The Annals of Applied Statistics
2022, Vol. 16, No. 4, 2648–2671
https://doi.org/10.1214/22-AOAS1607

TESTING FOR DIFFERENTIAL ABUNDANCE IN COMPOSITIONAL COUNTS DATA, WITH APPLICATION TO MICROBIOME STUDIES

BY BARAK BRILL[1,a], AMNON AMIR[2,c] AND RUTH HELLER[1,b]

[1]*Department of Statistics and Operations Research, Tel Aviv University,* [a]*barakbri@mail.tau.ac.il,* [b]*ruheller@gmail.com*
[2]*Microbiome center, The Chaim Sheba Medical Center,* [c]*amnonim@gmail.com*

Identifying which taxa in our microbiota are associated with traits of interest is important for advancing science and health. However, the identification is challenging because the measured vector of taxa counts (by amplicon sequencing) is compositional, so a change in the abundance of one taxon in the microbiota induces a change in the number of sequenced counts across all taxa. The data are typically sparse, with many zero counts present either due to biological variance or limited sequencing depth. We examine the case of Crohn's disease, where the microbial load changes substantially with the disease. For this representative example of a highly compositional setting, we show existing methods designed to identify differentially abundant taxa may have an inflated number of false positives. We introduce a novel nonparametric approach that provides valid inference, even when the fraction of zero counts is substantial. Our approach uses a set of reference taxa that are non-differentially abundant which can be estimated from the data or from outside information. Our approach also allows for a novel type of testing: multivariate tests of differential abundance over a focused subset of the taxa. Genera-level multivariate testing discovers additional genera as differentially abundant by avoiding agglomeration of taxa.

1. Introduction. The microbiome is the collection of microorganisms and bacteria which are part of the physiological activity of a host body or ecosystem (Hamady and Knight (2009)). It is of interest to associate change in microbial structure to disease and other environmental factors. Identifying the associations is difficult when the microbial load changes substantially with the trait, as is common in inflammatory bowl diseases. Our motivating example is the study of Vandeputte et al. (2017) that investigated the change in the microbial ecology of fecal samples in the presence of Crohn's disease. This change is associated with a change in the composition of the gut microbiome of patients along with a substantial decline in the *microbial load*, that is, the physical abundance of bacteria. A better understanding of the microbial changes in the gut may lead to a better understanding and treatment of Crohn's disease.

A common method of measuring the composition of the bacterial community is by sequencing the 16S rRNA gene which codes for a crucial part of the ribosome common to all living cells. The variable regions in the 16S rRNA gene are subject to mutations along genetic lineages. Due to these variations, 16S rRNA sequence patterns serve as a proxy for the taxonomic identification of their organism. The data are generated by collecting samples from different specimen, and the targeted variable regions are duplicated and amplified using polymerase chain reaction (PCR). Sequencing technology allows one to read the amplicons of the PCR procedure and list all sequences read for each sample. This list of sequences is then trimmed to a constant length of, for example, 150 base pairs (Nelson et al. (2014)), and the amount of each unique sequence in each sample is recorded. Due to errors in the sequencing,

Received August 2020; revised January 2022.
Key words and phrases. Compositional bias, analysis of composition, normalization, rarefaction, nonparametric tests.

not all unique sequences actually represent unique bacteria. In order to identify the bacteria actually present in each sample, two alternative methods can be used: Operational taxonomic units (OTUs) are sequences which differ up to a certain threshold, for example, 3% of base pairs out of 150 (Hamady and Knight (2009)). Amplicon sequence variants (ASVs) are the individual sequences (Callahan et al. (2016), Amir et al. (2017)) obtained after a denoising of the reads. The OTUs, or ASVs, represent the finest resolution of organism type identifiable from sequencing variants of the 16S rRNA gene. The data consist of the number of observed sequences for each OTU or ASV in each sample. The units of interest for analysis, referred to as taxa, are the single OTUs or ASVs, or coarser units that aggregate phylogenetically related OTUs or ASVs.

The first challenge in statistical analysis is that the number of sequenced reads, or *sequencing depth*, varies from sample to sample and is mostly an artifact of the sequencing procedure rather than a proxy to the sample's microbial load. Therefore, only the relative frequencies are informative, that is, the count data are compositional (Gloor et al. (2017), Mandal et al. (2015), Kumar et al. (2018)).

The second challenge is that the vector of taxa counts is sparse by nature, as not all taxa are measured in all samples. The percentage of zeros in the data ranges between 50% and 90% for many types of samples (Xu et al. (2015)). A taxon with zero counts can occur for two reasons: (1) low frequency in the sampled units, so the sample does not capture the very rare taxa, henceforth referred to as technical zeros, (2) taxa not shared by the entire population, henceforth referred to as structural zeros.

Additional challenges are the study size (the number of samples can be much smaller than the number of taxa, Nelson et al. (2014)) and the strong (yet unknown) dependence between taxa counts. Microbiome data, being compositional data, may feature both positive and negative correlations between components (Aitchison (1986), Hawinkel et al. (2019)). Due to the above challenges, it is difficult to design a valid inferential method for identifying the taxa that are associated with the trait. Statistical tests that ignore compositionality can lead to false positive findings, as demonstrated by the following example.

EXAMPLE. A toy example demonstrating the danger of ignoring compositionality. Suppose we have a binary trait, that is, two groups of samples. In terms of absolute abundance, we assume the first taxon out of m taxa increased its absolute abundance in the second group, compared to the first group, while all other taxa have the same absolute abundances across the two groups. The relative abundances of taxa are measured for each sample, the observed vector of counts for each sample is multinomial with N total counts, and a probability vector determined by the samples group identity. For this setting, denoting the probability vector for a sample that belongs to the first group as $\mathbf{P}$, the probability vector for a sample that belongs to the second group is, for some $w \in (0, 1)$, $(1 - w) \times \mathbf{P} + w \times \mathbf{e}_1$ (where $\mathbf{e}_1$ is the binary vector with a single entry of one in the first coordinate). An analysis that ignores compositionality will perform for each taxon a two sample test of equality of distributions based on the taxon counts in each group. Since the first taxon has an increased relative frequency in the first group, compared to the second group, and all other taxa have decreased relative frequency in the second group compared to the first group, for every taxon the counts indeed differ in their distribution across groups. Therefore, for large enough sample sizes the two-sample test will reject the null hypothesis for each taxon. As w increases, the number of zero counts increase for all but the first taxon in the second group, and the probability of a significant result by the two-sample test increases for all taxa. However, we are interested in detecting only the first taxon, since it is the only one driving the differences across groups. (In microbiome studies, unlike in this example, the probability vector varies within each group.)

When taxa with varying absolute abundances across study groups substantially affect the microbial load, an analysis for identifying the taxa driving the association with the trait has to take compositionality into account in order to avoid an inflation of false positives. The data in Vandeputte et al. (2017) show a 67% decrease in the median microbial load for samples taken from subjects with Crohn's disease, compared to healthy subjects, as measured by flow cytometry. Hence, in order to identify the taxa driving the association with Crohn's disease the analysis should adjust carefully for the effect of compositionality. The main difficulty is the fact that a change in the relative abundance of taxa, due to compositionality, induces a change in the prevalence of technical zeros across study groups. Therefore, even if the total microbial load is known (as in Vandeputte et al. (2017)), scaling is not possible due to the technical zeros.

In addition to Crohn's disease, there are other applications studying ecologies that show vast changes in microbial load which can benefit from the methodology developed in this work. The study of Vieira-Silva et al. (2019) shows that fecal samples taken from patients with ulcerative colitis and primary sclerosing cholangitis also experience a substantial decrease in the microbial load. The microbial load quantification method of Jian et al. (2020) exemplifies how the differences in microbial load across human fecal samples may experience up to a tenfold change. The human salivary microbiome is another microbiome ecology in the human body that experiences large changes in absolute load; for example, the salivary microbial load is reduced by 50% following teeth brushing (Morton et al. (2019)).

In the study of the microbiome in soil and water enviroments, the microbial load can also change subtantially between study groups. For example, Tkacz, Hortala and Poole (2018) provide a method to quantify the absolute microbial load, and show a 39% increase in microbial load between two soil types. A different microbial load quantification method, by Guo et al. (2020), shows the microbial load in saline soil changes more than threefold following soil fertilization. The microbial load quantification method of Jiang et al. (2019) shows the rhizosphere microbial load experiences up to a 2.8-fold increase in wet season compared to dry season in roots. In an ocean microbiome study, Sunagawa et al. (2015) show how the total microbial load of ocean microbiota changes with depth. Specifically, the microbial load observed in surface water is 15 times higher than that observed in the Mesopelagic zone. Kong et al. (2021) also studied the microbiome in surface water and show samples have large fluctuations in microbial load.

We aim to identify the taxa whose original ecosystem abundance has changed. Generally, the original ecosystem abundance of taxa cannot be reconstructed from the relative frequencies of taxa alone. Hence, a change in the absolute abundance of taxa may be undetectable if the ratios between relative abundances of taxa have remained fixed. A testable hypothesis is whether the absolute abundance of a taxon changed in a way that is different from the majority of changes for the taxa in the ecosystem, that is, a test of differential abundance with respect to a reference frame of taxa (Morton et al. (2019)). Our first step is to identify a set of reference taxa. The reference set of taxa can be identified from the data, if most taxa are nondifferentially abundant, by selecting the taxa whose respective ratios vary little across the data. Of course, if there is outside information on which taxa are nondifferentially abundant, it can be used to aid selection of reference taxa. Moreover, if the study has a very large sample size (e.g., the American Gut Project, McDonald et al. (2018)), then the dataset can be split and part of it sacrificed for identification of the reference set. Our next step is to tackle the question of how to test for differential abundance. Specifically, we are interested in testing whether the ratio of taxon abundance to reference set abundance is independent of the trait of interest.

Traditionally, the common practice of normalization by rarefaction has been used for tests involving all microbial taxa (Weiss et al. (2017)). Our work addresses an existing gap in the

microbiome statistical literature, regarding the difference in normalization techniques used for whole-microbiome analysis compared with differential abundance testing for a single taxon. We extend the idea of normalization by rarefaction and show that a subsampling, based normalization technique, is crucial not only when analyzing the collective microbiome composition but also when analyzing a single taxon, or a subset of taxa, for differential abundance.

In Section 1.1 we review methods for analysis of differential abundance in microbiome studies and point out limitations which this work aims to overcome. In Section 2 we present our motivating example, a Crohn's disease dataset that shows strong variability in microbial load. In Section 3 we formalize our analysis goal of detecting differential abundance. In Section 4 we describe our main result, a testing procedure for discovering the differentially abundant taxa that has guaranteed control over false positives. This test relies on the availability of a reference set of taxa, and we show how to estimate this reference set from the data. We also show how to test whether a group of taxa is independent of a trait. In Section 5 we conduct a simulation study with simulated datasets based on the data set presented in Section 2. Data generation aims to capture the data-specific characteristics discussed in Section 2. In Section 6 we analyze the data presented in Section 2 using both our method and the methods presented in Section 1.1. In Section 7 we conclude with final remarks.

1.1. *Review of methods for differential abundance analysis.* Let $\mathbf{X}$ be the m-dimensional vector of observed taxa counts. Let $C(\mathbf{X})$ be a function from $\mathbb{N}^m$ to the real positive line so that the analysis will associate the scaled counts, $\mathbf{X}/C(\mathbf{X})$, with the trait. Total sum scaling (TSS) normalization selects $C(\mathbf{X})$ to be the total number of counts in $\mathbf{X}$. The example of Section 1 demonstrates that the trait may be associated with a nondifferentially abundant normalized taxon, so a test of independence following normalization cannot be used to identify the differentially abundant taxa.

Paulson et al. (2013) suggest cumulative sum scaling (CSS). CSS normalization selects $C(\mathbf{X})$ so that the smallest q_{CSS} values in $\mathbf{X}$ sum to one, with q_{CSS} chosen adaptively from the data. As with TSS, this normalization does not resolve the bias in testing induced by compositionality, as shown in Mandal et al. (2015) as well as in our simulations in Section 5.

