PNAS RESEARCH ARTICLE | STATISTICS

LOCOM: A logistic regression model for testing differential abundance in compositional microbiome data with false discovery rate control

Yingtian Hu[a], Glen A. Satten[b], and Yi-Juan Hu[a,1]

Edited by Kenneth Lange, University of California, Los Angeles, CA; received December 20, 2021; accepted June 14, 2022

Compositional analysis is based on the premise that a relatively small proportion of taxa are differentially abundant, while the ratios of the relative abundances of the remaining taxa remain unchanged. Most existing methods use log-transformed data, but log-transformation of data with pervasive zero counts is problematic, and these methods cannot always control the false discovery rate (FDR). Further, high-throughput microbiome data such as 16S amplicon or metagenomic sequencing are subject to experimental biases that are introduced in every step of the experimental workflow. McLaren et al. [*eLife* 8, e46923 (2019)] have recently proposed a model for how these biases affect relative abundance data. Motivated by this model, we show that the odds ratios in a logistic regression comparing counts in two taxa are invariant to experimental biases. With this motivation, we propose logistic compositional analysis (LOCOM), a robust logistic regression approach to compositional analysis, that does not require pseudocounts. Inference is based on permutation to account for overdispersion and small sample sizes. Traits can be either binary or continuous, and adjustment for confounders is supported. Our simulations indicate that LOCOM always preserved FDR and had much improved sensitivity over existing methods. In contrast, analysis of composition of microbiomes (ANCOM) and ANCOM with bias correction (ANCOM-BC)/ANOVA-Like Differential Expression tool (ALDEx2) had inflated FDR when the effect sizes were small and large, respectively. Only LOCOM was robust to experimental biases in every situation. The flexibility of our method for a variety of microbiome studies is illustrated by the analysis of data from two microbiome studies. Our R package LOCOM is publicly available.

log ratio | sparse data | pseudocount | experimental bias | logit model

Significance

High-throughput sequencing of 16S gene or metagenomes provides an unprecedented opportunity to discover microbes associated with traits such as clinical outcomes or environmental factors. However, the microbial data are highly complex because they are compositional, sparse (50 to 90% zeros), high-dimensional, and in particular subject to ubiquitous experimental biases. Existing methods developed specifically for compositional analysis of the microbiome data cannot always control the false discovery rate and often require replacing zeros with a pseudocount. Our proposed method, logistic compositional analysis (LOCOM), always preserves the false discovery rate, has much improved sensitivity over existing methods, does not require pseudocounts, and thus can accelerate the search for microbial biomarkers for prognosis and diagnosis of diseases or microbial targets for drug discovery.

Microbiome association studies are useful for the development of microbial biomarkers for prognosis and diagnosis of a disease or for the development of microbial targets (e.g., pathogenic or probiotic bacteria) for drug discovery, by detecting the taxa that are most strongly associated with the trait of interest (e.g., a clinical outcome or environmental factor). Read count data from 16S amplicon or metagenomic sequencing are typically summarized in a taxa count (or feature) table. Because the total sample read count (library size) is an experimental artifact, only the relative abundances of taxa, not absolute abundances, can be measured (1). Thus, microbial data are compositional (constrained to sum to 1). Analysis of microbial associations is further encumbered by data sparsity (having 50 to 90% zero counts in the taxa count table), high-dimensionality (having hundreds to thousands of taxa), and overdispersion. In addition, most microbiome association studies have relatively small sample sizes; further complications arise as the traits of interest may be either binary or continuous, and the detected associations may need to be adjusted for confounding covariates. Finally, any method for detecting taxon–trait associations should control the false discovery rate (FDR) (2). The capability to handle all these features is essential for any statistical method to be practically useful.

There are (at least) two biological models for how microbial communities may change when comparing groups with different phenotypes or along a phenotypic gradient. In one model, a substantial proportion of the taxa in the community change; the concept community state types exemplifies this approach (see, e.g., refs. 3, 4). The null hypothesis of no differential abundance that is tested at a taxon is that the taxon relative abundance remains the same; i.e., any change in taxon relative abundance across conditions is of interest. Methods for testing this hypothesis include the linear decomposition model (LDM) (5) and direct application of nonparametric tests (e.g., the Wilcoxon rank-sum test) to relative abundance data or rarefied count data. In the other model, only a few key taxa are considered to change, while the other taxa show changes in relative abundance because of the compositional constraint (6, 7). Thus, the null hypothesis that is tested

Author contributions: G.A.S. and Y.-J.H. designed research; Y.H., G.A.S., and Y.-J.H. performed research; Y.H. analyzed data; and G.A.S. and Y.-J.H. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: yijuan.hu@emory.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2122788119/-/DCSupplemental.

Published July 22, 2022.

at a taxon is that the ratio of the relative abundances at the taxon against some null taxon is unchanged. Methods for testing this hypothesis include analysis of composition of microbiomes (ANCOM) (8), ANCOM with bias correction (ANCOM-BC) (9), ANOVA-Like Differential Expression tool (ALDEx2) (10), WRENCH (6), and Testing for Differential Abundance in Compositional Counts Data (DACOMP) (7). Because the hypothesis in the second model accounts for the compositional constraint that a change in relative abundance for one taxon necessarily implies a counterbalancing change in other taxa, it is generally referred to as compositional analysis (11).

Methods for compositional analysis are typically based on some form of log-ratio transformation of the read count data. The ratio can be formed against a reference taxon or the geometric mean of relative abundances of all taxa, referred to as additive log-ratio (alr) or centered log-ratio (clr) transformation, respectively (12). Thus, zero count data, which cannot be log-transformed, is the major challenge in using compositional methods on microbiome data. A common practice is to add a pseudocount, most frequently 1 or 0.5 or even smaller values, to the zeros or all entries of the taxa count table (8, 9, 12–15). However, there is no consensus on how to choose the pseudocount, and it has been shown that the choice of pseudocount can affect the conclusions of a compositional analysis (16, 17).

The most popular pseudocount-based method for compositional analysis is perhaps ANCOM (8), which has now evolved into ANCOM-BC (9). After adding 0.001 to all count data, ANCOM performs the alr transformation and treats the transformed data as the response of the linear regression model that includes the traits of interest and confounding variables as covariates. For each taxon, ANCOM uses all other taxa, one at a time, as the reference in forming the alr transformation, and then it employs a heuristic strategy to declare taxa that are significantly differentially abundant (outputting rankings of taxa instead of P values). ANCOM-BC first estimates sampling fractions that are different across samples and then models the log of read count data, in which zeros are replaced by pseudocount 1, through a linear regression model including the estimated sampling fraction as an offset term. This is essentially a normalization approach that first attempts to recover the absolute abundances of taxa and then test hypotheses about the absolute abundances. Unlike ANCOM, ANCOM-BC provides P values for individual taxa. Both ANCOM and ANCOM-BC are restricted to group comparisons and cannot handle continuous traits of interest, although adjustment for confounding covariates is supported.

Several methods have been developed that circumvent the use of pseudocount. ALDEx2 (10) first draws Monte Carlo samples of nonzero relative abundances from Dirichlet distributions (with parameters constructed from read count data plus a uniform prior 0.5). Then, the sampled relative abundances are clr transformed and tested against the traits of interest via linear regression to yield P values and adjusted P values by the Benjamini–Hochberg (BH) procedure (18), both of which are averaged over sampling replicates to give the final P values and adjusted P values. However, the sampling process adds noise to the data, which may cause loss of power. In addition, by using the clr transformation, ALDEx2 is designed to identify differential abundant taxa relative to the mean of all taxa, which may be sensitive to outliers. DACOMP (7) is a normalization approach that first selects a set of null reference taxa by a data-adaptive procedure and then normalizes read count data by rarefaction so that each taxon within the reference has similar counts across samples. However, the selected reference set may mistakenly contain causal taxa, which may compromise the performance of the normalization. In addition, adjustment for confounding covariates is not supported, although continuous traits of interest are allowed. WRENCH (6) is also a normalization approach that estimates group-specific compositional factors to bring the read counts of null taxa across groups to a similar level and employs differential expression analysis based on the negative binomial distribution (DESeq2) to detect differentially abundant taxa. It is limited to group comparisons without confounding covariates. A general method that can be used for all types of microbiome data without introduction of pseudocounts is thus an important goal.

It is also of interest to test differential abundance at the community (i.e., global) level, rather than taxon by taxon, using the compositional analysis approach. The most commonly used method for testing community-level hypotheses about the microbiome is permutational multivariate analysis of variance (PERMANOVA) (19), which is a distance-based version of ANOVA. For compositional analysis, use of the Aitchison distance is recommended (11), which is simply the Euclidean distance applied to the clr transformed data (20). Again, the clr transformation necessitates the use of pseudocount, so the choice of pseudocount may affect the outcome of the test.

Finally, it is of vital interest to develop a method that can provide valid inference even in the presence of experimental bias. Experimental bias is ubiquitous because each step in the sequencing experimental workflow (i.e., DNA extraction, PCR amplification, amplicon or metagenomic sequencing, and bioinformatics processing) preferentially measures (i.e., extracts, amplifies, sequences, and bioinformatically identifies) some taxa over others (1, 21–23). For example, bacterial species differ in how easily they are lysed and therefore how much DNA they yield during DNA extraction (24). As a result, the bias distorts the measured taxon relative abundances from their actual values.

We are particularly interested in the case of differential bias, where the bias of taxa that are associated with a trait is systematically different from the bias of null taxa. A concrete example of this is the differential bias between bacteria in the phyla *Bacteroidetes* and *Firmicutes*. *Bacteroidetes* are gram-negative, while *Firmicutes* are gram-positive. It is known that gram-positive bacteria have strong cell walls and are hence harder to lyse than gram-negative bacteria; thus, gram-positive bacteria may be underrepresented due to bias in the extraction step. The *Bacteroidetes–Firmicutes* ratio has been implicated in a number of studies of the gut microbiome (e.g., refs. 25, 26). Thus, studies that compare *Bacteroidetes* to *Firmicutes* may be affected by differential extraction bias. In some of our simulations, we consider the effect this kind of differential bias can have on the FDR.

