Testing for Differential Abundance in Compositional counts data, with Application to Microbiome Studies

Barak Brill[*], Amnon Amir[**], and Ruth Heller[*]

[*]Tel Aviv University, Tel Aviv, email for correspondence: barakbri@mail.tau.ac.il

[**]Sheba Medical Center, Tel Hashomer, affiliated with the Tel Aviv University

August 6, 2019

Abstract

In order to identify which taxa differ in the microbiome community across groups, the relative frequencies of the taxa are measured for each unit in the group by sequencing PCR amplicons. Statistical inference in this setting is challenging due to the high number of taxa compared to sampled units, low prevalence of some taxa, and strong correlations between the different taxa. Moreover, the total number of sequenced reads per sample is limited by the sequencing procedure. Thus, the data is compositional: a change of a taxon's abundance in the community induces a change in sequenced counts across all taxa. The data is sparse, with zero counts present either due to biological variance or limited sequencing depth (technical zeros). For low abundance taxa, the chance for technical zeros is non-negligible and varies between sample groups. Compositional counts data poses a problem for standard normalization techniques since technical zeros cannot be normalized in a way that ensures equality of taxon distributions across sample groups. This problem is aggravated in settings where the condition studied severely affects the microbial load of the host. We introduce a novel approach for differential abundance testing of compositional data, with a non-neglible amount of zero counts. Our approach uses a set of reference taxa, which are non-differentially abundant. We suggest a data-adaptive approach for identifying the set of reference taxa from the data. We show the usefulness of our method via simulations and real data from a Crohn's disease study and studies from the Human Microbiome Project. We also show that existing methods for differential abundance testing, including methods designed to address compositionality, do not provide control over the rate of false positive discoveries when the change in microbial load is vast.

A R software package, dacomp, implementing the novel methods suggested is publicly available.

Keywords: Compositional bias, Analysis of composition, Normalization, Rarefaction, Non-parametric tests.

1 Introduction

The microbiome is the collection of micro-organisms and bacteria which are part of the physiological activity of a host body or ecosystem [Hamady and Knight, 2009]. It is of scientific interest to associate change in microbial structure to disease and other environmental factors. For example, the human microbiome project [HMP, Gevers et al., 2012], examined changes in microbiome composition across sites of the human body. The earth microbiome project [Thompson et al., 2017] examined the association of microbiome samples from water, soil, sediment and plants with factors such as sampling location, temperature, pH and salinity.

The study of the human microbiome is of medical interest. A better understanding of the changes in the human microbiome can lead to a better diagnosis and treatment of certain diseases. For example, the study of Vandeputte et al. [2017] 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. A better understanding of the microbial changes in the gut may lead to a better understanding and treatment of Crohn's disease. Another example is the study of Teo et al. [2015], which investigated the change in the microbial ecology of the lower respiratory tract of children, in the presence of infection. Infections at the lower respiratory tract at a young age may contribute to the development of asthma at later years.

A common method of measuring the abundance of the bacterial community is by 16S sequencing. The 16S rRNA gene codes for a crucial part of the ribosome common to all living cells. The variable regions in the 16S rRNA gene, named V1,V2,...,V9, 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: a sequence of 100-150 base pairs indicates the taxon of bacteria from which it was sampled.

To conduct a microbiome study, samples from different specimen are collected. Targeted variable regions of the 16S gene are duplicated and amplified using PCR. Sequencing technology allows one to read the amplicons of the PCR procedure and list all sequences read for each sample. The list of sequences trimmed to a constant length of, e.g., 150 base pairs [Nelson et al., 2014] is recorded per sample.

Due to the high rate of mutations in the 16S region, sequences of the same region which are different up to a certain threshold, e.g., 3% of base pairs out of 150, are considered as a single cluster of sequences. These clusters are termed Operational Taxonomic Units (OTUs) and represent the finest resolution of organism type [Hamady and Knight, 2009]. The number of observed sequences for each OTU is recorded per sample.

Several challenges are encountered when trying to identify which taxa are associated with a condition of interest based on the OTU counts. The first challenge is that the number of sequenced reads varies from sample to sample, and is mostly an artifact of the sequencing procedure rather than a proxy to the sample's original abundance of bacteria. Therefore, comparison of relative frequencies is informative while comparison of actual counts between samples is not. This effect in data is referred to as compositionality [Gloor et al., 2017, Kumar et al., 2018, Mandal et al., 2015].

The second challenge is that the vector of OTU counts is sparse by nature, as not all OTUs 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]. Zero counts of an OTU occur for two reasons: (1) low frequency in the sampled units, so the sample does not capture the very rare OTUs, henceforth referred to as technical zeros; (2) OTUs not shared by the entire population, henceforth referred to as structural zeros.

The third challenge is the strong dependence between OTU counts. Intuitively, compositionality implies negative correlations ,"more of one OTU, less of the other". However, strong positive correlations between OTUs across subjects are also observed [Hawinkel et al., 2017].

The fourth challenge is that microbiome data usually features more OTUs than samples. For most studies, several dozens to a few hundred samples are measured. The number of OTUs, on the other hand, ranges between a few hundreds to several thousands [Nelson et al., 2014].

Due to dimensionality limitations, the high percentage of zero counts, along with the complex correlation structure and varying community membership, one cannot fully model the counts data. Statistical tests that ignore compositionality can lead to false positive findings when attempting to detect which taxa have changed their abundance between groups, as demonstrated by the following example.

Example 1: a toy example demonstrating the danger of ignoring compositionality. Suppose we have two groups of samples from multinomial distributions with the same number of total counts N, but the probability for counts are $\vec{P}$ for samples from X and $\vec{Q}$ for samples from Y, with $\vec{P}$ and $\vec{Q}$ m-dimensional vectors related by

$$\vec{Q} = (1-w)\cdot\vec{P} + w\cdot(1,0,...,0)\,, w\in(0,1)\,, \sum_{i=1}^{m} P_i = \sum_{i=1}^{m} Q_i = 1. \tag{1.1}$$

We observe the marginal distributions across the two groups: the first taxon has an increased relative frequency in the Y group compared to the X group, and all other taxa have decreased relative frequency in the Y group compared to the X

group. Therefore for large enough sample sizes, the two-sample test for equality of relative frequencies will reject the null hypothesis at each coordinate. However, we are interested in detecting only the first taxon, since it is the only one driving the observed differences across groups. Moreover, if $\vec{P}$ and $\vec{Q}$ are random vectors, as in microbiome studies, where each sample has its own microbiome, we still would like to detect only the first coordinate.

In this paper, we aim at constructing a method for statistical inference in a compositional setting which considers as true discoveries only 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, i.e., a test of differential abundance with respect to a reference frame of taxa [Morton et al., 2019]. In this work, we formulate a definition of differential abundance. We show that given a subset of the taxa, known to be non-differentially abundant, a test for a change in the unknown absolute abundance of a taxon can be constructed. A set of non-differentially abundant taxa can be identified from the data if most taxa are non-differentially abundant.

The structure of the paper is as follows. The remainder of the introduction reviews methods for OTU analysis in microbiome studies. In § 1.1 we describe normalization methods which account for the variable number of total counts across samples. After normalization it is possible to compare OTU counts distributions across groups. However, these normalization methods do not resolve our concern about an inflation of false positives due to the compositional nature of the data, as exemplified in Example 1. In § 1.2 we examine works that take compositionality into account, and point out limitations which this work aims to overcome. The analysis goal of differential abundance discovery is formalized in § 2. In § 3 we describe our main result, a valid discovery procedure with false positive rate guarantees. In § 4 we describe a simulation study comparing the methods presented in § 1.1-§ 1.2 to the method presented in § 3. In § 5 healthy subjects and subjects with Crohn's disease are compared in order to identify which of the taxa are differentially abundant. In § 6 we conclude with final remarks.

1.1 Review of methods that adjust for different sequencing depths across samples.

Let $\vec{X} = (X_1, ..., X_m)$ be a sample vector of counts. The simplest intuitive form of normalization is dividing each sample by its total number of reads. This form of normalization is henceforth referred to as Total Sum Scaling (TSS), and is denoted by $TSS\left(\vec{X}\right) = \vec{X}/\sum_{i=1}^{m} X_i$. TSS disregards the total number of sequenced reads per sample so even after normalization, different samples from the same sample group are not equally distributed due to different sequencing resolutions [Weiss et al., 2017].

The total number of reads per sample may vary greatly due to a small number of highly abundant taxa. This motivated Paulson et al. [2013b] to suggest cumulative sum scaling (CSS), a scaling factor which has smaller variance than the total number of reads per sample. The CSS transform is given by $CSS\left(\vec{X}\right) = \vec{X}/\sum_{i=1}^{q_{CSS}} X_{(i)}$, where $X_{(i)}$ denotes the ith smallest entry of $\vec{X}$ and q_{CSS} is an index chosen adaptively from the data.

Chen and Li [2013] used the Dirichlet Multinomial (DM) distribution to model microbiome sample counts. The model assumes each sample is generated by first sampling a vector of proportions from a Dirichlet distribution and then sampling from the multinomial distribution with the sampled vector of proportions. The DM distribution has been extended to allow more complex covariance structures via mixtures of Dirichlet components as proportion vectors [Holmes et al., 2012, O'Brien and Record, 2016, Shafiei et al., 2015].

A popular approach is to adapt for the microbiome the analysis for data on a simplex, which originated in the work of Aitchison [1986]. Consider the vector $\vec{u} = (u_1, ..., u_m)$ in the unit simplex $\mathcal{U} = \{\vec{u}|0 < u_j, \sum_{j=1}^{m} u_j = 1\}$. Let $g(\vec{u})$ be the geometric mean of the vector $\vec{u}$, $g(\vec{u}) = \left(\prod_{j=1}^{m} u_i\right)^{1/m}$. The central log-ratio (CLR) transformation of $\vec{u}$ is $CLR(\vec{u}) = log(\vec{u}/g(\vec{u})) \in \mathbb{R}^m$. Another transformation called the additive log-ratio (ALR) is given by selecting an entry of $\vec{u}$ as reference, e.g. the mth entry, and calculating $ALR(\vec{u}) = \left(log\left(\frac{u_1}{u_m}\right), log\left(\frac{u_2}{u_m}\right), ..., log\left(\frac{u_{m-1}}{u_m}\right)\right)$. $ALR(\vec{u})$ is a vector in $\mathbb{R}^{m-1}$. Following transformation to an unconstrained space via $CLR(u)$ or $ALR(u)$ analysis can proceed by traditional methods. Analysis results, e.g. parameter estimates and confidence intervals, can be transformed back to the constrained coordinate space of $\mathcal{U}$ via the inverse transformations of $CLR^{-1}(x)$ and $ALR^{-1}(x)$. These methods have been used for analyzing chemical compound compositions, where the dimension is smaller than the number of samples and measured quantities are continuous and positive.

For microbiome data, entries with zero counts are prevalent, and computation of the CLR and ALR transformations is thus impossible for $TSS\left(\vec{X}\right)$. A common practice is to add a psuedo-count of 1 to all vectors entries. While this makes the CLR and ALR transformations computable, the interpretation following pseudocount addition is less clear, since an addition of a psuedo-count has made all bacteria present in all samples. Moreover, zero counts in a vector entry can be either technical zeros or structural zeros with interpretation depending on the total number of reads in a vector: an entry of zero in a vector containing 10^3 reads total is more likely to be a technical zero compared to an entry of zero in a vector with 10^5 reads total. This interpretation is lost when psuedocounts are used.

Kaul et al. [2017] use ALR and CLR transformations for microbiome data by first classifying zeros in the data as 'structural' or 'technical'. Parameters for taxa proportions are estimated only using non zero entries and technical zeros, to which an addition of a psuedocount is plausible. For classifying zeros as either technical or structural, Kaul et al. [2017] assume there is no overdispersion of taxon proportions for samples of the same group, but this assumption is not realistic.

Another approach used to preprocess microbiome data for use with the ALR and CLR transformations is to treat all values of zeros in the data as missing values for the true taxa proportions, which are too low to be measured [Quinn et al., 2018]. The missing values are imputed in order to transform count vector into an unconstrained sample space.

Generalized linear models for the count distribution of a single OTU can be found in Paulson et al. [2013b] and Xu et al. [2015]. These models take into account the sequencing depth of each sampled unit for parameter estimation. However, compositionality constraints are not met by the parameter estimates of different OTU models, since model parameters are estimated separately for each OTU.

1.2 Review of methods that provide further adjustments for compositionality

In a very interesting work, Mandal et al. [2015] suggested a framework for analysis under composionality (ANCOM). The key reasonable assumption is that the effect of compositionaly is such that inter-taxa ratios are maintained for non differentially abundant taxa. Under this assumption, it is possible to identify if a taxon is differentially abundant. In Example 1, a single taxon has an increased number of counts in group Y compared to group X. We expect the number of observed counts for the other taxa to decrease in group Y. An increase in the frequency of a differentially abundant taxon due to an external effect, causes a reduction in relative frequencies for other taxa in a way that does not change their respective ratios.

For the two-sample case, the ANCOM procedure is as follows: Let $X_{i,j}, Y_{l,j}$, represent the counts for taxon $j \in \{1, ..., m\}$ in samples $i \in \{1, ..., n_X\}$ from group one and $l \in \{1, ..., n_Y\}$ from group two. The first step is to perform Wilcoxon rank sum tests, at level α, comparing the distributions of $\frac{X_{i,j}+1}{X_{i,k}+1}$ and $\frac{Y_{i',j}+1}{Y_{i',k}+1}$ for all taxa pairs. Let $W_{j,k}$ be the Wilcoxon test statistic for taxa j and k, with p-value $p_{j,k}$. Let the indicator function for its rejection be $I_{j,k} = \mathbb{1}\left(p_{j,k} \leq \alpha\right)$. The number of pairwise rejections consisting of taxon j is denoted by $\mathcal{W}_j = \sum_{k=1,k\neq j}^{m} I_{j,k}$.

By assumption, frequencies of non differentially abundant taxa maintain their respective ratios, so in a well powered study it is expected that the number of rejections per taxon, $\mathcal{W}_j$, will be relatively high for the differentially abundant taxa, as they changed their ratios compared to almost all other taxa. Let m_1 be the unknown number of differentially abundant taxa. For the taxa which are not differentially abundant, we expect $\mathcal{W}_j$ to be much lower if $m_1 << m$. Mandal et al. [2015] suggested declaring the set of indices $\{j|\mathcal{W}_j \geq \mathcal{W}^*\}$ as the set of differentially abundant taxa, where $\mathcal{W}^*$ is chosen according to the empirical distrubtion of $\mathcal{W}_j$'s, see Mandal et al. [2015] for details. This decision rule can lead to a inflation of false positive findings, especially when $m_1 = 0$. In § 4 we show settings in which the false positives are not controlled at any reasonable level.

