Statistical Theory and Methodology for the Analysis of Microbial Compositions, with Applications

by

Huang Lin

BS, Xiamen University, China, 2015

Submitted to the Graduate Faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2020

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Huang Lin

It was defended on

April 2nd 2020

and approved by

Shyamal Das Peddada, PhD, Professor and Chair, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Jeanine Buchanich, PhD, Research Associate Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Ying Ding, PhD, Associate Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Matthew Rogers, PhD, Research Assistant Professor, Department of Surgery, UPMC Children's Hospital of Pittsburgh

Hong Wang, PhD, Research Assistant Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Dissertation Director: Shyamal Das Peddada, PhD, Professor and Chair, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Statistical Theory and Methodology for the Analysis of Microbial Compositions, with Applications

Huang Lin, PhD

University of Pittsburgh, 2020

Abstract

Increasingly researchers are finding associations between the microbiome and human diseases such as obesity, inflammatory bowel diseases, HIV, and so on. Determining what microbes are significantly different between conditions, known as differential abundance (DA) analysis, and depicting the dependence structure among them, are two of the most challenging and critical problems that have received considerable interest. It is well documented in the literature that the observed microbiome data are relative abundances with excess zeros. These data are necessarily compositional; hence conventional DA methods are not appropriate as they significantly inflate the false discovery rate (FDR), and the standard notion of correlation often results in spurious correlation. To overcome such difficulties, in this dissertation, we develop a general statistical framework that can address a broad collection of problems encountered by researchers.

This dissertation work is organized as follows. In Chapter 1, we conduct a brief review of the literature of a variety of parameters used to characterize microbial composition. Specifically, we shall describe various concepts of diversity and differential taxa abundance.

In Chapter 2, an off-set based regression model, called the Analysis of Composition of Microbiomes with Bias Correction (ANCOM-BC), is introduced. The ANCOM-BC model not only successfully controls the FDR at the desired level but also maintains high power. Simulations and real data analysis were conducted to compare the performance of ANCOM-BC with other commonly used algorithms.

In Chapter 3, we extend ANCOM-BC for performing DA analysis when there are more than two ecosystems. We tested the method for a variety of alternative hypotheses. Similar simulation settings and real data were used to evaluate its performance.

Lastly, in Chapter 4, we introduce a distance correlation based methodology, called

Distance Correlation for Microbiome (DICOM), to untangle dependence structure among microbes within an ecosystem or across ecosystems (e.g., gut and oral microbiomes).

PUBLIC HEALTH SIGNIFICANCE: This dissertation proposes a general statistical framework for studying microbial compositions. The identified differentially abundant taxa and the constructed dependence network could provide medical experts more knowledge of changes in patients' microbiome. This information could contribute to developing precision medicine for better patient care.

Table of Contents

List of Tables

List of Figures

Preface

I would like to thank my advisors Dr. Peddada for his tremendous help and guidance in my dissertation. I really appreciate the opportunity of working with him and really enjoy the time we work together. I also want to express my gratitude to my committee members: Dr. Buchanich, Dr. Ding, Dr. Rogers, and Dr. Wang for their suggestions and comments for my dissertation work. Finally, I would like to thank all my family and friends for their constant support and encourage.

1.0 Introduction

Humans are estimated to have 45.6 million genes in oral and gut microbiome alone, which is about 2000-fold more genes than human genes (Tierney et al., 2019), therefore the microbiome is sometimes referred to as the "second genome", or another "organ" of human body (O'Hara and Shanahan, 2006; Relman and Falkow, 2001; Hurst, 2017). It is hence not surprising that numerous diseases such as obesity (Turnbaugh et al., 2009), inflammatory bowel diseases (Gevers et al., 2014) and HIV (Lozupone et al., 2013a) 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. A taxon is said to be differentially abundant between two ecosystems if its mean *absolute abundances* in the two are significantly different. Estimation of absolute abundance of a taxon in a unit volume of an ecosystem based on a random sample of specimen from the ecosystem is a function of several factors such as the library size (the total number of sequencing reads for all taxa in a sample), microbial load (total number of absolute abundances for all taxa in a unit volume of an ecosystem), and the faction of the sample obtained from the ecosystem.

To reduce the ambiguity, throughout this dissertation, we use *absolute abundance* to denote count of a taxon regardless it is in a unit volume of an ecosystem (e.g. a patient's intestine) or in a sample (e.g. a patient's stool sample); while *relative abundance* is proportion of the absolute abundance of a taxon relative to the total absolute abundance of all taxa, thus it is between 0 and 1. For ease of exposition, various terms used in the literature are summarized in Table 1. The notations described in statistical methods are summarized in Table 2.

