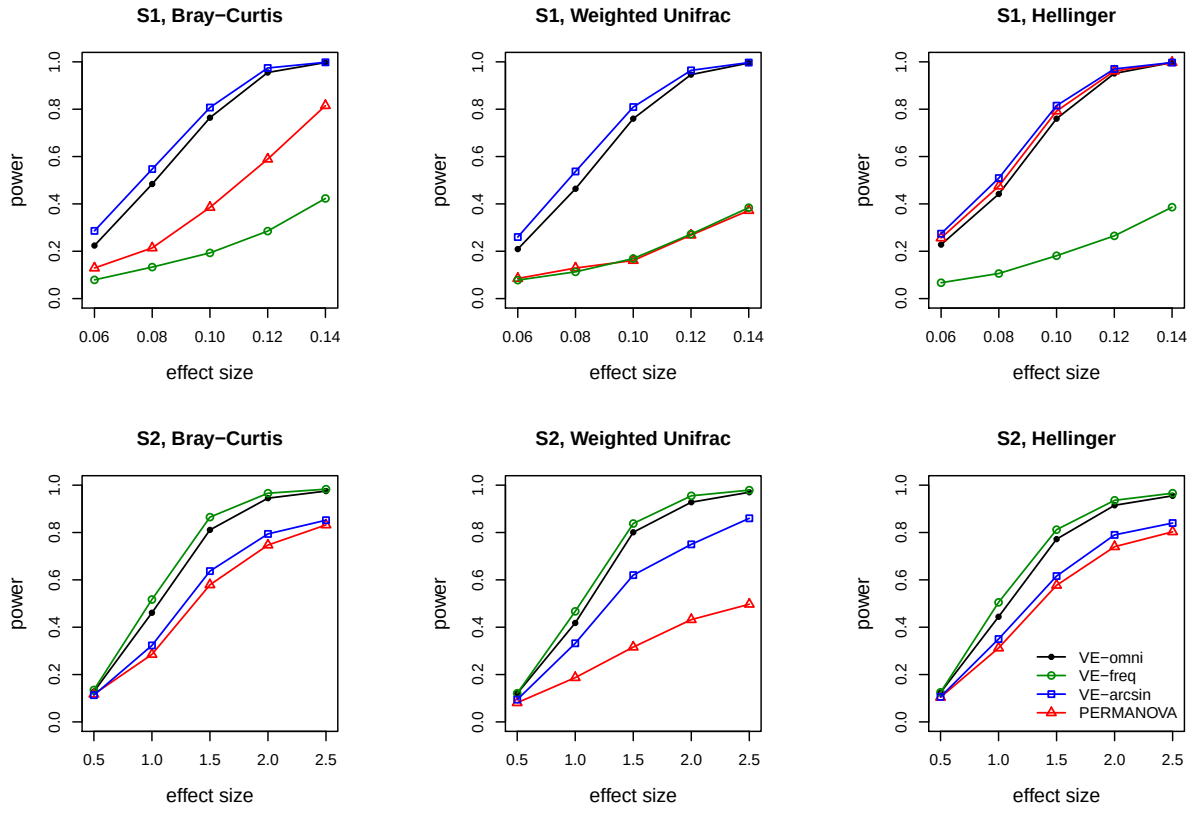

Fig. 1 Power for testing the global hypothesis at increasing effect size ($\boldsymbol{\beta}$) with unclustered data. Power estimates are based on 1000 replicates, overdispersion parameter of 0.02, mean library size of 10000, sample size (n) of 100, and nominal significance level of 0.05.

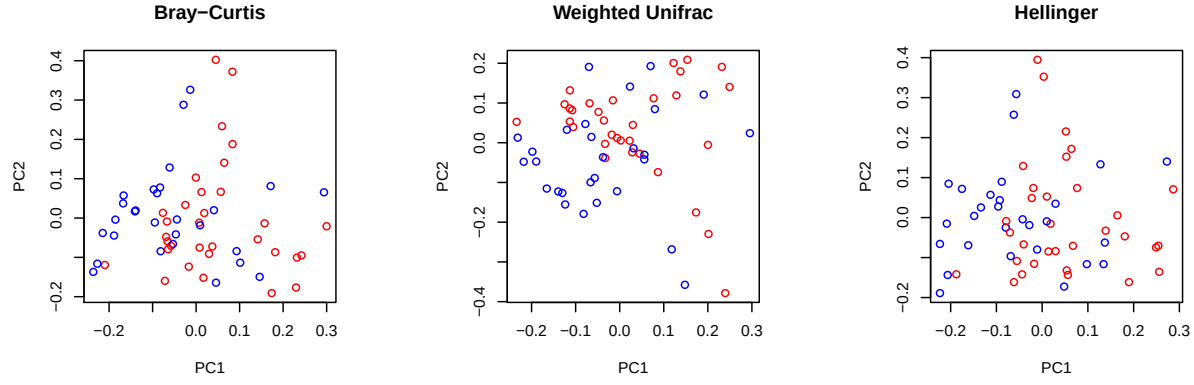

Fig. 2 Ordination plots from analysis of the URT microbiome data after removing the effects of gender and antibiotic use. The red and blue circles represent smokers and nonsmokers, respectively.