Kumar et al. [2018] addressed the problem of testing for differential abundance in microbial counts data. Similar to Mandal et al. [2015], Kumar et al. [2018] suggest that taxa not associated with the condition of interest have maintained the ratios of their respective proportions in each sample. To briefly describe the approach, we make use of the setup presented in Example 1. Kumar et al. [2018] observe that while the expected values of all coordinates differ across study groups, coordinate means across all taxa except the first taxon were lowered in group Y compared to group X by the same multiplicative factor. In Example 1, this ratio is given by the multiplier $1 - w$. Kumar et al. [2018] denote this common multiplicative factor of coordinate means by Λ. As a first step, Kumar et al. [2018] suggest estimating Λ using the data by a novel estimator, denoted by $\hat{\Lambda}$. As a second step, their proposed method suggests scaling taxa counts for group Y by a factor of $\left(\hat{\Lambda}\right)^{-1}$. Following the second step, coordinate means of taxa not associated with the disease should be similar across study groups. If a coordinate has substantial difference in means after normalization, it represents a taxon whose absolute abundance has changed with accordance to the condition of interest. The sparsity assumption that m_1 is small enough, say $m_1 < m/2$, will be essential also in the methodology we suggest. See Morton et al. [2019] for a recent work that suggests an analysis method when this sparsity assumption is not satisfied.

The CLR and ALR normalizations discussed in § 1.1 adjust for the different sequencing depths across samples. However, CLR and ALR normalizations do not resolve the challenge depicted in *Example 1*: after applying CLR or ALR normalizations to the data, the marginal distributions of counts differ across study groups for all taxa even though only the first coordinate is a discovery of interest. For the CLR transformation, this follows from the fact that normalizing observations by the geometric mean does not enforce equality of distributions for the transformed counts. For the ALR transformation, this follows from the fact that normalizing by a non-differentially abundant taxon as reference does not result in equality of distributions for the ratio of two counts when zero counts are probable, see § 2.1 for details.

The state-of-the-art ALDEx2 method and software package [Fernandes et al., 2013] uses a CLR type normalization with a geometric mean computed over a subset of the taxa. The suggested normalization method, named inter-quantile log ratio ($iqlr$) consists of a three step procedure. In the first step, The CLR transformation is applied to the data, after applying a pseudo-count of 0.5 to all data entries. In the second step, the variances of the CLR transformed counts are computed, by taxa. In the third step, a second CLR transformation is applied to the original data, but the geometric mean used for normalization is computed only using taxa whose variances (computed in the second step) fall between the first and third quantiles. In ALDEx2, P-values for tests of differential abundance are computed by averaging across multiple resamples of the data, with each observation resampled from a dirichlet distribution whose concentration parameters match the sample values. Using this algorithm, the rate of false positive discoveries resulting from a change in the probability of technical zeros across study groups is greatly reduced (see § 4.1 for numerical results). However, as with the ALR transformation above, the resulting statistics are not equally distributed across groups even if they involve only reference taxa, when zero counts are probable.

Another approach suggested is correcting for compositionality by flow-cytometric measurement. For example, Vandeputte et al. [2017] suggested a two step procedure based on measuring for each sample the number of bacteria per gram (via a flow-cytometer). The first step in the method of Vandeputte et al. [2017] is normalizing the data by rarefaction. Normalization by rarefaction is quite popular in microbiome studies, to address the problem of different samples from the same group having different probabilities for technical zeros [Weiss et al., 2017]. Rarefaction constitutes of selecting a subsample of the counts from each vector of microbial counts, so that the total number of remaining counts is identical for all vectors. The number of reads selected from each vector is known as the rarefaction depth. The second step in the method of Vandeputte et al. [2017] is to multiply each sample vector by the number of bacteria sampled per gram, as given by the flow cytometer. Once absolute abundances are reconstructed from the data, Vandeputte et al. [2017] argue that a change in the marginal distributions can be attributed to the changes in absolute abundances of bacteria. A major limitation of this approach is the fact that values of technical zeros in the data cannot be scaled to their original value in terms of abundance. Therefore the scaled marginal counts may have different distributions across groups even for taxa that are non differentially abundant, as we show in § 3.

2 Notation and goal

Let m be the number of taxa (OTUs). Let n_X (n_Y) be the number of samples in group X (Y). Let $X_{i,j}$ ($Y_{l,j}$) be the number of counts observed for taxon j in the ith (lth) sample of group X (Y). Let N_i^X (N_l^Y) denote the total number of counts sampled for subject i (l) of the first (second) group. Let $\mathcal{P}$ and $\mathcal{Q}$ be continuous distributions of dimension m over the unit simplex for the two groups, representing the relative abundances of taxa in the host ecosystem. For samples $i = 1, ..., n_X$ from the first group, we assume that $\vec{P_i}$ is sampled from $\mathcal{P}$ and the counts for each of N_i^X independent trials with probability $\vec{P_i}$ are observed. Similarly, for subjects $l = 1, ..., n_Y$ from group two, $\vec{Q_l}$ is sampled from $\mathcal{Q}$ and then the counts for each of N_l^Y independent trials are observed. Formally,

$$\vec{X_i}|\vec{P_i}, N_i^X \sim multinom\left(N_i^X, \vec{P_i}\right), \qquad \vec{P_i} \sim \mathcal{P} \quad, 0 \le P_{i,j}, \sum_{j=1}^{m} P_{i,j} = 1, \tag{2.1}$$

$$\vec{Y_l}|\vec{Q_l}, N_l^Y \sim multinom\left(N_l^Y, \vec{Q_l}\right), \qquad \vec{Q_l} \sim \mathcal{Q} \quad, 0 \le Q_{l,j}, \sum_{j=1}^{m} Q_{l,j} = 1.$$

Let X_j (Y_j)and P_j (Q_j) denote the entries for taxon j in general realizations of $\vec{X}$ ($\vec{Y}$) and $\vec{P}$ ($\vec{Q}$).

As in Mandal et al. [2015], we assume that the non differentially abundant taxa have not altered their relative proportions.

Definition 2.1. A set $B = (v_1, ..., v_r)$ is said to be a *reference set* if $\sum_{k=1}^{s} P_{v_k} > 0$ and $\sum_{k=1}^{s} Q_{v_k} > 0$ with probability 1 and

$$\frac{(P_{v_1}, P_{v_2}, ..., P_{v_s})}{\sum_{k=1}^{s} P_{v_k}} \overset{\mathrm{d}}{=} \frac{(Q_{v_1}, Q_{v_2}, ..., Q_{v_s})}{\sum_{k=1}^{s} Q_{v_k}}, \tag{2.2}$$

where $\overset{d}{=}$ indicates equality in distribution.

Let $\mathcal{B} \subset \{1, 2, ...m\}$ represent the largest reference set. We assume that the population relative frequency of non differentially abundant taxa is non zero:

$$Pr\left(\sum_{k\in\mathcal{B}} P_{v_k} > 0\right) = Pr\left(\sum_{k\in\mathcal{B}} Q_{v_k} > 0\right) = 1. \tag{2.3}$$

Our goal is to find all taxa which are differentially abundant,i.e., the complement of the set $\mathcal{B}$.

Suppose a *reference set of taxa* $B = (b_1, b_2, ..., b_r) \subset \mathcal{B}$ is known, with cardinality $|B| = r$. The null hypothesis to be tested for taxon j is that taxon j is not differentially abundant, i.e., it belongs to $\mathcal{B}$. Taxon j and the reference set B constitute a subset of size $r + 1$ of the set $\mathcal{B}$. Therefore, (2.2) holds for the indices $\{j\} \cup B$:

$$H_0^{(j)}: \quad \frac{(P_j, P_{b_1}, P_{b_2}, ..., P_{b_r})}{P_j + \sum_{k=1}^{r} P_{b_k}} \overset{d}{=} \frac{(Q_j, Q_{b_1}, Q_{b_2}, ..., Q_{b_r})}{Q_j + \sum_{k=1}^{r} Q_{b_k}}. \tag{2.4}$$

If $H_0^{(j)}$ is false, then taxon $j \notin \mathcal{B}$. Thus, we would like to identify all taxa for which the null hypothesis in (2.4) is false. However, we are unable to test this hypothesis directly, as $\vec{P}$ and $\vec{Q}$ are not observed.

If there is no overdispersion of the multinomial counts, i.e., $\mathcal{P}$, $\mathcal{Q}$ being degenerate distributions receiving values $\vec{P}, \vec{Q}$ with probability 1, $H_0^{(j)}$ is reduced to:

$$\frac{(P_j, P_{b_1}, P_{b_2}, ..., P_{b_r})}{P_j + \sum_{k=1}^{r} P_{b_k}} = \frac{(Q_j, Q_{b_1}, Q_{b_2}, ..., Q_{b_r})}{Q_j + \sum_{k=1}^{r} Q_{b_k}}. \tag{2.5}$$

Assuming (2.5) holds, a valid approach is to use Fisher's exact test over a 2 × 2 table comparing the counts in taxon j and the reference set B across the two groups. However, since in microbiome studies there is overdispersion, this approach is not valid, see § 4 for numerical examples.

An intuitive approach for testing (2.4) may be to reject (2.4) if the test of $\frac{(X_j, X_{b_1}, X_{b_2}, ..., X_{b_r})}{X_j + \sum_{k=1}^{r} X_{b_k}} \overset{d}{=} \frac{(Y_j, Y_{b_1}, Y_{b_2}, ..., Y_{b_r})}{Y_j + \sum_{k=1}^{r} Y_{b_k}}$ is rejected. However, these random variables are ill-defined since for microbiome counts data, zero counts are prevalent in the data, and the probability of $X_j + \sum_{k=1}^{r} X_{b_k}$ and $Y_j + \sum_{k=1}^{r} Y_{b_k}$ to receive a value of zero counts, is non-negligible. A common approach is to use pseudocounts. Alternatively, it is possible to replace a value of 0 in the denominator by a value of 1. We conclude this section by explaining why this approach for testing $H_0^{(j)}$ is not valid.

2.1 The effect of zero counts on the distribution of counts ratios

Since the probability of $(X_j, X_{b_1}, X_{b_2}, ..., X_{b_r}) = \vec{0}$ and $(Y_j, Y_{b_1}, Y_{b_2}, ..., Y_{b_r}) = \vec{0}$ is non negligible, we may consider testing $\frac{\left(X_j, X_{b_1}, X_{b_2}, ..., X_{b_r}\right)}{max\left(1, X_j + \sum_{k=1}^{r} X_{b_k}\right)} \overset{\mathrm{d}}{=} \frac{\left(Y_j, Y_{b_1}, Y_{b_2}, ..., Y_{b_r}\right)}{max\left(1, Y_j + \sum_{k=1}^{r} Y_{b_k}\right)}$. However, when $H_0^{(j)}$ in (2.4) is true, $E\left(\frac{\left(X_j, X_{b_1}, X_{b_2}, ..., X_{b_r}\right)}{max\left(1, X_j + \sum_{k=1}^{r} X_{b_k}\right)}\right) \neq E\left(\frac{\left(Y_j, Y_{b_1}, Y_{b_2}, ..., Y_{b_r}\right)}{max\left(1, Y_j + \sum_{k=1}^{r} Y_{b_k}\right)}\right)$ if the probability of $\{X_j + \sum_{k=1}^{r} X_{b_k} = 0\}$ and $\{Y_j + \sum_{k=1}^{r} Y_{b_k} = 0\}$ differ. This is a straightforward consequence of the representation of the conditional mean in the following proposition.

Proposition 1. *For* $\vec{X}|\vec{P} \sim$ *multinomial* $\left(N, \vec{P}\right)$,

$$E\left[\frac{X_j}{max\left(1, X_j + \sum_{k=1}^{r} X_{b_k}\right)}|\vec{P}\right] = \frac{P_j}{P_j + \sum_{k=1}^{r} P_{b_k}} \cdot Pr\left(X_j + \sum_{k=1}^{r} X_{b_k} > 0|\vec{P}\right)$$

Proof. $E\left[\frac{X_j}{max\left(1, X_j + \sum_{k=1}^{r} X_{b_k}\right)}|\vec{P}\right] =$
$E\left[\frac{X_j}{X_j + \sum_{k=1}^{r} X_{b_k}}|\vec{P}, X_j + \sum_{k=1}^{r} X_{b_k} > 0\right] \cdot Pr\left(X_j + \sum_{k=1}^{r} X_{b_k} > 0|\vec{P}\right) =$
$E\left\{E\left[\frac{X_j}{X_j + \sum_{k=1}^{r} X_{b_k}}|\vec{P}, X_j + \sum_{k=1}^{r} X_{b_k}\right]|\vec{P}\right\} \cdot Pr\left(X_j + \sum_{k=1}^{r} X_{b_k} > 0|\vec{P}\right) =$
$\frac{P_j}{P_j + \sum_{k=1}^{r} P_{b_k}} \cdot Pr\left(X_j + \sum_{k=1}^{r} X_{b_k} > 0|\vec{P}\right)$ □

If $H_0^{(j)}$ is true,

$$E\left[\frac{X_j}{max\left(1, X_j + \sum_{k=1}^{r} X_{b_k}\right)}\right] - E\left[\frac{Y_j}{max\left(1, Y_j + \sum_{k=1}^{r} Y_{b_k}\right)}\right] =$$

$$E\left[\frac{P_j}{P_j + \sum_{k=1}^{r} P_{b_k}} \cdot \left(Pr\left(X_j + \sum_{k=1}^{r} X_{b_k} > 0|\vec{P}\right) - Pr\left(Y_j + \sum_{k=1}^{r} Y_{b_k} > 0|\vec{Q}\right)\right)\right].$$

Clearly, if zero counts are more probable in group Y then in group X, this expected difference is positive. So $\frac{\left(X_j, X_{b_1}, X_{b_2}, ..., X_{b_r}\right)}{max\left(1, X_j + \sum_{k=1}^{r} X_{b_k}\right)}$ and $\frac{\left(Y_j, Y_{b_1}, Y_{b_2}, ..., Y_{b_r}\right)}{max\left(1, Y_j + \sum_{k=1}^{r} Y_{b_k}\right)}$ are not equally distributed even if $H_0^{(j)}$ is true.

Another common solution used to avoid division by zero is to add a pseudocount of 1 to each count before testing for equality of distributions. However, this can make the inequality in the distributions of $\frac{\left(X_j,X_{b_1},X_{b_2},\ldots,X_{b_r}\right)}{max\left(1,X_j+\sum_{k=1}^{r}X_{b_k}\right)}$ and $\frac{\left(Y_j,Y_{b_1},Y_{b_2},\ldots,Y_{b_r}\right)}{max\left(1,Y_j+\sum_{k=1}^{r}Y_{b_k}\right)}$ even more pronounced when $H_0^{(j)}$ is true, as we show using a simple example.