The next generation sequencing (NGS) technologies have made the analysis of high-dimensional microbiome data increasingly informative and feasible. There are two common approaches of sequencing performed to study the microbiome: (a) amplification and sequencing of targeted genetic elements such as 16S rRNA gene in bacteria, or (b) shotgun metagenomics. While 16S rRNA sequencing is cost-effective and is very widely used (Amato,

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.
Feature Table	A matrix summarizing observed microbial absolute abundances in the sample. Columns represent samples and rows stand for OTUs or SVs.
Library Size	The total number of (observed) absolute 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.

2017), its main drawback is that it can only identify bacteria. On the other hand, shotgun metagenomics surveys cover all given genomic DNAs, including DNAs from bacteria, viruses, and fungi. Additionally, shotgun metagenomic sequencing has greater taxonomy resolution (species - strains level of shotgun metagenomics vs. genus - species level of 16S sequencing), functional profiling, and it is less susceptible to biases that are inherent in targeted gene amplification. However, as of today, metagenomic sequencing is substantially more expensive

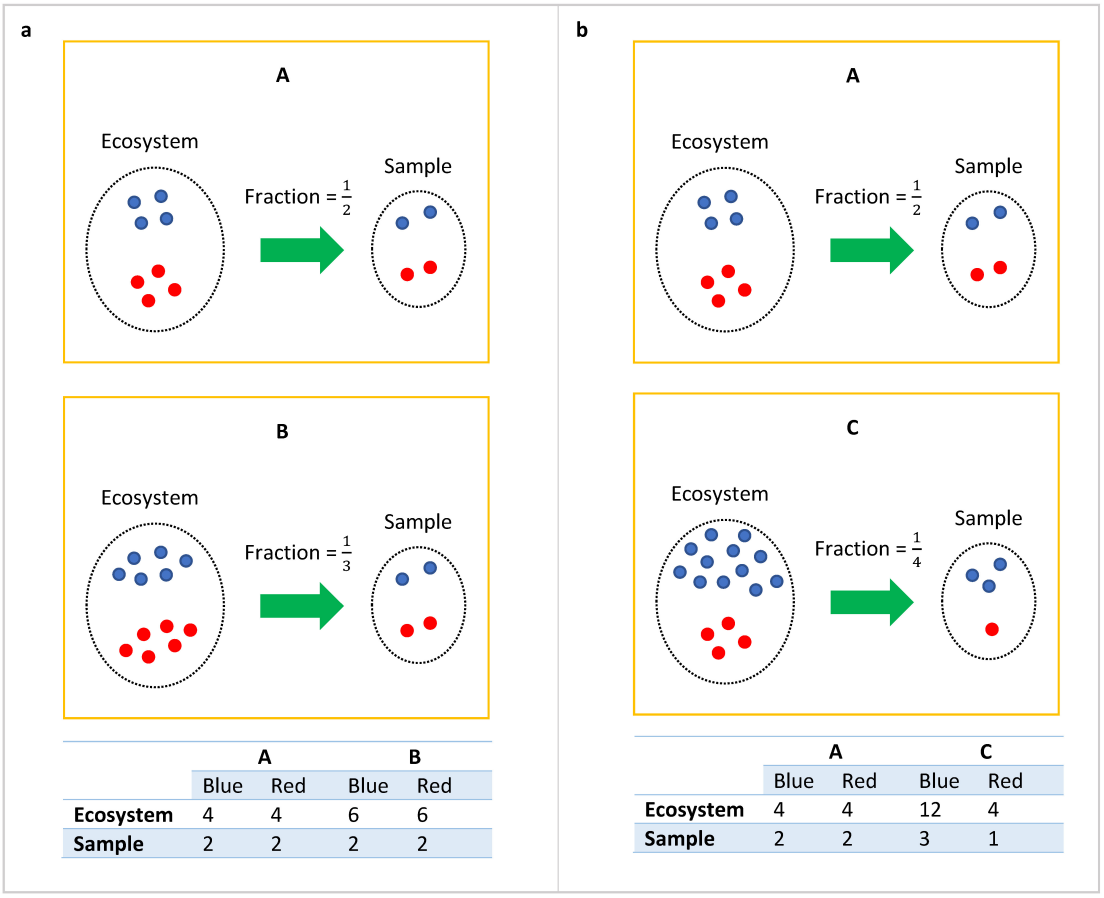
Table 2: Summary of notations.

