

REVIEW ARTICLE OPEN



Analysis of microbial compositions: a review of normalization and differential abundance analysis

Huang Lin ¹ and Shyamal Das Peddada ^{1,2}✉

Increasingly, researchers are discovering associations between microbiome and a wide range of human diseases such as obesity, inflammatory bowel diseases, HIV, and so on. The first step towards microbiome wide association studies is the characterization of the composition of human microbiome under different conditions. Determination of differentially abundant microbes between two or more environments, known as differential abundance (DA) analysis, is a challenging and an important problem that has received considerable interest during the past decade. It is well documented in the literature that the observed microbiome data (OTU/SV table) are relative abundances with an excess of zeros. Since relative abundances sum to a constant, these data are necessarily compositional. In this article we review some recent methods for DA analysis and describe their strengths and weaknesses.

npj Biofilms and Microbiomes (2020)6:60; <https://doi.org/10.1038/s41522-020-00160-w>

INTRODUCTION

Human oral and gut microbiome are estimated to have 45.6 million genes, which is ~2000-fold more genes than human genes¹, therefore the microbiome is sometimes referred to as the “second genome”, or another “organ” of human body^{2–4}. Hence it is not surprising that numerous diseases such as obesity⁵, inflammatory bowel diseases⁶ and HIV⁷ are associated or even caused by changes in the microbial ecosystem. For these reasons, understanding changes in the composition of microbiome under different conditions is important for studying human diseases.

For clarity, we begin by defining some important terms used in this paper and in the literature. The phrase absolute abundance of a taxon refers to the unobservable actual abundance of a taxon in a unit volume of an ecosystem, such as the gut. Accordingly, one could define absolute relative abundance of a taxon in a unit volume of an ecosystem as the ratio of the absolute abundance of the taxon to the total absolute abundance of all taxa in a unit volume of an ecosystem.

In practice, however, neither absolute abundance nor absolute relative abundance of a taxon in a unit volume of an ecosystem can be easily determined⁸. Although these parameters are typically not observable, the next-generation sequencing (NGS) technologies such as the 16S rRNA gene sequencing yield useful data for describing microbial compositions in an ecosystem. Following initial quality assessment/control steps, such as primer (s) removal, demultiplexing and quality filtering, the 16S amplicon sequences are either clustered into Operational Taxonomic Units (OTUs) representing the common working definition of bacterial species⁹ by OTU picking algorithms (e.g. UPARSE¹⁰), or grouped into Sequence Variants (SVs) using denoising algorithms (e.g. DADA2¹¹ and Deblur¹²). After the construction of OTU or SV, these observed counts are typically organized into a large matrix referred to as the feature table. Some researchers or software packages such as QIIME2¹³ represent samples by columns and features (OTUs or SVs) by rows, but this representation is not universal. The observed counts of features (OTUs or SVs) represent observed abundances of taxa in the sample. Since abundances in a feature table represent only relative information regarding each

taxa^{8,14–18}, these are compositional data and thus reside inside a simplex¹⁹. Some researchers refer to these frequencies as relative abundances of taxa in a sample. However, in our terminology, relative abundance of a taxon in the sample is the fraction of the taxon observed in the feature table relative to the sum of all observed taxa corresponding to the sample in the feature table. Thus, by our terminology, the relative abundances sum to 1. In a recent paper by Lin and Peddada²⁰, the authors refer to abundance of taxa in a feature table as “observed absolute abundance”, which is a confusing terminology and should be avoided. Instead they should have referred to it as “observed abundance”. Various terms used in this paper are summarized in Table 1. The notations described in statistical methods are summarized in Table 2.

We define a taxon to be differentially abundant between two ecosystems if its mean absolute abundance is different between two ecosystems. It is important to distinguish between absolute and relative abundances of taxa in a unit volume of an ecosystem. The choice of parameter for statistical analysis is important and needs to be clearly stated. Often researchers are interested in identifying taxa that are different in mean absolute abundance per unit volume between two or more ecosystems⁵. The mean absolute abundance may not be the only criterion of interest. Researchers may consider other criteria such as differential ranking⁸. Furthermore, there are instances such as niche apportionment, where researchers are interested in identifying taxa that are different in mean absolute relative abundance per unit volume between two or more ecosystems. Thus, the choice of statistical parameter depends upon the scientific question of interest.

For each taxon i within sample j , the sampling fraction is the ratio of the expected abundance of taxon i within the j th sample to its absolute abundance in a unit volume of an ecosystem (e.g. gut) where the sample was derived from. The sampling fraction is assumed to be constant for all taxa within the j th sample. Thus the sampling fraction for the j th sample is given by the following expression involving the conditional expectation of the observed abundance O_{ij} given the unobservable absolute abundance A_{ij} .

¹Department of Biostatistics, University of Pittsburgh, 130 De Soto Street, Pittsburgh, PA 15261, USA. ²Present address: Biostatistics and Bioinformatics Branch, NICHD, NIH, Bethesda, MD, USA. ✉email: SDP47@pitt.edu

Table 1. Definitions of key terminologies.

Term	Definition
Microbiota	Community of microscopic organisms.
Microbiome	Genes associated with the microbiota.
Amplicon	Product of PCR amplification.
High-throughput Sequencing	DNA sequencing approach that produces large amounts of sequence data rapidly at low cost.
OTU	Operational taxonomic unit: Group of DNA sequences with 97% similarity.
SV	Sequence variant: Individual DNA sequences recovered from a high-throughput marker gene analysis following the removal of spurious sequences generated during PCR amplification and sequencing.
Absolute abundance	Unobservable actual abundance of a taxon in a unit volume of an ecosystem.
Observed abundance	Observed counts of features (OTUs or SVs) in the feature table.
Relative abundance	The fraction of the taxon observed in the feature table relative to the sum of all taxa in the sample. It is between 0 and 1.
Feature Table	A matrix summarizing observed microbial abundances in the sample. Usually, columns represent samples and rows stand for OTUs or SVs.
Library Size	The total number of observed abundances for all taxa in a sample.
Microbial Load	The total number of (unobserved) absolute abundances for all taxa in a unit volume of an ecosystem.

Table 2. Summary of notations.

Notation	Description
m	Total number of taxa.
n	Total number of samples.
p	Total number of covariates.
i	Taxon index, $i = 1, 2, \dots, m$.
j	Sample index, $j = 1, 2, \dots, n$.
k	Index of covariates, $k = 1, 2, \dots, p$.
x_j	Covariates of interest for the j th sample. $x_j = (x_{j1}, \dots, x_{jp})^T$.
A_{ij}^a	Unobserved absolute abundance of i th taxon in a unit volume of ecosystem of j th sample.
A_j^a	Microbial load in a unit volume of ecosystem of j th sample. $A_j = \sum_{i=1}^m A_{ij}$.
y_{ij}^a	Unobserved absolute relative abundance of i th taxon in a unit volume of ecosystem of j th sample.
O_{ij}^a	Observed abundance of i th taxon in a random specimen taken from a unit volume of ecosystem of j th sample.
O_j^a	Library size of a random specimen taken from a unit volume of ecosystem of j th sample. $O_j = \sum_{i=1}^m O_{ij}$.
r_{ij}^a	Observed relative abundance of i th taxon in a random specimen taken from a unit volume of ecosystem of j th sample.
c_j^b	For the j th sample, c_j represents the proportion of its ecosystem (unobserved absolute abundance) in a random specimen (observed abundance), thus $c_j = \frac{E(O_{ij} A_{ij})}{A_{ij}}$. We shall refer to this constant as "sampling fraction".
y_{ij}^a	$\log(O_{ij})$.
d_j^b	Represents the effect of the scaling parameter c_j in log – scale

^aRandom variable.

^bParameter.

Definition 0.1 (Sampling fraction).

$$c_j = \frac{E(O_{ij}|A_{ij})}{A_{ij}}, \quad (1)$$

where

- (1) O_{ij} is the observed abundance of i th taxon in j th sample,
- (2) A_{ij} is the unobserved absolute abundance of i th taxon in the ecosystem of j th sample,
- (3) c_j is the sample-specific sampling fraction.

The problem underlying the differential abundance (DA) analysis of microbiome data is that while O_{ij} is known, c_j is unknown and can vary drastically from sample to sample. Consequently, the observed abundances are not comparable between samples. The goal of DA analysis described in this paper is to identify taxa whose mean absolute abundances, per unit

volume, of an ecosystem are significantly different with changes in the covariate of interest (e.g. study groups).

Similar to the toy example in ref. ²⁰, Fig. 1 is a toy example consisting of ecosystems of three subjects A, B, and C with each having two taxa, the blue and red taxa varieties. A false negative may occur when comparing the ecosystems of A and B. Clearly, the true absolute abundance of each taxon is 50% more in subject B's ecosystem as compared to subject A's. However, they each have the same library size (4 each) in their respective samples (e.g. stool samples). Without considering the differential sampling fractions, one would falsely conclude that none of the taxa are differentially abundant in the two ecosystems. This erroneous conclusion would be avoided if one recognizes that we have a larger sampling fraction in the sample obtained from A's ecosystem than from B's ($\frac{1}{2}$ vs. $\frac{1}{3}$). Similarly, we get a false positive result when comparing ecosystems of A and C. In their