Example 2: the effect of using pseudocounts. We consider a setting with $n_X = n_Y = 50$ samples and a constant sampling depth of $N_i^X = N_l^Y = N = 5000$ for all samples. We consider the population relative frequencies of taxa to be

$$\vec{P} = \left(1 - \frac{6}{N}, \frac{1}{N}, \frac{5}{N}\right), \tag{2.6}$$

$$\vec{Q} = (1-w)\cdot\vec{P} + w\cdot(1,0,0)\,,$$

where w may take values in $\{0.25, 0.33, 0.5\}$. The parameter w represents an increase in the total microbial load. For example, $w = 0.25$ represent a 33% increase in the total microbial load of samples in group Y compared to samples from group X. We test taxon 2 for differential abundance, with taxon 3 given as a reference, $B = \{3\}$. The null hypothesis $H_0^{(2)}$ is true.

Table 1 shows the unacceptably high type I error probability for the Wilcoxon rank sum test on 50 realizations of $\frac{X_2}{max(1,X_2+X_3)}$ and $\frac{Y_2}{max(1,Y_2+Y_3)}$, with and without the addition of pseudocounts. Figure 1 shows that the distribution of wilcoxon rank sum test P-value is stochastically smaller than the uniform distribution, so it is not a valid P-value for $H_0^{(2)}$.

Table 1: Probabilities for type I error when testing $\frac{X_2}{max(1,X_2+X_3)} \overset{\mathrm{d}}{=} \frac{Y_2}{max(1,Y_2+Y_3)}$ using the Wilcoxon rank sum test at $\alpha = 0.1$, in the setting defined by (2.6), for 3 values of w. Based on 10^4 simulations.

	$w = 0.25$	$w = 0.33$	$w = 0.5$
no pseudocount	0.19	0.17	0.38
pseudocount of 1	0.22	0.34	0.73

3 Testing for differential abundance

We now formulate a valid test for $H_0^{(j)}$. Let λ_j be the minimum total counts of the taxa with indices $j, b_1, \ldots, b_r$ across samples:

$$\lambda_j = \min\left\{\min_{i=1,\ldots,n_X}\left[X_{i,j} + \sum_{k=1}^{r} X_{i,b_k}\right], \min_{l=1,\ldots,n_Y}\left[Y_{l,j} + \sum_{k=1}^{r} Y_{l,b_k}\right]\right\}$$

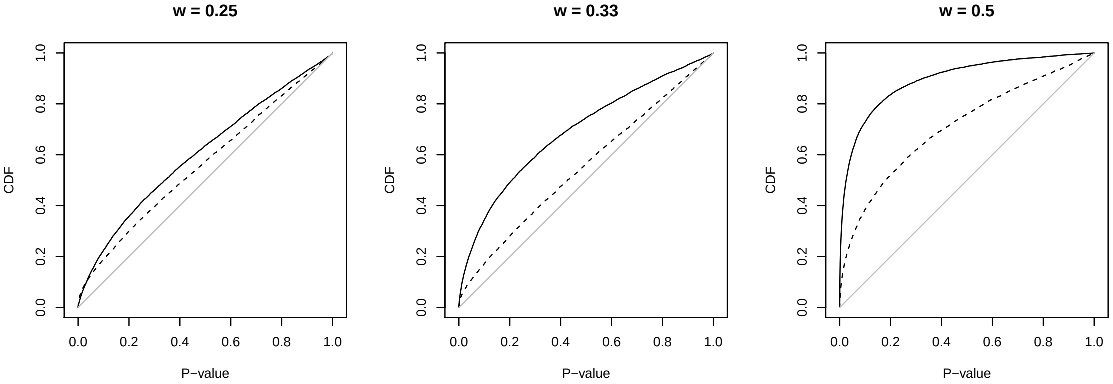


Figure 1: Cumulative Distribution Function (CDF) of the Wilcoxon rank sum test P-value, when testing $\frac{X_2}{max(1,X_2+X_3)} \overset{\mathrm{d}}{=} \frac{Y_2}{max(1,Y_2+Y_3)}$, in the setting defined by (2.6), without a pseudocount (dashed) and with a pseudocount of 1 (solid). The gray line depicts the CDF of the uniform distribution. Different subplots depict the CDF for different values of w.

Step I: Given λ_j,$X_{i,j}$ and $\sum_{k=1}^{r} X_{i,b_k}$, We sample $\tilde{X}_{i,j}$ from the hypergeometric distribution:

$$\tilde{X}_{i,j}| \left[\lambda_j, X_{i,j}, \sum_{k=1}^{r} X_{i,b_k}\right] \sim HG\left(\lambda_j, X_{i,j}, X_{i,j} + \sum_{k=1}^{r} X_{i,b_k}\right), \tag{3.1}$$

where $HG\left(t, z, z+w\right)$ is the distribution of the number of special items sampled when selecting t distinct items from a population of $z+w$ items, z of which are special. Similarly given λ_j,$Y_{l,j}$ and $\sum_{k=1}^{r} Y_{l,b_k}$, we sample $\tilde{Y}_{l,j}$:

$$\tilde{Y}_{l,j}| \left[\lambda_j, Y_{l,j}, \sum_{k=1}^{r} Y_{l,b_k}\right] \sim HG\left(\lambda_j, Y_{l,j}, Y_{l,j} + \sum_{k=1}^{r} Y_{l,b_k}\right).$$

Given λ_j, $\vec{P_i}$ and $\vec{Q_l}$, the subsampled counts $\tilde{X}_{i,j}$ and $\tilde{Y}_{l,j}$ have a binomial distribution:

$$\tilde{X}_{i,j}|\lambda_j, \vec{P_i} \sim Bin\left(\lambda_j, \frac{P_{i,j}}{P_{i,j} + \sum_k^r P_{i,b_r}}\right), \tag{3.2}$$

$$\tilde{Y}_{l,j}|\lambda_j, \vec{Q_l} \sim Bin\left(\lambda_j, \frac{Q_{l,j}}{Q_{l,j} + \sum_k^r Q_{l,b_r}}\right). \tag{3.3}$$

Step II: $H_0^{(j)}$ in (2.4) states that $\frac{P_{i,j}}{P_{i,j}+\sum_k^r P_{i,b_r}} \overset{\text{d}}{=} \frac{Q_{l,j}}{Q_{l,j}+\sum_k^r Q_{l,b_r}}$. From (3.2)-(3.3) it thus follows that if $H_0^{(j)}$ is true, the following null hypothesis is true:

$$\tilde{H}_0^{(j)} : \tilde{X}_{i,j} \overset{\text{d}}{=} \tilde{Y}_{l,j}. \tag{3.4}$$

This hypothesis can be tested using any two-sample test, e.g. the Wilcoxon rank sum test, on the counts $\{\tilde{X}_{i,j}, i = 1, ..., n_X\}$ and $\{\tilde{Y}_{l,j}, l = 1, ..., n_Y\}$.

Given the reference set B, the test of $H_0^{(j)}$ makes no assumptions on the distributions of $\mathcal{P}$, $\mathcal{Q}$, or on the structural zeros. The assumption free test comes at a price of first having to rarefy $X_{i,j}$ to $\tilde{X}_{i,j}$ and $Y_{i,j}$ to $\tilde{Y}_{i,j}$. Normalization by rarefaction has been criticized since only part of the data is used for inference [McMurdie and Holmes, 2014]. However, the alternative methods rely on parametric assumptions for modeling the data. Since little is known about the data generation mechanism, having no model assumptions is highly desired. 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 § 4-§ 5.

The value of λ_j is chosen so no samples are removed from the study for testing (3.4). Removing observations with total low count in taxons $j, b_1, ..., b_r$ below a fixed threshold may introduce a bias into the test, as we show in the following example.

Example 3: excluding samples based on reference size may induce bias. Consider taxon $j \in \mathcal{B}$ tested for differential abundance against a reference set B. We assume $\frac{P_j}{P_j+\sum_k^r P_{b_r}}$ and $\frac{Q_j}{Q_j+\sum_k^r Q_{b_r}}$ obtain the values 0.5 and 0.9 with equal probability. We observe a random sample, $n_X = n_Y = 16$. The total number of observed counts in taxa $\{j\} \cup B$ for samples in group Y is distributed $Pois\,(30)$. For samples in group X to total number of counts observed in taxa $\{j\} \cup B$ depends on $\frac{P_j}{P_j+\sum_k^r P_{b_r}}$: it is distributed $Pois\,(20)$ if $\frac{P_j}{P_j+\sum_k^r P_{b_r}} = 0.5$, and $Pois\,(40)$ otherwise. The sample sizes are $n_X = n_Y = 16$. Figure 2 shows that by subsampling to the minimum depth without exclusion of samples, the resulting samples appear to come from the same distribution, as expected (subplot B). However, if subsampling to a depth that requires samples below that depth to be excluded, the resulting samples no longer appear to come from the same distribution, potentially leading to spurious discovery claims (subplot C).

If the reference set of taxa has probability one of having a positive number of counts in all samples, it follows from Proposition 1 that $\frac{X_{i,j}}{X_{i,j}+\sum_{k=1}^r X_{i,b_k}}$ and $\frac{Y_{l,j}}{Y_{l,j}+\sum_{k=1}^r Y_{l,b_k}}$ have equal means. This motivates us to also consider testing whether there is a location shift in the distribution of $\frac{X_{i,j}}{X_{i,j}+\sum_{k=1}^r X_{i,b_k}}$ and $\frac{Y_{l,j}}{Y_{l,j}+\sum_{k=1}^r Y_{l,b_k}}$. We will refer to this

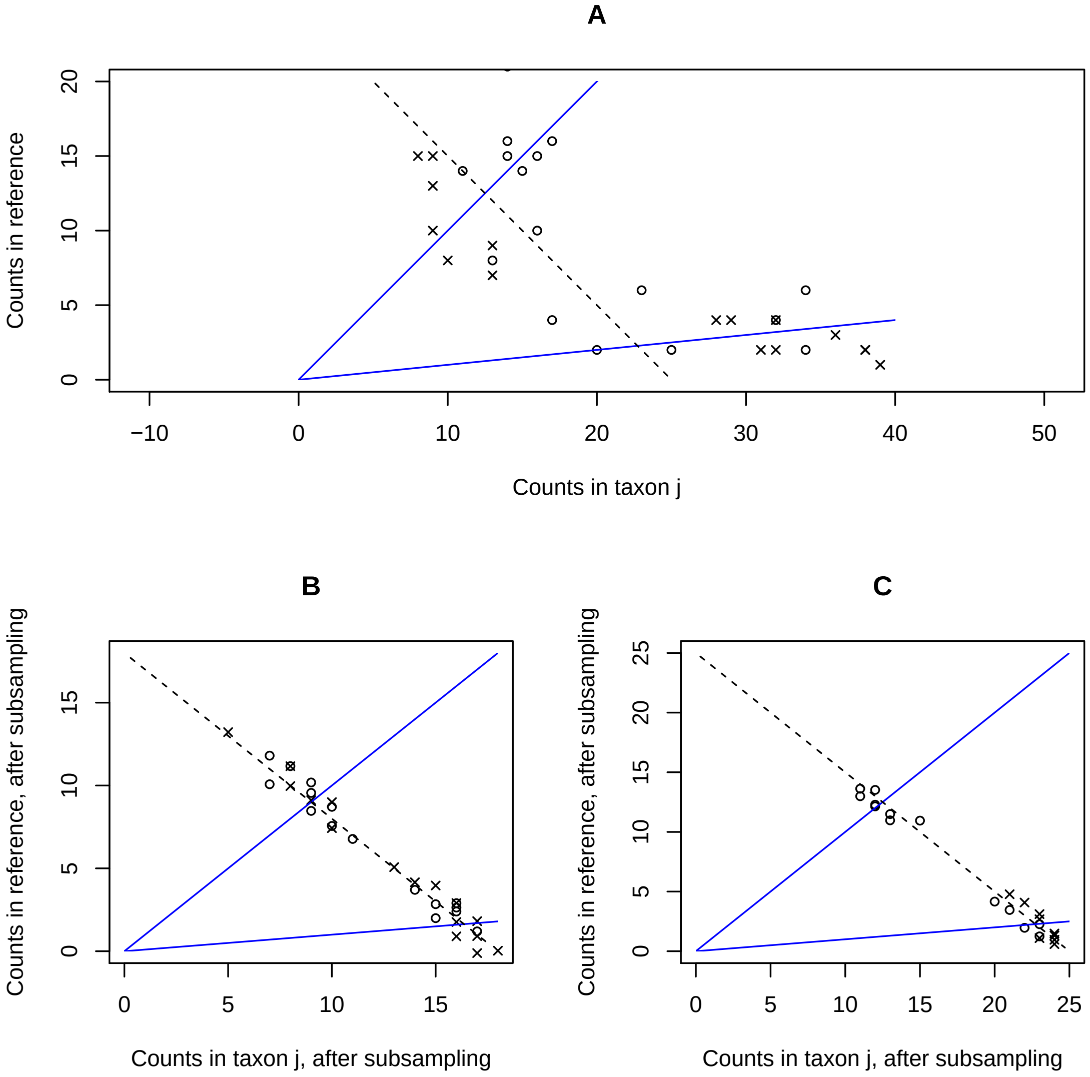


Figure 2: **(A)** The counts in taxon j vs. total counts in taxa set B. Crosses and circles represent samples from group X and Y, respectively. Blue lines form the two possible values of $\frac{P_j}{P_j+\sum_k^r P_{b_r}} \overset{\mathrm{d}}{=} \frac{Q_j}{Q_j+\sum_k^r Q_{b_r}}$, a ratio of 1 : 1 or 1 : 10. The dashed black line represents a total of 25 counts observed in taxon j and the reference set altogether. **(B)** Observations are subsampled to highest possible depth without removing samples, λ_j is shown by the black dashed line. To account for ties in the data, coordinates for the vertical axis were jittered. **(C)** Observations with less than 25 counts in taxa with indices $j \cup B$ were removed. The remaining observations, above the dashed line in subplot A, were subsampled to depth 25 and depicted in graph C.

method of differential abundance testing as ”normalization by ratio”. Since the distributions differ in spread, the test for location shift may not be valid. We compare this test to the necessary valid test of (3.4) in § 4-§ 5.

We note that samples with extremely low sampling depths may still be removed. This is justified if independence between the total number of counts per sample and the group labeling and thresholding over the total number of reads per sample is reasonable. Removal of samples with technical faults may be done prior to testing individual taxa for differential abundance.

Testing for differential abundance requires a set of reference taxa $\left(b_1, b_2, ..., b_{|B|}\right)$. The selection of reference taxa is discussed next.