In this article, we develop a method for compositional analysis of differential abundance, at both the taxon level and the global level, based on a robust version of logistic regression that we call logistic compositional analysis (LOCOM). Our method circumvents the use of pseudocount, does not require the reference taxon to be null, and does not require normalization of the data. Further, it is applicable to a variety of microbiome studies with binary or continuous traits of interest and can account for potentially confounding covariates. In *Method*, we give the motivation for using logistic regression as a way to minimize the effect of experimental bias in analyzing microbiome data and describe the details of our approach. In *Results*, we present simulation studies that compare the performance of LOCOM to other compositional methods. We also compare results from LOCOM and other methods in the analysis of two microbiome datasets. We conclude with *Discussion*.

Method

Let Y_{ij} be the read count of the jth taxon ($j = 1, \ldots, J$) in the ith sample ($i = 1, \ldots, n$) and N_i the library size of the ith sample. Because N_i can vary widely between samples, we focus on the relative abundance data as a form of normalized data. We denote by P_{ij} the observed relative abundance, given by Y_{ij}/N_i. We let X_i be a vector of q covariates including the (possibly multiple) traits of interest and other (confounding) covariates that we wish to adjust for but excluding the intercept.

Motivation. Our starting point is the model of McLaren et al. (1), as expanded by Zhao and Satten (27), which relates the expected value of the observed relative abundance, denoted by p_{ij}, to the true relative abundance we would measure in an experiment with no experimental bias, denoted by π_{ij}. In particular, this model assumes that

$$\log(p_{ij}) = \log(\pi_{ij}) + \gamma_j + \alpha_i, \qquad \textbf{[1]}$$

where γ_j is the taxon-specific bias factor that describes how the relative abundance is distorted by the bias and α_i is the sample-specific normalization factor that ensures the composition constraint $\sum_{j=1}^{J} p_{ij} = 1$. Following ref. 27, we further assume that the true relative abundance π_{ij} can be described by a baseline relative abundance π_j^0 that would characterize the true relative abundance of taxon j for a sample having $X_i = 0$ and a term that describes how the baseline relative abundance is changed in the presence of covariates $X_i \neq 0$. Then, we can replace Eq. **1** by

$$\log(p_{ij}) = \log(\pi_j^0) + X_i^{\mathrm{T}}\beta_j + \gamma_j + \alpha_i, \qquad \textbf{[2]}$$

where β_j describes the way the true relative abundance changes with covariates X_i and is our parameter of interest. The presence of bias factors in Eqs. **1** and **2** implies that inference based on the observed relative abundances P_{ij} may not give valid inference on β_j. It is clear that without knowing the bias factor γ_j, we cannot estimate $\log(\pi_j^0)$ as $\log(\pi_j^0)$ and γ_j always appear together as a sum.

We can examine Eq. **2** to see if there are any combinations of parameters that could potentially be estimated without knowing the bias factors. Analyzing log probability ratios such as $\log(p_{ij}/p_{ij'})$ removes the effect of α_i (which depends on bias factors through normalization) but does not remove the effect of γ_j. However, if we use Eq. **2** to write log odds ratios of observed relative abundances for two different taxa and two different samples, we find

$$\log\left(\frac{p_{ij}p_{i'j'}}{p_{ij'}p_{i'j}}\right) = (X_i - X_{i'})^{\mathrm{T}}(\beta_j - \beta_{j'}), \qquad \textbf{[3]}$$

which is independent of bias factors. This motivates the choice of logistic regression to analyze microbiome count data.

Note that testing $\beta_j - \beta_{j'} = 0$ in Eq. **3** corresponds to testing $p_{ij}/p_{ij'} = p_{i'j}/p_{i'j'}$, which is exactly the null hypothesis in a compositional analysis, e.g., in popular compositional models of the microbiome such as ANCOM and ALDEx2. As a result, logistic regression based on Eq. **3** is of interest even without the bias-removal motivation provided here.

Multivariate Logistic Regression Model. Eq. **3** implies a polychotomous logistic regression of the full $n \times J$ taxa count table. This is numerically difficult as the analysis of each taxon potentially requires all β_j parameters. Instead, we follow Begg and Gray (28) and analyze data using separate or individualized logistic regressions, each using data from just two taxa at a time. Rather than considering all possible pairs of taxa, we choose one taxon (without loss of generality, the Jth taxon) to be a reference taxon and compare all other taxa to the reference taxon. Then, if we define $\mu_{ij} = p_{ij}/(p_{ij} + p_{iJ})$, Eq. **2** implies

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \theta_j + X_i^{\mathrm{T}}(\beta_j - \beta_J), \qquad 1 \le j \le J-1, \quad \textbf{[4]}$$

where the intercepts $\theta_j = \left[\log(\pi_j^0) - \log(\pi_J^0)\right] + (\gamma_j - \gamma_J)$ are treated as nuisance parameters since estimation of the γ_j values is not possible when the π_j^0 values are not known. As written, the model is overparameterized because only the $J-1$ log odds ratios $\beta_j - \beta_J$ are identifiable. To make the full set of β_j values identifiable requires a constraint; we temporarily use $\beta_J = 0$ with the understanding that β_j then refers to an odds ratio that compares taxon j to the reference taxon J. According to ref. 28, the efficiency of individualized logistic regression highly depends on the prevalence (relative abundance) of the reference category, so we recommend that the reference taxon be a common taxon that is present in a large number of samples.

To avoid distributional assumptions in a standard logistic regression, we consider the score functions as estimating functions. When a taxon is rare and/or the sample size is small, it may occur that all (or nearly all) counts for that taxon are zero in one group (e.g., the case or control group), which is referred to as separation in the literature on logistic regression. It is known that the Firth bias correction (29), when applied to logistic regression (30), solves the problem of separation. Hence, we estimate (θ_j, β_j) by solving the Firth-corrected score equation

$$U_j(\theta_j, \beta_j) = \sum_{i=1}^{n}\left[Y_{ij} - M_{ij}\mu_{ij} + h_i(0.5 - \mu_{ij})\right]\begin{pmatrix}1\\ X_i\end{pmatrix} = 0,$$

where $M_{ij} = Y_{ij} + Y_{iJ}$ and h_i is the ith diagonal element of the weighted hat matrix $W_j^{\frac{1}{2}}X(X^{\mathrm{T}}W_jX)^{-1}X^{\mathrm{T}}W_j^{\frac{1}{2}}$ with the design matrix X (including a column of ones corresponding to the intercept) and the diagonal weight matrix $W_j = \mathrm{diag}\{M_{1j}\mu_{1j}(1-\mu_{1j}), \ldots, M_{nj}\mu_{nj}(1-\mu_{nj})\}$. Note that the zero count data Y_{ij} are used to build these equations in a natural and systematic way; in particular, zero counts get low weight rather than the upweighting of zero counts that can occur when taking the log of a small pseudocount value. We let $\widehat{\beta}_j$ denote the estimator of β_j obtained by solving the above equation.

Testing Hypotheses at Individual Taxa. Now we describe the formula for the null hypotheses we test to decide which taxa are null, i.e., have no effect. Write $\beta_j = (\beta_{j,1}, \beta_{j,-1})$, where $\beta_{j,1}$ is the coefficient for the trait of interest and $\beta_{j,-1}$ for the other covariates. We assume the trait of interest has only one component, but the approach can be generalized to test multiple traits simultaneously (*Discussion*). The naive formula $\beta_{j,1} = 0$ only implies that the effect of the trait on the jth taxon is the same as the effect of the trait on the reference taxon; thus, testing $\beta_{j,1} = 0$ only identifies null taxa when the reference taxon used in Eq. **4** is itself null.

As we have no a priori knowledge about whether the reference taxon is null or causal, we seek an approach that does not require such knowledge; in addition, we need a test for the reference taxon itself. To this end, we make the assumption that more than half of the taxa are null taxa, which has been frequently adopted in compositional methods (6, 7). With this assumption, we can expect $\mathrm{median}_{j'=1,\ldots,J}\{\beta_{j',1}\}$ to correspond to the value of $\beta_{j^*,1}$ for some taxon j^* that is null. If we then consider parameters $\widetilde{\beta}_{j,1} = \beta_{j,1} - \mathrm{median}_{j'=1,\ldots,J}\{\beta_{j',1}\} = \beta_{j,1} - \beta_{j^*,1}$ in place of parameters $\beta_j - \beta_J$ in Eq. **4**, then using $\widetilde{\beta}_{j,1} = 0$ as a null hypothesis does correspond to testing whether taxon j is null. Thus, we wish to test the null hypotheses

$$H_{j0} : \beta_{j,1} - \mathrm{median}_{j'=1,\ldots,J}\{\beta_{j',1}\} = 0.$$

Note that centering by the median can also be thought of as replacing the constraint $\beta_{J,1} = 0$ to identify $\beta_{j,1}$ for all j. To test these null hypotheses, we use the statistic

$$\mathbb{Z}_j = \widehat{\beta}_{j,1} - \mathrm{median}_{j'=1,\ldots,J}\{\widehat{\beta}_{j',1}\}.$$

Note that we use $\widehat{\beta}_{J,1} = 0$ both when calculating the median and obtaining $\mathbb{Z}_J$ for the reference taxon. Also note that the $\mathbb{Z}_j$ values are reminiscent of centered log-ratios, in which the log of the abundance is centered by the mean of the log abundances. Use of the median in place of the mean for our centering is advantageous as the mean is sensitive to large or outlying observations that do not affect the median. Since the odds ratios we estimate each use data from only two taxa, our method is subcompositionally coherent in the sense of ref. 31.