Kaul et al. (2017) suggest other scaling methods for use in microbiome studies based on the normalization of Aitchison (1982). But after transformation, the null hypothesis of independence between a taxon and the trait will be false also for nondifferentially abundant taxa since the scaling factors considered are functions of the differentially abundant taxa. Even taking $C(\mathbf{X})$ to be the count of a specific taxon, for example, the mth taxon, is problematic since typically for every taxon some samples have a zero count. A pseudocount is put in place of zero, but if the probability of zero counts changes with the trait, then $C(\mathbf{X})$ is associated with the trait, leading to a biased inference.

Kumar et al. (2018) suggest an alternative scaling approach, called Wrench, based on the assumption that taxa not associated with the condition of interest have maintained the ratios of their respective proportions in each sample. In Example 1, for example, while the expected values of all coordinates differ across study groups, coordinate means across all taxa, except the first taxon, are lower in the second group compared to the first group by the multiplicative factor $1 - w$. Kumar et al. (2018) suggest estimating the common multiplicative factor from the data for scaling taxa counts.

Fernandes et al. (2013) suggest ALDEx2, where the normalization factor, $C(\mathbf{X})$, is the geometric mean of the counts in a subset of the taxa. The counts are normalized with respect to taxa that are estimated to be nondifferentially abundant and then log-transformed for statistical inference, as detailed in Fernandes et al. (2013). In order to avoid division by zero, a pseudocount of 0.5 is added to all data entries. If the probability of zero counts changes with

the trait, then the inference may not be valid, but the bias is less severe than with the previous methods, as we show in Section 5.1.

Additional methods making use of auxiliary measurements to determine normalization factors include the approach of Vandeputte et al. (2017) which uses flow-cytometric measurements as a means to estimate the absolute microbial load of samples, the approach of Staemmler et al. (2016) which suggests artificially inserting bacteria of types nonendemic to the measured samples in predetermined abundance, and the use of spiked-in DNA sequences (Quinn et al. (2019)).

Mandal et al. (2015) suggest ANCOM which avoids the need of a "per-sample" scaling factor as follows for a binary trait. First, for every pair of taxa the Wilcoxon rank sum test is applied for testing the independence of the ratio between the two taxa and the binary trait. In order to avoid division by zero, ANCOM adds a pseudocount with a value of 1 to all taxa counts. Let $\mathcal{W}_j$ be the number of rejections by the Wilcoxon rank sum test, at a predefined level α, with taxon j in the numerator. Assuming that the ratio is more or less fixed for nondifferentially abundant taxa, in a well powered study we expect a higher value of $\mathcal{W}_j$ for a differentially abundant taxon j. The taxa with indices $\{j|\mathcal{W}_j \geq \mathcal{W}^*\}$ are declared differentially abundant where $\mathcal{W}^*$ is chosen adaptively, as detailed in Mandal et al. (2015).

Calgaro et al. (2020) show that existing methods for analyzing RNA sequencing data, such as Deseq2 (Love, Huber and Anders (2014)), may be useful for analyzing microbiome counts data. Deseq2 uses negative binomial GLMs to model the marginal distribution of counts. Calgaro et al. (2020) suggest using an extension of the Deseq2 model (suggested by Van den Berge et al. (2018)) that downweights the importance of different samples in estimation of model parameters according to the probability that they are structural zeros, using a zero-inflated negative-binomial model (Risso et al. (2018)). Their approach has good power, but it ignores compositionality, and it relies on parametric assumptions.

The methods described in this section suffer from the following three related limitations: (1) nondifferentially abundant taxa remain associated with the trait if the prevalence of zero counts varies with the trait, since zero counts cannot be scaled; (2) many of the methods, in order to apply transformations, use pseudocounts instead of the zero counts, which corresponds to microbial load not measured in practice, and (3) some methods ignore the compositional nature of the data, for example, TSS and CSS, while other methods, such as ANCOM and Wrench, provide inaccurate adjustment for it due to the two aforementioned limitations. We demonstrate in numerical experiments that the above approaches can lead to an unacceptably inflated rate of false positive discoveries.

2. The microbiome in Crohn's disease. Vandeputte et al. (2017) examined fecal samples from 29 subjects with Crohn's disease (CD) and 66 healthy controls. The V4 region of the 16S gene was amplified and sequenced from fecal samples, in addition to a microbial load count (number of bacteria per gram of fecal material). We picked sOTUs (a type of ASVs) using the method of Amir et al. (2017). We set sOTU length to the default value of 150 base pairs. Following sOTU selection, sOTUs appearing only in one sample were excluded from analysis, leaving 1569 sOTUs for analysis.

Table 1 provides summary statistics displaying key distributional differences between the group of healthy and the group of diseased subjects. The flow-cytometry measurements show that the total abundance of the microbiota is much lower for subjects with CD, with the median microbial load in the CD group being less than 33% than the median microbial load for the healthy group. With such a high change in microbial load, it is plausible that most taxa differ in their relative abundance with and without CD. Therefore, the inference has to take this into account.

TABLE 1

Summary statistics for the group of healthy and the group of diseased subjects in the study of Vandeputte et al. (2017). Row 4 shows the median and mean prevalence of taxa in each group, where a taxon's prevalence is the percentage of samples in the group with at least one read of that taxon

	Healthy	CD
Number of samples	66	29
Median 16S sequencing depth	22,871	18,874
Median microbial load (bacteria per gram)	$1.16 \cdot 10^{11}$	$3.76 \cdot 10^{10}$
Median taxa prevalence (mean)	0.09 (0.2)	0.03 (0.14)

The total number of reads in 16S samples is also associated with CD. We observe zero counts to be substantially more prevalent in the group of samples taken from subjects with CD, compared to healthy subjects, both by the mean and median prevalence of taxa.

In Section 6.1 we test for differential abundance using methods that address the compositional bias in testing, including normalization by the flow-cytometric measurements, as suggested in Vandeputte et al. (2017). We show that existing approaches can result in spurious discoveries that have limited replicability, whereas our novel approach provides discoveries that appear more replicable.

In Section 6.2 we extend our approach, which uses the reference set and a multivariate test, to identify association with genera that are masked using a univariate analysis. The number of reads observed under the focus subset of taxa is confounded with the research variable, either due to confounding sequencing depth or differentially abundant taxa found outside the focus set. The typical solution in whole microbiome testing is to adjust for confounding sequencing depth by rarefying the total number of reads (Weiss et al. (2017)). However, when considering a subset of taxa, the number of samples with zero total counts can be nonnegligible. For example, out of the 61 genera analyzed, 46 genera have at least one sample with zero counts, and 30 genera have at least 30 samples with zero counts. Samples with zero counts cannot be removed without invalidating the test (as detailed in Section 4). We avoid rarefaction to zero depth by using a set of reference taxa, and our tests result in detection of genera not discovered with a univariate analysis.

In Section 6.3 we extend the popular method of microbiome principal coordinates analysis (PCoA, Goodrich et al. (2014)) to analyze counts from a single genus. Our "within-genus" PCoA uses the reference taxa to perform a valid PCoA analysis for a subset of the taxa.

We also perform "within-genus" PCoA, after excluding any ASV-level discovery obtained by existing methods, to examine whether this genera subset appears to differ across groups, though the difference was undetected by the ASV-level analysis.

3. The setup and goal. Let m and n be the number of taxa and samples, respectively. For sample $i \in \{1, \ldots, n\}$, we denote by N_i the total number of counts sampled, by $\mathbf{Y}_i$ the measured (univariate or multivariate) trait, and by $\mathbf{X}_i$ the m-dimensional vector of observed taxa counts. Let $\mathbf{P}_i$ be the (unobserved) vector of the taxa population relative frequencies in sample i.

For simplicity, we omit the sample subscript i when addressing a single observation. For an m-dimensional binary vector with at least one entry of one, $\mathbf{s} = (s_1, \ldots, s_m)'$, we denote by $\mathbf{X}(\mathbf{s})$ and $\mathbf{P}(\mathbf{s})$ the subvectors of $\mathbf{X}$ and $\mathbf{P}$ of dimension $\mathbf{s}'\mathbf{s}$ (the number of nonzero entries in $\mathbf{s}$) containing the coordinates for which $\mathbf{s}$ is equal to 1. The sum of the entries in these subvectors is $\mathbf{s}'\mathbf{X}$ and $\mathbf{s}'\mathbf{P}$. We denote by $\mathbf{e}_j$ the m-dimensional binary vector with a single entry of one at coordinate j, so $\mathbf{P}(\mathbf{e}_j)$ and $\mathbf{X}(\mathbf{e}_j)$ are the population relative frequency and

observed count for taxon $j \in \{1, \ldots, m\}$. We observe n realizations of $(\mathbf{X}, \mathbf{Y})$, where $\mathbf{X}$ is the vector of multinomial counts, given the (unobserved) random vector $\mathbf{P}$ of population relative frequencies for the subject,

$$\mathbf{X}|\mathbf{P}, N \sim \text{multinom}(N, \mathbf{P}), \qquad \mathbf{P}(\mathbf{e}_j) \geq 0 \quad \text{for } j = 1, \ldots, m, \qquad \mathbf{1}'\mathbf{P} = 1.$$

We aim to identify the taxa that are associated with $\mathbf{Y}$, taking the compositional nature of the data into account. For this purpose we assume that there exists a group of taxa that may be associated with $\mathbf{Y}$ via their sum but are otherwise independent of $\mathbf{Y}$. Specifically, denoting the (unobserved) absolute abundance of the m taxa for the observation by $\boldsymbol{\mu}$, we have the relationship $\boldsymbol{\mu}/\mathbf{1}'\boldsymbol{\mu} = \mathbf{P}$. We assume that there is a subset vector $\boldsymbol{\mu}(\mathbf{s})$ that is independent of $\mathbf{Y}$, except possibly through a change in the total sum $\mathbf{s}'\boldsymbol{\mu}(\mathbf{s})$. The dependence on the sum may occur, for example, if an increase in other taxa (with relation to $\mathbf{Y}$) caused this subset of taxa to be less prevalent, but the relationship between the coordinates of this subset is unchanged with $\mathbf{Y}$. Therefore, $\boldsymbol{\mu}(\mathbf{s})/\mathbf{s}'\boldsymbol{\mu} = \mathbf{P}(\mathbf{s})/\mathbf{s}'\mathbf{P}$ is independent of $\mathbf{Y}$ (see Mandal et al. (2015) for a similar assumption). Such a group of taxa can serve as a *reference set*, defined below, for pointing toward the discoveries of interest. We use the symbol $\perp\!\!\!\perp$ to mean that two random vectors are mutually independent.

DEFINITION 3.1. A set of taxa with indices $\{b_1, \ldots, b_r\}$ is a *reference set* if for the m-dimensional indicator vector $\mathbf{b}$ with exactly r ones at entries $(b_1, \ldots, b_r)$, $\mathbf{b}'\mathbf{P} > 0$ with probability one, and

$$\frac{\mathbf{P}(\mathbf{b})}{\mathbf{b}'\mathbf{P}} \perp\!\!\!\perp \mathbf{Y}. \tag{3.1}$$

Our goal is to find all taxa that are differentially abundant, that is, taxa which vary with $\mathbf{Y}$, given the reference set, while taking compositionality into account. For a given *reference set* of r taxa, let $\mathbf{b}_j$ be the m-dimensional binary vector with entries of one in $\{b_1, b_2, \ldots, b_r\}$ and in j, where $j \notin \{b_1, b_2, \ldots, b_r\}$. We want to test the null hypothesis that taxon j is not differentially abundant,

$$H_0^{(j)}: \quad \frac{\mathbf{P}(\mathbf{b}_j)}{\mathbf{b}_j'\mathbf{P}} \perp\!\!\!\perp \mathbf{Y}. \tag{3.2}$$

If $H_0^{(j)}$ is false, then the normalized vector of relative frequencies which includes taxon j and the reference set varies with $\mathbf{Y}$, and this variability is not a consequence of a change in the relative abundance of the sum of the taxon and the reference set. Thus, we would like to identify all taxa for which the null hypothesis in (3.2) is false. We note that if $H_0^{(j)}$ is true, subcompositional coherence (Aitchison (1986)) implies that 3.1 holds for the reference set.