3.1 Choosing reference taxa

If domain knowledge exists regarding taxa which are not associated with the condition examined, it can be used to construct a reference set of taxa. One possible technique to generate such a reference set is through a spike-in of synthesized DNA [see Section ”Spike-in log-ratio normalization” in Quinn et al., 2018]. Otherwise, when the set of reference taxa to subsample against is not known a-priori, a data-adaptive method for finding a subset of $\mathcal{B}$ is needed.

Without external information, we need to both identify the reference taxa, and then test with respect to this reference set, using the same dataset. It is important to identify the reference taxa without invalidating the testing that follows. Ideally, we would like the statistic used for taxa selection to be independent of the test statistic for $\tilde{H}_0^{(j)}$ [Hommel and Kropf, 2005]. As a first principle, our statistic for selection of reference taxa should not use the group labels.

We define $SD_{j,k}$ to be:

$$SD_{j,k} = \underset{i=1}{\overset{n_X+n_Y}{\mathrm{sd}}} \left(log_{10} \left(\frac{Z_{i,j}+1}{Z_{i,k}+1} \right) \right), \tag{3.5}$$

where sd is the sample standard deviation taken over $n_X + n_Y$ values and $Z_{i,j}$, $i = 1, ..., n_X + n_Y$ is the number of counts for taxon j in the ith observation from the combined sample. The statistic for selection of reference taxa is the median:

$$S_j = \underset{k=1,k\neq j}{\overset{m}{\mathrm{median}}} \left(SD_{j,k}\right). \tag{3.6}$$

The reference set is:

$$Select\, B: \quad B = \{j | S_j \leq S_{crit}\}. \tag{3.7}$$

The appropriate value of S_{crit} may be application specific, see appendix B.1 for details. We examined the distribution of this statistic for samples from different body sites, and have found the default value of $S_{crit} = 1.3$ to perform quite well, allowing for abundant taxa to enter the reference set, so λ_j is not too small. If soil or water samples are to be examined, this threshold may need to be recalibrated. Following selection of the potential reference set, we proceed to add or remove reference taxa, depending on whether the minimal number of total reference counts per sample is too small or too large. If it is too small, e.g., less than 10, we increase S_{crit} until the minimum of 10 total reference counts is reached by all samples. If it is too large, e.g., more than 200, we reduce S_{crit} until the minimum of 200 is reached.

4 A Simulation study

We use simulations for comparing the power and control of type I error of our approach and competitors. The different methods presented in § 1.1-§ 1.2 include both data normalization and statistical inference. To compare the normalization methods of different approaches on equal grounds, we use the Wilcoxon rank sum test and Welch t-tests, with the exception of the method by Kumar et al. [2018] which uses the test defined by Love et al. [2014]. While different normalization methods may employ a variety statistical tests, the Wilcoxon rank sum test is inherently built into the framework of Mandal et al. [2015] and its software package.

The following methods are compared:

ANCOM - The method of Mandal et al. [2015], as implemented in version 1.1-3 of the ANCOM package, with default parameter values.

W-FLOW - Wilcoxon rank sum tests with the correction by Vandeputte et al. [2017].

W-CSS Wilcoxon rank sum tests with the CSS normalization of Paulson et al. [2013b], as implemented in the software package 'metaGenomeSeq' in R [Paulson et al., 2013a] in version 1.24-1.

W-TSS Wilcoxon rank sum tests with the TSS normalization.

DACOMP Differential abundance testing with compositionality adjustment using the method suggested in the paper, as detailed in § 3, with $S_{crit} = 1.3$, and the Wilcoxon rank sum test as the two sample test between $\tilde{X}$ and $\tilde{Y}$.

DACOMP-t Differential abundance testing with compositionality adjustment using the method suggested in the paper, with $S_{crit} = 1.3$, and Welch t-tests over $log\left(\tilde{X}+1\right)$ and $log\left(\tilde{Y}+1\right)$ as the two sample test.

DACOMP-ratio Wilcoxon rank sum tests using 'normalization by ratio' with $S_{crit} = 1.3$, as described in § 3. The method will be used to compare the possible loss of power due to subsampling counts with DACOMP, and the possible inflation of type I error due to mistreatment of technical zeros, as discussed in § 2.1.

ALDEx2-t and ALDEx2-W - The method of Fernandes et al. [2013] presented in § 1.2, based on Welch t-tests and Wilcoxon rank-sum tests, as implemented in version 1.16-0 of the ‘ALDEx2‘ package.

WRENCH The method of Kumar et al. [2018], implemented in version 1.2-0 of the ‘wrench‘ package, with default parameters. The method makes use of the tests of differential abundance implemented in the ‘deseq2‘ software package [Love et al., 2014].

HG - Fisher's exact test against a reference set, as described in § 2. The reference set of taxa was selected by computing S_j for all non differentially abundant taxa, and selecting as a reference set all non differentially abundant taxa with $S_j \leq 1.3$. For this method, differentially abundant taxa were restricted from entering the reference set, in order to demonstrate that the bias in testing with HG is due to a failure to account for over dispersion.

For all methods except ANCOM, the BH procedure [Benjamini and Hochberg, 1995] at level $q = 0.1$ was applied for multiplicity correction. We chose the BH procedure since it aims to control the False Discovery Rate [FDR, Benjamini and Hochberg, 1995], and although the theoretical guarantee is only for independence or positive dependence, empirical evidence and simulations suggest it controls the FDR for most dependencies encountered in practice, including microbiome applications [Jiang et al., 2017, Mandal et al., 2015]. For W-TSS, T-CSS and W-FLOW, the BH procedure was performed on all m p-values, as all taxa are tested for differential abundance. For variants of DACOMP, the number of hypotheses tested is smaller than the original number of taxa, as only $m - |B|$ taxa outside of the reference set are tested for differential abundance.

For the brevity of this section, results are presented only for the methods ANCOM, W-FLOW, W-CSS, DACOMP, DACOMP-ratio, ALDEx2-t and HG, which represent key approaches, See appendix A for results on the other methods.

All power and type I error estimations were performed using 100 simulated datasets for each setting. The chance of a differentially abundant taxa to erroneously enter the reference set B, for most simulated settings, was very small or identically 0, see appendix B.1 for details.

The simulations in this section make use of the reference selection method presented in § 3.1. In appendix B.2, we examine the validity of our approach with alternative reference selection methods.

4.1 Resampling from a microbiome dataset

The data used as a basis for this simulation is described in Vandeputte et al. [2017], as the 'Disease cohort' of the study. The V4 region of the 16S gene was amplified and sequenced from fecal samples of 66 healthy subjects. In addition, the number of bacteria per gram were measured using a flow cytometer. The method of Amir et al. [2017] was used for picking sOTUs (subOTUs, or OTUs with maximal disagreement between reads of at most one base pair) from the data. sOTU length was set to the default value of 150 base pairs. In total, 1722 sOTUs were selected. All sOTUs which appeared in less than 4 subjects were removed from the data, leaving $m = 1066$ sOTUs. The median number of reads across subjects was $N_{reads} \equiv 22449$ reads across the 1066 sOTUs.

For a simulated dataset, a total of $n_X = 60$ 'healthy' and $n_Y = 60$ 'sick' subjects were sampled. A healthy subject from group X was sampled in the following manner:

1. For generating the ith subject, a 16S and flow cytometric measurement from a real healthy subject was sampled. Let $\vec{u}_i^X$ be the vector of m sOTU measurements and $C_i^{X,flow}$ the flow-cytometric read obtained from the sampled subject.

2. Let N_i^X represent the total number of 16S reads observed for the ith subject. N_i^X was sampled from the Poisson distribution with parameter N_{reads}.

3. Let $\vec{v}_i^X$ denote the unobserved total abundances of taxa, $\vec{v}_i^X = C_i^{X,flow} \cdot \frac{\vec{u}_i^X}{\sum_{j=1}^m u_{i,j}^X}$. The observed count vector for sample i, $\vec{X}_i$, is sampled from the multinomial distribution with parameters N_i^X and $\vec{P}_i = \frac{\vec{v}_i^X}{\sum_{j=1}^m \vec{v}_{i,j}^X}$.

The 'sick' subjects from group Y was generated in a manner similar to steps 1-3 above, with the following changes: $m_1 \in \{10, 100\}$ differentially abundant taxa were selected at random. Each taxon j associated with the disease had a chance of 0.5 to experience an increase in its absolute abundance of bacteria in each 'sick' subject. The random number of bacteria added to the absolute abundance of the jth taxon was sampled, independently for each entry, from $N(\mu_{l,j}, \mu_{l,j})$ and rounded to the nearest integer, where $\mu_{l,j} = \lambda_{effect} \cdot C_l^{Y,flow} \cdot \delta_j / m_1$. The value of δ_j was sampled with equal probability from $\{0.5, 1.0, 1.5\}$ for each taxon and the value of λ_{effect} was set to $0, 0.5, 1.0, ..., 2.5$ in the different scenarios. In terms of simulation parameters, λ_{effect} represents the total effect of simulated condition over the microbial load of the host while δ_j sets the strength of association of a specific taxon with the simulated condition, and is fixed across different samples. Taxa not selected as differentially abundant maintain their absolute abundances across study groups.

Figure 3 shows the estimated FDR for each method, for the different scenarios. DACOMP is the only method controlling FDR across all scenarios considered. For the global null setting ($\lambda_{effect} = 0$), only ANCOM and HG do not control the FDR. For HG this is expected since we have overdispersion in the data. For ANCOM, we have observed that generally, under the global null, FDR is not controlled. The empirial decision rule used in ANCOM tends to declare the taxa with largest $\mathcal{W}_j$ as differentially abundant, even if $\mathcal{W}_j$ is small, i.e. may be non zero due to chance alone. In appendix C, we present additional scenarios with no differentially abundant taxa where ANCOM does not control FDR for several method parameter values. ANCOM and W-FLOW lack FDR control when λ_{effect} is large, i.e. $\lambda_{effect} \geq 2.0$. For ANCOM, this could be attributed either to the empirical decision rule being invalid or to mistreatment of technical zeros by using a psuedocount. For 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 smaller but non-neglible when the change in microbial load is large. The estimated FDR for $\lambda_{effect} = 3$ and $m_1 = 10$ is 0.17; for $\lambda_{effect} \in \{2, 2.5, 3\}, m_1 \in \{10, 100\}$ it is also above $q = 0.1$.

Figure 4 shows the estimated power for each method across the different simulated scenarios. HG was excluded, since its FDR was inflated in all settings. 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 $m_1 = 100, \lambda_{effect} = 0.5$; to 691, for $m_1 = 10, \lambda_{effect} = 3.0$ (maximal standard error across scenarios was 14.62). 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 $\lambda_{effect} = 2.5$ and

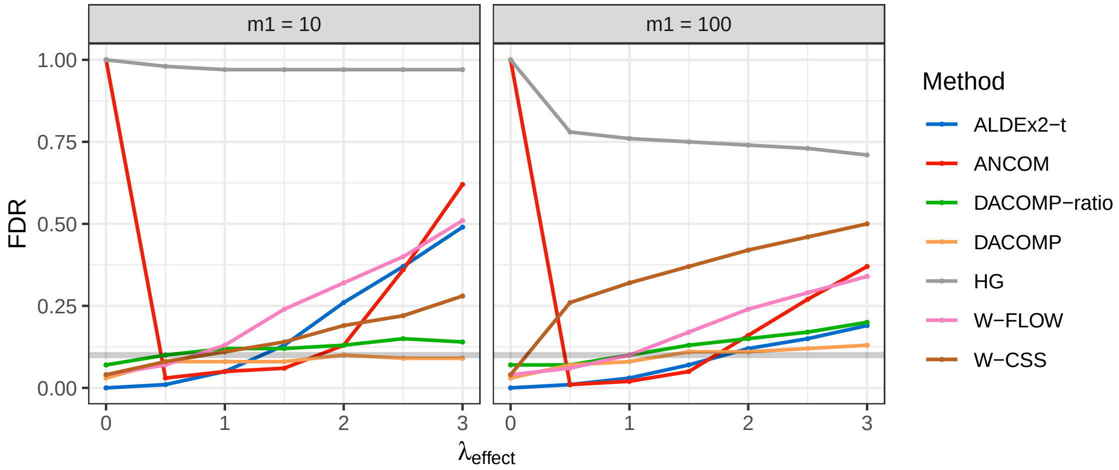


Figure 3: Estimated FDR versus λ_{effect} for DACOMP and competitors in the simulation settings of § 4.1. The maximal standard error is 0.04. BH procedure was used at $q = 0.1$, marked by the gray solid line.

$m_1 = 100$, W-CSS has 176 discoveries on average, but only 95 true discoveries, and a FDR of 0.46.

4.2 A setting where the most abundant taxa are not differentially abundant

We consider $\vec{P}$ to be constant, with the first m_A components having the values $\frac{p_A}{m_A}$. The remaining $m - m_A$ taxa have entries $\frac{1-p_A}{m-m_A}$:

$$\vec{P} = \left(\frac{p_A}{m_A}, ..., \frac{p_A}{m_A}, \frac{1-p_A}{m-m_A}, ..., \frac{1-p_A}{m-m_A}\right).$$

Subjects of the second group were multinomial samples with differentially abundant taxa selected from the taxa with relative frequencies $\frac{1-p_A}{m-m_A}$:

$$\vec{Q} = (1-w) \cdot \vec{P} + w \cdot (0, ..., 0, 1, ..., 1, 0, ..., 0),$$

where w is the proportion of signal added to the vector of relative frequencies and vector on the right term has m_1 entries with indices larger than m_A with a value of 1, rendering the corresponding taxa as differentially abundant. For each simulated dataset, $n_X = 20$ and $n_Y = 20$ subjects were sampled for each group as multinomial random vectors from $N_{reads} = 2500$ reads in each vector using $\vec{P}_X$ and $\vec{P}_Y$

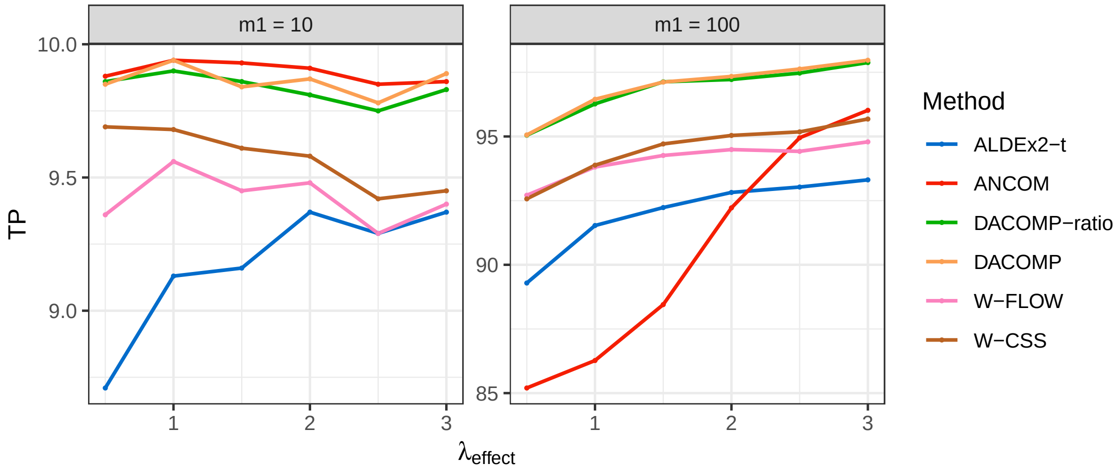


Figure 4: Estimated power versus λ_{effect} for DACOMP and competitors in the simulation settings of § 4.1. The maximal standard error is 0.73. The BH procedure was used at $q = 0.1$.

respectively. We examined simulations with $m = 300, m_{high} = 30, w = 0.35, m_1 \in \{120, 60\}, p_{high} \in \{0.9, 0.8, 0.7, 0.6, 0.5\}$.