In the simplest case testing a binary trait that takes values 0 and 1, with no other covariates, $\mathbb{Z}_j$ is invariant to different choices of the reference taxon. This is because in this simple case, all estimated (pairwise) log odds ratios are of the form $(\widehat{\beta}_{j,1} - \widehat{\beta}_{j',1}) = \log\{n_{1j}n_{0j'}/(n_{0j}n_{1j'})\}$, where $n_{xj} = \sum_{i:X_i=x} Y_{ij}$ and so are completely free of the reference taxon. This holds even if the Firth-corrected

estimator is used because in this simple case, the Firth-corrected estimator corresponds to adding 1/2 to each n_{xj} (29, 30); note that n_{xj} is an aggregated read count in a group of samples, and thus, this result is fundamentally different from the aforementioned pseudocount approach that adds a pseudocount to each read count of a sample. For the general case, we evaluate the dependence of $\mathbb{Z}_j$ on the reference taxon via simulations.

To avoid distributional assumptions in sparse microbiome data, we assess the significance of $\mathbb{Z}_j$ using the permutation scheme for logistic regression proposed by Potter (32), which is described as follows. The covariate vector X_i is partitioned into (T_i, C_i), where T_i denotes the trait of interest and C_i the other covariates. A linear regression of T_i on C_i and an intercept is fit to obtain the residual T_{ir}, which is then permuted to obtain $T_{ir}^{(b)}$ and to construct the new covariate vector $X_i^{(b)} = (T_{ir}^{(b)}, C_i)$. We follow the same procedure as for the observed dataset to obtain the estimate of $\beta_{j,1}$ from the bth permutation replicate, denoted by $\widehat{\beta}_{j,1}^{(b)}$, and the corresponding statistic $\mathbb{Z}_j^{(b)} = \widehat{\beta}_{j,1}^{(b)} - \text{median}_{j'}\{\widehat{\beta}_{j',1}^{(b)}\}$. We adopt Sandve's sequential stopping rule (33) with a minor modification to stop the permutation procedure, which is described below. For each taxon j, after the Bth permutation we store the (cumulative) number of times that $\mathbb{Z}_j^{(b)}$ falls on the left (i.e., is less than) and right side (i.e., is greater than) of $\mathbb{Z}_j$, which we denote by L_j and R_j, respectively. We count the number of rejections to be $2\min(L_j, R_j)$. The P value based on B permutations is given by $p_j = [2\min(L_j, R_j) + 1]/(B + 1)$, and the q value is calculated according to (33). The permutation procedure is continued until every taxon either has a q value below the nominal FDR level or has accumulated a number of rejections exceeding a prespecified value (e.g., 100). This stopping rule is slightly different from Sandve's in that we obtain $\widehat{\beta}_{j,1}^{(b)}$ for every taxon at every permutation, rather than stopping permutation early for some taxa, because the median calculation requires $\widehat{\beta}_{j,1}^{(b)}$ from all taxa.

Testing the Global Hypothesis. The global null hypothesis is that there are no differentially abundant taxa; i.e., H_{j0} holds for every taxon. Given the P values at individual taxa, it is straightforward to construct a global test statistic by combining the individual P values. Here we adopt the harmonic-mean approach to combining P values proposed by Wilson (34), which is more robust to the dependence structure among taxa than Fisher's method and has more focus on the smallest P value(s) (i.e., more power for scenarios with sparse, strong signals) than Fisher's method. The harmonic mean of the p_j values is $J/\left(\sum_{j=1}^{J} p_j^{-1}\right)$, for which smaller values correspond to stronger evidence against the null hypothesis. To have a test statistic with the "usual" directionality, we choose $\mathbb{Z}_{\text{global}} = \sum_{j=1}^{J} p_j^{-1}$. We use all permutation replicates generated for taxon-level tests, say B replicates, to assess the significance of $\mathbb{Z}_{\text{global}}$. At the bth replicate, the test statistic is $\mathbb{Z}_{\text{global}}^{(b)} = \sum_{j=1}^{J}\{p_j^{(b)}\}^{-1}$, where $p_j^{(b)}$ is the P value of taxon j for this null replicate. Following ref. 35, we calculate the null P value $p_j^{(b)}$ using the rank statistic to be $p_j^{(b)} = 2B^{-1}\min\left\{\left[\text{rank}\left(\mathbb{Z}_j^{(b)}\right) - 0.5\right], \left[B - \text{rank}\left(\mathbb{Z}_j^{(b)}\right) + 0.5\right]\right\}$, where $\text{rank}\left(\mathbb{Z}_j^{(b)}\right)$ is the rank of $\mathbb{Z}_j^{(b)}$ among B such statistics. Let R_{global} be the number of times that $\mathbb{Z}_{\text{global}}^{(b)}$ falls on the right-hand side of $\mathbb{Z}_{\text{global}}$. Then, the global P value is given by $\left(R_{\text{global}} + 1\right)/(B + 1)$.

Results

Simulation Studies. We used simulation studies to evaluate the performance of LOCOM and compare its performance to other compositional analysis packages. We based our simulations on data on 856 taxa of the upper respiratory tract (URT) microbiome; these taxa correspond to the "OTUs" in the original report on these data by Charlson et al. (36). We considered both binary and continuous traits of interest and both binary and continuous confounders, as well as the case of no confounder. We mainly focused on two causal mechanisms. For the first mechanism (referred to as M1), we randomly sampled 20 taxa (after excluding the most abundant taxon) whose mean relative abundances were greater than 0.005 as observed in the URT data (i.e., ranking among the top 40 most abundant taxa) to be causal (i.e., associated with the trait of interest). For the second mechanism (referred to as M2), we selected the top five most abundant taxa (having mean relative abundance 0.105, 0.062, 0.054, 0.050, and 0.049) to be causal. In some cases, we also considered two variations of M1, one randomly sampling 500 taxa (again excluding the top one) to be causal to create a scenario that violated our assumption that more than half of the taxa are null and one randomly sampling 20 rare taxa (whose mean relative abundances were between 0.001 and 0.002) to be causal, which are referred to as M1-500 and M1-rare, respectively. For simulations with a confounding covariate, we assumed the confounder was associated with 20 taxa under M1 (10 sampled at random from the 20 causal taxa and 10 from the null taxa) and 5 taxa under M2 (2 from the 5 causal taxa and 3 from the null taxa). We simulated most data without adding experimental bias but did conduct one set of simulations having differential experimental bias. We focused on datasets having 100 observations but also simulated some datasets with 50 or 200 observations.

To be specific, we let T_i denote the trait and C_i the confounder for the ith sample. To generate a binary trait, we selected an equal number of samples with $T_i = 1$ and $T_i = 0$. When a binary confounder was present, we drew C_i from the Bernoulli distribution with probability 0.2 in samples with $T_i = 0$ and from the Bernoulli distribution with probability 0.8 in samples with $T_i = 1$. When a continuous confounder was present, we drew C_i from the uniform distribution $U[-1, 1]$ in samples with $T_i = 0$ and $U[0, 2]$ in samples with $T_i = 1$. To generate a continuous trait, we sampled it from $U[-1, 1]$ when there was no confounder. When there was a binary confounder, we used the aforementioned data generated for a binary trait and a continuous confounder but exchanged the roles of trait and confounder. When there was a continuous confounder, we generated T_i from $U[-1, 1]$ and a third variable Z_i from $U[-1, 1]$ independently of T_i and then constructed the confounder $C_i = \rho T_i + \sqrt{1-\rho^2} Z_i$, where ρ was fixed at 0.5.

To simulate read count data for the 856 taxa, we first sampled the baseline (when $T_i = 0$ and $C_i = 0$) relative abundances $\pi_i^{(0)} = \left(\pi_{i1}^{(0)}, \pi_{i2}^{(0)}, \ldots, \pi_{iJ}^{(0)}\right)$ of all taxa for each sample from the Dirichlet distribution $Dirichlet(\overline{\pi}, \theta)$, where the mean parameter $\overline{\pi}$ and overdispersion parameter θ took the estimated mean and overdispersion (0.02) from fitting the Dirichlet-multinomial (DM) model to the URT data. We formed the relative abundances p_{ij} for all taxa by spiking the j' th causal taxon with an $\exp(\beta_{j',1})$ fold change and the j'' th confounder-associated taxon with an $\exp(\beta_{j'',2})$ fold change and then renormalizing the relative abundances, so that

$$p_{ij} = \frac{\exp\left(\gamma_j + \beta_{j,1} T_i + \beta_{j,2} C_i\right)\pi_{ij}^{(0)}}{\sum_{j'=1}^{J} \exp\left(\gamma_{j'} + \beta_{j',1} T_i + \beta_{j',2} C_i\right)\pi_{ij'}^{(0)}}, \qquad [5]$$

where γ_j was the bias factor for the jth taxon. Note that $\beta_{j,1} = 0$ for null taxa, $\beta_{j,2} = 0$ for confounder-independent taxa, and $\gamma_j = 0$ for all taxa for data without experimental bias. In most cases, for simplicity, we set $\beta_{j,1} = \beta$ for all causal taxa, and thus, β is a single parameter that we refer to as the effect size; we refer to $\exp(\beta)$ as the fold change. In some cases, we also considered the more general scenario when different values were sampled for different $\beta_{j,1}$. We fixed $\beta_{j,2} = \log(2)$ for all confounder-associated taxa. When there was no confounder, we simply dropped the term $\beta_{j,2} C_i$ (or equivalently, set $\beta_{j,2} = 0$ for all j values) in calculating p_{ij}. In cases with differential experimental bias, we drew γ_j from $N(0, 0.8^2)$ for noncausal taxa and from $N(1, 0.8^2)$

for causal taxa; thus, the bias-related fold changes varied roughly between 0.2 and 5 for most (95%) noncausal taxa and between 0.55 and 13.5 for most causal taxa, which are within a reasonable range according to ref. 1. Finally, we generated the taxon count data for each sample using the multinomial model with mean $p_i = (p_{i1}, p_{i2}, \ldots, p_{iJ})$ and library size sampled from $N(10000, (10000/3)^2)$ and left-truncated at 2,000.