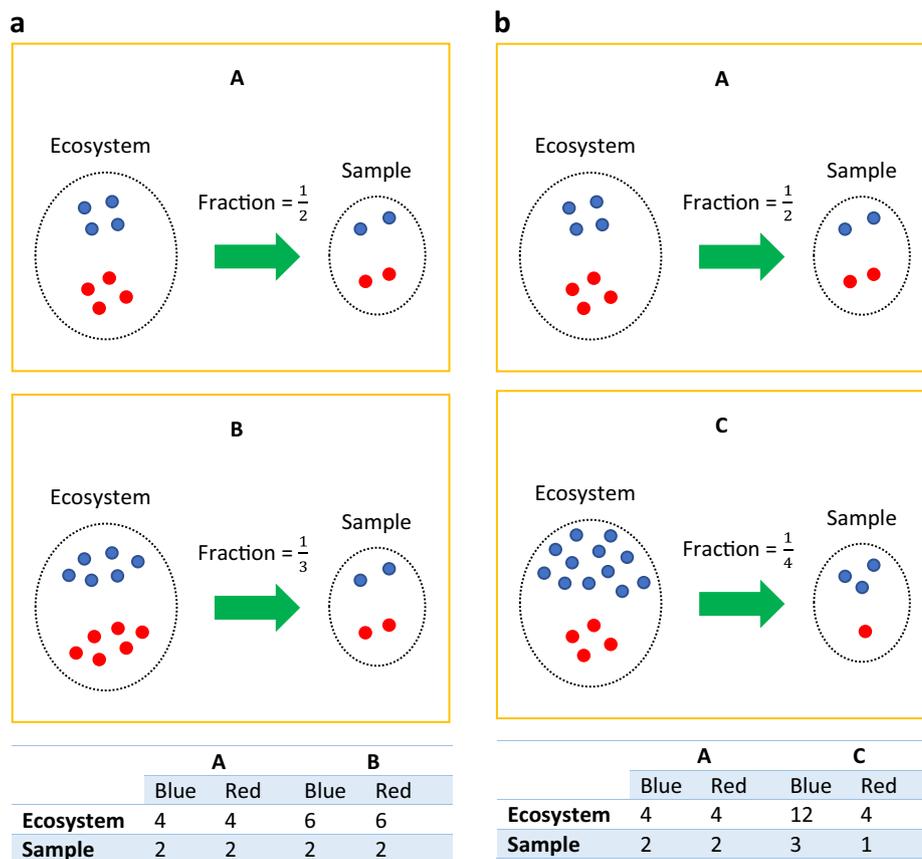


Fig. 1 Microbiome data is represented by relative abundances, thus differential abundance analysis should account for the bias introduced by across-sample variations in sampling fractions. Sampling fraction is defined as the ratio of expected abundance in a sample to the corresponding absolute abundance in the ecosystem, which could be empirically estimated by the ratio of library size to the microbial load. **a** Differences in sampling fractions introduce false negatives. In this toy example, the microbial load for subject A in a unit volume of the ecosystem (e.g. a unit volume of gut) is 8 (4 blue + 4 red), while for subject B it is 12 (6 blue + 6 red). However, the samples taken from subject A and B have the same library size 4 (2 blue + 2 red), same observed abundance as well as the same relative abundance of blue and red taxa. Thus, one may mistakenly conclude that the blue and red taxa are not differentially abundant between two ecosystems, which is not the case in the two ecosystems. This false negative conclusion is caused by differences in the sampling fractions in the two samples. The sampling fraction in sample A is 1/2 and for B it is 1/3. **b** Differences in sampling fractions introduces false positives. Consider another subject C, who has the microbial load of 16 (12 blue + 4 red) in a unit volume of ecosystem. Given the same library size in sample C (3 blue + 1 red) as sample A, one may mistakenly conclude that both blue and red are differentially abundant between ecosystems A and C, while in fact, only the blue taxon is differentially abundant. Thus a normalization method must account for differences in sampling fractions to avoid such erroneous conclusions.

ecosystems, blue is more abundant in C than in A (12 vs. 4), and both have the same amounts of red taxa (4 vs. 4). However, given that samples from A and C have same the library size, one may mistakenly conclude that both blue (2 vs. 3) and red taxa (2 vs. 1) are differentially abundant between A and C.

An important characteristic of a feature table is that it is typically sparse, sometimes as many as ~90% are zero entries²¹, which creates a challenge for analyzing rare taxa. A quick and simple strategy to deal with excess zeros is to add a small positive constant (e.g. 1) called pseudo-count^{14,22} to each cell of the feature table. The addition of a pseudo-count becomes necessary when using methods of analysis that require log transformation of the observed counts. Even though adding a pseudo-count is simple and widely used, the choice of the pseudo-count is ad hoc. Studies have shown that differential abundance or clustering results could be sensitive to the choice of pseudo count^{23,24}. Although different values of pseudo counts have been discussed in the literature^{23–26}, to the best of our knowledge, there is no consensus on how to choose the optimal value. Other strategies involve modeling zero counts by some probability models^{21,27}. However, these methods may not be valid if the underlying

assumptions do not hold. Instead of modeling zeros by parametric distributions, ANCOM-II²⁸ attempts to provide a general framework to classify and identify zeros into three different types, which includes outlier zeros caused by some extraneous reasons such as the wrong data entry, structural zeros because of the nature of the study groups, i.e. some bacteria are not expected to belong to certain environments (e.g. a desert) but in others (e.g. a rain forest), and sampling zeros owing to insufficient library size. In our opinion, the zero counts problem is still an open problem and requires further investigation.

NORMALIZATION METHODS

As we described intuitively in the introduction, an important obstacle for performing DA analysis is the unknown sampling fraction corresponding to each sample. It is critical to normalize the data to eliminate any bias due to differences in the sampling fractions. Thus, the primary objective of normalization is to transform the observed data so that expected differences in the mean absolute abundances between two ecosystems is not confounded by the differences in the sampling fractions. Failure to normalize the data will result in a

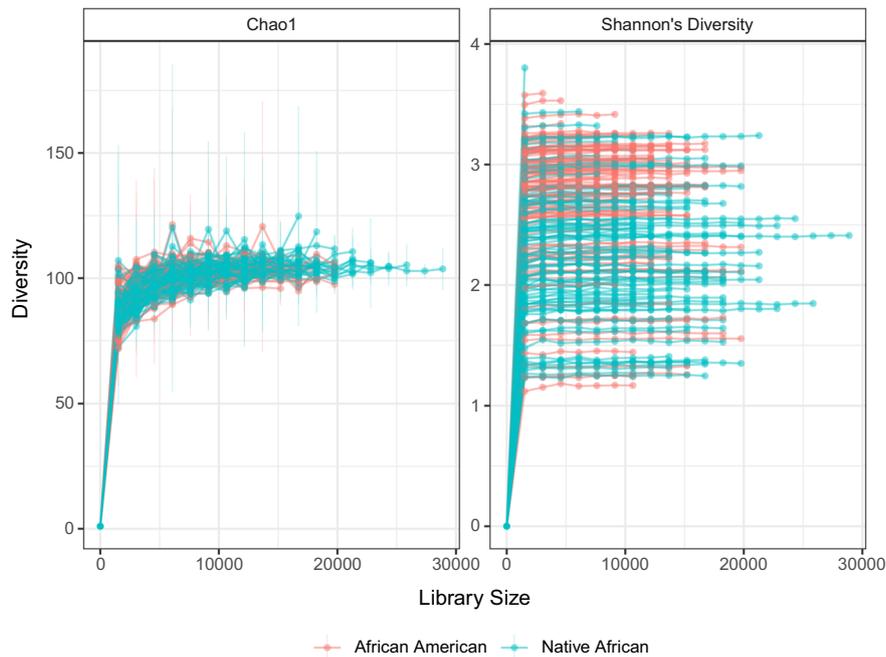


Fig. 2 Rarefaction curves using the diet swap data⁶⁸ at the genus level. The number of genera is 130, and the sample size is 222 (African American = 123, Native African = 99). x axis denotes the library size, and y axis represents the corresponding alpha diversity. Data are presented as mean values \pm standard error (SE). It shows that regardless of the choice of diversity measures, as the increase of library size, the rarefaction curve starts to “level out” suggesting that the diversity of the samples has been fully observed.

systematic bias that increases the false discovery rate (FDR) and also possible loss of power in some cases.

Rarefying

A traditional microbiome analysis workflow often involves rarefying^{29–31}, or subsampling to a given depth, a practice in the field of ecology long before its use in microbiome surveys³². Samples are rarefied to deal with differences in library sizes. Note that the terms rarefying and rarefaction are used interchangeably in microbiome literature³³. Rarefying was first recommended for microbiome data to deal with rare taxa³⁴, which impact some measures of alpha and beta diversities³³. Generally, the rarefying process includes the following steps:

- (1) Determine the minimum library size (O_{\min}). Samples with library sizes smaller than O_{\min} will be discarded,
- (2) Subsample taxa without replacement so that all samples have the same library size O_{\min} .

One way to select the minimum library size is to create rarefaction curves³⁵. Rarefaction curves represent diversity as a function of library size (Fig. 2). If lines of the plot appear to “level out” (i.e., approach a slope of zero) at certain library size along the x -axis, it indicates the diversity of the samples has been fully observed; otherwise, increasing the minimum library size would result in additional features. Originally, rarefaction curves were based on alpha diversities^{35,36}. However, lately researchers have considered beta diversities^{37,38} as well. Although rarefying is well established and widely used in practice, in recent years there has been some discussion on the effects of rarefying on statistical tests for differential abundance analysis^{33,39,40}. Some concerns discussed in the literature include:

- (1) The omission of available valid data,
- (2) The introduction of artificial uncertainty in the subsampling step,
- (3) The arbitrary selection of the minimum library size,
- (4) Challenges in estimating over-dispersion parameter.

Scaling

Scaling is another popular method used for normalizing microbiome data. The basic idea is to divide the observed abundance in the feature table by a “scaling factor” or “normalization factor” to eliminate biases resulting from unequal sampling fractions. More precisely, scaling is defined as follows.

Definition 0.2 (Scaling microbiome data).

$$\tilde{O}_{ij} = \frac{O_{ij}}{s_j}, \quad (2)$$

where

- (1) \tilde{O}_{ij} is the normalized observed abundance for taxon i within sample j ,
- (2) s_j is the scaling/normalization factor for sample j .

Comparing with the definition of sampling fraction (Eq. (1)), it is clear that an ideal scaling method should have scaling factor close to the unknown sampling fraction c_j , i.e. $s_j \approx c_j$ or is approximately proportional to c_j , i.e. $s_j \approx c_j \times c$ for all j , where c is a constant.