More generally, we can consider testing a group of taxa together. Let $\mathbf{e_j}$ be the m-dimensional binary vector with entries of one in $\mathbf{j}$. Let $\mathbf{b_j}$ be the m-dimensional binary vector with entries of one in $\{b_1, b_2, \ldots, b_r\}$ and in $\mathbf{j}$, where $\mathbf{j}$ is a vector of indices satisfying $\mathbf{j} \cap \{b_1, b_2, \ldots, b_r\} = \varnothing$. We want to test the null hypothesis that none of the $\mathbf{j}$ taxa is differentially abundant,

$$H_0^{(\mathbf{j})}: \quad \frac{\mathbf{P}(\mathbf{b_j})}{\mathbf{b_j}'\mathbf{P}} \perp\!\!\!\perp \mathbf{Y}. \tag{3.3}$$

If $H_0^{(\mathbf{j})}$ is false, then the normalized vector of relative frequencies, which includes $\mathbf{j}$, and the reference set varies with $\mathbf{Y}$, and this variability is not a consequence of a change in the relative abundance of the sum of the taxa in $\mathbf{j}$ and the reference set.

We are unable to test these null hypotheses directly, since $\mathbf{P}$ is not observed. A simplified analysis, which assumes that $\mathbf{P}$ is fixed across observations with the same $\mathbf{Y}$ value, allows application of well-known tests, but such tests will have an inflated type I error probability when $\mathbf{P}$ varies across observations. For example, for a binary $\mathbf{Y}$ and a fixed $\mathbf{P}$ for each value of $\mathbf{Y}$, a test of whether $\mathbf{P}(\mathbf{b_j})/\mathbf{b_j'P}$ is identical across the two groups can be conducted by constructing a 2×2 table, tabulating the number of counts in j (or $\mathbf{j}$) and the reference set $\mathbf{b}$ across the two study groups, and performing the Fisher exact test. In Section 5 we show that the type I error probability of the Fisher exact test can be much higher than the nominal level when the assumption that $\mathbf{P}$, given $\mathbf{Y}$, is fixed is violated.

Another simplified analysis may replace the unobserved $\mathbf{P}$ with the observed $\mathbf{X}$ in the test of (3.2), thus rejecting (3.2) if the test of $\mathbf{X}(\mathbf{b}_j)/\mathbf{b}_j'\mathbf{X} \perp\!\!\!\perp \mathbf{Y}$ is rejected. This analysis is possible only if $\mathbf{b}_j'\mathbf{X}$ is nonzero for all samples. In Section 2 of the Supplementary Material (Brill, Amir and Heller (2022)), we show that adding a pseudocount may lead to severe inflation in the level of the test when the total number of counts observed in $\mathbf{X}(\mathbf{b}_j)$ depends on $\mathbf{b}_j'\mathbf{P}$ (since all differentially abundant taxa determine the support of $\mathbf{X}(\mathbf{b}_j)/\mathbf{b}_j'\mathbf{X}$). For example, for low values of $\mathbf{b}_j'\mathbf{X}$, the ratio $\mathbf{X}(\mathbf{b}_j)/\mathbf{b}_j'\mathbf{X}$ will be more affected by the addition of pseudocounts, feature less possible values, and will appear "more discrete," even if $\mathbf{P}(\mathbf{b}_j)/\mathbf{b}_j'\mathbf{P}$ does not depend on $\mathbf{Y}$.

Next, we present our suggested approach which takes the variability of $\mathbf{P}$ into account and requires that the sum of the taxa in the reference set be positive for all samples (so no addition of a pseudocount is necessary).

4. Testing for differential abundance. In Section 4.1 we detail our reference selection procedure for data analysis; in Section 4.2 we present our test for differential abundance, assuming our selected reference set is valid (i.e., satisfies (3.1)). In Section 4.3 we suggest a diagnostic test for whether the reference set is valid and discuss the implication of rejection by this diagnostic test.

4.1. *Choosing the reference taxa* $(b_1, \ldots, b_r)$. If domain knowledge exists regarding taxa which are independent of the trait, they can serve as the reference set. For example, spike-in of synthesized DNA (see Section "Spike-in log-ratio normalization" in Quinn et al. (2019)) or bacteria not endemic to the ecosystem studied (Staemmler et al. (2016)).

When there is no external knowledge about taxa that are independent of the trait, we need to both identify the reference taxa and then test with respect to this reference set, using the same dataset. If a large number of samples is available, the data can be split into two parts, the first part for reference selection and the second part for testing. The reference selection procedure may include all taxa that appear least associated with the trait in the first part.

It is also possible to select a reference set and then test for differential abundance, using the same dataset, without sacrificing part of the data for reference set selection, under the following sparsity condition: that the absolute abundance of the vast majority of taxa is independent of the trait (or, more precisely, that the ratio of absolute abundance of one taxon over the other is independent of the trait for most taxa pairs). If the score statistics used for reference set selection are independent of the test statistic used for testing (3.2) and (3.3), the selection step does not invalidate the inference (Hommel and Kropf (2005)). As a first principle our statistic for selection of reference taxa does not use the trait values. One approach used for identifying a compositional reference frame is by selecting taxa with low log-ratio variances (VLR, Aitchison (1986)), as suggested in Wu et al. (2017) and Quinn et al. (2019). We want to take a similar approach and suggest inserting a taxon with index j to the reference set if the log-ratio variances $V((\mathbf{P}_i(\mathbf{e}_j))/(\mathbf{P}_i(\mathbf{e}_k)))$ are relatively small for many of the possible

$k \in \{1, \ldots, m\} \setminus j$. Since the $\mathbf{P}_i$'s are not observed, the score for selection of taxon j is based on the variance of log-ratios of counts and defined as follows:

$$S_j = \text{median}_{\{k:k\neq j, k=1,\ldots,m\}}(\text{SD}_{j,k}), \quad \text{SD}_{j,k} = \text{sd}\left\{\log_{10}\left(\frac{\mathbf{X}_i(\mathbf{e}_j)+1}{\mathbf{X}_i(\mathbf{e}_k)+1}\right) : i = 1, \ldots, n\right\},$$

where sd is the sample standard deviation taken over n values. The score S_j will tend to have higher values for differentially abundant taxa; see Section 6.1 of the Supplementary Material (Brill, Amir and Heller (2022)) for a comparison of the distributions of S_j's in differentially and nondifferentially abundant taxa.

We suggest thresholding the scores, so the reference set is $B(S_{\text{crit}}) = \{j | S_j \leq S_{\text{crit}}\}$. The minimal number of counts observed in the reference set $B(S_{\text{crit}})$ is $\lambda_{\min}(S_{\text{crit}}) = \min_{i=1,\ldots,n} \mathbf{b}' X_i$, where $\mathbf{b}$ is the binary vector with entry one for taxa in the reference set $B(S_{\text{crit}})$. The value of S_{crit} is set so the minimum count for each sample is appropriate in the following sense. We want the minimum count to be large enough for good power when testing (3.2) (or (3.3)), but we also want it to be no greater than necessary from power considerations in order to minimize the risk of contamination of the reference set with differentially abundant taxa. A brief discussion of a power analysis for choosing S_{crit} is provided in Section 7. In our simulations in Section 5 and Section 4 of the Supplementary Material (Brill, Amir and Heller (2022)), we select S_{crit} to be the lowest value such that $\lambda_{\min}(S_{\text{crit}}) \geq 100$.

The definition of $\text{SD}_{j,k}$ uses a pseudocount of one which is perhaps a natural but somewhat arbitrary value. We expect the choice of pseudocount value to have little impact on the procedure (as long as the pseudocount value is not too small). For example, the interquantile log ratio (IQLR, Wu et al. (2017), Quinn et al. (2019)) adds a pseudocount of 0.5 to the counts before computing log ratio variances. For any selected value of the pseudocount, the diagnostic procedure of Section 4.3 can be used to assess the validity of the selected reference set.

Figure 1(A) shows the distribution of the S_j's and $\lambda_{\min}(S_{\text{crit}})$ as a function of S_{crit} for the data presented in Section 2. We set $S_{\text{crit}} = 1.163$, which is the value where the function $\lambda_{\min}(S_{\text{crit}})$ for the CD data shows a jump, with $\lambda_{\min}(S_{\text{crit}})$ jumping from 56 to 126 counts. The selected reference set has 1197 taxa.

We end this section with two final remarks. The first remark regards the important observation that, in order to avoid biased inference, none of the samples should be excluded from the test of differential abundance (e.g., due to zero entries). Our approach for selecting a reference set is guided by this observation. The implication is that all samples must have a positive number of counts in the reference. In Section 3 of the Supplementary Material (Brill,

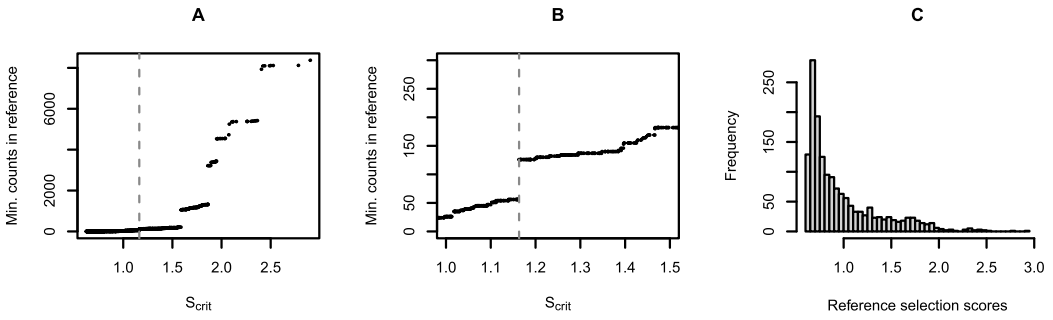


FIG. 1. *For the CD data presented in Section 2: the minimum number of reads (across samples) in a reference set $B(S_{\text{crit}})$ as a function of S_{crit} for all possible score values (left panel) and for scores between 1 and 1.5 (middle panel), and the histogram of the 1569 S_j's (right panel). The vertical line at $S_{\text{crit}} = 1.163$ is the threshold used for the CD data analysis in Section 6.*

Amir and Heller (2022)), we provide examples that show that if samples are excluded from analysis based on the the total number of reads in the reference, then the test for differential abundance may be biased. Samples with extremely low sampling depth (technical faults) may be removed from the entire analysis if it is reasonable to assume $\mathbf{Y}$ or $\mathbf{P}$ are independent of the total number of counts per sample.

The second remark regards the effect of using naive approaches for reference set selection. In Section 6.3 of the Supplementary Material (Brill, Amir and Heller (2022)), we examine approaches selecting the reference set to be the 50 most abundant taxa or selecting 50 taxa at random. We show these approaches can result in biased inference with an unacceptably high inflation of false positives.

4.2. *Discovering differentially abundant taxa.* Let $\{b_1, \ldots, b_r\}$ be the selected reference set of taxa for sample $\{(\mathbf{X}_i, \mathbf{Y}_i) : i = 1, \ldots, n\}$. One approach to consider is rejection of the null hypothesis (3.2) if the null hypothesis of independence between $\mathbf{X}(\mathbf{e}_j)/\mathbf{b}_j'\mathbf{X}$ and $\mathbf{Y}$ is rejected, using an appropriate level α test for the data

$$\{(\mathbf{X}_i(\mathbf{e}_j)/\mathbf{b}_j'\mathbf{X}_i, \mathbf{Y}_i) : i = 1, \ldots, n\}.$$

We will refer to this method of differential abundance testing as *normalization by ratio*. Our motivation for considering this test for differential abundance is the fact that if the null hypothesis (3.2) is true, then the conditional expectation of $\mathbf{X}(\mathbf{e}_j)/\mathbf{b}_j'\mathbf{X}$, given $\mathbf{P}$, is independent of $\mathbf{Y}$, as follows from the result formally stated in the next proposition.