Notation	**Description**
i	Taxon index, $i = 1, 2, \ldots, m$.
j	Sample index, $j = 1, 2, \ldots, n$.
k	Index of fixed effects, $k = 1, 2, \ldots, p$.
l	Index of Random effects, $l = 1, 2, \ldots, q$.
x_{jk}	The k^{th} fixed effect of interest for the j^{th} sample.
z_{jl}	The l^{th} random effect of interest for the j^{th} sample.
A_{ij}‡	Unobserved absolute abundance of i^{th} taxon in a unit volume of ecosystem of j^{th} sample.
$A_{\cdot j}$‡	Microbial load in a unit volume of ecosystem of j^{th} sample. $A_{\cdot j} = \sum_{i=1}^{m} A_{ij}$.
γ_{ij}‡	Unobserved relative abundance of i^{th} taxon in a unit volume of ecosystem of j^{th} sample.
O_{ij}‡	Observed absolute abundance of i^{th} taxon in a random specimen taken from a unit volume of ecosystem of j^{th} sample.
$O_{\cdot j}$‡	Library size of a random specimen taken from a unit volume of ecosystem of j^{th} sample. $O_{\cdot j} = \sum_{i=1}^{m} O_{ij}$.
r_{ij}‡	Observed relative abundance of i^{th} taxon in a unit volume of ecosystem of j^{th} sample.
c_j†	For the j^{th} sample, c_j represents the proportion of its ecosystem (unobserved absolute abundance) in a random specimen (observed absolute abundance), thus $c_j = \frac{E(O_{ij}\|A_{ij})}{A_{ij}}$. We shall refer to this constant as "sampling fraction".
y_{ij}‡	$\log(O_{ij})$.
d_j†	$\log(c_j)$.

† Parameter;

‡ Random variable.

than 16S sequencing, and it may not be deep enough to detect the 16S rRNA genes of rare species in a community (Shah et al., 2011).



	A		B	
	Blue	Red	Blue	Red
Ecosystem	4	4	6	6
Sample	2	2	2	2

	A		C	
	Blue	Red	Blue	Red
Ecosystem	4	4	12	4
Sample	2	2	3	1

Figure 1: The bias introduced by cross-sample variations in sampling fractions.

The microbiome data are intrinsically compositional because the observed 16S rRNA gene data provides information in the form of relative abundances regardless of the microbial load of ecosystems(Fernandes et al., 2014; Mandal et al., 2015; Gloor and Reid, 2016; Gloor et al., 2016, 2017; Morton et al., 2017, 2019). Thus, they are constrained by a simplex (Aitchison, 1982). 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 ecosystems (Morton et al., 2019). Second, not all samples have the same sampling fraction.

For each taxon i within sample j, the sampling fraction is the ratio of the expected absolute abundance of taxon i within the j^{th} sample (e.g. a stool sample) to its absolute abundance in a unit volume of the ecosystem (e.g. gut) where the sample was derived from. The sampling fraction is constant for all taxa i within the j^{th} sample. Thus the sampling fraction for the j^{th} sample is given by the following.

Definition 1.0.1 (Sampling fraction)**.**

$$c_j = \frac{E(O_{ij}|A_{ij})}{A_{ij}}, \tag{1.1}$$

where

(1) O_{ij} is the observed absolute 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 the differential abundance (DA) analysis of microbiome data is that while O_{ij} is known, A_{ij} is unknown and can vary drastically from sample to sample. Consequently, the observed absolute abundances are not comparable between samples. The goal of DA analysis is to identify taxa whose absolute abundances, per unit volume, of the ecosystem (A_{ij}) are significantly different with changes in the covariate of interest (e.g. the group effect).

Consider the toy example in Figure 1, suppose the ecosystems (e.g. gut) of subject A, B and C consist of only two taxa, the blue and red taxa. 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 ecosystems, blue is more abundant in C than in A (12 vs. 4), and both have same amounts of red taxa (4 vs. 4). However, given that samples

from A and C have same library sizes, one may mistakenly conclude that both blue (2 vs. 3) and red taxa (2 vs. 1) are differently abundant between A and C. A third characteristic of feature table is that it is typically sparse, with as many as ∼90% zero entries (Paulson et al., 2013), 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 (Mandal et al., 2015; Xia et al., 2013) to each cell of the feature table. Even though adding a pseudo-count is simple and also widely used, the choice of the pseudo-count is often ad-hoc. Other strategies involve modeling zero counts by some probability models (Paulson et al., 2013; Chen and Li, 2016). However, these methods may not be valid if the underlying parametric assumption does not hold. Instead of modeling zeros by parametric distributions, ANCOM-II[512017Kaul et al.] 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 experimental groups, i.e. some bacteria are not expected to belong to certain environments (e.g. a desert) but in others (e.g. a rainforest), and *sampling zeros* owing to insufficient library size. In our opinion the zero counts problem is still an open problem and requires further investigation.