In order to evaluate the robustness of our simulation results, we changed our simulation procedure in the following ways. First, we replaced the compositional model of Eq. **5** by a model that generates microbiome–trait associations by assigning differential relative abundances only at causal taxa (which corresponds to the first biological model introduced at the beginning and constitutes a misspecified model for LOCOM and other compositional analyses). Second, we replaced the DM model by a Poisson log-normal mixture (PLNM) model (which can impose any prespecified correlation structure across taxa) for generating read count data. Both replacement models are described in our LDM paper (supplementary text S2 in ref. 5). We also followed the LDM paper by basing our simulations on its association scenarios, which were denoted by S1 and S2. Scenario S1 assumed a large number of causal taxa (428 taxa in the LDM paper, which we modified here to 500 to create violation of our assumption that fewer than half the taxa are causal). Scenario S2 chose the top 10 most abundant taxa to be causal; here we will refer to the two scenarios as S1-500 and S2. Note that the data simulated using the PLNM model appeared to be less overdispersed and less sparse compared to data simulated using the DM model.

We applied two versions of LOCOM: one used the most abundant null taxon as the reference, which is referred to as LOCOM-null, and one used the most abundant causal taxon as the reference, referred to as LOCOM-causal. Both versions use the median of $\widehat{\beta}_{j',1}$ values to compute the test statistic. Of course, in a real application, we would not know whether or not the reference taxon we had chosen was null or causal; we differentiate these two versions of LOCOM here to show that LOCOM is robust to whether the reference taxon is null or causal. In practice, when the most abundant taxon is chosen as the reference, the results from LOCOM would correspond to LOCOM-null in M1 and to LOCOM-causal in M2.

For testing the global hypothesis, we compared LOCOM to PERMANOVA (the adonis2 function in the vegan R package) based on the Aitchison distance, which is referred to as PERMANOVA-half and PERMANOVA-one corresponding to adding pseudocount 0.5 and 1, respectively, to all cells. The type I error and power of the global tests were assessed at the nominal level 0.05 based on 5,000 and 1,000 replicates of data, respectively.

For testing individual taxa, we compared LOCOM to ANCOM, ANCOM-BC, ALDEx2, DACOMP, and WRENCH. However, ANCOM, ANCOM-BC, and WRENCH cannot handle continuous traits; DACOMP and WRENCH cannot adjust for other covariates. Prior to analysis, we removed taxa having fewer than 20% presence (i.e., present in fewer than 20% of samples) in each simulated dataset. For ANCOM and ANCOM-BC, we also considered their own filtering criterion with 10% presence as the cutoff and refer to these methods as ANCOM^o and ANCOM-BC^o. In the case with a binary trait only, we considered two additional pseudocount-based methods, Wilcox-alr-half and Wilcox-alr-one, which add pseudocounts 0.5 and 1, respectively, to all cells, form the alr using the most abundant null taxon as the reference, perform the Wilcoxon rank-sum test at individual log ratios, and correct multiple comparisons using the BH procedure. Because the reference was selected to be a taxon known to be null, these methods are not applicable to real studies but were included in the simulations here to assess the properties of the pseudocount approach to testing individual taxa. In the case with a binary trait only, we also applied the Wilcoxon test directly to relative abundance data, i.e., data with total-sum scaling (TSS); although not a compositional method, this is commonly used in microbiome studies. The sensitivity (proportion of truly causal taxa that were detected) and empirical FDR were assessed at nominal level 20% based on 1,000 replicates of data. We chose a relatively high nominal FDR level because the numbers of causal taxa in both M1 and M2 were small. In some cases, we also considered a lower nominal FDR level of 10%.

Simulation Results. The type I errors of the global tests for all simulation scenarios are summarized in *SI Appendix*, Table S1. In all scenarios, LOCOM-null and LOCOM-causal yielded type I error rates that were close to the nominal level and generally closer for sample size 200 than 100. Note that in cases when there was a confounder, there was substantial inflation of type I error when the confounder was not accounted for (*SI Appendix*, Table S2), demonstrating that LOCOM is effective in adjusting for confounders. The PERMANOVA tests also controlled type I error. In cases without any confounder, the zero data were similarly distributed across trait values under the (global) null, so the effect of adding pseudocount is nondifferential. In cases with a confounder, the taxa associated with the confounder caused the zeros to be differentially distributed across trait values, so that adding pseudocount had a differential effect for different trait values; however, this difference was adjusted by including the confounder as a covariate in the model. Note that although the pseudocount approach did not lead to invalid global tests, it did lead to invalid tests at individual taxa (in the presence of causal taxa), as indicated in the empirical FDR of Wilcox-alr-one and Wilcox-alr-half (e.g., Fig. 1).

Figs. 1–4 present power of the global tests and sensitivity and empirical FDR of the individual taxon tests, for a binary or continuous trait without and with a binary confounder, in scenarios M1 and M2 without experimental bias. The results for cases with a continuous confounder are deferred to *SI Appendix*, Figs. S1 and S2, which show similar patterns of results to their counterparts with a binary confounder (Figs. 2 and 4). The results in Figs. 1–4 all have sample size 100 and FDR level 20%. To explore the effects of changing sample size and FDR level, we restricted to the two most important scenarios, one with a binary trait with no confounder in which all methods are applicable and one with a binary trait and a binary confounder, which is very common in real data. We changed the sample size to 50 (*SI Appendix*, Figs. S3 and S4) or 200 (*SI Appendix*, Figs. S5 and S6), then changed the nominal FDR level to 10% (*SI Appendix*, Figs. S7 and S8). In general, those results show similar patterns to their counterparts with sample size 100 and nominal FDR level 20%.

In the simplest scenario with a binary trait and no confounder (Fig. 1 and *SI Appendix*, Figs. S3 and S5), LOCOM-null and LOCOM-causal yielded identical type I error and power; in fact, the two methods gave identical P values for every dataset in this case, which corroborated our claim that the test is invariant to different reference taxa. In other scenarios, LOCOM-null and LOCOM-causal produced similar results, although the one using the more abundant taxon as the reference (LOCOM-null in M1 and LOCOM-causal in M2) tended to be more powerful and more sensitive. The aforementioned figures (Figs. 1–4 and *SI Appendix*, Figs. S1–S8) show that the LOCOM tests yielded (almost) the highest power for testing the global hypothesis; LOCOM always controlled the FDR for testing individual taxa

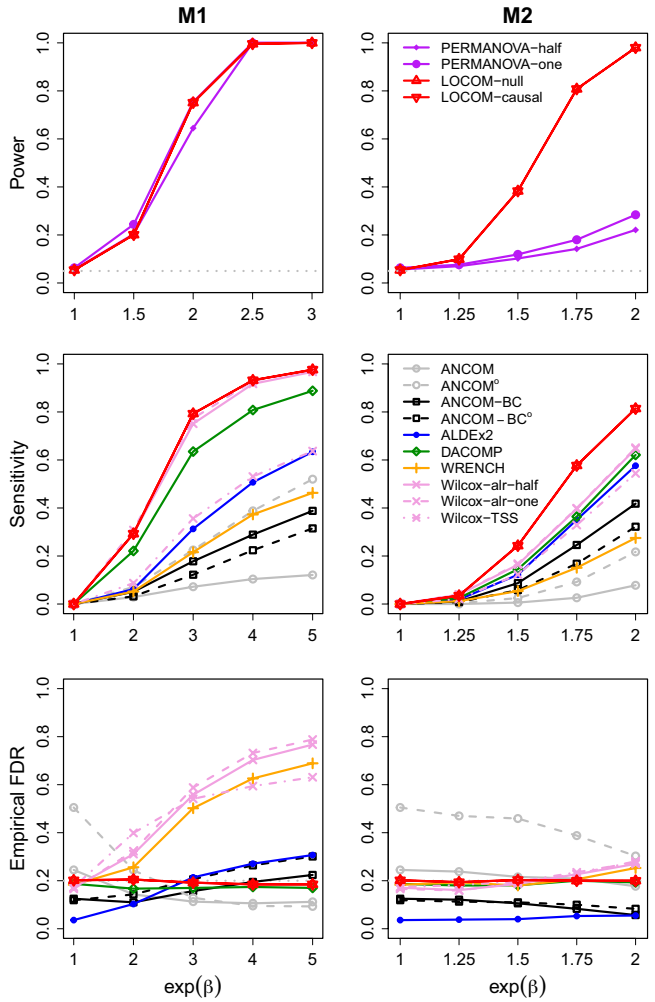

Fig. 1. Simulation results for data ($n = 100$) with a binary trait (and no confounder). The power at $\exp(\beta) = 1$ corresponds to the type I error. The gray dotted line indicates the nominal type I error 0.05 in the first row and the nominal FDR 20% in the third row.

(even with the sample size 50) and had the highest sensitivity among methods that also controlled the FDR.