Table 2 shows the estimated FDR, for each simulated setting, by m_1 and p_{high}. We note that the only methods providing FDR control across all scenarios are DACOMP and HG. Since there is no overdispersion in the data, these two methods are theoretically valid (but DACOMP is also valid when there is overdispersion). ANCOM and ALDEx2-t provide FDR control for settings with $m_1 = 60$, but not for settings with $m_1 = 120$. ANCOM's loss of FDR control for settings with $m_1 = 120$, is related to the loss of power: As described in § 1.2, the method of Mandal et al. [2015] makes use of the test statistics $W_{j,k}$. Implicitly, it is assumed that if taxon j or k are differentially abundant, the p-value of $W_{j,k}$ will be smaller than α, e.g., $\alpha = 0.1$, with high probability. If this assumption is violated, the highest values of $\mathcal{W}_j$ may not be obtained by the differentially abundant taxa. The setting generated demonstrates this effect. ANCOM fails to identify the differentially abundant taxa, and points to the most abundant taxa which are non differentially abundant as associated with the disease. W-CSS provides FDR control for only two of the scenarios considered. For DACOMP-ratio, the estimated FDR for $m_1 = 60, p_{high} \in \{0.7, 0.8, 0.9\}$ is higher than $q = 0.1$.

In terms of power, all methods discovered all differentially abundant taxa, expect for: ANCOM discovered 33,30,28,34 and 28 taxa in the settings of rows 1-5, respectively; ALDEx2-t discovered 120,118,116,115 and 111 taxa in the settings of rows 1-5, respectively. The maximum standard error for average number of taxa discovered is 3.71.

Table 2: Estimated FDR of DACOMP and competitors (Columns 3-8) for the simulations where the most abundant taxa are not differentially abundant. Column 1-2 give the number of differentially abundant taxa and the value of the parameter p_A, respectively. BH procedure was used at $q = 0.1$. For DACOMP, $S_{crit} = 1.3$. The maximum standard error a table entry is 0.03. For DACOMP-ratio, the maximum standard error across table entries is 0.004.

m_1	p_A	ALDEx2-t	ANCOM	DACOMP-ratio	DACOMP	HG	W-CSS
120	0.90	0.3	0.34	0.1	0.07	0.07	0.54
120	0.80	0.46	0.25	0.09	0.07	0.07	0.57
120	0.70	0.5	0.28	0.08	0.07	0.06	0.48
120	0.60	0.52	0.24	0.07	0.06	0.05	0.38
120	0.50	0.54	0.24	0.07	0.06	0.06	0.34
60	0.90	0	0.01	0.14	0.09	0.1	0.59
60	0.80	0.02	0.01	0.12	0.09	0.09	0.44
60	0.70	0.02	0.01	0.11	0.08	0.08	0.43
60	0.60	0.01	0.01	0.1	0.08	0.07	0.1
60	0.50	0	0.01	0.09	0.07	0.07	0.08

4.3 Cases with no compositionality

We wish to asses the potential loss of power by using a method that adjusts for compositionality, when adjustment for compositionality is in fact unnecessary for valid inference. Taxon counts are considered as an independent sample from a negative binomial distribution where the mean is μ and the variance is given by $\mu + \frac{\mu^2}{5}$.

Simulated data for group X consisted of $m = 1000$ taxa sampled as independent negative binomial variables, with 50 highly abundant taxa with a mean of 200, 150 medium abundance taxa with a mean of 20 and 800 taxa with low abundance having a mean of 1. For simulating group Y, 10 taxa with high abundance, 10 taxa with medium abundance, and 30 taxa with low abundance were selected as differentially abundant. Out of each abundance group (means of 1,20,200), of the differentially abundant taxa half had their means reduced by 75% and half had their means increased by 75%. Therefore the distribution of non differentially abundant taxa is the same in the two groups. Sample size was $n_X = n_Y \in \{15, 20, 25, 30\}$.

Table 3 describes the estimated FDR for DACOMP and competitors across the different methods. W-CSS, ANCOM ALDEx2 and both DACOMP variants control the FDR at the required rate. For W-CSS and DACOMP-ratio, this result is expected since all non differentially abundant taxa have maintained their marginal distributions across study groups. HG does not provide FDR control due to overdispersion in the data.

Table 4 describes the average number of differentially abundant taxa discovered by each setting. W-CSS discovers the highest number of differentially abundant taxa, followed by DACOMP-ratio. ANCOM and ALDEx2-t have a comparable number of discoveries across all sample sizes considered. DACOMP has higher power compared to both ANCOM and ALDEx2-t, and lower power than DACOMP-ratio. The difference in power between W-CSS and DACOMP-ratio to other competitors results mainly from having better power to detect taxa with low counts that are differentially abundant.

Table 3: Estimated FDR of DACOMP and competitors (Columns 2-7) for simulations with no compositionality, for various sample sizes (Column 1). BH procedure was applied at level $q = 0.1$. The maximum standard error of a table entry is 0.01.

n_X:n_Y	ALDEx2-t	ANCOM	DACOMP-ratio	DACOMP	HG	W-CSS
15:15	0	0	0.11	0.06	0.72	0.09
20:20	0	0.01	0.09	0.07	0.7	0.09
25:25	0	0	0.11	0.08	0.7	0.09
30:30	0	0	0.1	0.08	0.68	0.11

Table 4: Average number of differentially abundant taxa discovered by DACOMP and competitors that controlled FDR (Columns 2-6) for simulations with no compositionality, for different sample sizes (Column 1). BH procedure was applied at level $q = 0.1$. The maximum standard error of a table entry is 0.42.

n_X:n_Y	ALDEx2-t	ANCOM	DACOMP-ratio	DACOMP	W-CSS
15:15	10.39	11	15.44	12.2	17.34
20:20	13.58	12.42	19.62	14.63	22.18
25:25	15.48	13.91	23.38	16.78	26.91
30:30	16.91	16.43	28.27	19.4	31.14

5 Comparing fecal samples from healthy subjects and subjects with Crohn's Disease

Vandeputte et al. [2017] examined 16S profiles taken from fecal samples to investigate changes in gut microbiome with Crohn's Disease (CD). 95 subjects were studied, 29 of which have CD. All subjects had 16S profiling for their fecal samples taken along with a microbial load count, given in number of bacteria per gram of fecal material.

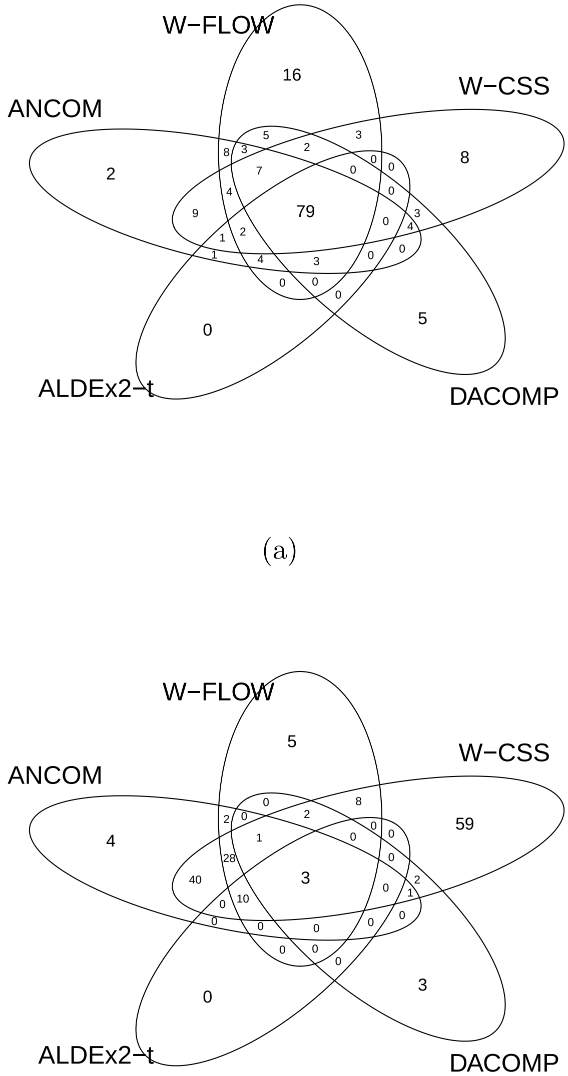
16S profiling of subjects had a median value of 20437 reads per sample, The median microbial load across study groups was $1.16 \cdot 10^{11}$ and $3.76 \cdot 10^{10}$ bacteria per gram for healthy subjects and subjects with CD, respectively. The ratio of median microbial load between groups was 3.08, indicating a vast change in the total abundance of microbial ecosystem when CD is present. With such a high change in microbial load, it is implausible to assume most taxa have not altered their absolute abundance across study groups in the presence of CD.

To analyze the data, sOTUs were picked using the method of Amir et al. [2017]. The count matrix to be analyzed consisted of 1980 sOTUs across 95 subjects. All sOTUs that appeared in at most one subject were removed, leaving 1569 sOTUs in the data set.

Table 5 shows the number of discoveries for each method, along with the number of discoveries shared by the different methods. DACOMP, ALDEx2-W and ALDEx-t agree more or less in terms of discoveries, with all three methods discovering substantially less taxa than ANCOM, W-FLOW and W-CSS.

Figure 5 depicts the number of discoveries shared by each method, except for ALDEx2-W (which relatively agrees with ALDEx2-t). Discoveries are examined separately for taxa which have on average per subject at least 10 counts or less than 10 counts, henceforth known as 'abundant' and 'rare' taxa. The 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 11 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 ANCOM and W-FLOW but not DACOMP are found among the rare taxa.

Table 6 shows the number of discoveries by DACOMP for several values of S_{crit} alongside the obtained reference size and the number of discoveries shared with other methods. For the values of S_{crit} described in the table, as S_{crit} increases, more taxa enter the selected set of references. As a result, less taxa are tested and discovered



(a)

(b)

Figure 5: Graphical representation of discoveries shared by different methods. The top panel shows discoveries of 'abundant' taxa with at least 10 counts, on average, per sample. The bottom panel shows discoveries of 'rare' taxa with less than 10 counts, on average, per sample.

Table 5: Number of discoveries by each method, for the data of Vandeputte et al. [2017]. Diagonal entries show the number of discoveries by the method. Off-diagonal entries show the number of discoveries shared by two methods. Results for DACOMP are presented for $S_{crit} = 1.3$. BH procedure was applied at the 0.1 level.

Method Name	ANCOM	W-FLOW	W-CSS	DACOMP	ALDEx2-W	ALDEx2-t
ANCOM	216	154	189	101	127	103
W-FLOW		195	149	105	116	101
W-CSS			276	104	118	95
DACOMP				123	89	85
ALDEx2-W					127	101
ALDEx2-t						103

as differentially abundant.

For $S_{crit} \in \{1.2, 1.3, 1.4\}$, DACOMP-ratio discovered 200,163, and 147 taxa as differentially abundant, respectively. The difference in the number of discoveries between DACOMP and DACOMP-ratio may result from two reasons: (1) a reduction in power due to subsampling step done in DACOMP or (2) DACOMP-ratio not controlling the rate of false positive discoveries. Investigating whether the additional findings by DACOMP-ratio are true is interesting but outside the scope of this work.

Table 6: Number of discoveries by S_{crit} for DACOMP. Columns 2-4 show for each value of S_{crit} the number of discoveries, shared discoveries with Mandal et al. [2015] and the normalization method of Vandeputte et al. [2017] and the number of OTUs in the selected reference set B, respectively.

S_{crit}	Discoveries	Shared, ANCOM	Shared, W-FLOW	Shared, ALDEx2-t	$\|B\|$
1.2	149	121	122	92	1221
1.3	123	101	105	85	1288
1.4	108	93	98	79	1335

6 Final Remarks

We present a novel method for detecting differential abundance of taxa in a compositional regime, while accounting for the discrete nature of counts and in particular the many zeros. The method presented in this work stands out from previous works in that it avoids the need to define the generating model of relative frequencies and specifically, the generating model of zeros. This method can also be used for testing the association of a continuous phenotype with the microbial composition. Moreover, generally, the method can be used with a multivariate phenotype. We present the details in appendix D.

Our methodology depends on the particular rarefied sample that resulted in one draw. A non-valid method for using several rarefied samples would be to average test statistics across multiple rarefactions of the data, or to average the rarefied draws themselves. To see why, consider the case where the tested taxon j is not differentially abundant, and the total number of counts available in taxa $\{j\} \cup B$ for samples in group X is stochastically smaller than for samples in group Y., i.e., less counts are available to sample from in one group compared to the other. Hence, counts in samples belonging to group X are more likely to be resampled across multiple rarefactions of the data compared to counts from group Y. Therefore, the bivariate distribution of two rarefied draws taken from a single sample is different across study groups. For example, multiple draws from a sample in group X will have a higher correlation compared to multiple draws from a sample in group Y. We showed that by not rarefying, i.e., applying the DACOMP-ratio approach, we typically gain power at the price of a small inflation in the type 1 error probability.

The procedure presented in § 3.1 selects a set of non-differentially abundant taxa to be used for testing under the assumption that most taxa are not differentially abundant. If the fraction of differentially abundant taxa is not small, a preset threshold of S_{crit} may not suffice to ensure that no differentially abundant taxa were selected as reference. This is a topic for future research.

For brevity, we presented only one tissue analysis. Analysis of other tissues is carried out in appendix E. We analyzed the the differential abundance of taxa across adjacent body sites in the human body using data from the Human Microbiome Project [Gevers et al., 2012]. 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. Interestingly, the same reference selection procedure is found to operate well across all pairwise comparisons of body sites, despite different sequencing depths, prevalence of zero counts and the number of differentially abundant taxa discovered.