PROPOSITION 1. *If* $\mathbf{X} \mid N, \mathbf{P} \sim \text{multinom}(N, \mathbf{P})$, *then*

$$E\left\{\frac{\mathbf{X}(\mathbf{e}_j)}{\max(1, \mathbf{b}_j'\mathbf{X})} \,\middle|\, \mathbf{P}\right\} = \frac{\mathbf{P}(\mathbf{e}_j)}{\mathbf{b}_j'\mathbf{P}} \Pr(\mathbf{b}_j'\mathbf{X} > 0 \mid \mathbf{P}).$$

See Section 1 of the Supplementary Material (Brill, Amir and Heller (2022)) for the proof. However, even if (3.2) is true, that is, $\mathbf{P}(\mathbf{b}_j)/\mathbf{b}_j'\mathbf{P} \perp\!\!\!\perp \mathbf{Y}$, $\mathbf{b}_j'\mathbf{X}$ may depend on $\mathbf{Y}$, so the spread of $\mathbf{X}_i(\mathbf{e}_j)/\mathbf{b}_j'\mathbf{X}_i$ may depend on $\mathbf{Y}$ when the null hypothesis (3.2) is true. Therefore, this approach can be approximately valid at best. The next example demonstrates the limits of the "normalization by ratio" approximation.

EXAMPLE. Inflated false positive error rates when using the normalization by ratio approach. We consider a binary trait $\mathbf{Y}$, and $m = 3$ taxa, with taxa relative abundances distributed $\mathbf{P} \sim \text{Dirichlet}(90, 0.1, 10)$ in the first group (with $\mathbf{Y} = 0$) and $\mathbf{P} \sim \text{Dirichlet}((1 - w) \times 90 + w, (1 - w) \times 0.1, (1 - w) \times 10)$ in the second group (with $\mathbf{Y} = 1$). We observe 50 samples from each group. The sequencing depth is 10^4 for all samples. We test the second taxon for differential abundance with the third taxon serving as reference, that is, whether $\mathbf{P}(\mathbf{e}_2)/(\mathbf{P}(\mathbf{e}_2) + \mathbf{P}(\mathbf{e}_3)) \perp\!\!\!\perp \mathbf{Y}$. This null hypothesis is true for all values of w since the distribution of $P(\mathbf{e}_2)/(P(\mathbf{e}_2) + P(\mathbf{e}_3))$ is Beta(0.1, 10), that is, it is independent of $\mathbf{Y}$ (due to properties of the beta dist.; see p. 59 of Forbes et al. (2011)). However, $\mathbf{X}(\mathbf{e}_2)/(\mathbf{X}(\mathbf{e}_2) + \mathbf{X}(\mathbf{e}_3)) \perp\!\!\!\perp \mathbf{Y}$ is false, since the distribution of $X(\mathbf{e}_2)/(X(\mathbf{e}_2) + X(\mathbf{e}_3))$ differs between the two groups for $w > 0$; see Figure 2. Testing using the normalization by ratio approach, by applying the Wilcoxon rank-sum test or the Welch t-test, leads to an inflated number of false positives. The inflation increasing with w, when testing at level $\alpha = 0.1$ using the Wilcoxon rank-sum test, the estimated false positive rates for $w = 0, 0.5, 0.75$, are 0.1, 0.4, and 0.82, respectively; when testing using the Welch t-test, the estimated false positive rates are 0.1, 0.12, and 0.18, respectively.

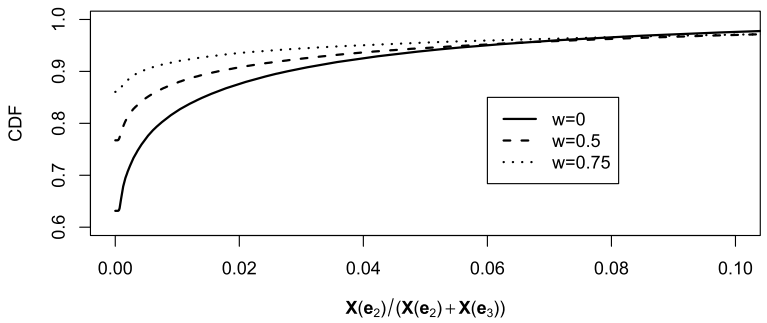


FIG. 2. *In the example of Section 4.2, the cumulative distribution function (CDF) of* $\mathbf{X}(\mathbf{e}_2)/(\mathbf{X}(\mathbf{e}_2)+\mathbf{X}(\mathbf{e}_3))$ *in the second group for different values of* w. *The CDF for* $w=0$ *coincides with the CDF of* $\mathbf{X}(\mathbf{e}_2)/(\mathbf{X}(\mathbf{e}_2)+\mathbf{X}(\mathbf{e}_3))$ *in the first group.*

A valid testing approach using the reference set relies on a key observation that if the null hypothesis (3.2) (or (3.3)) is true, then rarefying the subvectors $\mathbf{X}(\mathbf{b}_j)$ (or $\mathbf{X}(\mathbf{b_j})$) to an equal number of counts implies that the rarefied subvectors will be independent of the trait $\mathbf{Y}$. Therefore, rejection of the hypothesis of independence between these rarefied subvectors and $\mathbf{Y}$ will lead to rejection of (3.2) (or (3.3)) with the desired nominal type I error level control guarantee. Specifics follow.

We start by describing our approach for testing (3.2). For taxon $j \in \{1,\dots,m\}$, the conditional distribution of $\mathbf{X}(\mathbf{e}_j)$, given $\mathbf{b}_j'\mathbf{X}$ and $\mathbf{P}$, is binomial with parameters $\mathbf{b}_j'\mathbf{X}$ and $\mathbf{P}(\mathbf{e}_j)/\mathbf{b}_j'\mathbf{P}$. We can eliminate the dependence on the total count, $\mathbf{b}_j'\mathbf{X}$ (which may depend on $\mathbf{Y}$), by rarefying to a depth $\lambda_j = \min_{i=1,\dots,n} \mathbf{b}_j'\mathbf{X}_i > \lambda_{\min}(S_{\text{crit}})$ (step 2 in the algorithm below). The distribution of the rarefied count is binomial with parameters λ_j and $\mathbf{P}(\mathbf{e}_j)/\mathbf{b}_j'\mathbf{P}$; see Lemma 1 in Section 1 of the Supplementary Material (Brill, Amir and Heller (2022)) for details. Therefore, if (3.2) is true, the rarefied count for $\mathbf{X}(\mathbf{e}_j)$ is independent of $\mathbf{Y}$, even if $\mathbf{b}_j'\mathbf{X}$ depends on $\mathbf{Y}$, so the steps toward a valid test are as follows:

1. Compute the minimum total counts of the taxon and the reference set, $\lambda_j = \min_{i=1,\dots,n} \mathbf{b}_j'\mathbf{X}_i$.
2. For each observation $i = 1,\dots,n$, sample a count from the conditional hypergeometric distribution (given the observed count for taxon j and the total count in the reference set) with parameters λ_j, $\mathbf{X}_i(\mathbf{e}_j)$, $\mathbf{b}_j'\mathbf{X}_i$, where hypergeom$(t,z,z+w)$ is the distribution of the number of special items sampled (without replacement) when selecting t distinct items from a population of $z+w$ items, z of which are special. The sampled count is denoted by $Z_{i,j}$.
3. Test the null hypothesis of independence between the rarefied count $Z_{i,j}$ and $\mathbf{Y}_i$ using an appropriate α level test of independence on the n pairs of observations $\{(Z_{i,j},\mathbf{Y}_i) : i = 1,\dots,n\}$.

PROPOSITION 2. *If the null hypothesis* (3.2) *is true, then the aforementioned testing procedure has level* α.

See Section 1 of the the Supplementary Material (Brill, Amir and Heller (2022)) for a proof.

The dimension and possible values of $\mathbf{Y}$ dictate which test is performed. For a univariate $\mathbf{Y}$, the choice is among tests for equality of distributions, if $\mathbf{Y}$ is categorical, and among tests of independence between random variables if $\mathbf{Y}$ is continuous. For a multivariate $\mathbf{Y}$, the choice is among tests of independence between a univariate random variable and a multivariate vector (Gretton et al. (2008), Székely and Rizzo (2009), Heller, Heller and Gorfine (2013)).

The procedure for the test of (3.3) is similar to the procedure for the test of (3.2). It uses the multivariate hypergeometric (MHG) distribution defined as follows. Let $\mathbf{U} \sim \mathrm{MHG}(\lambda, \mathbf{v}, M)$ denote a sample from the multivariate hypergeometric distribution, so $\mathbf{U}$ is a random vector of dimension p formed by counting the number of balls of types 1, ..., p, when sampling λ balls, without replacement, from an urn containing M balls, out of which the number of balls of type i is v_i, for $i = 1, \ldots, p$, and $\mathbf{v} = (v_1, \ldots, v_p)$. For our test we use the MHG distribution as follows. Let $\lambda_{\mathbf{j}} = \min_{i=1,\ldots,n} \mathbf{b}_{\mathbf{j}}'\mathbf{X}_i$. For the ith observation, sample $\mathbf{Z}_{i,\mathbf{j}} \sim \mathrm{MHG}(\lambda_{\mathbf{j}}, \mathbf{X}_i(\mathbf{e}_{\mathbf{j}}), \mathbf{b}_{\mathbf{j}}'\mathbf{X}_i)$. A test of independence between $\mathbf{Z}_{i,\mathbf{j}}$ and $\mathbf{Y}_i$ is a valid test for (3.3), using the same reasoning as in the proof for Proposition 2.

If $\mathbf{Y}$ is categorical, the popular PERMANOVA test (Anderson (2001)) can be used on the vector of rarefied counts, as in our multivariate analysis of the data in Section 6.2. If $\mathbf{Y}$ is multivariate, the choice is among tests of independence between two random vectors (Gretton et al. (2008), Székely and Rizzo (2009), Heller, Heller and Gorfine (2013)). The choice of test should take into account the number of taxa examined (in particular, whether this number is greater than the sample size).

Another testing procedure for (3.3) we consider is using *normalization by ratio*. This procedure rejects the null hypothesis using a level α multivariate test for the data,

$$\big\{\big(\mathbf{X}_i(\mathbf{e}_{\mathbf{j}})/\mathbf{b}_{\mathbf{j}}'\mathbf{X}_i, \mathbf{Y}_i\big) : i = 1, \ldots, n\big\}.$$

We end this section with a remark about the rarefying step 2 toward a valid test. Regarding the criticism that normalization by rarefaction leads to an inefficient analysis since only part of the data are used for inference (McMurdie and Holmes (2014)). We view rarefaction favorably despite this fact, since it enables inference with no parametric assumptions on the distribution of $\mathbf{P}$ or on the structural zeros. Arguably, the potential power loss, due to rarefaction, is worth the gain in assurance that the correctness of discoveries does not hinge on model assumptions and sequencing resolution. We support our argument via examples and extensive simulations in Sections 5–6.

4.3. *A diagnostic check for contamination in the reference set.* We can test that none of the taxa in the reference set are differentially abundant as follows:

1. For each reference taxon, test whether it is differentially abundant with respect to the reference set, excluding that taxon, using the test described in Section 4.2. Let $p_{b_1}, \ldots, p_{b_r}$ denote the resulting p-values for the reference set with indices $\{b_1, \ldots, b_r\}$.
2. Test the intersection hypothesis that none of the taxa in the reference set are differentially abundant using Simes combination test, so contamination is established using a level α test if

$$p_{\mathrm{Simes}} = \min_{j=1,\ldots,r} \frac{r \times p_{(j)}}{j} \le \alpha,$$

where $p_{(1)} \le \cdots \le p_{(r)}$ are the sorted $p_{b_1}, \ldots, p_{b_r}$.