Some commonly used normalization methods include Cumulative-Sum Scaling (CSS) implemented in metagenome-Seq²¹, Median (MED) in DESeq2⁴¹, Upper Quartile (UQ)⁴² and Trimmed Mean of M-values (TMM)⁴³ in edgeR⁴⁴ and Wrench⁴⁵, and Total-Sum Scaling (TSS) which simply transforms the abundance table (feature table) into relative abundance table, i.e. scale by each sample’s library size. The authors of the user manual of edgeR⁴⁶ state that to deal with the “RNA composition” effect, one should multiply the normalization factors with the corresponding library size to account for “effective library size”. Hence, Lin and Peddada²⁰ also considered modified versions of UQ and TMM, denoted by “ELib-UQ” (Effective library size using UQ) and “ELib-TMM” (Effective library size using TMM) in their simulation studies. Since the literature is often not explicit regarding the mathematical formulas used by various methods, we provide some useful formulas in Table 3.

TSS is known to have a bias in differential abundance estimates^{33,39,42,47} since a few preferentially sampled

Table 3. Summary of different normalization methods.

Method	Sampling fraction estimate
ANCOM-BC	$\log(\hat{\epsilon}_j^{\text{ANCOM-BC}}) = \frac{1}{m} \sum_{i=1}^m (y_{ij} - x_j^T \hat{\beta}_i)$
CSS	$\hat{c}_j^{\text{CSS}} = \frac{s_j^{j+1}}{N}$
MED	$\hat{c}_j^{\text{MED}} = \text{median}_{i:O_{ij}^i > 0} \frac{O_{ij}}{O_j^i}$
UQ	$\hat{c}_j^{\text{UQ}} = \text{UQ}_{i:O_{ij} > 0} \left(\frac{O_{ij}}{O_j^i} \right)$
TMM	$\log_2(\hat{c}_j^{\text{TMM}}) = \frac{\sum_{i \in G^*} w_{ij} M_{ij}}{\sum_{i \in G^*} w_{ij}}$
Elib-UQ	$\hat{c}_j^{\text{Elib-UQ}} = O_j \hat{c}_j^{\text{UQ}}$
Elib-TMM	$\hat{c}_j^{\text{Elib-TMM}} = O_j \hat{c}_j^{\text{TMM}}$
Wrench	$\hat{c}_j^{\text{Wrench}} = \frac{1}{m} \sum_{i=1}^m b_{ij} \frac{r_{ij}}{\bar{r}_i}$
TSS	$\hat{c}_j^{\text{TSS}} = O_j$

$\hat{\beta}_i$ is obtained from ANCOM-BC algorithm.
 N = an approximately chosen normalization constant.
 $s_j^l = \sum_{i:O_{ij} \leq q_j^l} O_{ij}$.
 $q_j^l = l^{\text{th}}$ quantile of sample j .
 $O_j^i = \left(\prod_{j=1}^n O_{ij} \right)^{\frac{1}{n}}$.
 $\text{UQ}(X)$ denotes the upper quartile of X .
 $M_{ij} = \log_2 \left(\frac{O_{ij}}{O_j^i} \right) - \log_2 \left(\frac{O_{ij'}}{O_j'^i} \right)$, where j' is the reference sample.
 $w_{ij} = \frac{O_{ij} - O_{ij'}}{O_j^i - O_j'^i} + \frac{O_{ij'} - O_{ij}}{O_j'^i - O_j^i}$, where j' is the reference sample.
 G^* represents a set of taxa that were not considered as extreme data for fold-change (M values) and average intensity (A values). Refer to Robinson and Oshlack⁴³ for details.
 b_{ij} represents the taxon-specific weight. Refer to Kumar et al.⁴⁵ for details.

measurements (e.g. taxa, genes) will have an undue influence on the relative abundance data. Change in the abundance of a single taxon can alter the relative abundances of all taxa. Generally, the FDR generated from TSS-based analyses is unacceptably large. The CSS²¹ in metagenomeSeq modifies TSS in a sample-specific manner to reduce biases resulting from preferentially sampled taxa. CSS assumes that observed abundances of samples should be roughly independent and identically distributed up to a specific quantile l . Thus, instead of normalizing each sample by its library size (which is also known as total sum), CSS selects the scaling factor to be the cumulative sum of observed abundances for each sample up to the l th quantile. This quantile is determined adaptively in a data-driven way, which relies on the change point of the distribution of cumulative sum switching from stability to instability. The Median normalization (MED) method used in DESeq2⁴¹ assumes that the taxon of median absolute abundance is not differentially abundant. Although it may be a valid assumption in gene expression studies where a large proportion of genes are not differentially expressed, it may not be a valid assumption in microbiome studies. Depending upon the application, a very large proportion of taxa may be differentially abundant between two or more study groups, especially when the data are analyzed at higher taxonomic classification levels (e.g. phylum, order, etc.). The Upper Quartile normalization (UQ) and the TMM used in edgeR have similar issues as MED in DESeq2. UQ assumes that the upper quartile of the observed abundances for each library is able to capture the invariant segment of the count distribution. However, choosing the most effective quantile is nontrivial^{21,42,44,47–49}. Similar to MED, TMM is based on the hypothesis that most taxa are not differentially abundant. The scaling factor is calculated using a weighted trimmed mean of log abundance ratios by first trimming (by default) the taxa belong to upper and lower 30% M values (taxon-wise log-fold-change) or 5%

A values (abundance level). Wrench⁴⁵ assumes that the observed abundances are from a hurdle Log-Gaussian distribution. A robust location estimate of the Gaussian distribution leads to the desired scaling factor for each sample. However, Wrench currently implements strategies for categorical variable only, and the estimated scaling factor is essentially the average of ratios of relative abundances across taxa, which implicitly requires that a large proportion of taxa do not change across study groups, or the effect sizes of differentially abundant taxa are not too large.

One must exercise caution when using scaling methods. Most importantly, a scaling method is likely to overestimate or underestimate the fraction of zero counts depending on the corresponding library size of each sample^{49,50}. This problem becomes more obvious for microbiome data since its feature table is typically sparse.

Recently a new method called Analysis of Compositions of Microbiome with Bias Correction (ANCOM-BC) was introduced by Lin and Peddada²⁰ to address the problem of unequal sampling fractions. ANCOM-BC assumes that the observed abundance in a feature table is, in expectation, proportional to the unobservable absolute abundance of a taxon in a unit volume of the ecosystem. This proportion is defined as the sampling fraction and is allowed to vary from sample to sample. ANCOM-BC accounts for sampling fraction by introducing a sample-specific offset term in a linear regression model that is estimated from the observed abundance data. The offset term serves as the bias correction. Statistical properties of this approach have also been discussed in²⁰.

Extensive simulation studies using Poisson-Gamma model as well as some based on real data, were performed in²⁰ to evaluate the performance of various normalization methods. Results reported in Fig. 3 of this article are similar to those provided in²⁰, but in the present simulation study we have three groups, which are denoted by G_1 , G_2 , and G_3 (see Supplementary Information for simulation settings). We compared all normalization methods using the centered residuals between true and estimated sampling fractions in log scale.

Definition 0.3 (Centered Residual).

$$h_j = d_j - t_j - \frac{1}{n} \sum_{j=1}^n (d_j - t_j) \quad (3)$$

where

- (1) d_j (see Table 2)
- (2) $t_j = \log s_j$.

As noted at the beginning of this subsection, for each sample j , a reasonable scaling method should estimate scaling factors close to the true sampling fractions with possibly a constant shift across all samples. Not all scaling methods are expected to achieve this goal since many normalization methods were proposed solely to address the differences in library sizes (e.g. TSS). Failure to correct for differences in sampling fractions would lead to undesirable systematic bias in the test statistic, which can be identified by fitting a simple linear regression between centered residual h_j and the covariate of interest, such as x_{jk} (e.g. study groups):

$$h_j = a_0 + a_1 x_{jk} + e_j. \quad (4)$$

The existence of systematic bias due to differences in sampling fractions may be determined by testing the null hypothesis $H_0: a_1 = 0$ against the alternative $H_1: a_1 \neq 0$ or simply by drawing box plots of the centered residuals, as commonly done in linear regression diagnostics (Fig. 3). For an ideal normalization method, the box plot should display no pattern with respect to the covariate of interest, and the centered residuals should be randomly distributed around 0. As can be seen in the box plots provided in Fig. 3, except for ANCOM-BC, UQ, and TMM methods, for all other methods the groups G_1 , G_2 , and G_3 cluster separately, indicating that in the estimation of sampling fractions, scaling