7 Supplementary Material

The methods presented in this paper for differential abundance testing and reference selection are available as an R package on Github (github.com/barakbri/dacomp). The `dacomp` package supports differential abundance testing for studies with $k \geq 2$ study groups, a continuous phenotype, and a multivariate phenotype. Source code and instructions describing how to reproduce the results in this paper are found on (github.com/barakbri/CompositionalAnalysis_CodeBase).

An additional PDF file with supplementary material contains the following Sections: appendix A contains simulations results for additional competitor methods, for the settings described in § 4; appendix B contains further examination of the reference selection procedure, discusses how S_{crit} was set, and reviews alternative reference selection procedures; appendix C contains a simulation analyzing the control of false positive discoveries by ANCOM when $m_1 = 0$; appendix D reformulates the method presented § 3 to a general setting, when testing for association with continuous and multivariate phenotypes ; finally, appendix E describes an analysis of differential abundance in the Human Microbiome Project (HMP) across pairs of body sites in the human body.

References

John Aitchison. The statistical analysis of compositional data. 1986.

Amnon Amir, Daniel McDonald, Jose A Navas-Molina, Evguenia Kopylova, James T Morton, Zhenjiang Zech Xu, Eric P Kightley, Luke R Thompson, Embriette R Hyde, Antonio Gonzalez, et al. Deblur rapidly resolves single-nucleotide community sequence patterns. *MSystems*, 2(2):e00191–16, 2017.

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

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

Jun Chen and Hongzhe Li. Variable selection for sparse dirichlet-multinomial regression with an application to microbiome data analysis. *The annals of applied statistics*, 7(1), 2013.

Andrew D Fernandes, Jean M Macklaim, Thomas G Linn, Gregor Reid, and Gregory B Gloor. Anova-like differential expression (aldex) analysis for mixed population rna-seq. *PLoS One*, 8(7):e67019, 2013.

Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200): 675–701, 1937.

Dirk Gevers, Rob Knight, Joseph F Petrosino, Katherine Huang, Amy L McGuire, Bruce W Birren, Karen E Nelson, Owen White, Barbara A Methé, and Curtis Huttenhower. The human microbiome project: a community resource for the healthy human microbiome. *PLoS biology*, 10(8):e1001377, 2012.

Gregory B Gloor, Jean M Macklaim, Vera Pawlowsky-Glahn, and Juan J Egozcue. Microbiome datasets are compositional: and this is not optional. *Frontiers in microbiology*, 8:2224, 2017.

Micah Hamady and Rob Knight. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome research*, 19(7): 1141–1152, 2009.

Stijn Hawinkel, Federico Mattiello, Luc Bijnens, and Olivier Thas. A broken promise: microbiome differential abundance methods do not control the false discovery rate. *Briefings in bioinformatics*, 2017.

Ruth Heller and Yair Heller. Multivariate tests of association based on univariate tests. In *Advances in Neural Information Processing Systems*, pages 208–216, 2016.

Ruth Heller, Yair Heller, and Malka Gorfine. A consistent multivariate test of association based on ranks of distances. *Biometrika*, 100(2):503–510, 2012.

Ian Holmes, Keith Harris, and Christopher Quince. Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PloS one*, 7(2):e30126, 2012.

Gerhard Hommel and Siegfried Kropf. Tests for differentiation in gene expression using a data-driven order or weights for hypotheses. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 47(4):554–562, 2005.

L Jiang, A Amir, JT Morton, R Heller, E Arias-Castro, and R Knight. Discrete false-discovery rate improves identification of differentially abundant microbes. msystems 2: e00092-17. 2017.

Abhishek Kaul, Siddhartha Mandal, Ori Davidov, and Shyamal D Peddada. Analysis of microbiome data in the presence of excess zeros. *Frontiers in microbiology*, 8: 2114, 2017.

William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

M Senthil Kumar, Eric V Slud, Kwame Okrah, Stephanie C Hicks, Sridhar Hannenhalli, and Hector Corrada Bravo. Analysis and correction of compositional bias in sparse sequencing count data. *BMC genomics*, 19(1):799, 2018.

Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550, 2014.

Siddhartha Mandal, Will Van Treuren, Richard A White, Merete Eggesbø, Rob Knight, and Shyamal D Peddada. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial ecology in health and disease*, 26(1):27663, 2015.

Paul J McMurdie and Susan Holmes. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS computational biology*, 10(4):e1003531, 2014.

James T Morton, Clarisse Marotz, Alex Washburne, Justin Silverman, Livia S Zaramela, Anna Edlund, Karsten Zengler, and Rob Knight. Establishing microbial composition measurement standards with reference frames. *Nature communications*, 10(1):2719, 2019.

Michael C Nelson, Hilary G Morrison, Jacquelynn Benjamino, Sharon L Grim, and Joerg Graf. Analysis, optimization and verification of illumina-generated 16s rrna gene amplicon surveys. *PloS one*, 9(4):e94249, 2014.

John D. O'Brien and Nicolas Record. The power and pitfalls of dirichlet-multinomial mixture models for ecological count data. *bioRxiv*, 2016. doi: 10.1101/045468. URL https://www.biorxiv.org/content/early/2016/03/24/045468.

Joseph N. Paulson, Mihai Pop, and Hector Corrada Bravo. *metagenomeSeq: Statistical analysis for sparse high-throughput sequncing.*, 2013a. URL http://www.cbcb.umd.edu/software/metagenomeSeq. Bioconductor package.

Joseph N Paulson, O Colin Stine, Héctor Corrada Bravo, and Mihai Pop. Differential abundance analysis for microbial marker-gene surveys. *Nature methods*, 10(12): 1200, 2013b.

Thomas P Quinn, Ionas Erb, Greg Gloor, Cedric Notredame, Mark F Richardson, and Tamsyn M Crowley. A field guide for the compositional analysis of any-omics data. *bioRxiv*, page 484766, 2018.

Maria L Rizzo, Gábor J Székely, et al. Disco analysis: A nonparametric extension of analysis of variance. *The Annals of Applied Statistics*, 4(2):1034–1055, 2010.

Mahdi Shafiei, Katherine A Dunn, Eva Boon, Shelley M MacDonald, David A Walsh, Hong Gu, and Joseph P Bielawski. Biomico: a supervised bayesian model for inference of microbial community structure. *Microbiome*, 3(1):8, 2015.

C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. ISSN 00029556. URL http://www.jstor.org/stable/1412159.

Gábor J Székely and Maria L Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5(16.10), 2004.

Shu Mei Teo, Danny Mok, Kym Pham, Merci Kusel, Michael Serralha, Niamh Troy, Barbara J Holt, Belinda J Hales, Michael L Walker, Elysia Hollams, et al. The infant nasopharyngeal microbiome impacts severity of lower respiratory infection and risk of asthma development. *Cell host & microbe*, 17(5):704–715, 2015.

Luke R Thompson, Jon G Sanders, Daniel McDonald, Amnon Amir, Joshua Ladau, Kenneth J Locey, Robert J Prill, Anupriya Tripathi, Sean M Gibbons, Gail Ackermann, et al. A communal catalogue reveals earth's multiscale microbial diversity. *Nature*, 551(7681):457, 2017.

Doris Vandeputte, Gunter Kathagen, Kevin D'hoe, Sara Vieira-Silva, Mireia Valles-Colomer, João Sabino, Jun Wang, Raul Y Tito, Lindsey De Commer, Youssef Darzi, et al. Quantitative microbiome profiling links gut community variation to microbial load. *Nature*, 551(7681), 2017.

Brandie D Wagner, Charles E Robertson, and J Kirk Harris. Application of two-part statistics for comparison of sequence variant counts. *PloS one*, 6(5):e20296, 2011.

Sophie Weiss, Zhenjiang Zech Xu, Shyamal Peddada, Amnon Amir, Kyle Bittinger, Antonio Gonzalez, Catherine Lozupone, Jesse R Zaneveld, Yoshiki Vázquez-Baeza, Amanda Birmingham, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5(1):27, 2017.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

Lizhen Xu, Andrew D Paterson, Williams Turpin, and Wei Xu. Assessment and selection of competing models for zero-inflated microbiome data. *PloS one*, 10(7): e0129606, 2015.

A Simulation results for additional methods

In this appendix we present results for the simulation study discussed in § 4, for the following methods: W-TSS, ALDEx2-W, DACOMP-t and WRENCH. Method description and details are at the start of § 4. In terms of Power and FDR, unless stated otherwise, W-TSS was similar to W-CSS, ALDEx2-W was similar to ALDEx2-t and DACOMP-t was similar to DACOMP.

Figure 6 shows the estimated FDR for the additional methods listed above, obtained the simulation settings discussed in § 4.1. Similar to DACOMP, DACOMP-t is shown to control the false discovery rate at $q = 0.1$. W-TSS does not control the false discovery rate for all scenarios with $\lambda_{effect} >= 1.0$, since it provides marginal inference alone. ALDEx2-W failed to control for false positives at $\lambda_{effect} >= 1.5$. When comparing ALDEx2-W to ALDEx2-t in terms of FDR control, FDR rates for ALDEx2-W were significantly higher, e.g., for $\lambda_{effect} = 2$ with $m_1 = 100$, the estimated FDR for ALDEx2-W was 0.22 but was only 0.12 for ALDEx2-t. For WRENCH, FDR was not controlled under the global null, $\lambda_{effect} = 0$, or for $m_1 = 10$. For $m_1 = 100, \lambda_{effect} > 0$, the standard error for WRENCH's FDR estimates were smaller than 0.01, indicating that the observed FDR levels, approximately 0.16 across all values of $\lambda_{effect} > 0$, are significantly different from $q = 0.1$. The lack of FDR control could be related to the warning message about failure of algorithm convergence. The method was run with the default parameters. Other parameter estimates for WRENCH may produce better FDR control.

Figure 7 shows the power of DACOMP-t and alternative methods, for the scenarios of § 4.1. All method variants (DACOMP-t,ALDEx2-W, W-TSS) gave results similar to their respective variants, as described in the beginning of the section.

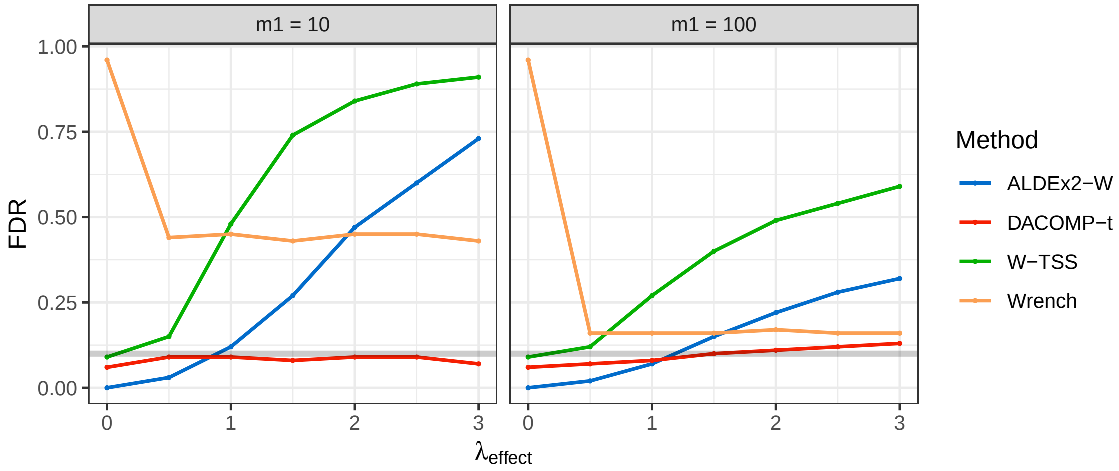


Figure 6: Estimated FDR of DACOMP and competitors for the simulation settings of § 4.1. The Y axis represents estimated FDR, the X axis represents λ_{effect}, the increase in percents in the microbial load of a sample with the simulated condition, e.g. a value of 1.0 means a 100% increase in the microbial load. The maximal standard error is 0.04. BH procedure was used at $q = 0.1$. The dashed gray line marks a distance of twice the maximal standard error from $q = 0.1$.

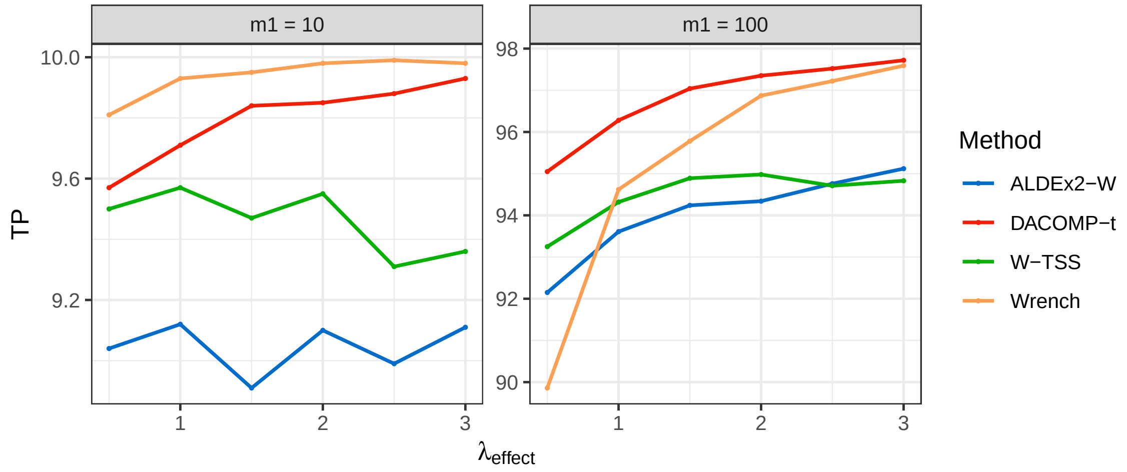


Figure 7: Estimated power of DACOMP and competitors for the simulation settings of § 4.1. The Y axis represents average number of true discoveries, the X axis represents λ_{effect}, the increase in percents in the microbial load of a sample with the simulated condition, e.g. a value of 1.0 means a 100% increase in the microbial load. The maximal standard error is 0.73. BH procedure was used at $q = 0.1$.

Table 7 shows the estimated FDR for DACOMP-t and competitors, for the simulation scenarios discussed in § 4.2. DACOMP-t is the only method shown to provide FDR control across all scenarios, similar to DACOMP in Table 2. In terms of power, for the scenario in rows 1-5, all methods discovered all differentially abundant taxa, except for ALDEx2-W which discovered 120,117,115,113,109 of the differentially abundant taxa; . The maximum standard error for average number of taxa discovered is 3.71.

Table 8 presents the estimated FDR for DACOMP-t and alternative methods, for the scenarios described in § 4.3. For the scenarios of § 4.3, non-differentially abundant taxa maintained their marginal distributions of counts across study groups. Hence, all tests presented in Table 8 provide valid FDR control.

Table 9 shows the number of true positive discoveries for DACOMP-t and alternative methods, for the scenarios of § 4.3. WRENCH is shown to provide the highest power, even higher than W-TSS.