Although Simes test is only valid for p-values that are independent or have a type of positive dependence (Simes (1986)), we prefer it over the valid Bonferroni test, $r \times p_{(1)} \le \alpha$, since it is uniformly more powerful than Bonferroni. It is possible to generate the null permutation distribution of the test statistic p_{Simes} and thus obtain a valid p-value, instead of the approximation p_{Simes}, using more demanding computational resources.

Due to lack of power, we may observe $p_{\mathrm{Simes}} > \alpha$, even though the reference set is contaminated. In our simulations we observed this approach to be well powered for detecting contaminations in the reference set and that undetected contaminations did not result in an inflation of false positives in the set of differential abundance discoveries; see Section 6.2 of the Supplementary Material.

On the other hand, if the study is very well powered, we may observe $p_{\text{Simes}} < \alpha$, even though the reference set is not really contaminated but just because tiny effects are detected with a large enough sample size. We suggest adding the following post hoc analysis to our pipeline if $p_{\text{Simes}} < \alpha$:

1. Shrink the reference set to include less "differentially abundant" looking taxa. For example, if the original reference set was selected using S_{crit}, which was the lowest possible value, so that $\lambda_{\min}(S_{\text{crit}}) \geq 100$, use, for example, 90 instead of 100 as the threshold for reference set selection.
2. Test for contamination by computing p_{Simes} on the shrunk reference set. If $p_{\text{Simes}} < \alpha$, return to the previous step and shrink the reference set further. Repeat until $p_{\text{Simes}} > \alpha$ or if the minimal count is a prespecified minimal size, say 20 counts (the minimal size may be reached for studies that are very well powered).

In Section 6.2 of the Supplementary Material (Brill, Amir and Heller (2022)), we conduct a simulation study in order to investigate the situation where differentially abundant taxa enter the reference set. We empirically show our approach to be quite robust to differentially abundant taxa entering the reference set in terms of false positive rate control and statistical power, even for excessively high values of $\lambda_{\min}(S_{\text{crit}})$, for example, 3000 or 7000. We empirically show that the diagnostic check for detecting contamination in the reference is well powered and that by shrinking the reference set by lowering S_{crit}, we are able to remove contaminations from the reference set. The post hoc analysis with shrunken reference sets is shown to provide good false positive rate control and power.

5. A simulation study. A simulation study was performed to compare the power and error rate control of various tests for discovering the differentially abundant taxa. We focus on settings where $\mathbf{Y}$ is a binary variable indicating group membership. Our data generation procedure, described in Section 5.1, uses the data of Vandeputte et al. (2017), so methods are compared in a setting that resembles our data set of interest.

The new procedures, described in Section 4, for differential abundance testing with compositionality adjustment are denoted by DACOMP, DACOMP-t, and DACOMP-ratio. These tests use a reference set that is adaptively chosen from the data, as described in Section 4.1, with S_{crit} chosen to be the minimal value such that $\lambda_{\min}(S_{\text{crit}}) \geq 100$. The chance that differentially abundant taxa erroneously enter the reference set was negligible in the vast majority of our simulated settings; see Section 6.4 of the Supplementary Material (Brill, Amir and Heller (2022)) for details for details. DACOMP and DACOMP-t follow the procedure in Section 4.2, and they differ only with regard to the two-sample test carried out in Step 3: Wilcoxon rank-sum test for DACOMP, and Welch two sample t-test on transformed counts, $\log(Z_{i,j} + 1)$, for DACOMP-t. DACOMP-t uses the pseudocount in order to log-transform the counts $Z_{i,j}$, which may be right skewed; however, the use of the pseudocount does not effect the validity of the test, since the $Z_{i,j}$'s are already independent of the group labeling if the null hypothesis is true. DACOMP-ratio follows the *normalization by ratio* approach in Section 4.2 with Wilcoxon rank-sum test as the two-sample test. Since the normal approximation for the distribution of the Wilcoxon rank-sum test statistic may be inaccurate in the presence of many ties in the data (caused by the many zeros), and the t approximation for the distribution of Welch's t statistic may be inaccurate if the distribution of $Z_{i,j}$ has heavy tails, P-values for all DACOMP variants (as well as W-TSS,W-CSS, and W-Flow mentioned below) were computed using permutations, as suggested in Tsagris et al. (2020). Tsagris et al. (2020) showed that, when the P-value of the Welch t-test is computed using permutations, the test is powerful against a wide range of scenarios in addition to the normal shift setting.

We compare the new procedures to the following: ANCOM (Mandal et al. (2015)), as implemented in version 1.1-3 of the *ANCOM* package; W-FLOW, Wilcoxon rank sum tests

with the correction by Vandeputte et al. (2017); W-CSS and W-TSS, Wilcoxon rank sum tests with the CSS and TSS normalization, respectively, with W-CSS, as implemented in the software package *metaGenomeSeq* in R (Paulson, Pop and Bravo (2013)) in version 1.24-1; ALDEx2-t and ALDEx2-W (Fernandes et al. (2013)), using the two-sample Welch t-test and Wilcoxon rank-sum test, respectively, as implemented in version 1.16-0 of the *ALDEx2* package; WRENCH (Kumar et al. (2018)), implemented in version 1.2-0 of the *wrench* package, with default parameters (it makes use of the tests of differential abundance implemented in the "deseq2'' software package (Love, Huber and Anders (2014))); ZINB-WAVE, the method combination shown in Van den Berge et al. (2018), using the parameter configuration of Calgaro et al. (2020); HG, Fisher's exact test against a reference set. The reference set for HG was the oracle set that includes all nondifferentially abundant taxa with $S_{\text{crit}} = 1.3$ in order to demonstrate that the test is biased due to a failure to account for over dispersion (rather than due to the reference set being contaminated with signal).

For error control we chose the false discovery rate (FDR, Benjamini and Hochberg (1995)). ANCOM carries out its own multiplicity correction aimed at FDR control. For all other methods we applied the Benjamini–Hochberg (BH) procedure (Benjamini and Hochberg (1995)) at level $q = 0.1$. We chose the BH procedure since empirical evidence and simulations suggest it controls the FDR for most dependencies encountered in practice, including microbiome applications (Jiang et al. (2017)), even though the theoretical guarantee is only for independence or a type of positive dependence. The family of tests is smaller for the new DACOMP tests than for the other tests, since the taxa in the reference set are not tested for differential abundance.

We display the results for ANCOM, W-FLOW, W-CSS, DACOMP, DACOMP-ratio, ALDEx2-t, and HG which represent key approaches. The other results are detailed in Section 4 of the Supplementary Material (Brill, Amir and Heller (2022)), since they are slight variations on the key approaches or have a higher inflation of false positives. We also considered in Section 4 of the Supplementary Material (Brill, Amir and Heller (2022)) the following additional settings: (a) a setting where sequencing depth varies across groups (as discussed, e.g., in Weiss et al. (2017)) for which we show that only DACOMP and DACOMP-t provide adequate control over false positives; (b) a (unrealistic) setting where the total microbial load of the differentially abundant taxa is identical across study groups, so marginal methods provide a valid method of testing, since there is no bias due to compositionality, for which we show that the loss of power when using DACOMP is small, and (c) a setting where only the rare taxa are differentially abundant, causing a severe inflation of false positives for some competitor methods. The simulation results are based on 100 replications.

5.1. *Data generation*. As a basis for the simulation, we use the 16S samples and flow cytometric measurements obtained from the 66 healthy subjects in the data set described in Section 2. All sOTUs which appeared in less than four subjects were removed from the data, leaving $m = 1066$ sOTUs. The median number of reads across subjects was $N_{\text{reads}} = 22449$ reads.

For a simulated dataset, a total of 60 "healthy" and 60 "sick" subjects were sampled. The absolute abundance vector for the "healthy" ith sample, was generated by the following steps: (1) pick at random a healthy subjects from the original CD data, $k \in \{1, \ldots, 66\}$, and let $\mathbf{u}_k$ and C_k^{flow} denote the 16S vector of counts and the flow cytometric measurement measured in Vandeputte et al. (2017) for this sample; (2) set the (unobserved) absolute abundances to be $\boldsymbol{\mu}_i^H = C_k^{\text{flow}} \times \mathbf{u}_k / \mathbf{1}'\mathbf{u}_k$, H standing for "healthy."

In order to simulate the absolute abundance in the "sick" study group, $m_1 \in \{10, 100\}$ taxa were chosen to have increased absolute abundances compared to the 'healthy" study group, with the effect size determined by a value of $\delta_j \in \{0.5, 1.0, 1.5\}$ chosen at random, with equal

probabilities, for each of the taxa. Absolute abundance for "sick" subjects were simulated as follows: (1) simulate taxa absolute abundances for sick subjects, denoted by $\boldsymbol{\mu}_i^S$, in the same way as for healthy subjects; (2) update the absolute abundance of taxa associated with the disease. Specifically, if taxon j was chosen as differentially abundant, we updated its absolute abundance via: $\boldsymbol{\mu}_i^S(\mathbf{e}_j) \leftarrow \boldsymbol{\mu}_i^S(\mathbf{e}_j) + \text{Ber}(0.5) \cdot N(\rho_{i,j}, \rho_{i,j})$, where $\rho_{i,j} = \Delta_{\text{ml}} \times C_i^{S,\text{flow}} \times \delta_j / m_1$, $C_i^{S,\text{flow}}$ is the sum of $\boldsymbol{\mu}_i^S$ before any simulated increments in the second step, and Ber marks a Bernoulli random variable. The expected increase in the host microbial load, due to the simulated conditioned, is determined by $\Delta_{\text{ml}} \in \{0, 0.5, 1.0, \ldots, 3.0\}$. For example, with $\Delta_{\text{ml}} = 1.0$ there is an expected increase of 100% in the total host microbial load, distributed over the m_1 differentially abundant taxa. Clearly, the resulting absolute abundance vector of taxa, $\boldsymbol{\mu}_i^S$, differs in distribution from $\boldsymbol{\mu}_i^H$ only in the m_1 coordinates where counts were added, and only for these coordinates the null hypothesis (3.2) is false.

Finally, the observed counts were sampled using the multinomial distribution, with taxa proportions based on the $\boldsymbol{\mu}_i^H$'s and $\boldsymbol{\mu}_i^S$'s (after normalizing their proportions to a sum of 1) and with the number of reads in each count vector selected at random from the Poisson distribution with parameter N_{reads}. The sums of $\boldsymbol{\mu}_i^H$'s and $\boldsymbol{\mu}_i^S$'s were reported as simulated flow-cytometric measurements.

5.2. *Results*. Results in this section demonstrate the validity and good power properties of our method, DACOMP, and the potential inflation of false positives of other methods. In numerical comparisons we show that with DACOMP-ratio we can gain power but at a price of an inflation in the type 1 error probability. However, this inflation is typically small in comparison with the inflation incurred by other methods.