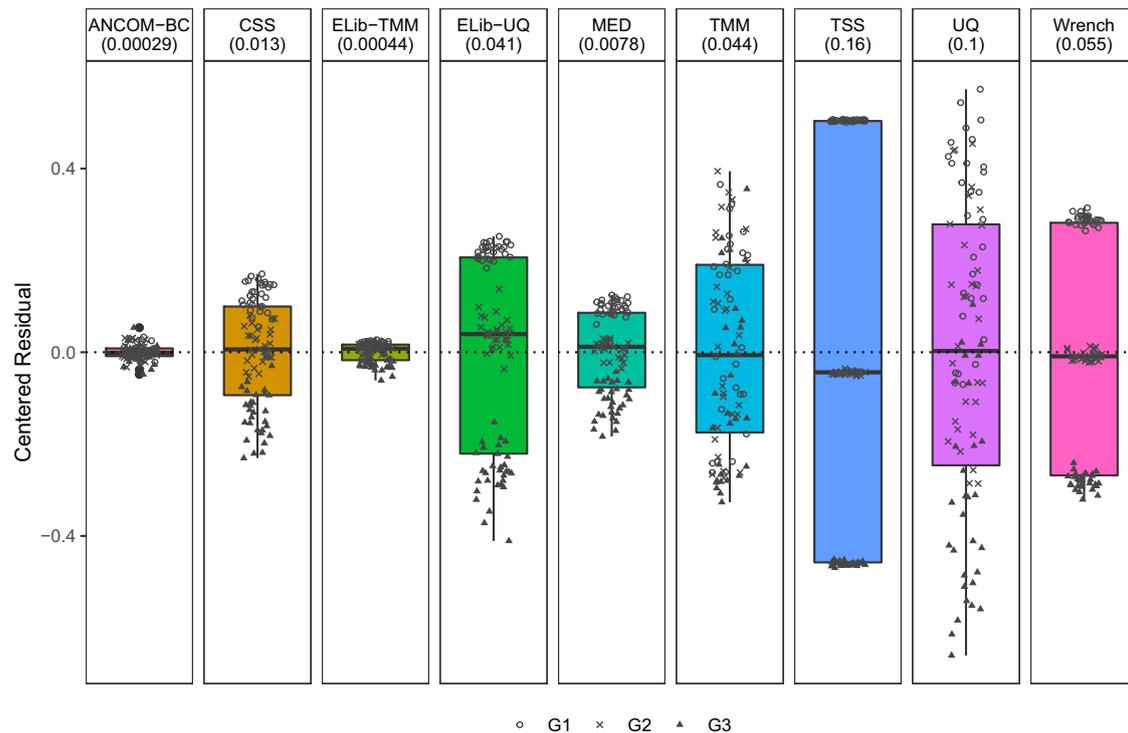


Fig. 3 Box plot of residuals between true sampling fraction and its estimate for each sample. In the box plot, the lower and upper hinges correspond to the first and third quartiles (the 25th and 75th percentiles). The median is represented by a solid line within the box. The upper whisker extends from the hinge to the largest value (maxima) no further than 1.5 times Interquartile Range (IQR, distance between the first and third quartiles) from the hinge, the lower whisker extends from the hinge to the smallest value (minima) at most 1.5 times IQR of the hinge. Data beyond the end of the whiskers are called “outlying” points. $N = 90$ samples examined over three study groups (denoted by circle, cross, and triangle, with 30 samples per group) and the data points are overlaid in each box. Each facet title indicates the normalization method and its variance is provided within parenthesis. The microbial absolute abundances in the ecosystem are generated from the log-normal distribution. By comparing residuals across different groups, an ideal box-plot should display a narrow height (i.e. smaller variability) and samples from different groups should be inter-mixed and not display any systematic separation. We note that all existing methods have larger variances compared to ANCOM-BC, and TSS has the largest variance. Except ANCOM-BC, UQ, and TMM, we see from the plot that circles, cross, and triangles are systematically separated, which indicates that ELib-UQ, ELib-TMM, CSS, MED, and TSS do not account for systematic bias due to differences in sampling fractions across groups.

factors estimated by these methods (with the exception of ANCOM-BC, UQ, and TMM) systematically differ by group labels. Furthermore, the box plot of ANCOM-BC had the shortest width, suggesting that it not only successfully estimates the true sampling fractions and eliminates bias due to its variability, but it also has the smallest variance which is not the case with other methods. This has a direct effect on the type I error and FDR as seen later in this paper and demonstrated in²⁰.

Log-ratio based methods

As an alternative to the above class of methods, several methods have been proposed in the literature that are inspired by Aitchison’s methodology for compositional data. These methods do not explicitly perform normalization such as the ones described above, since they convert the observed abundances to log-ratios within each sample. Thus, within each sample, by taking log-ratios of all taxa with respect to some common reference taxon or some suitable function of all taxa, these methods are intrinsically eliminating the effect of the sampling fraction. This class of methods include DR⁸, ANCOM¹⁴, and ALDEx2⁵¹. ALDEx2 uses a pre-specified taxon as a reference taxon and transforms the observed abundances to log ratios of the observed abundance each taxon relative to the reference taxon. Such a log-transformation of observed abundance data is called the additive log transformation (alr). Mathematically, it is defined as follows:

Definition 0.4 (additive log-ratio transformation (alr)¹⁹, $S^m \rightarrow \mathbb{R}^{m-1}$).

$$\text{alr}(O_j) = \left[\log \left(\frac{O_{1j}}{O_{ij}} \right), \dots, \log \left(\frac{O_{mj}}{O_{ij}} \right) \right]. \quad (5)$$

Thus, the alr transformation converts observed m dimensional observed abundance vector, representing the m taxa, that are in a simplex (i.e. sum to a constant), to a $m - 1$ dimensional data in the Euclidean space. A challenge with alr, and hence ALDEx2, is that the user needs to pre-specify the reference taxon. While this might be easy to do in some applications, it is generally a challenge when the number of taxa m is large such as when we are interested in performing DA analysis at the genus level. Although ANCOM is also based on alr transformation, it overcomes the above deficiency because it repeatedly applies the alr transformation by taking each of the m taxa to be a reference taxon one at a time. Thus, for each taxon, it performs $m - 1$ regressions. Hence, it overall fits $m(m - 1)$ regression models.

To avoid the above challenges due to alr transformation, rather than using a pre-specified taxon as the reference taxon, one may consider the center of mass of all taxa as the reference. Thus, within each sample, for each taxon, the log-ratios are computed relative to the geometric mean of all taxa. This transformation is called the clr transformation. More precisely, it is defined as follows:

Definition 0.5 (centered log-ratio transformation (clr)¹⁹, $\mathbb{S}^m \rightarrow \mathbb{U}^m$).

$$\text{clr}(\mathbf{O}_j) = \left[\log \left(\frac{O_{1j}}{g(\mathbf{O}_j)} \right), \dots, \log \left(\frac{O_{mj}}{g(\mathbf{O}_j)} \right) \right], \quad (6)$$

where

- (1) $g(x)$ is the geometric mean of x ,
- (2) $U^m = \{(u_1, \dots, u_m) \in \mathbb{R}^m: u_1 + \dots + u_m = 0\}$ is a hyperplane in \mathbb{R}^m .

This transformation to a real space again makes the implementation of unconstrained statistical methods possible. clr transformation is an isometry, but sum of the transformed values equals to 0, leading to a degenerate distribution.

The alr transformation is not isometric and clr is not an isomorphism. The isometric log-ratio transformation (ilr)²⁵ (also known as balance) is both an isomorphism and an isometry, and consequently orthonormal coordinates can be defined using this transformation.

Definition 0.6 (isometric log-ratio transformation (ilr), $\mathbb{S}^m \rightarrow \mathbb{R}^{m-1}$).

$$\text{ilr}(\mathbf{O}_j) = \text{clr}(\mathbf{O}_j)\Psi^T, \quad (7)$$

where Ψ is a $(m-1, m)$ orthonormal basis.

There are multiple ways to construct orthonormal bases. Typically, if a bifurcating tree is given then we can construct a basis from the internal nodes in the tree. Each element in the ilr transformed data is of the following form:

$$b_l = \sqrt{\frac{|l_L||l_R|}{|l_L| + |l_R|}} \log \left[\frac{g(l_L)}{g(l_R)} \right], \quad (8)$$

where

- (1) b_l is the balance at internal node l ,
- (2) l_L is the set of relative abundances contained in the left subtree at internal node l ,
- (3) l_R is the set of relative abundances contained in the right subtree at internal node l ,
- (4) $|l_L|$ is the number of taxa contained in l_L ,
- (5) $|l_R|$ is the the number of taxa contained in l_R ,
- (6) $g(x)$ is the geometric mean of x .

METHODS OF DIFFERENTIAL ABUNDANCE ANALYSIS

A number of procedures have been introduced and used in the literature for identifying differentially abundant taxa. One common approach is to apply a nonparametric test (e.g. the Mann–Whitney/Wilcoxon rank-sum test for two sample classes; the Kruskal–Wallis test for multiple sample classes) after normalizing the feature table. Unfortunately, these standard nonparametric tests do not take into account the compositional structure of microbiome data.

RNA-seq based methods: edgeR and DESeq2

As alternatives to standard nonparametric tests, many parametric models have been proposed in the literature based on transcriptomics data, such as the RNA-Seq data, for testing differences across study groups. Among them, DESeq2⁴¹ and edgeR⁴⁴ are two popular methods. These methods model the observed abundances using negative binomial (NB) distribution after normalizing data with corresponding scaling methods to account for differences in sampling fractions. Thus O_{ij} are modeled using the a negative binomial distribution as follows:

$$O_{ij} \sim \text{NB}(s_j \mu_i, \phi_i), \quad (9)$$

where

- (1) s_j is the scaling factor for sample j ,
- (2) μ_i is the mean absolute abundance (in ecosystem) for taxon i ,
- (3) ϕ_i is the dispersion parameter for taxon i .

Introduction of the dispersion parameter ϕ_i is inspired by mean-variance dependence in count data (e.g. RNA-Seq, microbiome data), and recognizing that the variance is typically larger than mean especially when the mean value is large. Thus, the variance of the observed abundance is modeled as follows:

$$\text{Var}(O_{ij}) = s_j \mu_i + \phi_i s_j^2 \mu_i^2. \quad (10)$$

The NB distribution is more appropriate for modeling these types of count data than the Poisson distribution because it provides greater flexibility in modeling the variance. We remind the readers that by conditioning independent Poisson random variables on the total count results in multinomial distribution^{52,53}.

The estimation of the dispersion parameter is critical for both edgeR as well as DESeq2. Based on the assumption that taxa with similar observed abundances also share similar variances, edgeR estimates the taxon-wise dispersion by conditional maximum likelihood⁵⁴, and then shrinks the dispersion estimate for each taxon towards a common estimate of taxa with similar observed abundances using an empirical Bayes procedure⁵⁵. Similarly, DESeq2 first estimates the taxon-wise dispersion by maximum likelihood estimation, and then fits the dispersion trend combining all individual estimates, and finally shrinks the taxon-wise dispersion estimates towards the values predicted by the trend curve using an empirical Bayes approach.