Table 7: Estimated FDR of DACOMP-t and competitors (Columns 3-6) for the simulations where the most abundant taxa are not differentially abundant. Column 1-2 give the number of differentially abundant taxa and the value of the parameter p_A, respectively. BH procedure was used at $q = 0.1$. The maximum standard error a table entry is 0.03.

m_1	p_A	ALDEx2-W	DACOMP-t	W-TSS	WRENCH
120	0.90	0.42	0.07	0.34	0.46
120	0.80	0.55	0.07	0.43	0.56
120	0.70	0.57	0.07	0.49	0.57
120	0.60	0.57	0.06	0.53	0.57
120	0.50	0.57	0.06	0.55	0.58
60	0.90	0	0.09	0.52	0.46
60	0.80	0.01	0.09	0.64	0.65
60	0.70	0.01	0.08	0.71	0.71
60	0.60	0.01	0.08	0.74	0.73
60	0.50	0	0.08	0.76	0.75

Table 8: Estimated FDR of DACOMP-t and competitors (Columns 2-5) for simulations with no compositionality, for various sample sizes (Column 1). BH procedure was applied at level $q = 0.1$. The maximum standard error of a table entry is 0.01.

n_X:n_Y	ALDEx2-W	DACOMP-t	W-TSS	WRENCH
15:15	0	0.05	0.09	0.06
20:20	0	0.08	0.09	0.06
25:25	0	0.09	0.09	0.06
30:30	0	0.07	0.1	0.07

Table 9: Average number of differentially abundant taxa discovered by DACOMP-t and competitors that controlled FDR (Columns 2-5) for simulations with no compositionality, by sample size (Column 1). BH procedure was applied at level $q = 0.1$. The maximum standard error of a table entry is 0.42.

n_X:n_Y	ALDEx2-t	DACOMP-t	W-TSS	WRENCH
15:15	11.48	12.35	16.97	18.22
20:20	13.52	14.81	22.5	22.82
25:25	15.6	16.73	27.08	27.78
30:30	17.18	19.84	30.86	31.78

B Further examination of the reference selection procedure

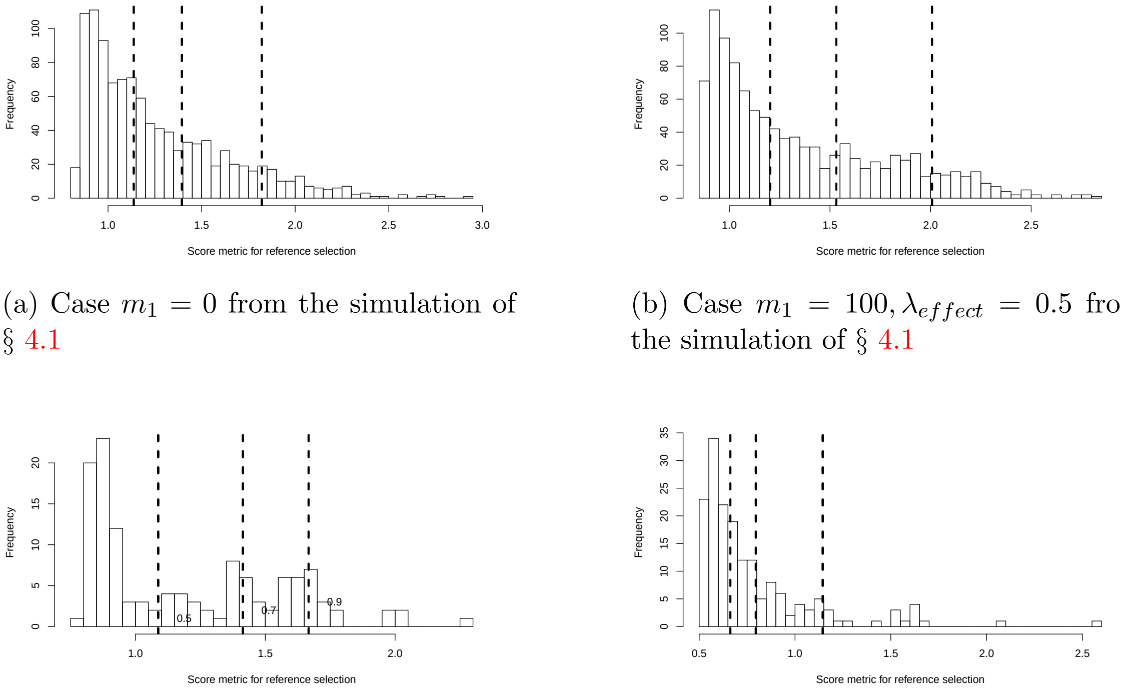
In this appendix we further examine the reference selection procedure suggested in § 3.1 and alternative reference selection procedures. In appendix B.1 we detail how the tuning parameter of S_{crit} was selected and examine the chance of a differentially abundant taxon to erroneously be inserted into the selected reference set. In appendix B.2 we examine the FDR of naive reference selection methods, e.g., picking the reference set of taxa at random. In appendix B.3 we propose a procedure for checking the validity of a reference set of tax.

B.1 Selecting S_{crit}

The data adaptive method for reference selection presented in § 3.1 has a single tuning parameter, S_{crit}. Taxa with a reference score below the parameter S_{crit} constitute the reference set. The value of S_{crit} was set to 1.3 in § 4 - § 5 after observing the distribution of reference scores in real and simulated data.

Figure 8 shows the distribution of reference scores for several real and simulated data sets. Values of S_{crit} in the range $[1.0, 1.4]$ select roughly 60-70% of taxa as a reference set. The remaining portion of taxa exhibit reference scores which are substantially higher than 1.3, and are not valid candidates to form the reference set B. Subplot (d) shows a relatively large portion of taxa with reference scores below 1.3. However, the comparison in subplot (d) is between the left and right Retroauricular creases. If any taxa are differentially abundant between the two sites, it is plausible to believe their number is small. Hence, the extreme values of the distribution in subplot D hint at $S_{crit} = 1.3$ as a plausible threshold as well.

Table 10 shows the mean number of differentially abundant taxa inserted into the reference set. This includes the scenarios of § 4.2-§ 4.3, where the signal present in differentially abundant taxa is much smaller than § 4.1. For most scenarios, no differentially abundant taxa have entered the reference set. We see that for some scenarios, some of the differentially abundant taxa have entered the reference set, however this occurred in a small fraction of the cases, with the mean number of differentially abundant included in the reference set being less than 1. Moreover, control of the FDR was not compromised in those settings.



(a) Case $m_1 = 0$ from the simulation of § 4.1

(b) Case $m_1 = 100, \lambda_{effect} = 0.5$ from the simulation of § 4.1

(c) Comparison of Hard Palate and Subgingival Plaque in appendix E

(d) Comparison of left and right Retroauricular Crease in appendix E

Figure 8: Histograms for reference scores computed in selected simulations and data analyses. Median and 0.7, 0.9 percentiles for reference scores are presented using vertical dashed lines in each subplot.

Table 10: Mean number of differentially abundant (DA) taxa inserted into the reference set, by simulation scenario. The standard error across 72 simulated data sets is given in brackets.

Simulation case	Mean number of DA taxa in B
§ 4.1, $m_1 = 0$	0 (0)
§ 4.1, $m_1 = 10, \lambda_{effect} = 0.5$	0 (0)
§ 4.1, $m_1 = 100, \lambda_{effect} = 0.5$	0.49 (0.05)
§ 4.1, $m_1 = 10, \lambda_{effect} = 1.0$	0 (0)
§ 4.1, $m_1 = 100, \lambda_{effect} = 1.0$	0.67 (0.05)
§ 4.1, $m_1 = 10, \lambda_{effect} = 1.5$	0 (0)
§ 4.1, $m_1 = 100, \lambda_{effect} = 1.5$	0.74 (0.06)
§ 4.1, $m_1 = 10, \lambda_{effect} = 2.0$	0 (0)
§ 4.1, $m_1 = 100, \lambda_{effect} = 2.0$	0.8 (0.06)
§ 4.1, $m_1 = 10, \lambda_{effect} = 2.5$	0 (0)
§ 4.1, $m_1 = 100, \lambda_{effect} = 2.5$	0.9 (0.07)
§ 4.1, $m_1 = 10, \lambda_{effect} = 3.0$	0 (0)
§ 4.1, $m_1 = 100, \lambda_{effect} = 3.0$	0.98 (0.06)
§ 4.2, Case 1	0 (0)
§ 4.2, Case 2	0 (0)
§ 4.2, Case 3	0 (0)
§ 4.2, Case 4	0 (0)
§ 4.2, Case 5	0 (0)
§ 4.2, Case 6	0 (0)
§ 4.2, Case 7	0 (0)
§ 4.2, Case 8	0 (0)
§ 4.2, Case 9	0 (0)
§ 4.2, Case 10	0 (0)
§ 4.3, Case 1	0.41 (0.07)
§ 4.3, Case 2	0.46 (0.08)
§ 4.3, Case 3	0.39 (0.07)
§ 4.3, Case 4	0.46 (0.08)

B.2 Examining naive approaches for reference selection

The reference selection method presented in § 3.1 aims to find a subset of $\mathcal{B}$, the full set of non-differentially abundant taxa. In this subsection, we show how selecting references at random, or while disregarding (2.2), can lead to lack of FDR control by the method presented in § 3.

We examine two possible alternative approaches for the reference selection method § 3.1. The first approach picks taxa at random for the reference set. The second approach picks the most abundant taxa as a reference set. Taxon abundance is computed by the total number of counts observed in a taxon across all subjects. In order to evaluate these approaches, we performed the following evaluation: for a given simulation setting, e.g., the 5th setting presented in § 4.2, 200 data sets were sampled. For each realized dataset, two reference set of taxa were selected using the approaches stated above. The method proposed in § 3 was used to detect differentially abundant taxa using the selected reference sets. The BH procedure was applied for FDR control at level $q = 0.1$.

Table 11 presents the estimated FDR of the two alternative reference selection methods by scenario. Both procedures are observed to select a large number of differentially abundant taxa into the reference set B. As a result, the procedure of § 3 lacks FDR control.

Table 11: Estimated FDR for naive reference selection methods, across selected scenarios. RAND stands for picking 50 taxa at random as B. ABUND stands for picking the 50 most abundant taxa as differentially abundant. Entries significantly higher than 0.1 are marked with a *.

Scenario	RAND	ABUND
§ 4.1, $m_1 = 100, \lambda_{effect} = 0.5$	0.19*	0.06
§ 4.1, $m_1 = 10, \lambda_{effect} = 2.5$	0.34*	1.00*
§ 4.2, Case 5	0.53*	0.17*
§ 4.3, Case 4	0.19*	0.09

B.3 Checking the validity of a reference set of taxa

The selected set of references B should be a part of the complete set of non-differentially abundant taxa $\mathcal{B}$. For a set $B \subset \mathcal{B}$, the relation given by (2.2) should hold:

$$H_0^B : \frac{(P_{b_1}, P_{b_2}, ..., P_{b_r})}{\sum_{k=1}^{r} P_{b_k}} \overset{\text{d}}{=} \frac{(Q_{b_1}, Q_{b_2}, ..., Q_{b_r})}{\sum_{k=1}^{r} Q_{b_k}}, \tag{B.1}$$

where $|B| = r$. Tests for (B.1) test the validity of the reference B: if B is comprised solely of non differentially abundant taxa, then (B.1) holds. A simple test for (B.1) is the following: (1) From each sample, select the sub-vector of reference taxa given by the indices $(b_1, b_2, ..., b_r)$ (2) Rarefy all sub-vectors of reference taxa across samples to uniform depth (3) Test for equality of distributions over the rarefied sub-vectors, using a multivariate test for equality of distributions, e.g., the tests of Anderson [2001] or Heller et al. [2012]. This procedure is assumption-free, and only requires selection of a distance metric for computing pairwise distances between samples. We will denote this procedure as a reference validation procedure, or RVP.

In order to examine the validity of this procedure, we conduct a simulation study. If the proposed RVP is valid, and no differentially abundant taxa have entered the reference set, the probability of the RVP to reject its null hypothesis should match the nominal Type I error rate used for testing. A higher probability to reject the RVP's null hypothesis indicates a problem in the proposed procedure. For a given simulation setting, e.g. case 1 from § 4.2, we sample 1000 datasets. For each sampled dataset, we select references according to the method presented in § 3.1 with $S_{crit} = 1.3$. We carry out the reference validation procedure suggested above, for all data realizations in which the reference set of taxa contains no differentially abundant taxa. As a multivariate test for equality of distributions, we use several options for each sampled data set: the HHG test of Heller et al. [2012], the PERMANOVA test of Anderson [2001], and the DISCO test of Rizzo et al. [2010]. As a distance metric to be used by the suggested tests, we use the L2 and L1 distances and the Bray-Curtis dissimilarity metric. Overall, 9 variations of the above procedure are considered. Multivariate tests are performed at level $\alpha = 0.1$. For this simulation study, we considered only the settings whose effect size was either the smallest or the largest in the respective subsection, specifically: simulation cases from § 4.1 with $\lambda_{effect} \in \{0.5, 3.0\}$, and simulation cases 1, 5, 6, 10 from § 4.2. The simulation settings of § 4.3 have a non-zero chance for selecting a reference set with a single taxon. For a reference set of taxa comprised of a single taxon, the RVP cannot be carried out. Hence, the settings of § 4.3 are excluded from this simulation study.

Table 12 describes the probability estimates of the RVP test to reject the null hypothesis, based on different multivariate tests, distance metrics and simulations cases. Most table entries are within 2 standard errors of the nominal error rate, with the exception of the probability estimates obtained for the HHG test in cases 5 and 10 of § 4.2, and the HHG based test with the L2 distance metric for the simulation setting with $m_1 = 100, \lambda_{effect} = 3.0$ in § 4.1. The inflated false-positive rate in some of the scenarios indicates that while the reference set of taxa was selected without considering the group labeling of observations, the counts vectors are not independent of the group labeling. This dependence is discovered when using a multivariate test of independence with a distance metric between count vectors. This inflation in T1E could be avoided if the data used in the RVP is independent of the data used for selecting the reference set of taxa. However, while counts vectors for reference taxa not exactly independent of the group labeling, we found empirically that the procedure of § 3 provides adequate FDR control in these settings.

Table 12: Probability to reject the null hypothesis in the RVP procedure proposed in appendix B.3. Column 1 describes the simulation setting. Columns 2-10 describe the chance to reject the null hypothesis according to multivariate test used (HHG, DISCO, PERMANOVA) and distance metric (L2 and L1 distances, and the Bray-Curtis dissimilarity metric). The maximal standard error for a table entry is 0.02. Testing is done at level $\alpha = 0.1$. Probability estimates significantly different from 0.1 are marked in grey.