The competing methods generally have limited application to the scenarios we considered and significantly inferior performance to LOCOM. PERMANOVA had similar power to the LOCOM global test in M1 but lost substantial power to LOCOM in M2 (e.g., Figs. 1–4), likely because the Aitchison distance used by PERMANOVA may not be efficient in capturing sparse signals (only five causal taxa in M2), whereas the harmonic mean P value combination method that LOCOM uses focuses on the strongest signal(s). For testing individual taxa, ALDEx2 is the only method that is applicable to all scenarios we considered; however, it tended to lose control of FDR when the effect size β was large (e.g., Figs. 1 and 2), and it had much lower sensitivity than LOCOM in all cases. ANCOM and ANCOM-BC are only applicable for testing binary traits, with or without confounders. ANCOM easily lost control of FDR when the effect size was small, especially with their own, less stringent filtering criterion (e.g., Figs. 1 and 2). ANCOM-BC tended to lose control of FDR when the effect size was large, especially when there was a confounder (e.g., Fig. 2). Both ANCOM and ANCOM-BC had substantially lower sensitivity than LOCOM when they controlled the FDR. DACOMP is applicable for testing both binary and continuous traits but does not allow adjustment of any confounder. In scenarios without a confounder, DACOMP had good control of FDR, and while the sensitivity of DACOMP tended to be the largest among all competing methods, it was noticeably lower than that of LOCOM (e.g., Figs. 1 and 3). WRENCH is only applicable to one scenario (with a binary trait and no confounder) in which case it had inflated FDR and nevertheless low sensitivity (e.g., Fig. 1). The pseudocount methods, Wilcox-alr-half and Wilcox-alr-one, almost always produced inflated FDR, especially when the effect size was large so that zeros at null taxa were more differentially distributed across trait values (e.g., Fig. 1). As expected, the Wilcox-TSS method had inflated FDR in simulations based on the compositional model in Eq. **5** (e.g., Fig. 1) but controlled the FDR in simulations based on differential relative abundances (Fig. 6).

Results for simulated data with differential experimental bias (and a binary trait and no confounder) are shown in Fig. 5. These simulations showed that, while LOCOM and DACOMP were unaffected by differential bias, all other methods were sensitive to differential bias and yielded significantly inflated FDR in the presence of such bias.

Results for simulations based on the differential relative abundance model and the PLNM model are shown in Fig. 6. In this setting, Wilcox-TSS is the most appropriate method. Indeed, it always controlled the FDR and yielded the highest sensitivity (except for the pseudocount methods which had inflated FDR).

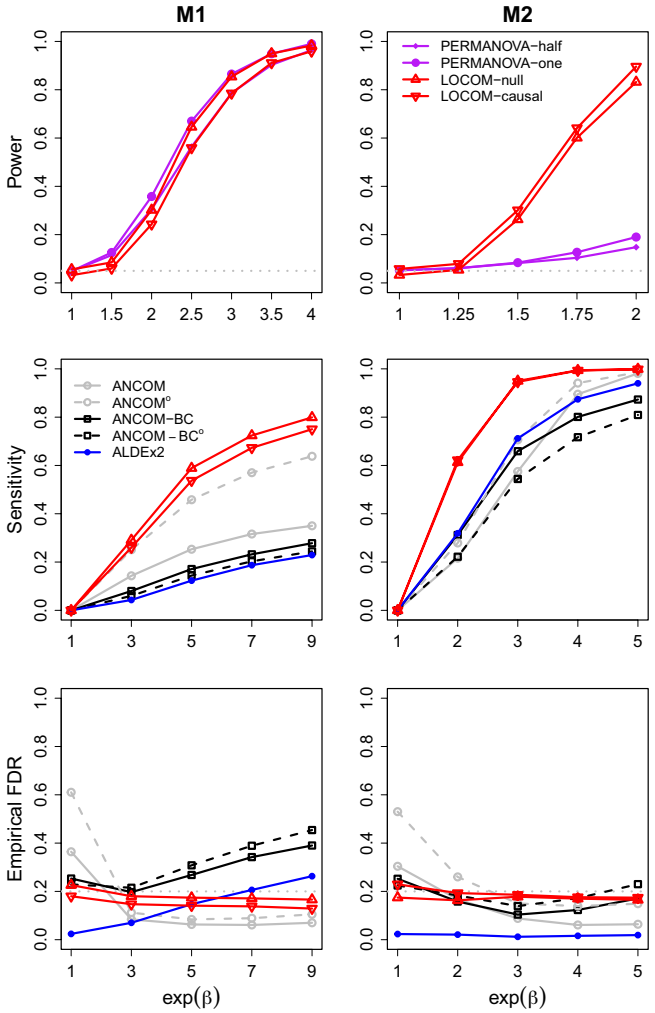

Fig. 2. Simulation results for data ($n = 100$) with a binary trait and a binary confounder.

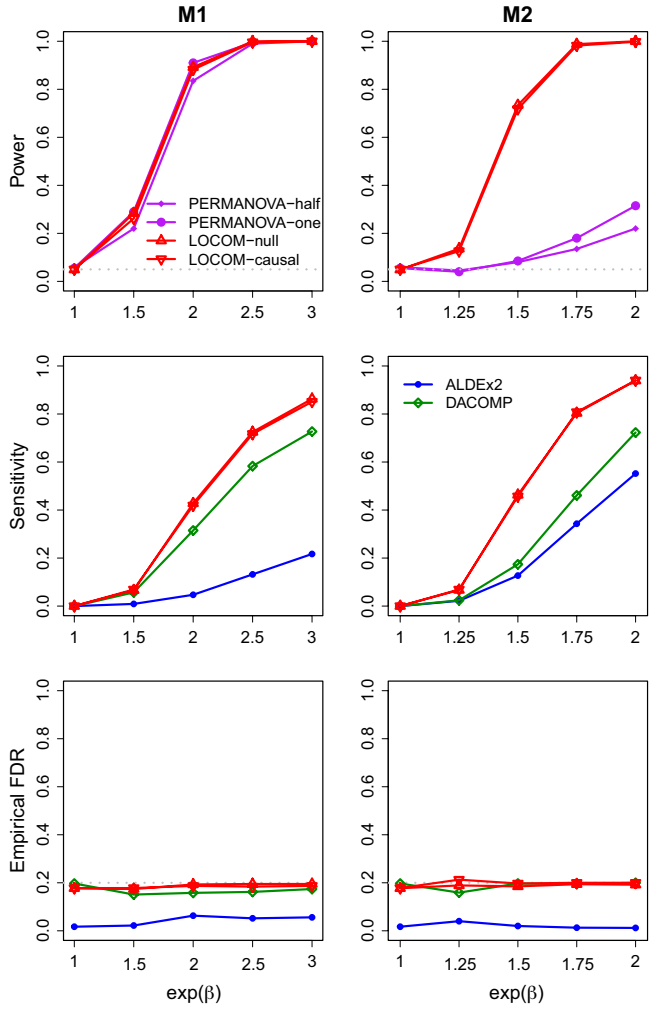

Fig. 3. Simulation results for data ($n = 100$) with a continuous trait (and no confounder).

Interestingly, LOCOM controlled the FDR in both S1-500 and S2, even when S1-500 assumed 500 taxa to be causal; the reason might be that most causal taxa in this setting had very weak signals and act almost like null taxa. Note that LOCOM generated similar sensitivity to the gold standard (Wilcox-TSS). The PERMANOVA tests had higher power than the LOCOM global tests in S1-500 likely because the signals were very dense there.

Results for simulated data generated under M1-500 and M1-rare are shown in *SI Appendix*, Fig. S9. When our assumption that more than half of the taxa are null was violated (M1-500), LOCOM lost control of the FDR as expected. However, the FDR inflation of LOCOM appears to be smaller than most competing methods, and LOCOM maintained good sensitivity. When the causal taxa were all rare (M1-rare), LOCOM still yielded the highest sensitivity while controlling the FDR, although the absolute sensitivity values were low.

Results for simulated data with heterogeneous $\beta_{j,1}$ values are displayed in *SI Appendix*, Fig. S10. The patterns we observed with heterogeneous $\beta_{j,1}$ values were similar to those seen in the analogous simulations with homogeneous $\beta_{j,1}$ values (Fig. 2).

URT Microbiome Data. The data for our first example were generated as part of a study to examine the effect of cigarette smoking on the oropharyngeal and nasopharyngeal microbiome (36). We focused on the left oropharyngeal microbiome in this analysis. The 16S sequence data were summarized into a taxa count table consisting of data from 60 samples and 856 taxa. The trait of interest was a binary variable for smoking status, which divided the participants into 28 smokers and 32 nonsmokers. Other covariates include gender and antibiotic use within the last 3 mo. There was an imbalance in the proportion of males by smoking status (75% in smokers, 56% in nonsmokers), indicating a potential confounding effect of gender. Since there were only three samples who used antibiotics within the last 3 mo, we excluded these samples from our analysis and adjusted for gender only. We adopted the same filter (20% presence) as in the simulation studies, which resulted in 111 taxa for downstream analysis. We applied LOCOM with the most abundant taxon (having mean relative abundance 10.5% before filtering and 11.4% after filtering) as the reference. Given the need to adjust for gender, we only applied ANCOM, ANCOM-BC, and ALDEx2 as a comparison. The nominal FDR was set at 10%.