While both methods are generally very reasonable and appropriate for gene expression data, they seem to perform poorly for microbiome data. This is largely because, as stated earlier, the normalization methods used by these two methods intrinsically assume that a very small fraction of taxa are differentially abundant. This assumption is not necessarily valid for microbiome data. As a consequence, the test statistics used by these methods are intrinsically biased under the null hypothesis. As demonstrated analytically as well as empirically in Lin and Peddada²⁰, and reproduced here empirically using similar log-normal distribution based simulation settings (Fig. 4, see Supplementary Information for simulation settings), the bias in the test statistic results in inflated FDRs for these methods. What is worse, because of the bias, as the sample size increases, the FDR increases for these methods²⁰. Similar phenomena were reported by Weiss et al.³⁹. When dealing with population studies, it is important to recognize that there is variability within subject and there is variability between subjects in the population. In simple terms, observed abundance of a taxon from a subject may vary from stool sample to stool sample obtained from the same subject. This is within subject variation. Hence when calculating variability in measurements of random subject, one needs to take into account variation within as well as between subjects. This results in over-dispersion³³. While it is important to account for this over-dispersion, it does not correct the intrinsic bias due to differential sampling fractions noted above. RNA-seq inspired methods do not perform well for microbiome data even after correcting for the over-dispersion parameter.

MetagenomeSeq

Instead of using a negative binomial model, an alternative mixture model based on zero-inflated Gaussian (ZIG) is implemented in metagenomeSeq²¹, where excess zeros due to both sampling zeros and structural zeros are accounted by a probability mass, and the Gaussian distribution modeling the non-zero observed

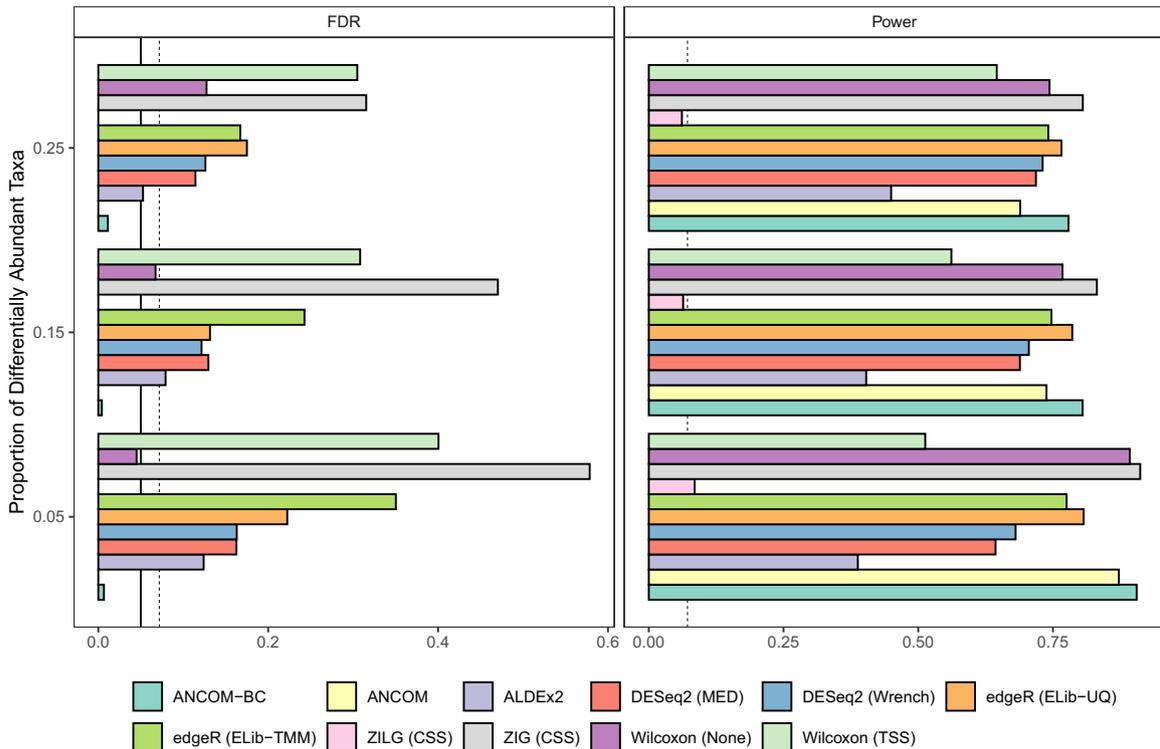


Fig. 4 False Discovery Rate (FDR) and power comparisons using synthetic data. The FDR and power of various differential abundance (DA) analyses (two-sided) are shown in **a**, **b**, respectively. The number of taxa is set to be 200, and the sample size equals to 60, with 30 samples per group. The microbial absolute abundances in the ecosystem are generated from the log-normal distribution. The y-axis denotes patterns of proportion of differential abundant taxa, ranging from 0.05 to 0.25. The solid vertical line is the 5% nominal level of FDR, and the dashed vertical line denotes 5% nominal level plus one standard error (SE). Legend on the bottom indicates the color for each method and the normalization method is provided within parenthesis. By default, ANCOM-BC implements Holm-Bonferroni⁷⁰ method and other DA methods implement BH procedure⁵⁷ to adjust for multiple comparisons. By detecting differentially abundant taxa between two groups, results show that only ANCOM and ANCOM-BC control the FDR under the nominal level (5%) while maintaining power comparable to other methods. Gaussian model version of metagenomeSeq has highly inflated FDR, while the log Gaussian version has substantial loss of power, sometimes well below 5%.

abundances. The framework can be summarized as follows:

$$\begin{aligned}
 y_{ij} &= \log_2(O_{ij} + 1), \\
 f_{\text{zig}}(y_{ij}, O_{ij}, \mu_j, \sigma_j^2) &= \pi_j(O_{ij})I_{\{0\}}(y_{ij}) + [1 - \pi_j(O_{ij})]\phi(y_{ij}, \mu_j, \sigma_j^2), \\
 \mu_j &= \eta_j \log_2\left(\frac{s_j^{j+1}}{N}\right) + \beta_j^T \mathbf{x}_j,
 \end{aligned}
 \tag{11}$$

where

- (1) N is a normalization constant,
- (2) \hat{l} is determined by CSS normalization,
- (3) $q_j^{\hat{l}}$ is the \hat{l} th quantile of observed abundances for sample j ,
- (4) $s_j^j = \sum_{i:O_{ij} \leq q_j^{\hat{l}}} O_{ij}$.

However, as shown in our benchmark simulations (Fig. 4) as well as in other previously published simulation studies^{14,33,39}, although metagenomeSeq has marginally higher powers than most of the other DA methods, it is subject to unreasonably high FDRs even though the observed abundances are normalized by their built-in scaling method (CSS). Furthermore, the problem of FDR inflation gets worse when sample size or the effect size (i.e. fold change of mean absolute abundances) increases^{20,39}. It is also worth pointing out that metagenomeSeq was the only method, among all parametric models, that increases FDR when applied to rarefied data^{33,39}. This is likely due to its zero-inflated model which requires the input of precise library sizes to capture the zero proportions.

Note that the authors of metagenomeSeq modified their procedure and recommended replacing zero-inflated Gaussian

(ZIG) mixture model by zero-inflated Log-Gaussian (ZILG) mixture model for DA analysis. Although switching to zero-inflated Log-Gaussian distribution improves the FDR control, the procedure becomes extremely conservative, with FDR close to zero and a substantial loss of power in our simulations (Fig. 4) and in ref.²⁰.

ALDEx2

It is based on the original version of ANOVA-Like Differential Expression (ALDEx) analysis⁵⁶. It was proposed as a compositional data analysis tool that is applicable to three different types of data: RNA-Seq, ChIP-Seq, and 16S rRNA gene sequencing⁵¹. By acknowledging these high-throughput sequencing data are fundamentally compositional, the methodology of ALDEx2 can be summarized as follows:

- (1) The observed abundances are converted to relative abundances by Monte Carlo (MC) sampling from the Dirichlet distribution with the addition of a uniform prior. The MC sampling is repeated for K times ($K = 128$ times by default), thus essentially, for each taxon i in sample j , the observed abundance O_{ij} is represented by a vector of MC samples of relative abundances $(r_{ij}^{(1)}, \dots, r_{ij}^{(K)})^T$,
- (2) Within each sample j and each MC Dirichlet realization k , $k = 1, \dots, K$, the relative abundance vector $(r_{ij}^{(k)}, \dots, r_{mj}^{(k)})^T$ is clr transformed,
- (3) Significance test (Welch's t -test or Wilcoxon test) is performed on each taxon in the vector of clr transformed

values. Since there are a total of K MC Dirichlet samples, each taxon will result in K p -values.

- (4) Each resulting p -value is corrected using the B-H⁵⁷ procedure, and the expected adjusted p -value for each taxon is reported by taking the empirical mean of K adjusted p -values.

The ALDEx2 was designed to identify differential abundances of features (genes, taxa, or genomic segments), relative to the geometric mean abundance, between two or more groups. As reported in the simulation study described in this paper (Fig. 4) ALDEx2 not only generally exceeds the nominal level of FDR (5%), but also has substantially smaller power as compared to competing DA methods. Similar results were also reported in Morton et al.⁸.

ANCOM

Analysis of composition of microbiomes (ANCOM)¹⁴ is an alr based methodology, which accounts for the compositional structure of microbiome data. Given a total of m taxa, ANCOM relies on two assumptions as follows.

Assumption 0.1: The mean log absolute abundance (in the ecosystem) of 2 taxa are not different.

Assumption 0.2: The mean log absolute abundance (in the ecosystem) of all m taxa do not differ by the same amount between two study groups. For example, suppose the absolute abundance of m taxa for a subject in group 1 (C-section born babies) are A_1, A_2, \dots, A_m and suppose the absolute abundance of taxa for a subject in group 2 (vaginally born babies) are B_1, B_2, \dots, B_m . Then $B_i \neq CA_i$, for all $i = 1, 2, \dots, m$. Thus, not all taxa are changing by the same constant C .

Note that the first assumption made by ANCOM is substantially weaker than the assumptions made by DESeq2 and edgeR, which require very "few" taxa to be differentially abundant.

Under the above assumptions, together with the fact that ANCOM performs all possible DA analyses by successively using each taxon as a reference taxon, the authors proved that one can test the null hypothesis regarding mean log absolute abundance in a unit volume of an ecosystem using relative abundances.