Scenario	HHG			ENERGY			PERMANOVA		
	L2	L1	BC	L2	L1	BC	L2	L1	BC
§ 4.1, $m_1 = 10, \lambda_{effect} = 0.5$	0.09	0.10	0.11	0.11	0.11	0.12	0.11	0.11	0.11
§ 4.1, $m_1 = 100, \lambda_{effect} = 0.5$	0.10	0.11	0.10	0.11	0.11	0.12	0.12	0.13	0.12
§ 4.1, $m_1 = 10, \lambda_{effect} = 3.0$	0.10	0.10	0.09	0.10	0.09	0.09	0.10	0.09	0.09
§ 4.1, $m_1 = 100, \lambda_{effect} = 3.0$	0.16	0.10	0.12	0.13	0.10	0.10	0.10	0.11	0.11
§ 4.2, Case 1	0.08	0.09	0.09	0.04	0.04	0.04	0.03	0.03	0.03
§ 4.2, Case 5	0.20	0.19	0.19	0.12	0.11	0.12	0.11	0.12	0.11
§ 4.2, Case 6	0.08	0.09	0.08	0.04	0.04	0.04	0.04	0.03	0.04
§ 4.2, Case 10	0.20	0.20	0.20	0.11	0.10	0.10	0.11	0.11	0.10

C Simulations for control of type I error under the global null

In order to estimate the control over false discoveries in ANCOM under the global null, i.e. no differentially abundant taxa, we simulated datasets with taxon counts independently sampled from $pois\left(\mu\right)$ across m taxa. We considered two equal groups ,$n_X, n_Y \in \{50, 100\}$, $m \in \{50, 100\}$, and $\mu \in \{30, 60\}$.

ANCOM has several parameters used in its empirical decision rule. One of the parameters, `multcorr` specifies the type of multiple comparison correction used. ANCOM is highly sensitive to changes in this parameter. `multcorr` may receive one of three values, as follows:

- `multcorr = 3` : The matrix of P-values used, $P_{j,k}$ as defined in § 1.2, is not corrected for multiplicity. This is the default software parameter.
- `multcorr = 2` : The values of $P_{j,k}$ are substituted row-by-row, by their adjusted P-values given by the BH procedure.
- `multcorr = 1` : The values of $P_{j,k}$ are substituted by their adjusted P-values as given by the BH procedure. Correction for multiplicity is done across all table entries.

Testing and multiplicity correction was done at $\alpha = q = 0.05$. All other ANCOM parameters were set to default values. Table 13 gives the estimate of erroneously rejecting the global null hypothesis for ANCOM across the different settings and values of `multcorr`. The main result is that ANCOM fails to control the false positive rate across all scenarios under the global null, with parameters `multcorr` = 2 and `multcorr` = 3.

Table 13: Probability estimates of ANCOM to erroneously declare taxa as differentially abundant. Counts data generated as independent $pois\,(\mu)$, for m taxa, and equal sample sizes $n_X = n_Y$. Columns 4-6 give T1E estimates by value for parameter 'multcorr'. T1E level was set in software to $\alpha = 0.05$ Estimates are across 200 repetitions, maximum standard error is 0.035.

μ	m	n_X, n_Y	`multcorr = 1`	`multcorr = 2`	`multcorr = 3`
30	50	50	0.00	0.36	1.00
60	50	50	0.00	0.36	0.99
30	100	50	0.00	0.51	1.00
60	100	50	0.00	0.54	1.00
30	50	100	0.00	0.38	1.00
60	50	100	0.00	0.30	0.99
30	100	100	0.00	0.48	1.00
60	100	100	0.00	0.49	1.00

D Testing for association with a continuous or multivariate phenotype

We reformulate the methodology presented in § 2 - § 3 for a general vector of phenotypes measured for the different samples and other study designs. In the general setting, we denote the observed counts for sample i as $\vec{X}_i$, $i = 1, ..., n$ indexing all available samples. Let $X_{i,j}$ denote the number of counts observed taxon j in sample i. Let $\vec{z}_i$ denote the In addition to the microbial counts, we assume a p-dimensional vector of phenotypes measured for sample i.

Similar to (2.1), we assume each observed sample is a realization of a multinomial random variable whose vector of relative proportions depends on the measured phenotypes:

$$\vec{X}_i|\vec{P}_i, N_i \sim multinom\left(N_i, \vec{P}_i\right), \qquad \vec{P}_i|\vec{z}_i \sim \mathcal{P}\left(\vec{z}_i\right) \quad ,0 \leq P_{i,j}, \sum_{j=1}^{m} P_{i,j} = 1,$$

where $\mathcal{P}\left(\vec{z}_i\right)$ indicates that the distribution of $\vec{P}_i$ depends on $\vec{z}_i$. Furthermore, we assume (2.2) holds. The distribution of the ratios of relative abundances between non differentially abundant taxa is not affected by the measured phenotypes, i.e., it holds that for any two possible vectors of phenotypes, $\vec{a}_1$ and $\vec{a}_2$, and for any $(v_1, v_2, ..., v_s) \subseteq \mathcal{B}$, $s \in \{2, ..., |\mathcal{B}|\}$, such that $\sum_{k=1}^{s} P_{i,v_k} > 0$ with probability 1 :

$$\frac{(P_{i,v_1}, P_{i,v_2}, ..., P_{i,v_s})}{\sum_{k=1}^{s} P_{i,v_k}} | \left[\vec{z}_i = \vec{a}_1\right] \overset{\mathrm{d}}{=} \frac{(P_{i,v_1}, P_{i,v_2}, ..., P_{i,v_s})}{\sum_{k=1}^{s} P_{i,v_k}} | \left[\vec{z}_i = \vec{a}_2\right]. \tag{D.1}$$

Similar to (2.3), the sum of population relative frequency of all non differentially abundant taxa is assumed to be non zero with probability 1.

The test for differential abundance requires a set of reference taxa $B = (b_1, b_2, ..., b_r) \subset \mathcal{B}$. Let λ_j denote the minimal number of reads available in taxa $j \cup B$ across samples. Given B and λ_j, the test of (D.1) is as follows. Let $\tilde{X}_{i,j}$ denote the rarefied counts as in (3.1). Next, test the independence hypothesis:

$$H_0^{(j)} : \tilde{X}_j \perp\!\!\!\perp \vec{Z}, \tag{D.2}$$

where $\tilde{X}_j$ is the subsampled number of counts, and $\vec{Z}$ is the phenotype vector. For example, if $\vec{Z}$ is a univariate continuous variable, a test based on the Spearman correlation coefficient [Spearman, 1904] can be used to test (D.2). For a multivariate $\vec{Z}$, non-parametric tests of independence such as Heller et al. [2012] and Heller and Heller [2016] can be used. The Kruskal-Wallis test [Kruskal and Wallis, 1952] can be used for associating taxa to a univariate phenotype with a $k \geq 2$ categories.

When testing taxa for differential abundance in a study with block design, an additional categorical value denoted by d_i is available for each sample. For a study with D blocks, d_i receives values in $1, 2, ..., D$. Hypothesis (D.2) is extended to conditional independence given the block identity of the sample:

$$H_0^{(j)} : \tilde{X}_j \perp\!\!\!\perp \vec{Z}|D, \tag{D.3}$$

A paired design is represented in this formulation by $n/2$ blocks, with each block having exactly two observations with differing phenotype vectors. The Wilcoxon sign rank test [Wilcoxon, 1945] or the Friedman test [Friedman, 1937], can be used to test (D.3)

E Comparing adjacent body sites in the Human Microbiome Project

The Human Microbiome Project [Gevers et al., 2012] is a joint collaboration aimed at studying the behavior of microbial ecologies across the human body. 16S profiles of 300 subjects were sampled at 15-18 body sites, with sampling locations being in the oral cavity, skin sites across the body, airways, vagina and fecal samples. We wish to analyze the differences in microbiome composition at adjacent body sites. The OTU table and taxonomy available from by the link given in Kumar et al. [2018] contains 4788 samples and 45383 OTUs. Since OTU picking was done for all body sites combined, many OTUs are prevalent at a small portion of body sites. See Kumar et al. [2018] for a comprehensive comparison of normalization approaches with this dataset.

OTUs in the data are associated with a taxon in the known common taxonomy of Kingdom-Phylum-Class-Order-Family-Genus-Species. Some OTUs are associated with a known species of bacteria while others are associated with a high level taxon such as a Genera or Family. Moreover, several OTUs may be linked to the same taxonomic affiliation as a single species may have several known 16S sequences.

To reduce the dimensionality of the data, OTUs counts were aggregated to the Genus level. All OTUs with the same Genus affinity were aggregated to the same vector index. OTUs whose taxonomic affiliation was higher than Genus, were aggregated by their closest affinity, i.e. all OTUs which had Family identification avaiable at most and were identified with the same Family were aggregated to a taxon representative of the Family. 664 Genera (or above) taxa were present in the data after aggregation.

For each pair of body sites, each subject had two samples, one in each body site. In order to avoid across sample dependencies only one of the samples per subject, selected at random, was considered for analysis.

Genera which appeared in less than 2.5% of the subjects were removed. Some samples contained an irregular low number of reads due to technical faults. Therefore, at each pairwise comparison of body sites, the 10% of samples with the lowest number of reads (in sample) were removed. An alternative way to filter technical faults would have been to set a minimal number of counts required of a valid sample. However, sampling locations exhibit different sequencing depths, and that would require a specific cutoff value for technical faults for each body sampling location.

Table 14 describes the number of discoveries in each pairwise comparison of body sites. In general, samples taken from skin sites and the vagina have reads concentrated at a smaller number of OTUs, compared with samples taken from the oral cavity. This can be seen by the number of taxa considered in the comparisons inside the oral cavity compared with comparisons between skin sites. As observed taxa are more abundant in the oral cavity, more differentially abundant taxa are observed in pairwise comparisons by all methods. W-CSS, a method for marginal inference, has more discoveries compared to ANCOM and DACOMP, across most pairwise comparisons. This is not suprising since W-CSS does not control for compositionality. When comparing ANCOM and DACOMP across the oral cavity, many of the discoveries of DACOMP are shared by ANCOM. Most discoveries of DACOMP are also shared with ALDEx2-t.

Table 14: Pairwise comparison of adjacent body sites in the Human Microbiome Project. For each pair of body sites (columns 1-2), the number of taxa (genera or above) considered for differential abundance between the two sites (coloumn 3), the number of differentially abundant taxa discovered by ANCOM, W-CSS,ALDEx2-t and DACOMP (columns 4-7), the number of discoveries shared by ANCOM and DACOMP (column 8) and the size of reference set (column 9). The BH procedure was applied at level $q = 0.1$. For DACOMP S_{crit} was set to 1.3.

Site 1	Site 2	NR.Taxa	ANCOM	W-CSS	ALDEx2-t	DACOMP	Shared	$\|B\|$
Saliva	Tongue_dorsum	111	36	83	26	24	18	67
Saliva	Hard_palate	147	30	46	25	29	24	92
Saliva	Buccal_mucosa	145	49	63	31	31	28	106
Saliva	Attached_Keratinized_gingiva	138	67	93	47	38	38	93
Saliva	Palatine_Tonsils	146	39	64	28	41	30	86
Saliva	Throat	156	25	47	20	16	15	126
Saliva	Supragingival_plaque	123	39	95	39	25	22	84
Saliva	Subgingival_plaque	133	44	86	37	32	27	88
Tongue_dorsum	Hard_palate	106	29	54	20	36	18	65
Tongue_dorsum	Buccal_mucosa	102	52	64	36	29	26	70
Tongue_dorsum	Attached_Keratinized_gingiva	91	54	59	38	29	27	60
Tongue_dorsum	Palatine_Tonsils	98	23	40	17	25	15	53
Tongue_dorsum	Throat	110	16	54	7	14	10	67
Tongue_dorsum	Supragingival_plaque	101	60	68	43	34	32	61
Tongue_dorsum	Subgingival_plaque	102	67	76	50	40	36	55
Hard_palate	Buccal_mucosa	142	38	53	28	39	32	87
Hard_palate	Attached_Keratinized_gingiva	137	51	74	33	46	42	81
Hard_palate	Palatine_Tonsils	131	29	52	26	35	18	84
Hard_palate	Throat	149	37	36	16	20	19	122
Hard_palate	Supragingival_plaque	119	59	87	47	34	27	83
Hard_palate	Subgingival_plaque	126	55	83	45	36	35	80
Buccal_mucosa	Attached_Keratinized_gingiva	125	36	60	16	32	31	76
Buccal_mucosa	Palatine_Tonsils	129	48	60	34	31	26	91
Buccal_mucosa	Throat	146	49	58	34	25	22	114
Buccal_mucosa	Supragingival_plaque	115	40	73	28	30	24	84
Buccal_mucosa	Subgingival_plaque	127	42	72	33	32	30	87
Attached_Keratinized_gingiva	Palatine_Tonsils	117	51	56	34	30	27	79
Attached_Keratinized_gingiva	Throat	143	48	56	33	28	27	112
Attached_Keratinized_gingiva	Supragingival_plaque	101	47	64	31	32	27	66
Attached_Keratinized_gingiva	Subgingival_plaque	116	50	66	37	31	30	81
Palatine_Tonsils	Throat	145	17	19	3	13	10	117
Palatine_Tonsils	Supragingival_plaque	106	50	67	40	30	26	65
Palatine_Tonsils	Subgingival_plaque	120	50	62	42	38	36	72
Throat	Supragingival_plaque	150	57	91	39	36	31	110
Throat	Subgingival_plaque	144	69	108	50	42	38	98
Supragingival_plaque	Subgingival_plaque	103	30	44	19	17	12	64
Right_Antecubital_fossa	Left_Retroauricular_crease	244	5	50	2	17	1	244
Right_Antecubital_fossa	Right_Retroauricular_crease	190	6	44	2	0	0	188
Right_Antecubital_fossa	Left_Antecubital_fossa	286	2	0	0	0	0	269
Right_Antecubital_fossa	Anterior_nares	209	19	50	7	10	7	190
Left_Retroauricular_crease	Right_Retroauricular_crease	172	1	0	0	0	0	166
Left_Retroauricular_crease	Left_Antecubital_fossa	198	5	81	2	1	0	196
Left_Retroauricular_crease	Anterior_nares	202	8	11	7	14	7	183
Right_Retroauricular_crease	Left_Antecubital_fossa	200	7	54	2	1	1	198
Right_Retroauricular_crease	Anterior_nares	200	8	15	9	15	6	180
Left_Antecubital_fossa	Anterior_nares	209	11	54	8	7	6	189
Vaginal_introitus	Posterior_fornix	120	5	9	2	2	0	105
Vaginal_introitus	Mid_vagina	129	2	1	0	0	0	106
Posterior_fornix	Mid_vagina	96	5	5	0	0	0	89