As shown in Table 1, the global P value of LOCOM is 0.0045, which indicates a significant difference in the overall microbiome profile between smokers and nonsmokers after adjusting for gender. At the taxon level, LOCOM, ALDEx2, ANCOM, and ANCOM-BC detected six, zero, two, and two taxa, respectively; Fig. 7 displays a Venn diagram of these sets of taxa, and *SI Appendix*, Table S3 lists information on the six taxa detected by LOCOM. Fig. 8 shows the distributions of relative abundance across four covariate groups cross-classified by smoking status and

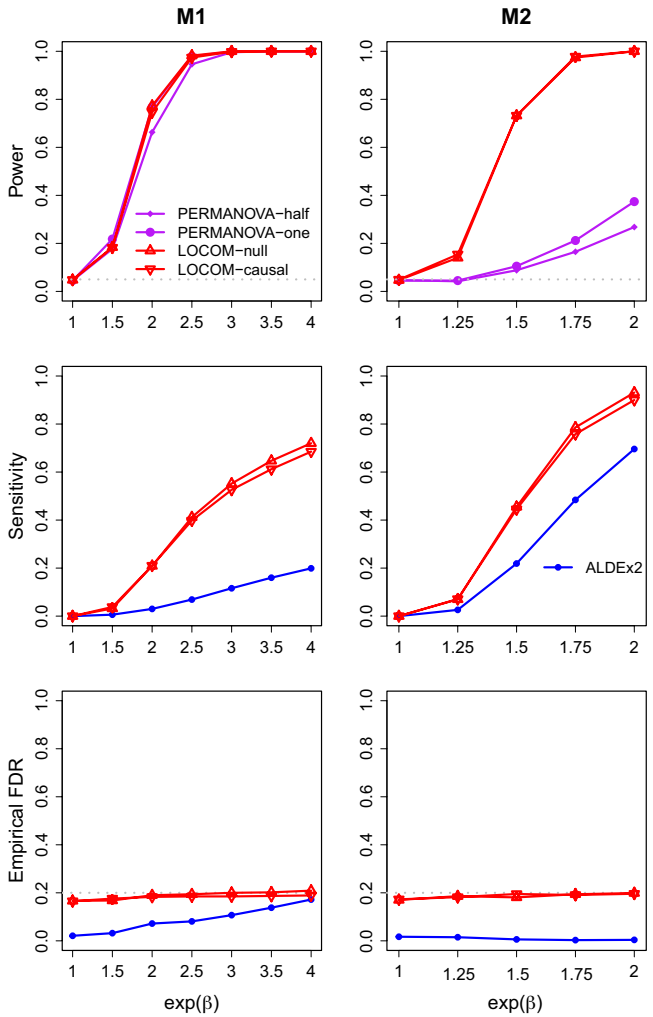

Fig. 4. Simulation results for data ($n = 100$) with a continuous trait and a binary confounder.

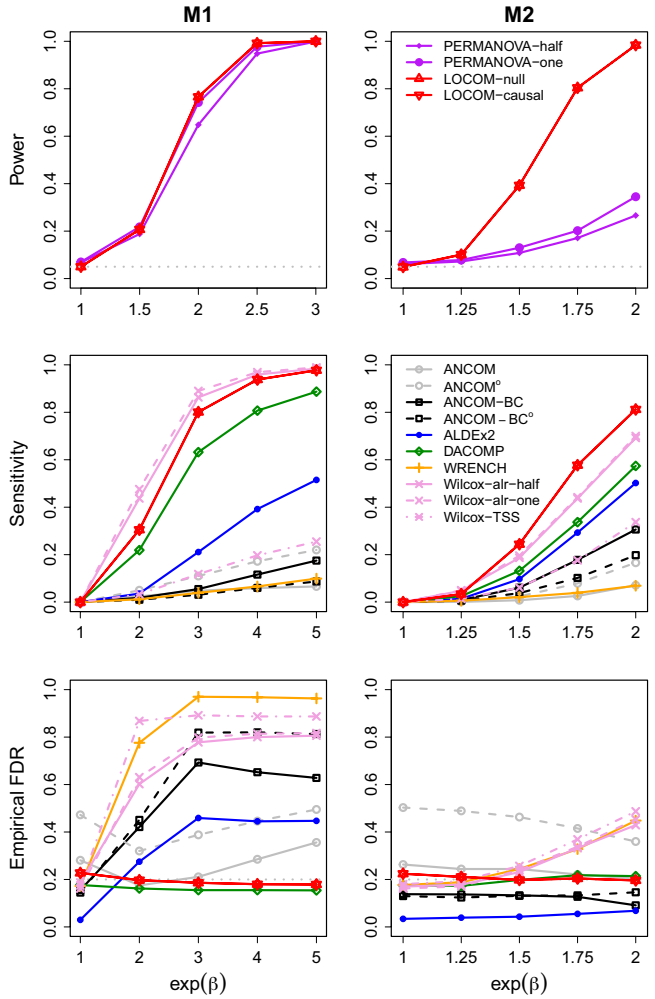

Fig. 5. Simulation results for data ($n = 100$) with differential experimental bias in the binary trait setting (no confounder).

gender, for taxa detected by LOCOM, ANCOM, and ANCOM-BC, as well as for two null taxa. One null taxon is the taxon with the median $\widehat{\beta}_{j,1}$ value. The other is the average of a group of null taxa for improved stability. The two null taxa both had lower relative abundance in smokers than in nonsmokers, among either females or males. The six taxa detected by LOCOM all had the opposite trend (i.e., higher relative abundance in smokers than in nonsmokers), indicating that these taxa are likely to be real signals (i.e., overgrew in smokers). The taxon detected by ANCOM only also had the opposite trend to the null taxa, but it was not detected by LOCOM because the adjusted P value (0.137) by LOCOM did not meet the nominal FDR. The taxon detected by ANCOM-BC only had a similar trend as the null taxa, suggesting that this taxon may actually be a null taxon; indeed, the adjusted P value by LOCOM is 0.674. Note that the difference in relative abundance distributions between smokers and nonsmokers at null taxa may be considered as the counterbalancing change that the null taxa underwent in response to the changes at the causal taxa.

The original analysis of this dataset (36) reported that *Megasphaera* and *Veillonella* spp. were most enriched in the left oropharynx of smokers compared to nonsmokers. Later, a large study of oral microbiome (from oral wash samples) in 1,204 American adults (37) reported enrichment of *Atopobium*, *Streptococcus*, and *Veillonella* in smokers compared to nonsmokers. More recently, a shotgun metagenomic sequencing study of salivary microbiome in Hungary population (38) reported enrichment of *Prevotella* and *Megasphaera* in smokers compared to nonsmokers. Thus, all six taxa detected by LOCOM have been implicated in the literature, even if we only consider the latter two independent studies. These taxa were largely missed by ANCOM and ANCOM-BC.

PPI Microbiome Data. 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 inflammatory bowel disease (IBD) (39). The PPI microbiome data from 196 IBD patients were summarized in a taxa count table with 7,000 taxa classified at the genus level. The gene expression data at 33,297 host transcripts, as well as clinical metadata such as antibiotic use (yes/no), inflammation score (0 to 9), and disease type (familial adenomatous polyposis [FAP] and non-FAP) were also available. The data also included nine gene principal components (gPCs) that together explained 50% of the total variance in host gene expression. Here we included all nine gPCs as multiple traits of interest into one model while adjusting for the three potentially confounding covariates. We filtered out taxa based on our previous filtering criterion, which resulted in 507 taxa to be included in the analysis. We

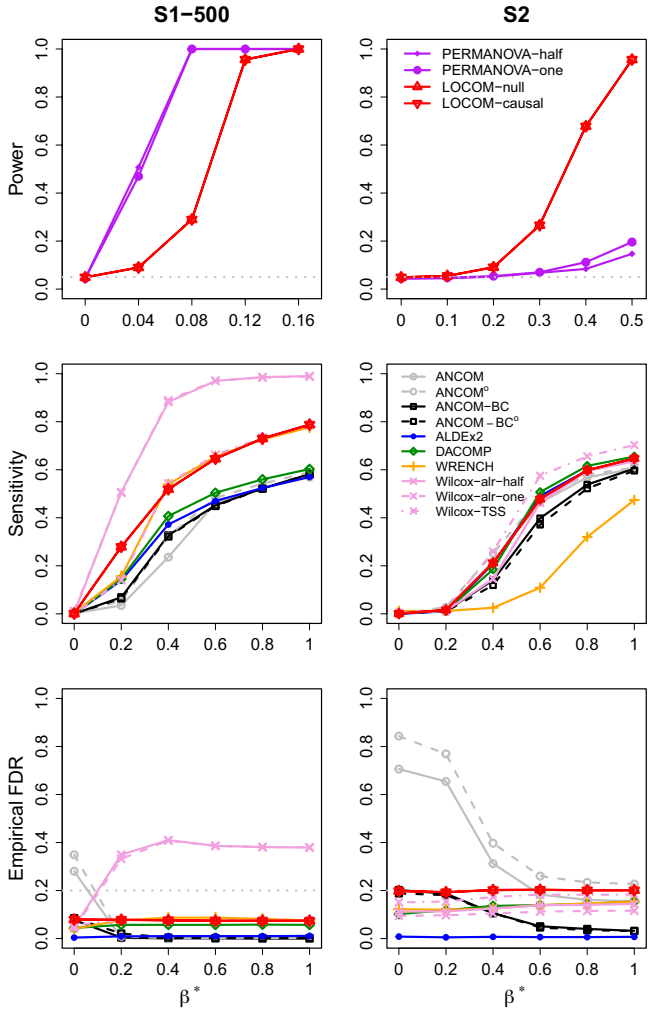

Fig. 6. Simulation results for data ($n = 100$) generated from the differential relative abundance model and the PLNM model in the binary trait setting (no confounder). Here β^* corresponds to the effect size β used in the LDM paper (5); S1-500 and S2 correspond to scenarios S1 and S2 in the LDM paper, except that in S1-500, there are 500 causal taxa.

Table 1. Results from analysis of the two real datasets

Trait	Global *P* value: LOCOM	Number of detected taxa			
		LOCOM	ALDEx2	ANCOM	ANCOM-BC
URT microbiome data					
Smoking	0.0045	6	0	2	2
PPI microbiome data					
gPC1	0.70	0	0	NA	NA
gPC2	0.020	2	0	NA	NA
gPC3	0.018	2	0	NA	NA
gPC4	0.16	0	0	NA	NA
gPC5	0.0070	32	0	NA	NA
gPC6	0.59	0	0	NA	NA
gPC7	0.11	0	0	NA	NA
gPC8	0.21	0	0	NA	NA
gPC9	0.11	0	0	NA	NA

ANCOM and ANCOM-BC are not applicable for testing continuous traits.

applied LOCOM with the most abundant (8.2%) taxon as the reference. Given the continuous traits of interest and the three covariates, we only considered ALDEx2 for comparison. The nominal FDR was set at 10%.