For the i th taxon and j th sample, ANCOM uses standard ANOVA model formulation:

$$\log \frac{r_{ij}^{(g)}}{r_{i'j}^{(g)}} = \alpha_{i'} + \beta_{i'}^{(g)} + \sum_k x_{jk} \beta_{i'k} + \epsilon_{i'j}^{(g)}, \quad (12)$$

where

- (1) i' is the reference taxon, $i' \neq i = 1, 2, \dots, m$,
- (2) $g = 1, 2, \dots, G$ is the number of study groups.

By virtue of Assumption 0.1 and Assumption 0.2, to test whether a taxon i is differentially abundant according to a factor of interest with G levels, it is equivalent to test:

$$H_{0(i'')} : \beta_{i'}^{(1)} = \dots = \beta_{i'}^{(G)} = 0,$$

$$H_{1(i'')} : \text{Not all } \beta_{i'}^{(g)} \text{ equal to } 0,$$

for every $i \neq i'$.

P -values from $\frac{m(m-1)}{2}$ distinct null hypotheses $H_{0(i'')}$, $i \neq i'$ are adjusted using a multiple testing correction procedure such as the Benjamini-Hochberg (BH) procedure⁵⁷ or Bonferroni correction^{58,59}. For each taxon, the number of rejections, denoted by W_i , is counted, and ANCOM makes use of the empirical distribution of $\{W_1, W_2, \dots, W_m\}$ to determine the cut-off value of significant taxon. The rule of thumb is, when the value of W_i is large, then it is more likely that taxon i is differentially abundant. The authors recommend using 70th percentile of the W distribution as the empirical cut-off value. However, the ANCOM outputs results from different cutoffs such as the 60th to 90th percentile and lets the user select the threshold of their interest.

As shown in the simulation studies (Fig. 4) as well as in^{14,20}, using the 70th percentile of W distribution as the cut-off, ANCOM successfully controls the FDR under the nominal level (5%) while maintaining adequate power. However, ANCOM can be computationally intensive since for each taxon, it performs alr transformation using all remaining taxa. The computation time scales up quadratically with the number of taxa. Additionally, the statistical decision made by ANCOM depends on the quantile of its test statistic W , rather than p -values, which some researchers find it difficult to interpret.

DR

Differential Ranking (DR)⁸ exploits the fact that the ranks of relative differentials (i.e. log ratio between absolute relative abundances) are identical to the ranks of absolute differentials (i.e. log ratio between absolute abundances). They estimate relative differentials using a linear regression where relative abundances are alr transformed. The regression coefficients corresponding to different taxa are ranked in order to determine the most important to the least important taxa.

The DR model can be summarized as follows:

$$\beta_{ik} \sim N(0, \mu_\beta),$$

$$\mathbf{r}_j = \text{alr}^{-1}(\beta_j^T \mathbf{x}_j), \quad (13)$$

$$\mathbf{A}_j \sim \text{Multinomial}(\mathbf{r}_j),$$

where

- (1) \mathbf{x}_j is the vector of covariates of interest (e.g. study groups) for the j th sample,
- (2) \mathbf{r}_j is the vector of observed relative abundances for the j th sample,
- (3) \mathbf{A}_j is the vector of absolute abundances in the ecosystem for the j th sample.

The model parameters are estimated using a maximum a posteriori (MAP) estimation by stochastic gradient descent.

To understand the implementation of the DR procedure, consider a simple example where the true absolute relative abundance is known. Suppose there are only two samples belonging to two groups (e.g. control vs treatment) and the unobserved absolute abundance is linearly related with the group effect in log scale, i.e.:

$$\log A_{ij} = \alpha_{i0} + \alpha_{i1} I(j \in \text{group } 1). \quad (14)$$

Suppose sample j_1 is in group 1 and sample j_2 is in group 2, then from (Eq. 14) we have

$$\log A_{ij_1} - \log A_{ij_2} = \alpha_{i1}. \quad (15)$$

Denoting the true absolute relative abundances by y_{ij} and $y_{i'j}$ one can write down the DR model (Eq. 13) as:

$$\log \frac{y_{ij}}{y_{i'j}} = \log \frac{A_{ij}}{A_{i'j}} = \beta_{i0} + \beta_{i1}. \quad (16)$$

where i' is the reference taxon. Thus,

$$\begin{aligned} \log \frac{y_{ij_1}}{y_{i'j_1}} - \log \frac{y_{ij_2}}{y_{i'j_2}} &= \log \frac{A_{ij_1}}{A_{i'j_1}} - \log \frac{A_{ij_2}}{A_{i'j_2}} \\ &= \log A_{ij_1} - \log A_{ij_2} - (\log A_{i'j_1} - \log A_{i'j_2}) \\ &= \beta_{i1}. \end{aligned} \quad (17)$$

Comparing (Eq. 15) with (Eq. 17), it is clear that although $\beta_{i1} \neq \alpha_{i1}$, due to the bias term $\log A_{i'j_1} - \log A_{i'j_2}$. However, since the bias term is constant for taxon i , the rank of β_{i1} is same as the rank of α_{i1} .

Thus, unlike typical DA methods in which the estimated coefficient reflects the change in absolute abundances, the interpretation of DR results requires care because it is based on the ranks. Due to the presence of the microbial load bias

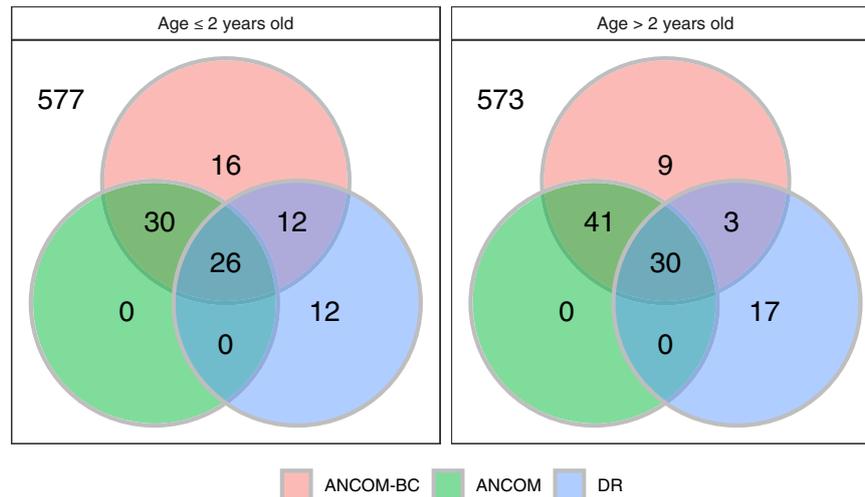


Fig. 5 Venn diagrams representing consistency of differentially abundant genera identified by ANCOM-BC, ANCOM, and DR. The global gut microbiota data⁶⁰ is used to make the Venn diagram. The dataset contains 673 genera, with subjects from Malawi (MA, $n_1 = 114$), USA (US, $n_2 = 317$), and Venezuela (VEN, $n_3 = 99$). We compare the absolute abundance of different genera for (1) Subjects who are less than or equal to 2 years old with sample size (MA:US:VEN) = (47:50:27) and (2) Subjects who are greater than 2 years old with sample size (MA:US:VEN) = (36:260:70). ANCOM-BC and ANCOM generally have large overlap of significant genera.

($\log A_{i_j_1} - \log A_{i_j_2}$ in the above example), a positive valued coefficient from DR model does not necessarily mean that the absolute abundance has increased. Similarly, a zero valued coefficient does not imply the absolute abundance of the corresponding taxon has not changed. Nevertheless, based on the ranks of coefficients, one can focus on taxa with high or low ranks since they are the ones that are potentially increasing or decreasing the most in absolute abundances relative to other taxa.

Note that since different reference taxon in the air transformation of DR model will lead to the same result regarding the ranks, DR is robust to the choice of reference taxon.

ANCOM-BC

Analysis of compositions of microbiomes with bias correction (ANCOM-BC)²⁰ models the observed abundances using an offset-based log-linear model.

$$y_{ij} = d_j + \beta_i^T \mathbf{x}_j + \epsilon_{ij}, \quad (18)$$

where

- (1) $y_{ij} = \log O_{ij}$ is the log observed abundance,
- (2) d_j (see Table 2)

In this set-up, the zero counts are handled using the methodology described in Kaul et al.²⁸. This formulation explicitly tests the hypothesis regarding differential absolute abundance of individual taxon while estimating sample-specific sampling fractions and correcting the bias appropriately. As demonstrated in our simulation studies, ANCOM-BC not only controls the FDR very well, but also competes very well with other methods in terms of power (Fig. 4). Furthermore, unlike any of the existing methods, ANCOM-BC provides valid confidence intervals for differential abundance of individual taxa between two study groups and also provides a valid p -value²⁰. Since it has a linear regression framework, it allows for repeated measurement designs as well as covariate adjustments. ANCOM-BC can also be extended to describe patterns of differential abundance in multiple study groups such as time course or dose-response studies²⁰.

As a benchmark analysis, we also compared significant genera identified by ANCOM-BC, ANCOM, and DR using the global gut microbiota data⁶⁰. This data set consists of 11,905 OTUs obtained from fecal samples of subjects in the USA ($n = 317$), Malawi ($n = 114$), and Venezuela ($n = 99$). We first subdivided the data into

two age strata “ ≤ 2 years” and “ > 2 years”. This stratification was performed because it is expected that microbial composition of infants changes when they switch over from breast milk (or formula milk) to solid food⁷. The sample sizes in the two age categories (≤ 2 years, > 2 years) for Malawi (MA), USA (US) and Venezuela (VEN) are (47, 36), (50, 260), and (27, 70), respectively. Note that samples with missing values of age were discarded in the downstream analysis. Without a hard threshold available for DR, as suggested in the original paper⁸, we investigated the highest/lowest ranks of genera by selecting the top 25 and bottom 25 genera in terms of rank order of regression parameter estimates. As seen in Fig. 5, the three methods generally have a large number of overlapping genera, with ANCOM-BC and ANCOM having more taxa in common that are differentially abundant. While implementing ANCOM, we used the 70th percentile of the distribution of W as the cut-off. Note that the DR method was applied with all hyper-parameters of the multinomial model set to their default values in the algorithm which can be further tuned.