Figure 3 shows the estimated FDR and power for each method, for the different scenarios. DACOMP is the only method controlling FDR across all scenarios considered. For the global null setting ($\Delta_{\text{ml}} = 0$), only ANCOM and HG do not control the FDR. For HG this is expected since $\mathbf{P}|\mathbf{Y}$ is not constant in each study group. For ANCOM we have observed that generally, under the global null, FDR is not controlled. In Section 7 of the Supplementary Material (Brill, Amir and Heller (2022)), we present additional scenarios with no differentially abundant taxa where ANCOM does not control the FDR. ANCOM and W-FLOW lack FDR control when $\Delta_{\text{ml}} \geq 2.0$. For ANCOM this could be attributed either to the empirical decision rule being invalid or to mistreatment of technical zeros by using a pseudocount. For

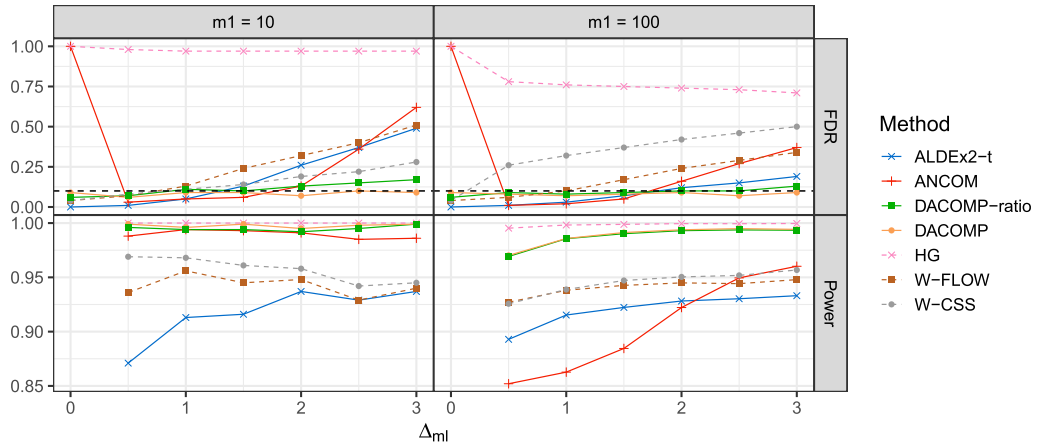


FIG. 3. *Estimated FDR and power vs.* Δ_{ml} *for DACOMP and competitors in the simulation settings of Section 5.1. The level of the BH procedure was* $q = 0.1$ *(in dashed black). The maximal standard error for FDR and power was* 0.04 *and* 0.0072, *respectively.*

W-FLOW the lack of FDR control can be attributed to mistreating technical zeros as well: W-FLOW uses a multiplicative factor to correct for compositional bias, providing no solution for technical zeros. ALDEx2-t provides FDR control for $m_1 = 100$ but not for $m_1 = 10$. For DACOMP-ratio, the inflation is largest with $\Delta_{\text{ml}} = 3$, with a maximum realized FDR value of 0.17.

For $m_1 = 10$, the power is close to one for all methods. For $m_1 = 100$, DACOMP has the highest statistical power, despite being the only valid procedure. The increase in power results mainly from excluding the reference set of taxa from testing: the mean size of selected reference sets across scenarios varied from 506 for $m_1 = 100$ and $\Delta_{\text{ml}} = 0.5$, to 691 for $m_1 = 10$ and $\Delta_{\text{ml}} = 3.0$ (the standard error was <15). While DACOMP has the highest expected number of true discoveries, its expected number of discoveries is substantially lower, as other methods do not provide adequate FDR control. For example, for the case where $\Delta_{\text{ml}} = 2.5$ and $m_1 = 100$, W-CSS has 176 discoveries, on average, but only 95 true discoveries.

6. Analysis of the Crohn's disease data. We proceed to analyze the data set presented in Section 2. DACOMP is particularly suitable for this analysis, as it is the only method that provides valid inference when the total number of reads is associated with the sample's group identity; see additional details in Section 4.2 of the Supplementary Material (Brill, Amir and Heller (2022)). Our units for analysis were the 1569 ASVs present in at least two subjects. The reference set for analysis was selected, as described in Section 4.1, and passed the diagnostic check of Section 4.3 with $p_{\text{Simes}} = 0.14$; see Section 5.6 of the Supplementary Material (Brill, Amir and Heller (2022)) for an analysis with an alternative reference set selection that uses the flow cytometric measurements available in this dataset.

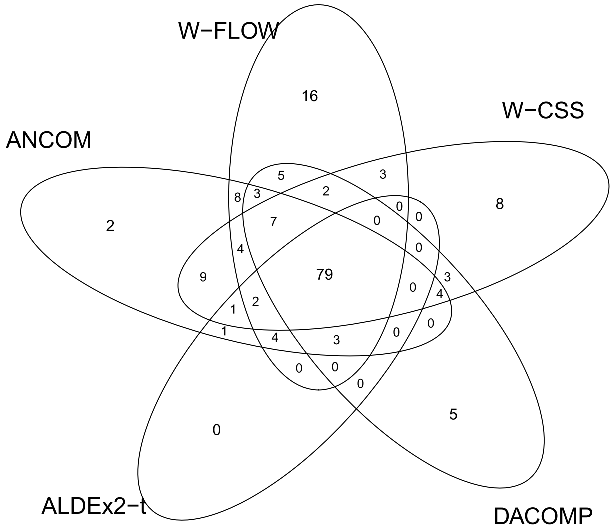
6.1. *A univariate analysis.* Table 2 shows the number of taxa discovered for each method along with the number of discoveries shared by the different methods. Procedures DACOMP and ALDEx-t have a similar number of discoveries, and it is substantially lower than ANCOM, W-FLOW, and W-CSS. W-FLOW uses an additional flow-cytometric measurement, yet it has a lower number of discoveries than ANCOM and W-CSS.

Of course, a reduced number of discoveries does not indicate that there are less true discoveries. Arguably, the discoveries with DACOMP are more trustworthy from the theoretical guarantees in Section 4 and the empirical evidence, when truth is known in Section 5, as well as from the following analysis that examines the agreement between methods. Figures 4 and 5 depict the number of discoveries shared by each method, for abundant and "rare" taxa, respectively. Abundant and "rare" taxa were defined to be taxa which have, on average per subject, at least 10 counts or less than 10 counts, with 264 of 1569 taxa being abundant. The

TABLE 2

Number of discoveries by each method, for the data of Vandeputte et al. (2017). The number of discoveries by each method on the diagonal and shared with the other methods on the off-diagonal entries. For DACOMP and DACOMP-ratio $S_{\text{crit}} = 1.163$

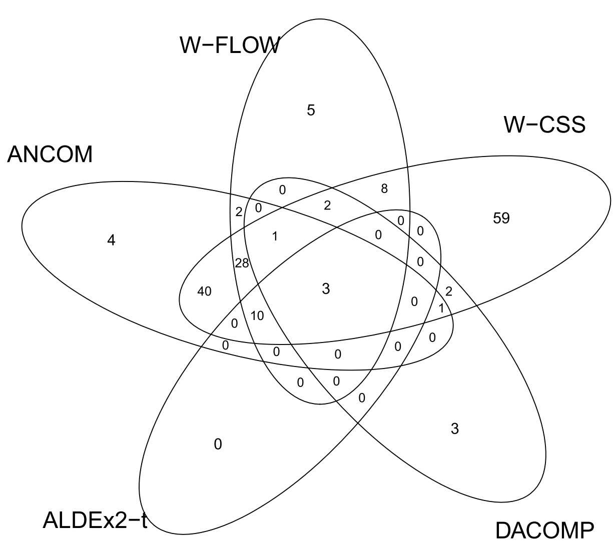
Method	ANCOM	W-FLOW	W-CSS	ALDEx2-t	DACOMP	DACOMP-ratio
ANCOM	216	159	189	103	127	151
W-FLOW		217	157	102	131	149
W-CSS			277	95	123	166
ALDEx2-t				103	95	102
DACOMP					151	143
DACOMP-ratio						213



(a)

FIG. 4. *Graphical representation of discoveries shared by different methods for the* 264 *taxa with at least* 10 *counts*, *on average*, *per sample*.

methods compared agree fairly well on which of the abundant taxa are differentially abundant. For rare taxa, methods have higher disagreement. For abundant taxa the majority of discoveries are shared by all methods. Only 12 of the discoveries made by DACOMP are not shared with W-FLOW which uses flow-cytometric measurements. For rare taxa, W-CSS has 59 unique discoveries not shared by any other method. No other method discovers such a high number of differentially abundant taxa that are rare. The majority of taxa, discovered by both ANCOM and W-FLOW but not DACOMP, are found among the rare taxa, 40 out of 58 taxa. When comparing discoveries of "rare" taxa by the different methods to W-Flow,



(a)

FIG. 5. *Graphical representation of discoveries shared by different methods for the* 1305 *"rare" taxa with less than* 10 *counts*, *on average*, *per sample*.

ANCOM, and W-CSS discover 45 and 102 "rare" taxa, respectively, in addition to the ones discovered by W-Flow; on the other hand, DACOMP and ALDEx2-t discover either six or no additional "rare" taxa as differentially abundant, compared to W-FLOW. We conclude that discoveries involving abundant taxa are replicated across different methods (relying on different assumptions), while discoveries involving rare taxa are not replicated.

6.2. *A multivariate analysis*. In order to identify the genera which are differentially abundant, the sOTUs were assigned taxonomy level data using a taxonomy classifier (Wang et al. (2007)), as implemented in the `assignTaxonomy` function of the `dada2` (Callahan et al. (2016)) software package, Version 1.18. The classifier used the Green Genes taxonomic training set (version 13.8, DeSantis et al. (2006)) as a reference database, and parameter `minBoot` $= 80$, 955 sOTUs were assigned to 61 genera that contained more than a single sOTU, with a median genus size of 5 sOTUs.

For a specific genus that includes the taxa with indices $\mathbf{j}$, we tested the null hypothesis (3.3) by applying the PERMANOVA test in order to discover whether the rarefied counts, $\mathbf{Z}_{i,\mathbf{j}}$, for the DACOMP approach, or $\mathbf{X}_i(\mathbf{e_j})/\mathbf{b}_\mathbf{j}'\mathbf{X}_i$, for the DACOMP-ratio approach, are associated with CD status. Our metric was the robust Mahalanobis distance detailed in Chapter 8.3 of Rosenbaum (2010) which is robust to outliers and takes the correlation among counts into account. We also tested (3.2) by treating each genus as a taxon, where the observed taxon count is the sum of sOTU counts in the genus. We applied the BH procedure at level 0.1 on the 61 genera. From Table 3 we see that the number of discoveries is about the same using the multivariate or univariate analysis at the genera level with an overlap of 2/3 discoveries (using DACOMP, and separately, using DACOMP-ratio).

6.3. *Examining signals using "within-genus" PCoA analysis*. In order to examine multivariate differential abundances, we extend the method of the principal coordinates analysis (PCoA, Goodrich et al. (2014)) for graphical ordination of microbiome data samples. Classic PCoA (Torgerson (1952)) receives as input a distance or dissimilarity matrix between data points and outputs a two-dimensional plot with points in the graph corresponding to data samples. Point locations are set so that relative distances between points in the graph approximate the relative distances given by the original distance or dissimilarity matrix. Microbiome PCoA analysis typically begins with rarefying 16S count vectors to constant depth, followed by plotting the PCoA graph using the Unifrac distance or Bray–Curtis (BC) dissimilarity measure. The "within-group" and "between-group" distances in the graph are then examined for meaningful biological patterns on the ecosystem level.

TABLE 3

Number of genera discovered (out of 61) as differentially abundant using the PERMANOVA test in the DACOMP approach (Multi) and the DACOMP-ratio approach (Multiratio), and using the Wilcoxon test at the genera level in the DACOMP approach (Uni) and the DACOMP-ratio approach (Uniratio). The number of discoveries by each method on the diagonal and shared with the other methods on the off-diagonal entries

DACOMP:	Multi	Multiratio	Uni	Uniratio
Multi	20	18	15	17
Multiratio		33	16	21
Uni			22	22
Uniratio				34

We examine the distribution of counts across genus **j**, compared to the total number of counts in both genus **j** and the reference taxa, by constructing PCoA plots using a distance metric between $\mathbf{Z}_{i,\mathbf{j}}$'s. Our approach naturally fits this analysis since distances are computed after accounting for the varying number of reads available under the subset of taxa analyzed; hence, no additional rarefaction steps are needed. We term this approach "within-genus" PCoA analysis. We use the BC distance measure; see discussion on alternative distance metrics in Section 5.4 of the Supplementary Material (Brill, Amir and Heller (2022)). In order to avoid a setting where some $\mathbf{Z}_{i,\mathbf{j}}$'s contain only zeros and the BC metric cannot be computed, we append a coordinate to $\mathbf{Z}_{i,\mathbf{j}}$'s which contains the count $\lambda_j - \sum_{j\in\mathbf{j}} Z_{i,j}$. For this setting, where the BC measure is computed over subvectors with identical sums, values for the BC measure are equivalent to the ℓ_1 distances up to a constant scaling factor.