The results of PPI data analysis are presented in Table 1. LOCOM discovered that gPC2, gPC3, and gPC5 had significant associations with the overall microbial profiles at the $\alpha = 0.05$ level. LOCOM detected 2, 2, and 32 taxa as associated with gPC2, gPC3, and gPC5, respectively, at the 10% FDR level and did not detect any taxa for the gPCs that were not found to be associated with the microbiome by the global test. Among the 32 taxa associated with gPC5, 15 belong to the genus *Escherichia* (*SI Appendix*, Table S3), which appeared frequently in the literature of IBD according to a highly cited review article (40). ALDEx2 failed to detect any taxa.

Discussion

We have presented LOCOM, a compositional approach for testing differential abundance in the microbiome data, at both the taxon level and the global level. The global statistic is an aggregate of P values from tests of individual taxa, so results from the taxon level and global tests are coherent. LOCOM allows both binary and continuous traits of interest, can test multiple traits simultaneously, and can adjust for confounding covariates. In our simulations, the taxa detected by LOCOM always preserved FDR, while those identified by the competing methods did not, even though LOCOM had clearly superior sensitivity. In addition, LOCOM also provided a global test that always controlled the type I error and had good power compared to PERMANOVA. In analysis of the URT microbiome data, we demonstrated that the taxa detected by LOCOM were likely to be real signals, while those detected by ANCOM and/or ANCOM-BC but not LOCOM may be false positives. In analysis of the PPI microbiome data, since global and taxon-specific tests were coherent, LOCOM identified significant taxa only for gene principal components that were globally significant.

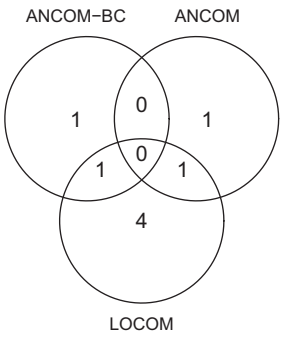

Fig. 7. Taxa detected to be differentially abundant in the URT data.

Like many compositional methods (e.g., DACOMP and WRENCH), LOCOM adopts the assumption that more than half of the taxa in the community are null. This assumption may not be valid in some cases, for example, in testing higher taxonomic levels such as the class or phylum level. In theory, when this assumption does not hold, LOCOM, which always compares each taxon with the median taxon (with the median effect size estimate $\widehat{\beta}_{j,1}$), would find differences at truly null taxa. In our simulations, however, we found that when most causal taxa had very weak signals, LOCOM still controlled the FDR (Fig. 6 and *SI Appendix*, Fig. S9).

We showed both theoretically and with simulation studies that LOCOM is unaffected by experimental bias, even when bias factors are differentially distributed between causal and noncausal taxa. While some competing compositional methods (ANCOM and DACOMP) share this robustness, others (ANCOM-BC, ALDEx2, and WRENCH) do not. The problem in ALDEx2 may be related to the choice of centering; in general, the centered log ratio will not be robust when there are cells with zero counts, since this centering will depend on the set of taxa seen in each sample even if a pseudocount is used. Thus, the centering may not cancel out when comparing log ratios from different samples, leaving these comparisons affected by the particular bias factors that characterize the data being analyzed. Note that any compositional method should perform well when the bias is nondifferential since the centering will be the same on average in each sample.

It is possible to generalize LOCOM to test a trait with more than one component, such as a categorical trait with more than two levels. While ordered categories could be handled in the framework presented here by assigning an appropriate score to each category and then treating this score as a continuous variable, a categorical trait with K unordered categories would presumably require testing $K - 1$ components to fully describe the variable. Within the framework presented here, we could then compare some summary (e.g., max or mean) of these test statistics to their equivalent value in the null permutations. Although this better analysis would require some software development and simulation testing, a simpler proposal could provide results within the existing framework, by calculating separate (marginal) P values for each of the $K - 1$ components and then combining these P values

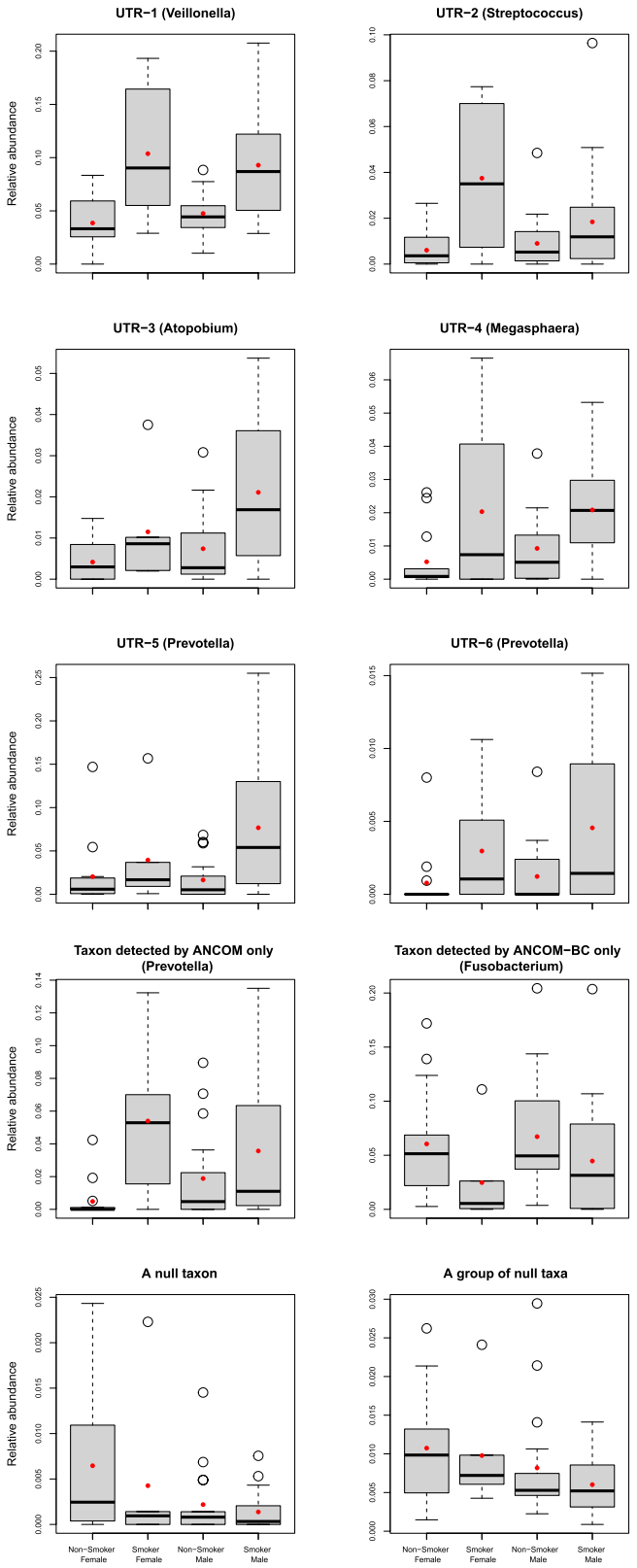

Fig. 8. Distributions of relative abundances for taxa in the URT data. The red dots represent the means. The six taxa in the first, second, and third rows were detected by LOCOM; among these, URT-1 was also detected by ANCOM-BC, and URT-5 was also detected by ANCOM. In the fifth row, a null taxon corresponds to the taxon (*Shigella*) with the median $\widehat{\beta}_{j,1}$ value. A group of null taxa include the taxon with the median $\widehat{\beta}_{j,1}$ value and 20 taxa with $\widehat{\beta}_{j,1}$ values closest to (10 less than and 10 greater than) the median; their relative abundances were averaged.

into a single test statistic, e.g., by using the harmonic mean statistic we used to form our global test. Choosing these $K-1$ components to be orthogonal may be helpful here. We hope to modify LOCOM to incorporate multicomponent traits such as multicategory variables in future work.

Our filtering criterion to exclude taxa with fewer than 20% presence in the sample worked well for the extensive simulation studies we conducted. In fact, a compositional analysis performs best when nonnull taxa are relatively common throughout all samples. Analyses that look for the effect of rare taxa should probably be focused on a presence–absence analysis (41, 42) or on a method based directly on relative abundances.

The compositional null hypothesis considered here is also appropriate in other experimental settings, such as studies of gene expression. This hypothesis corresponds to the scenario that a small number of microbes have "bloomed" while the absolute counts of the others have not changed; this is the reason we made the assumption that more than half of the taxa are null taxa, which is commonly made in other compositional methods. In the gene expression experiment, we often see only a few genes that are differentially expressed; the majority of genes have the same expression in cases and controls. However, it is not completely clear that the compositional hypothesis is applicable to microbiome data because, unlike genes, microbes interact with each other: not only do they compete for resources, but they also change their environment in ways that favor some microbes and suppress others. For example, *Lactobacilli* generally make lactic acid, which changes the pH of the environment. This suppresses microbes that do not thrive in an acidic environment while encouraging growth of microbes that do. Because the microbiota are a community, it is not unreasonable to expect that potentially every taxon changes between cases and controls. The community change null hypothesis may also be reasonable because when comparing the alpha diversity with causal taxa spiked in to a case group, the control group would have a lower alpha diversity (i.e., lower evenness); if this change in alpha diversity is meaningful, then the community change null hypothesis is appropriate. Note that unlike the compositional null, the community change null hypothesis will consider all taxon relative abundances to be potentially changed if extra counts of a small number of taxa are "spiked in." When the community change null hypothesis seems more reasonable than the compositional null hypothesis, then a method that applies directly to relative abundance data such as the LDM is more appropriate. However, the LDM when applied to relative abundance data are not invariant to experimental bias the way LOCOM is; in fact, hypotheses based on differences in relative abundances typically require tests based on unbiased data to be valid.