Balance-based methods

A variety of methods have been proposed in the literature that are based on balances described earlier in this paper. Some examples include gneiss¹⁸, phylofactorization^{61,62}, PhILR⁶³, and selbal⁶⁴. Although the balance-based methods were not explicitly designed for performing formal statistical DA analyses for individual taxon, they are often used for that purpose.

To overcome the challenges posed by the compositional structure of 16S rRNA data for identifying individual differentially abundant taxa, gneiss¹⁸ was developed to identify taxa distribution across different covariates with the help of balances. The balances (Eq. (8))^{65,66} are useful to infer meaningful properties of sub-communities. Gneiss aims to associate the effect of parameter of interest to the matrix of balances:

Definition 0.7 (gneiss model).

$$b_{jl} = \beta_l^T \mathbf{x}_j, \quad (19)$$

where

- (1) b_{jl} represents the balance for sample j at node l ,
- (2) $\beta_l = (\beta_{l1}, \dots, \beta_{lp})^T$ represents a vector of coefficients,
- (3) $\mathbf{x}_j = (x_{j1}, \dots, x_{jp})^T$ represents the measures for covariates.

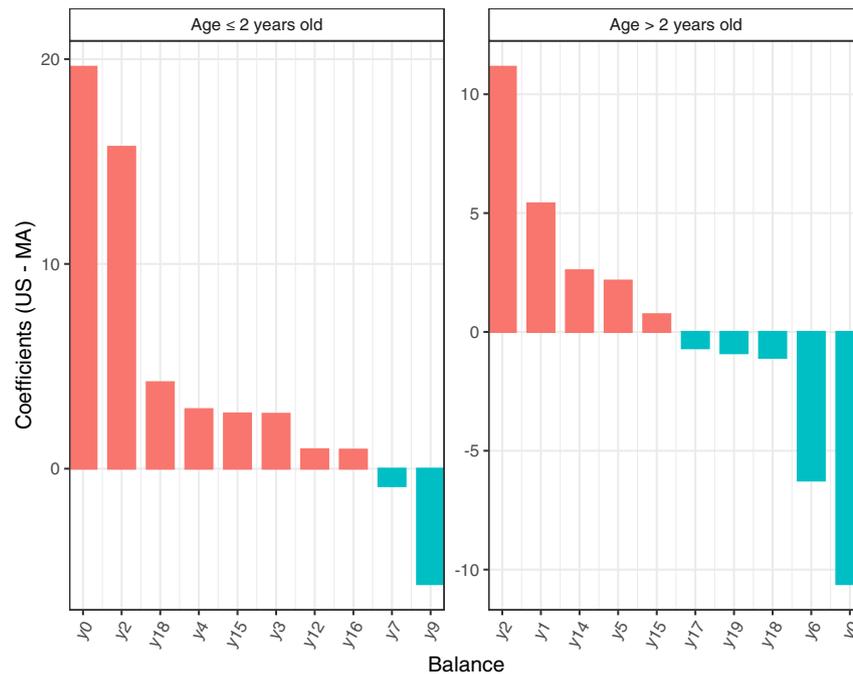


Fig. 6 Waterfall plot visualizing coefficient (US: MA) for top 20 balances identified by gneiss using the global gut microbiota data⁶⁰. We subset the dataset to subjects whose nationality is Malawi (MA) and USA (US), with sample size (MA:US) = (47:50) for the group of age ≤ 2 years old, and (MA:US) = (36:260) for the group of age > 2 years old. The columns of the plot represent coefficients, and the rows of the plot represent balances. BH procedure⁵⁷ was applied to correct for multiple comparisons, and coefficients with FDR corrected p -values < 0.05 are discarded.

Gneiss methodology is very flexible and can be broadly used for determining niches of microbes in various sub-communities. Thus, it is a very useful method for discovering niche differentiation in microbes.

Similar to gneiss, phylofactorization^{61,62} is not designed for the DA analysis as defined in this paper, but it focuses on the comparison between clades with a clear phylogenetic interpretation. It is based on a greedy algorithm which sequentially selects edges, instead of nodes or splits in a phylogeny, whose ilr basis element maximizes a pre-specified objective function (e.g. the percentage of variation explained). Therefore, besides comparing sister clades, phylofactorization compares the relative abundances between all other clades.

We illustrate gneiss using the global gut data⁶⁰ discussed earlier in this paper using Malawi (MA, $n_1 = 114$) and the USA (US, $n_2 = 317$) data. Gneiss identified different trends among various balances (Fig. 6). For example, balance y_0 is detected to increase in US as compared to MA for subjects who are ≤ 2 years old; It is in a reverse direction for subjects who are > 2 years old. One caveat to keep in mind is that the components of balances are not necessarily the same across different data sets. The first balance y_0 for the younger generation (age ≤ 2 years old) consists of 642 taxa in the numerator (the left subtree) and 31 taxa in the denominator (the right subtree); On the other hand, y_0 for the older group (age > 2 years old) has 655 taxa in the numerator and 18 taxa in the denominator. It is important to note gneiss is not designed to infer changes in abundance for each individual taxon, however, it can answer questions such as whether the absolute abundances of taxa in the numerator of y_0 on average have increased or decreased as compared to those in the denominator.

LEfSe

Linear Discriminant Analysis Effect Size (LEfSe)⁶⁷ is specifically designed for group comparisons of microbiome data with a

particular focus on detecting change in relative abundance between two or more groups of samples with biological consistency. Important statistical and computational steps implemented in LEfSe are as follows:

1. For each taxon, test whether its observed abundances in different groups are differentially distributed using Kruskal–Wallis test.
2. (Optional, only if subgroups are defined) Discard taxa which are not statistically significant in step 1 (e.g. p -value > 0.05). The pairwise Wilcoxon test is then applied to retain taxa. A taxon is not retained for further consideration if it is not significant in every pairwise comparison (e.g. p -value > 0.05 for at least one pairwise comparison) or if the signs of test statistics are not equal among all comparisons.
3. After feature selection, a Linear Discriminant Analysis (LDA) model is built with the group label as the dependent variable and observed abundance of taxa selected in above step, subgroup label, and demographic features as independent variables. This model is used to calculate the effect size for each taxon. This effect size serves as the average of each taxon's variability and discriminatory power.
4. Finally, the LDA score for each taxon is obtained by computing the logarithm (base 10) of the effect size after being scaled in the $[1, 10^6]$ interval. The rank for each taxon is assigned based on the corresponding LDA score and further feature selection could be achieved by setting a threshold (e.g. 2.0) for LDA scores.

By its construction, LEfSe method is more a discriminant analysis method rather than a DA method. Unlike the DA analysis methods discussed earlier in this paper, LEfSe is more focused on investigating the relationship among microbial profiles and an outcome or phenotype (Step 3). More precisely, LEfSe tries to quantify the magnitude of the effect size of such associations between microbial profiles (e.g. a set of taxa) and the outcome of interest.

DISCUSSION

Microbiome studies are becoming very popular in biomedical sciences. As new scientific questions emerge, so do new statistical and computational methods of analysis. This is a very rapidly growing area of research with new statistical methods being developed on a regular basis. Hence an up-to-date comprehensive review of the statistical methods in the field is a challenging problem. This is particularly true with methods for DA analysis. A number of methods exist in the literature and each method has its own strengths and weaknesses. One of the challenges in evaluating the performance of various methods is that not all methods are designed to test statistical hypotheses regarding the same parameter. Some methods are designed for testing hypotheses regarding the relative abundance, while others are designed for testing hypothesis regarding absolute abundance. If a simulation study is designed for testing hypothesis regarding absolute abundance then methods for relative abundance parameter may show an inflated FDR and vice versa. A related problem is that often researchers use the terms “relative abundance” and “absolute abundance in a unit volume” interchangeably. This makes the simulation studies difficult to interpret. Therefore journals and researchers should make the terminology precise. In this paper, simulation studies were set-up to compare FDR and power of various methods when testing hypotheses regarding absolute abundance of taxa in a unit volume of a tissue.

We performed simulation studies using the log-normal distribution for modeling abundances. Consistent with the findings of²⁰, ANCOM and ANCOM-BC control the FDR at the desired nominal level for most configurations while competing well with all procedures in terms of the overall power. The only situations where ANCOM as well as ANCOM-BC fail to control FDR is when the sample sizes are very small, such as $<10^{20}$. All other methods considered in this paper tend to inflate FDR for all sample sizes and their FDR gets worse with the sample size increases²⁰. This is because, under the null hypothesis, each of these methods is biased away from zero. This bias increases with sample size. Hence the FDR increases with sample size.

While ANCOM and ANCOM-BC have very similar operating characteristics in terms of FDR and power, ANCOM-BC is computationally simpler and faster to implement because unlike ANCOM it requires only m linear regression fits rather than $\frac{m \times (m-1)}{2}$ models fits needed by ANCOM. Secondly, unlike ANCOM, ANCOM-BC provides individual p-values and confidence intervals of pairwise difference in mean abundance for each taxon. Among the methods available today, ANCOM-BC is the only procedure that provides valid p-values and confidence intervals. Furthermore, since ANCOM-BC is based on a regression model framework, it can easily be extended to repeated measures/longitudinal data covariate adjustments.

DATA AVAILABILITY

DNA sequences from the global gut microbiota study⁶⁰ can be found in MG-RAST <https://www.mg-rast.org/index.html> server under search string “mgp401” for Illumina V4-16S rRNA; feature table, metadata, and taxonomy of the diet swap data⁶⁸ is available in the microbiome⁶⁹ R package <http://microbiome.github.com/microbiome>.

CODE AVAILABILITY

All datasets and analysis scripts can be found under <https://github.com/FrederickHuangLin/Microbiome-Review-Code-Archive>.