"Within-genus" PCoA plots allow for an unsupervised analysis of the data, as the next example shows. Using the multivariate tests, described in Section 6.2, it is possible to identify genera that have species positively and negatively associated with the trait. PCoA analysis can be used for visualizing these multivariate associations. In Section 6.2, five genera were discovered by the multivariate test alone. In the PCoA plots the point clouds for the two study groups differ; see Figure 6 for "within-genus" PCoA plots for two such genera. These genera

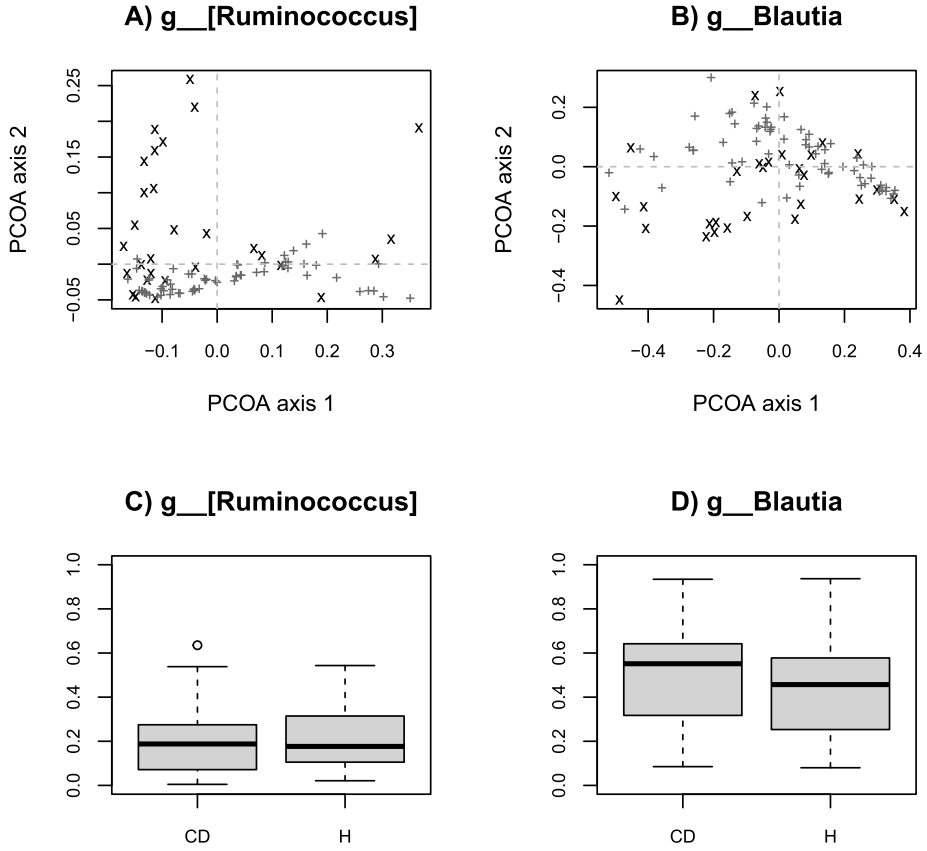


FIG. 6. *Top row*: *"within-genus" PCoA* (*using BC dissimilarity*) *for two genera found differentially abundant by multivariate DACOMP analysis over genus ASVs in Section* 6.2 *but not discovered as differentially abundant by a univariate test using the sum of genus counts. Black "x's" and grey "+" represent samples from the CD and control groups, respectively. The origin is depicted by two blue dashed lines. "Within-genus" PCoA done using BC dissimilarity*; *see additional details in Section* 6.3. *Bottom row*: *boxplots for the ratio of genus count to the total number of counts found under both the reference and genus in each study group. The genera "g__[Ruminococcus]" (left column) and "g__Blautia" (right column) contain* 18 *and* 30 *taxa, respectively, of which only five and* 12 *ASV-level taxa discoveries were made.*

are particularly interesting from a biological perspective since they show ASVs from the same genus behave differently in the presence of CD. For some genera this behaviour has also been observed in other studies. For example, the genus "Blautia" is known to include ASVs that increase as well as ASVs that decrease in relative abundance in the presence of CD. We provide two examples of such ASVs in Section 5.5 of the Supplementary Material (Brill, Amir and Heller (2022)). Similar effects are documented for these ASVs in our analysis of the data. In Section 5.1 of the Supplementary Material (Brill, Amir and Heller (2022)), we provide the "within-genus" PCoA plots for all genera discovered as differentially abundant in Section 6.2.

"Within-genus" PCoA accounts for variation in the number of reads observed under $\mathbf{j}$, due to both variations in the total sequencing depth and variations in the number of counts measured for differentially abundant taxa outside of $\mathbf{j}$. A naive PCoA analysis for a specific genus would begin by rarefying all samples to constant depth, followed by computing distances between the subvectors corresponding to genus ASVs, for example, using the ℓ_1 distance metric. This naive analysis would disregard variations in the number of counts observed under $\mathbf{X}(\mathbf{e_j})$ due to differentially abundant taxa outside of $\mathbf{j}$. In Section 5.2 of the Supplementary Material (Brill, Amir and Heller (2022)), we show "within-genus" and naive PCoA plots for the genera "Clostridium" and "Oscillospira;" two examples were our "within-genus" PCoA analysis that has higher sensitivity to differential abundance effects in the data compared to the naive analysis.

When performing "within-genus" PCoA analysis, an interesting question is whether signals at the genus level result from specific ASV level discoveries alone or are a property of the genus. Our "within-genus" PCoA approach allows researchers to visually explore this. In Section 5.3 of the Supplementary Material (Brill, Amir and Heller (2022)), we reconduct our "within-genus" PCoA analysis after excluding from the data ASVs discovered as differentially abundant by any of the methods considered in Table 2. Figure 15 in Section 5.3 of the Supplementary Material (Brill, Amir and Heller (2022)) shows the genera "Bacteroids" and "Lachnospira" as examples for differentially abundant genera that present a strong signal in the "with-genus" PCoA plot, even after removing all ASV-level discoveries from the data. This result hints that the number of differentially abundant taxa in these genera is greater than identified by the univariate analyses of Section 6.1.

7. Final remarks. A crucial step in our approach is the identification of an appropriate reference set. In Section 4.1 we provide a data adaptive method which avoids using the trait values explicitly for reference selection. Since the validity of the test is ensured only if (3.1) holds for the selected reference set, and the larger the reference set, the larger the chance of differentially abundant taxa contaminating it, we suggest selecting the smallest possible reference set for which the subsequent analysis has good power.

Researchers may conduct a power analysis for setting the minimal total count in the reference set. A numerical simulation approach can be taken, for example, using methods for simulating microbiome data as in Hawinkel et al. (2019) for setting this parameter for anticipated effect sizes.

The DACOMP procedure for testing (3.2) and (3.3) relied on a single draw of counts for analysis of the data. It may be tempting to "derandomize" the analysis by considering several rarefied samples, instead of relying on a single draw; however, simplistic approaches, such as averaging test statistics across multiple rarefactions of the data or averaging the rarefied draws themselves, will result in a nonvalid test. To see why, consider the case where the tested taxon j is not differentially abundant, and the total number of counts available in the taxa with indices $\{j, b_1, \ldots, b_r\}$ for samples with group label $\mathbf{Y} = 0$ is stochastically smaller than for samples with group label $\mathbf{Y} = 1$, that is, $\mathbf{b}_j'\mathbf{X}_i$ tends to be smaller if $\mathbf{Y}_i = 0$ than if

$\mathbf{Y}_i = 1$. Hence, counts in samples with a trait of $\mathbf{Y} = 0$ are more likely to be resampled across multiple rarefactions of the data compared to counts from samples with $\mathbf{Y} = 1$. Therefore, the bivariate distribution of two rarefied draws taken from a single sample is different across different values of $\mathbf{Y}$. Specifically, multiple draws from a sample with $\mathbf{Y} = 0$ will have a higher correlation compared to multiple draws from a sample with $\mathbf{Y} = 1$. Constructing valid inference methods that consider the dependence between subsequent draws is a challenge for future research.

We note that, when analyzing counts in a block design, the limitation of using a shared rarefaction depth for analysis is no longer needed, and DACOMP rarefaction can be performed independently in each block. For example, when testing for gut taxa differential abundance w.r.t. BMI, while conditioning on the host diet and with several diets recorded in the data, a different value of λ_j can be selected for each diet group. For this setting, computation of valid P-values can be performed by permuting trait values within blocks. We plan to address the tasks of reference selection and differential abundance testing in block designs in future research.

We provide empirical evidence that our approach is useful in a study of Crohn's disease, where the compositional effect is large. In addition, we analyze in Section 8 of the Supplementary Material (Brill, Amir and Heller (2022)) the differential abundance of taxa across 49 pairs of adjacent body sites in the human body, using data from the Human Microbiome Project (Gevers et al. (2012)), where DACOMP discovers a considerable number of taxa as differentially abundant. Adjacent body sites in the oral cavity, throat, and skin are more likely to have similar microbial loads and most pairs of taxa maintaining their ratio of abundances across body sites. Hence, alternative methods are found in high agreement with DACOMP. Specifically, we show there are little to no discoveries lost when normalizing by rarefaction.

In Section 9 of the Supplementary Material (Brill, Amir and Heller (2022)), as a second example for data set with a large change in the microbial load across study groups, we analyze data from a stool sample dilution experiment (Staemmler et al. (2016)). Fecal samples in the study were first diluted at different ratios and then "spiked-in" with a known load of three types of bacteria. Unlike previous examples, for this data set the "ground truth" for differential abundance is known. Moreover, the traits examined are continuous: the dilution factor and the microbial load spiked-in. Therefore, we tested (3.2) using Spearman's correlation test, and we showed that DACOMP detects the true differentially abundant taxon and that some of the other methods have an inflation of false positives. Instead of Spearman's correlation test, other tests of independence can also be used with our software.

Other fields of study that gather data by sequencing PCR amplicons also make use of statistical methods aimed at analyzing compositional data, for example, RNA-seq (Quinn et al. (2019)), metabolomics (Kalivodova et al. (2015)), shotgun sequencing techniques for microbiome data (Luz (2019)), and study of immune system response (Vieira (2020)). Adapting DACOMP and DACOMP-ratio to such datasets is an interesting direction for future work.

Acknowledgments. The authors would like to thank the anonymous referees, an Associate Editor, and the Editor for their constructive comments that improved the quality of this paper.

Funding. The work of Barak Brill and Ruth Heller was funded by ISF Grant 1049/16.

SUPPLEMENTARY MATERIAL

Additional simulations and data analyses (DOI: 10.1214/22-AOAS1607SUPPA; .pdf). An additional PDF file with supplementary material containing simulations results for additional competitor methods, further examination of the reference selection procedure, and additional data analyses.

Detailed list of discoveries for the Crohn's disease data (DOI: 10.1214/22-AOAS1607SUPPB; .zip). A XLSX file containing the list of discoveries, by method, for the univariate and multivariate analyses described in Section 6.

R-package for the DACOMP method (DOI: 10.1214/22-AOAS1607SUPPC; .zip). An R language implementation for the DACOMP method.

Code for reproducing results (DOI: 10.1214/22-AOAS1607SUPPD; .zip). Scripts and data used for the simulations and analyses performed in this paper.

QIIME2 plugin for DACOMP (DOI: 10.1214/22-AOAS1607SUPPE; .zip). A QIIME2 plugin allowing DACOMP to be run from within the QIIME 2 software for analysis of 16S counts data.