Like LOCOM, ANCOM is based on comparing pairs of taxa. However, ANCOM yielded lower sensitivity than LOCOM in our simulations (e.g., Figs. 1 and 2). There are several possible reasons. First, LOCOM analyzes the count data using logistic regression, which downweighs zero counts, while ANCOM analyzes alr-transformed count data using linear regression, which makes data with zero or very small counts more influential; the former is based on transformation of parameters (i.e., true relative abundances), while the latter is based on transformation of data. Second, ANCOM's approach of adding pseudocounts further introduces noise and possibly bias to the data. Third, LOCOM uses the most abundant taxon as the reference, while ANCOM looks at all possible pairs of taxa, which can lead to unstable log ratios when both taxa are rare. Finally, ANCOM's strategy to declare differentially abundant taxa uses an arbitrary cutoff which may not be well calibrated.

We have implemented our method in the R package LOCOM, which is available on GitHub at https://github.com/yijuanhu/LOCOM in formats appropriate for Macintosh or Windows. LOCOM is computationally efficient for data with small sample sizes but can take longer for larger sample sizes. For example, using parallel computing (by parallelizing permutation replicates) with 4 cores of a MacBook Pro laptop (1.4 GHz Quad-Core Intel Core i5, 8GB memory), it took 11 s to analyze a simulated dataset with 100 samples, 11 s to analyze the URT data, and 40 min to analyze the PPI data. In considering this last timing, it should be noted that the analysis considered nine traits simultaneously in the presence of three confounding covariates and as such is more complex than the typical microbiome analysis. In addition, LOCOM could be further parallelized by splitting the data into subsets with sets of taxa that only share the reference taxon and then combining the values of $\beta_{j,1}$ from each dataset (care should be taken to use the same seed for each analysis so that the same set of permutations is used).

Data Availability. Previously published data were used for this work (36, 39). The R package LOCOM is publicly available at GitHub (https://github.com/yijuanhu/LOCOM) (43).

ACKNOWLEDGMENTS. This research was supported by the NIH awards R01GM141074 (Y.-J.H. and G.A.S.) and R01GM116065 (Y.-J.H. and G.A.S.).

Author affiliations: [a]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322; and [b]Department of Gynecology and Obstetrics, Emory University School of Medicine, Atlanta, GA 30322

1. M. R. McLaren, A. D. Willis, B. J. Callahan, Consistent and correctable bias in metagenomic sequencing experiments. *eLife* **8**, e46923 (2019).
2. S. Hawinkel, F. Mattiello, L. Bijnens, O. Thas, A broken promise: Microbiome differential abundance methods do not control the false discovery rate. *Brief. Bioinform.* **20**, 210–221 (2019).
3. M. Arumugam *et al.*; MetaHIT Consortium, Enterotypes of the human gut microbiome. *Nature* **473**, 174–180 (2011).
4. O. Koren *et al.*, A guide to enterotypes across the human body: Meta-analysis of microbial community structures in human microbiome datasets. *PLOS Comput. Biol.* **9**, e1002863 (2013).
5. Y. J. Hu, G. A. Satten, Testing hypotheses about the microbiome using the linear decomposition model (LDM). *Bioinformatics* **36**, 4106–4115 (2020).
6. M. S. Kumar *et al.*, Analysis and correction of compositional bias in sparse sequencing count data. *BMC Genomics* **19**, 799 (2018).
7. B. Brill, A. Amir, R. Heller, Testing for differential abundance in compositional counts data, with application to microbiome studies. arXiv [Preprint] (2019). https://arxiv.org/abs/1904.08937. Accessed 30 March 2020.
8. S. Mandal *et al.*, Analysis of composition of microbiomes: A novel method for studying microbial composition. *Microb. Ecol. Health Dis.* **26**, 27663 (2015).
9. H. Lin, S. D. Peddada, Analysis of compositions of microbiomes with bias correction. *Nat. Commun.* **11**, 3514 (2020).
10. A. D. Fernandes *et al.*, Unifying the analysis of high-throughput sequencing datasets: Characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* **2**, 15 (2014).
11. G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, J. J. Egozcue, Microbiome datasets are compositional: And this is not optional. *Front. Microbiol.* **8**, 2224 (2017).
12. J. Aitchison, *The Statistical Analysis of Compositional Data* (Chapman and Hall, London, 1986).
13. J. N. Paulson, O. C. Stine, H. C. Bravo, M. Pop, Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* **10**, 1200–1202 (2013).
14. N. Zhao, X. Zhan, K. A. Guthrie, C. M. Mitchell, J. Larson, Generalized Hotelling's test for paired compositional data with application to human microbiome studies. *Genet. Epidemiol.* **42**, 459–469 (2018).
15. M. B. Sohn, H. Li, Compositional mediation analysis for microbiome studies. *Ann. Appl. Stat.* **13**, 661–681 (2019).
16. P. I. Costea, G. Zeller, S. Sunagawa, P. Bork, A fair comparison. *Nat. Methods* **11**, 359 (2014).
17. J. N. Paulson, H. C. Bravo, M. Pop, Reply to: "A fair comparison". *Nat. Methods* **11**, 359–360 (2014).
18. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).
19. B. H. McArdle, M. J. Anderson, Fitting multivariate models to community data: A comment on distance-based redundancy analysis. *Ecology* **82**, 290–297 (2001).
20. J. Aitchison, C. Barceló-Vidal, J. A. Martín-Fernández, V. Pawlowsky-Glahn, Logratio analysis and compositional distance. *Math. Geol.* **32**, 271–275 (2000).
21. J. P. Brooks, Challenges for case-control studies with microbiome data. *Ann. Epidemiol.* **26**, 336–341.e1 (2016).
22. L. W. Hugerth, A. F. Andersson, Analysing microbial community composition through amplicon sequencing: From sampling to hypothesis testing. *Front. Microbiol.* **8**, 1561 (2017).
23. J. Pollock, L. Glendinning, T. Wisedchanwet, M. Watson, The madness of microbiome: Attempting to find consensus "best practice" for 16s microbiome studies. *Appl. Environ. Microbiol.* **84**, e02627-17 (2018).
24. P. I. Costea *et al.*, Towards standards for human fecal sample processing in metagenomic studies. *Nat. Biotechnol.* **35**, 1069–1076 (2017).
25. D. Mariat *et al.*, The Firmicutes/Bacteroidetes ratio of the human microbiota changes with age. *BMC Microbiol.* **9**, 123 (2009).
26. F. Magne *et al.*, The firmicutes/bacteroidetes ratio: A relevant marker of gut dysbiosis in obese patients? *Nutrients* **12**, 1474 (2020).
27. N. Zhao, G. A. Satten, "A log-linear model for inference on bias in microbiome studies" in *Statistical Analysis of Microbiome Data*, S. Datta, S. Guha, Eds. (Springer-Verlag, New York, 2021), pp. 221–247.
28. C. B. Begg, R. Gray, Calculation of polychotomous logistic regression parameters using individualized regressions. *Biometrika* **71**, 11–18 (1984).
29. D. Firth, Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38 (1993).
30. G. Heinze, M. Schemper, A solution to the problem of separation in logistic regression. *Stat. Med.* **21**, 2409–2419 (2002).
31. J. Atchinson, "Concise guide to compositional data analysis" in *In2do Compositional Data Analysis Workshop CoDaWork October* (2005), vol. 5, pp. 17–21.
32. D. M. Potter, A permutation test for inference in logistic regression with small- and moderate-sized data sets. *Stat. Med.* **24**, 693–708 (2005).
33. G. K. Sandve, E. Ferkingstad, S. Nygård, Sequential Monte Carlo multiple testing. *Bioinformatics* **27**, 3235–3241 (2011).
34. D. J. Wilson, The harmonic mean *p*-value for combining dependent tests. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 1195–1200 (2019).
35. P. H. Westfall, S. S. Young, *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment* (John Wiley & Sons, 1993).
36. E. S. Charlson *et al.*, Disordered microbial communities in the upper respiratory tract of cigarette smokers. *PLoS One* **5**, e15216 (2010).
37. J. Wu *et al.*, Cigarette smoking and the oral microbiome in a large study of American adults. *ISME J.* **10**, 2435–2446 (2016).
38. R. Wirth *et al.*, A case study of salivary microbiome in smokers and non-smokers in Hungary: Analysis by shotgun metagenome sequencing. *J. Oral Microbiol.* **12**, 1773067 (2020).
39. X. C. Morgan *et al.*, Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease. *Genome Biol.* **16**, 67 (2015).
40. J. Ni, G. D. Wu, L. Albenberg, V. T. Tomov, Gut microbiota and IBD: Causation or correlation? *Nat. Rev. Gastroenterol. Hepatol.* **14**, 573–584 (2017).
41. Y. J. Hu, A. Lane, G. A. Satten, A rarefaction-based extension of the LDM for testing presence-absence associations in the microbiome. *Bioinformatics* **37**, 1652–1657 (2021).
42. Y. J. Hu, G. A. Satten, A rarefaction-without-resampling extension of permanova for testing presence-absence associations in the microbiome. *Bioinformatics*, btac399. https://doi.org/10.1093/bioinformatics/btac399 (20 June 2022).
43. Y.-J. Hu, LOCOM. GitHub. https://github.com/yijuanhu/LOCOM. Deposited 23 March 2022.