Received: 3 March 2020; Accepted: 16 October 2020;

Published online: 02 December 2020

REFERENCES

1. Tierney, B. T. et al. The landscape of genetic content in the gut and oral human microbiome. *Cell Host Microbe* **26**, 283–295 (2019).
2. O'Hara, A. M. & Shanahan, F. The gut flora as a forgotten organ. *EMBO Rep.* **7**, 688–693 (2006).
3. Relman, D. A. & Falkow, S. The meaning and impact of the human genome sequence for microbiology. *Trends Microbiol.* **9**, 206–208 (2001).
4. Hurst, G. D. Extended genomes: symbiosis and evolution. *Interface Focus* **7**, 20170001 (2017).
5. Turnbaugh, P. J. et al. A core gut microbiome in obese and lean twins. *Nature* **457**, 480 (2009).
6. Gevers, D. et al. The treatment-naive microbiome in new-onset crohn's disease. *Cell Host Microbe* **15**, 382–392 (2014).
7. Lozupone, C. A. et al. Alterations in the gut microbiota associated with hiv-1 infection. *Cell Host Microbe* **14**, 329–339 (2013).
8. Morton, J. T. et al. Establishing microbial composition measurement standards with reference frames. *Nat. Commun.* **10**, 2719 (2019).
9. Schloss, P. D. The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16s rna gene-based studies. *PLoS Comput. Biol.* **6**, e1000844 (2010).
10. Edgar, R. C. Uparse: highly accurate otu sequences from microbial amplicon reads. *Nat. Methods* **10**, 996 (2013).
11. Callahan, B. J. et al. Dada2: high-resolution sample inference from illumina amplicon data. *Nat. Methods* **13**, 581 (2016).
12. Amir, A. et al. Deblur rapidly resolves single-nucleotide community sequence patterns. *MSystems* **2**, e00191–16 (2017).
13. Bolyen, E. et al. Reproducible, interactive, scalable and extensible microbiome data science using qiime 2. *Nat. Biotechnol.* **37**, 852–857 (2019).
14. Mandal, S. et al. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb. Ecol. Health Dis.* **26**, 27663 (2015).
15. Gloor, G. B. & Reid, G. Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. *Can. J. Microbiol.* **62**, 692–703 (2016).
16. Gloor, G. B., Wu, J. R., Pawlowsky-Glahn, V. & Egozcue, J. J. It's all relative: analyzing microbiome data as compositions. *Ann. Epidemiol.* **26**, 322–329 (2016).
17. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* **8**, 2224 (2017).
18. Morton, J. T. et al. Balance trees reveal microbial niche differentiation. *MSystems* **2**, e00162–16 (2017).
19. Aitchison, J. The statistical analysis of compositional data. *J. Royal Stat. Soc. Ser. B.* 139–177 (1982).
20. Lin, H. & Peddada, S. D. Analysis of compositions of microbiomes with bias correction. *Nat. Commun.* **11**, 1–11 (2020).
21. Paulson, J. N., Stine, O. C., Bravo, H. C. & Pop, M. Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* **10**, 1200 (2013).
22. Xia, F., Chen, J., Fung, W. K. & Li, H. A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics* **69**, 1053–1063 (2013).
23. Costea, P. I., Zeller, G., Sunagawa, S. & Bork, P. A fair comparison. *Nat. Methods* **11**, 359 (2014).
24. Paulson, J. N., Bravo, H. C. & Pop, M. Reply to: "a fair comparison". *Nat. Methods* **11**, 359 (2014).
25. Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G. & Barcelo-Vidal, C. Isometric logratio transformations for compositional data analysis. *Math. Geol.* **35**, 279–300 (2003).
26. Greenacre, M. Measuring subcompositional incoherence. *Math. Geosci.* **43**, 681–693 (2011).
27. Chen, E. Z. & Li, H. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics* **32**, 2611–2617 (2016).
28. Kaul, A., Mandal, S., Davidov, O. & Peddada, S. D. Analysis of microbiome data in the presence of excess zeros. *Front. Microbiol.* **8**, 2114 (2017).
29. Navas-Molina, J. A. et al. Advancing our understanding of the human microbiome using qiime. In *Methods in Enzymology*, Vol. 531, 371–444 (Elsevier, 2013).
30. Hughes, J. B. & Hellmann, J. J. The application of rarefaction techniques to molecular inventories of microbial diversity. *Methods Enzymol.* **397**, 292–308 (2005).
31. Koren, O. et al. A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. *PLoS Comput. Biol.* **9**, e1002863 (2013).
32. Gotelli, N. J. & Colwell, R. K. Estimating species richness. *Biol. Divers. Front. Meas. Assess.* **12**, 39–54 (2011).
33. McMurdie, P. J. & Holmes, S. Waste not, want not: why rarefying microbiome data

is inadmissible. *PLoS Comput. Biol.* **10**, e1003531 (2014).

34. Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J. & Knight, R. Unifrac: an effective distance metric for microbial community comparison. *ISME J.* **5**, 169 (2011).

35. Gotelli, N. J. & Colwell, R. K. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecol. Lett.* **4**, 379–391 (2001).

36. Brewer, A. & Williamson, M. A new relationship for rarefaction. *Biodivers. Conserv.* **3**, 373–379 (1994).

37. Horner-Devine, M. C., Lage, M., Hughes, J. B. & Bohannan, B. J. A taxa–area relationship for bacteria. *Nature* **432**, 750 (2004).

38. Jernvall, J. & Wright, P. C. Diversity components of impending primate extinctions. *Proc. Natl Acad. Sci. USA* **95**, 11279–11283 (1998).

39. Weiss, S. et al. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5**, 27 (2017).

40. Beule, L. & Karlovsky, P. Improved normalization of species count data in ecology by scaling with ranked subsampling (srs): application to microbial communities. *PeerJ* **8**, e9593 (2020).

41. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for rna-seq data with *DESeq2*. *Genome Biol.* **15**, 550 (2014).

42. Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC Bioinform.* **11**, 94 (2010).

43. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biol.* **11**, R25 (2010).

44. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. *edgeR*: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

45. Kumar, M. S. et al. Analysis and correction of compositional bias in sparse sequencing count data. *BMC Genomics* **19**, 799 (2018).

46. Chen, Y., McCarthy, D., Robinson, M. & Smyth, G. K. *edgeR*: differential expression analysis of digital gene expression data user's guide. <http://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf> (2014).

47. Dillies, M.-A. et al. A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Brief. Bioinforma.* **14**, 671–683 (2013).

48. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).

49. Agresti, A. & Hitchcock, D. B. Bayesian inference for categorical data analysis. *Stat. Methods Appl.* **14**, 297–330 (2005).

50. Friedman, J. & Alm, E. J. Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* **8**, e1002687 (2012).

51. Fernandes, A. D. et al. Unifying the analysis of high-throughput sequencing datasets: characterizing rna-seq, 16s rna gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* **2**, 15 (2014).

52. Steel, G. et al. Relation between poisson and multinomial distributions. <https://ecommons.cornell.edu/bitstream/handle/1813/32480/BU-39-M.pdf?sequence=1> (1953).

53. Taddy, M. Multinomial inverse regression for text analysis. *J. Am. Stat. Assoc.* **108**, 755–770 (2013).

54. Smyth, G. K. & Verbyla, A. P. A conditional likelihood approach to residual maximum likelihood estimation in generalized linear models. *J. R. Stat. Soc. Ser. B* **58**, 565–572 (1996).

55. Robinson, M. D. & Smyth, G. K. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **23**, 2881–2887 (2007).

56. Fernandes, A. D., Macklaim, J. M., Linn, T. G., Reid, G. & Gloor, G. B. Anova-like differential expression (aldex) analysis for mixed population rna-seq. *PLoS ONE* **8**, e67019 (2013).

57. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.: Ser. B* **57**, 289–300 (1995).

58. Dunn, O. J. Estimation of the means of dependent variables. *Annal. Math. Stat.* 1095–1111 (1958).

59. Dunn, O. J. Multiple comparisons among means. *J. Am. Stat. Assoc.* **56**, 52–64 (1961).

60. Yatsunencko, T. et al. Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).

61. Washburne, A. D. et al. Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. *PeerJ* **5**, e2969 (2017).

62. Washburne, A. D. et al. Phylofactorization: a graph partitioning algorithm to identify phylogenetic scales of ecological data. *Ecol. Monogr.* **89**, e01353 (2019).

63. Silverman, J. D., Washburne, A. D., Mukherjee, S. & David, L. A. A phylogenetic transform enhances analysis of compositional microbiota data. *Elife* **6**, e21887 (2017).

64. Rivera-Pinto, J. et al. Balances: a new perspective for microbiome analysis. *MSystems* **3** (2018).

65. Egozcue, J. J. & Pawłowsky-Glahn, V. Groups of parts and their balances in compositional data analysis. *Math. Geol.* **37**, 795–828 (2005).

66. Pawłowsky-Glahn, V. & Egozcue, J. J. Exploring compositional data with the coda-dendrogram. *Austrian J. Stat.* **40**, 103–113 (2011).

67. Segata, N. et al. Metagenomic biomarker discovery and explanation. *Genome Biol.* **12**, R60 (2011).

68. O'Keefe, S. J. et al. Fat, fibre and cancer risk in african americans and rural africans. *Nat. Commun.* **6**, 6342 (2015).

69. Lahti, L., Shetty, S., Blake, T. & Salojarvi, J. Tools for microbiome analysis in r. version 2.1.28. <https://microbiome.github.io/tutorials/> (2017).

70. Holm, S. A simple sequentially rejective multiple test procedure. *Scand J. Stat.* 65–70 (1979).

ACKNOWLEDGEMENTS

This research was funded by the Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261, USA.

AUTHOR CONTRIBUTIONS

This research work was conceived by S.D.P. All numerical calculations were performed by H.L. Both authors contributed equally in writing the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at <https://doi.org/10.1038/s41522-020-00160-w>.

Correspondence and requests for materials should be addressed to S.D.P.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020