REFERENCES

AITCHISON, J. (1982). The statistical analysis of compositional data. *J. Roy. Statist. Soc. Ser. B* **44** 139–177. MR0676206

AITCHISON, J. (1986). *The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability*. CRC Press, London. MR0865647 https://doi.org/10.1007/978-94-009-4109-0

AMIR, A., MCDONALD, D., NAVAS-MOLINA, J. A., KOPYLOVA, E., MORTON, J. T., XU, Z. Z., KIGHTLEY, E. P., THOMPSON, L. R., HYDE, E. R. et al. (2017). Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems* **2** e00191-16.

ANDERSON, M. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology* **26** 32–46.

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BRILL, B., AMIR, A. and HELLER, R. (2022). Supplement to "Testing for differential abundance in compositional counts data, with application to microbiome studies." https://doi.org/10.1214/22-AOAS1607SUPPA, https://doi.org/10.1214/22-AOAS1607SUPPB, https://doi.org/10.1214/22-AOAS1607SUPPC, https://doi.org/10.1214/22-AOAS1607SUPPD, https://doi.org/10.1214/22-AOAS1607SUPPE

CALGARO, M., ROMUALDI, C., WALDRON, L., RISSO, D. and VITULO, N. (2020). Assessment of single cell rna-seq statistical methods on microbiome data. *BioRxiv*.

CALLAHAN, B. J., MCMURDIE, P. J., ROSEN, M. J., HAN, A. W., JOHNSON, A. J. A. and HOLMES, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13** 581–583.

DESANTIS, T. Z., HUGENHOLTZ, P., LARSEN, N., ROJAS, M., BRODIE, E. L., KELLER, K., HUBER, T., DALEVI, D., HU, P. et al. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72** 5069–5072.

FERNANDES, A. D., MACKLAIM, J. M., LINN, T. G., REID, G. and GLOOR, G. B. (2013). ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-seq. *PLoS ONE* **8** e67019.

FORBES, C., EVANS, M., HASTINGS, N. and PEACOCK, B. (2011). *Statistical Distributions*. Wiley, Hoboken, NJ. MR2964192

GEVERS, D., KNIGHT, R., PETROSINO, J. F., HUANG, K., MCGUIRE, A. L., BIRREN, B. W., NELSON, K. E., WHITE, O., METHE, B. A. et al. (2012). The human microbiome project: A community resource for the healthy human microbiome. *PLoS Biol.* **10** e1001377.

GLOOR, G. B., MACKLAIM, J. M., PAWLOWSKY-GLAHN, V. and EGOZCUE, J. J. (2017). Microbiome datasets are compositional: And this is not optional. *Front. Microbiol.* **8** 2224.

GOODRICH, J. K., DI RIENZI, S. C., POOLE, A. C., KOREN, O., WALTERS, W. A., CAPORASO, J. G., KNIGHT, R. and LEY, R. E. (2014). Conducting a microbiome study. *Cell* **158** 250–262.

GRETTON, A., FUKUMIZU, K., TEO, C., SONG, L., SCHÖLKOPF, B. and SMOLA, A. (2008). A kernel statistical test of independence. In *Advances in Neural Information Processing Systems* 20, Red Hook, NY, USA 585–592. Max-Planck-Gesellschaft, Curran.

GUO, X., ZHANG, X., QIN, Y., LIU, Y.-X., ZHANG, J., ZHANG, N., WU, K., QU, B., HE, Z. et al. (2020). Host-associated quantitative abundance profiling reveals the microbial load variation of root microbiome. *Plant Commun.* **1** 100003.

HAMADY, M. and KNIGHT, R. (2009). Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Res.* **19** 1141–1152.

HAWINKEL, S., MATTIELLO, F., BIJNENS, L. and THAS, O. (2019). A broken promise: Microbiome differential abundance methods do not control the false discovery rate. *Brief. Bioinform.* **20** 210–221.

HELLER, R., HELLER, Y. and GORFINE, M. (2013). A consistent multivariate test of association based on ranks of distances. *Biometrika* **100** 503–510. MR3068450 https://doi.org/10.1093/biomet/ass070

HOMMEL, G. and KROPF, S. (2005). Tests for differentiation in gene expression using a data-driven order or weights for hypotheses. *Biom. J.* **47** 554–562. MR2190478 https://doi.org/10.1002/bimj.200410118

JIAN, C., LUUKKONEN, P., YKI-JÄRVINEN, H., SALONEN, A. and KORPELA, K. (2020). Quantitative pcr provides a simple and accessible method for quantitative microbiota profiling. *PLoS ONE* **15** e0227285.

JIANG, L., AMIR, A., MORTON, J. T., HELLER, R., ARIAS-CASTRO, E. and KNIGHT, R. (2017). Discrete false-discovery rate improves identification of differentially abundant microbes. *mSystems* **2** e00092-17.

JIANG, S.-Q., YU, Y.-N., GAO, R.-W., WANG, H., ZHANG, J., LI, R., LONG, X.-H., SHEN, Q.-R., CHEN, W. et al. (2019). High-throughput absolute quantification sequencing reveals the effect of different fertilizer applications on bacterial community in a tomato cultivated coastal saline soil. *Sci. Total Environ.* **687** 601–609.

KALIVODOVA, A., HRON, K., FILZMOSER, P., NAJDEKR, L., JANECKOVA, H. and ADAM, T. (2015). PLS-DA for compositional data with application to metabolomics. *J. Chemom.* **29** 21–28.

KAUL, A., MANDAL, S., DAVIDOV, O. and PEDDADA, S. D. (2017). Analysis of microbiome data in the presence of excess zeros. *Front. Microbiol.* **8** 2114.

KONG, J., LIU, X., WANG, L., HUANG, H., OU, D., GUO, J., LAWS, E. A. and HUANG, B. (2021). Patterns of relative and quantitative abundances of marine bacteria in surface waters of the subtropical northwest Pacific Ocean estimated with high-throughput quantification sequencing. *Front. Microbiol.* **11** 599614.

KUMAR, M. S., SLUD, E. V., OKRAH, K., HICKS, S. C., HANNENHALLI, S. and BRAVO, H. C. (2018). Analysis and correction of compositional bias in sparse sequencing count data. *BMC Genomics* **19** 799.

LOVE, M. I., HUBER, W. and ANDERS, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15** 550.

LUZ, C. M. (2019). Statistical analysis of metagenomics data. *Genomics Inform.* **17** e6.

MANDAL, S., TREUREN, W. V., WHITE, R. A., EGGESBØ, M., KNIGHT, R. and PEDDADA, S. D. (2015). Analysis of composition of microbiomes: A novel method for studying microbial composition. *Microb. Ecol. Health Dis.* **26** 27663.

MCDONALD, D., HYDE, E., DEBELIUS, J. W., MORTON, J. T., GONZALEZ, A., ACKERMANN, G., AKSENOV, A. A., BEHSAZ, B., BRENNAN, C. et al. (2018). American gut: An open platform for citizen science microbiome research. *mSystems* **3** e00031-18.

MCMURDIE, P. J. and HOLMES, S. (2014). Waste not, want not: Why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* **10** e1003531.

MORTON, J. T., MAROTZ, C., WASHBURNE, A., SILVERMAN, J., ZARAMELA, L. S., EDLUND, A., ZENGLER, K. and KNIGHT, R. (2019). Establishing microbial composition measurement standards with reference frames. *Nat. Commun.* **10** 2719.

NELSON, M. C., MORRISON, H. G., BENJAMINO, J., GRIM, S. L. and GRAF, J. (2014). Analysis, optimization and verification of Illumina-generated 16S RRNA gene amplicon surveys. *PLoS ONE* **9** e94249.

PAULSON, J. N., POP, M. and BRAVO, H. C. (2013). metagenomeSeq: Statistical analysis for sparse high-throughput sequncing. Bioconductor package.

PAULSON, J. N., STINE, O. C., BRAVO, H. C. and POP, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* **10** 1200–1202.

QUINN, T. P., ERB, I., GLOOR, G., NOTREDAME, C., RICHARDSON, M. F. and CROWLEY, T. M. (2019). A field guide for the compositional analysis of any-omics data. *GigaScience* **8** giz107.

RISSO, D., PERRAUDEAU, F., GRIBKOVA, S., DUDOIT, S. and VERT, J.-P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* **9** 284.

ROSENBAUM, P. R. (2010). *Design of Observational Studies. Springer Series in Statistics*. Springer, New York. MR2561612 https://doi.org/10.1007/978-1-4419-1213-8

SIMES, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73** 751–754. MR0897872 https://doi.org/10.1093/biomet/73.3.751

STAEMMLER, F., GLAESNER, J., HIERGEIST, A., HOLLER, E., WEBER, D., OEFNER, P. J., GESSNER, A. and SPANG, R. (2016). Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. *Microbiome* **4** 28.

SUNAGAWA, S., COELHO, L. P., CHAFFRON, S., KULTIMA, J. R., LABADIE, K., SALAZAR, G., DJAHANSCHIRI, B., ZELLER, G., MENDE, D. R. et al. (2015). Structure and function of the global ocean microbiome. *Science* **348** 1261359.

SZÉKELY, G. J. and RIZZO, M. L. (2009). Brownian distance covariance. *Ann. Appl. Stat.* **3** 1236–1265. MR2752127 https://doi.org/10.1214/09-AOAS312

TKACZ, A., HORTALA, M. and POOLE, P. S. (2018). Absolute quantitation of microbiota abundance in environmental samples. *Microbiome* **6** 110.

TORGERSON, W. S. (1952). Multidimensional scaling. I. Theory and method. *Psychometrika* **17** 401–419. MR0054219 https://doi.org/10.1007/BF02288916

TSAGRIS, M., ALENAZI, A., VERROU, K.-M. and PANDIS, N. (2020). Hypothesis testing for two population means: Parametric or non-parametric test? *J. Stat. Comput. Simul.* **90** 252–270. MR4038806 https://doi.org/10.1080/00949655.2019.1677659

VANDEPUTTE, D., KATHAGEN, G., D'HOE, K., VIEIRA-SILVA, S., VALLES-COLOMER, M., SABINO, J., WANG, J., TITO, R. Y., DE COMMER, L. et al. (2017). Quantitative microbiome profiling links gut community variation to microbial load. *Nature* **551** 507–511.

VAN DEN BERGE, K., PERRAUDEAU, F., SONESON, C., LOVE, M. I., RISSO, D., VERT, J.-P., ROBINSON, M. D., DUDOIT, S. and CLEMENT, L. (2018). Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biol.* **19** 24.

VIEIRA, M. C. (2020). Evolution of adaptability and the immune response to influenza and HIV. Ph. D. thesis, The Univ. Chicago.

VIEIRA-SILVA, S., SABINO, J., VALLES-COLOMER, M., FALONY, G., KATHAGEN, G., CAENEPEEL, C., CLEYNEN, I., VAN DER MERWE, S., VERMEIRE, S. et al. (2019). Quantitative microbiome profiling disentangles inflammation- and bile duct obstruction-associated microbiota alterations across PSC/IBD diagnoses. *Nat. Microbiol.* **4** 1826–1831.

WANG, Q., GARRITY, G. M., TIEDJE, J. M. and COLE, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73** 5261–5267.

WEISS, S., XU, Z. Z., PEDDADA, S., AMIR, A., BITTINGER, K., GONZALEZ, A., LOZUPONE, C., ZANEVELD, J. R., VAZQUEZ-BAEZA, Y. et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5** 27.

WU, J. R., MACKLAIM, J. M., GENGE, B. L. and GLOOR, G. B. (2017). Finding the centre: Corrections for asymmetry in high-throughput sequencing datasets. Available at arXiv:1704.01841.

XU, L., PATERSON, A. D., TURPIN, W. and XU, W. (2015). Assessment and selection of competing models for zero-inflated microbiome data. *PLoS ONE* **10** e0129606.