1.1 Measures of diversity

1.1.1 Alpha diversity

The alpha diversity (α-diversity) is a measure of diversity within a sample (Whittaker, 1960, 1972). One of the simplest and widely used measures to represent diversity is *richness* which is the number of taxa present in a sample (Magurran, 2013). Whereas, *evenness* is a measure of relative abundance of different taxa that make up the richness in that sample. Low values of evenness indicates that a small number of taxa dominate the composition, and high values indicates that relative abundances of different taxa are somewhat evenly distributed. For example, consider the guts of two 10-day old babies A and B. Suppose the stool sample from A has 10% Actinobacteria, 15% Bacteroidetes, 5% Proteobacteria and 70%

Firmicutes, and suppose the stool sample from B has 1% Actinobacteria, 0% Bacteroidetes, 5% Proteobacteria and 94% Firmicutes. In this example, eye-balling the relative abundances, we may conclude that baby A has higher evenness than baby B. Such eye-ball comparisons are not feasible when there are a large number of taxa. A variety of alpha diversity measures are available to quantify abundance and evenness. They can be broadly classified into two types, those that take into account the phylogenic relationships and those that do not. Some widely used non-phylogeny based metrics are Chao1 (Chao, 1984), Shannon's diversity (Shannon, 1948) and Gini-Simpson's index (Simpson, 1949). Among these, the latter two are commonly used in practice. These indices, take into account taxa's richness, relative abundance and evenness(Morris et al., 2014).

A popular phylogeny based metric is Faith's Phylogenetic Diversity (PD)(Faith, 1992) which is defined as follows.

Definition 1.1.1 (Phylogenetic diversity (PD))**.** As a quantitative measure of phylogenetic diversity, "PD" has been defined as the minimum total length of all the phylogenetic branches required to span a given set of taxa on the phylogenetic tree.

Unlike non-phylogeny based metrics, PD is a measure of biodiversity which incorporates phylogenetic difference among taxa. Phylogenetic patterns among taxa reflect general patterns of taxa variation at the level of genes or other features (Faith and Baker, 2006). Larger PD values are expected to correspond to greater feature diversity.

For convenience, the mentioned alpha diversity measures are summarized in Table 3. In Figure 2, we show different alpha diversity measures using the same data set, which is based on the two-week diet swap study between western (USA) and traditional (rural Africa) diets (O'Keefe et al., 2015). Note that since Shannon's and Gini-Simpson's diversity indices are based on similar principles, they are expected to be highly correlated in any given data set.

1.1.2 Beta diversity

Beta diversity (β-diversity) provides a measure of between-sample diversity, or distance or dissimilarity (Whittaker, 1960). When more than two samples are used, the beta diversity is calculated for every pair of samples to generate a distance/dissimilarity matrix. Similar

Table 3: Formulas for calculating alpha diversities.

Category	**Metric**	**Formula**
Non-phylogeny	Observed species	Count of unique taxa in a sample
	Chao1	$S_j = \begin{cases} m + \frac{f_1^2}{2f_2} & f_2 > 0 \\ m + \frac{f_1(f_1-1)}{2} & f_2 = 0 \end{cases}$
	Shannon's diversity index	$H_j = -\sum_{i=1}^{m} r_{ij} \log r_{ij}$
	Gini-Simpson's diversity index	$D_j = 1 - \sum_{i=1}^{m} r_{ij}^2$
Phylogeny	Phylogenetic diversity (PD)	Minimum total branch length of the phylogenetic tree that incorporates all taxa in a sample

Note that:
(1) f_1 = number of singletons (taxon appear once) in the sample,
(2) f_2 = number of doubletons (taxon appear twice) in the sample.

to alpha diversity, the beta diversity can be categorized into non-phylogeny based metrics, such as Bray-Curtis dissimilarity (Bray and Curtis, 1957), Jaccard distance (Jaccard, 1901), and phylogeny based metrics such as Unweighted UniFrac (Lozupone and Knight, 2005) and Weighted UniFrac (Lozupone et al., 2007). For simplicity of exposition, suppose we have two samples, i.e. sample A and B, the mentioned beta diversity measures are summarized in Table 4.

Among non-phylogeny based beta diversities, Bray-Curtis dissimilarity is constructed using the observed absolute abundance (count data) and it ranges from 0 to 1, with 0 corresponding to the case when A and B have identical observed absolute abundance of all

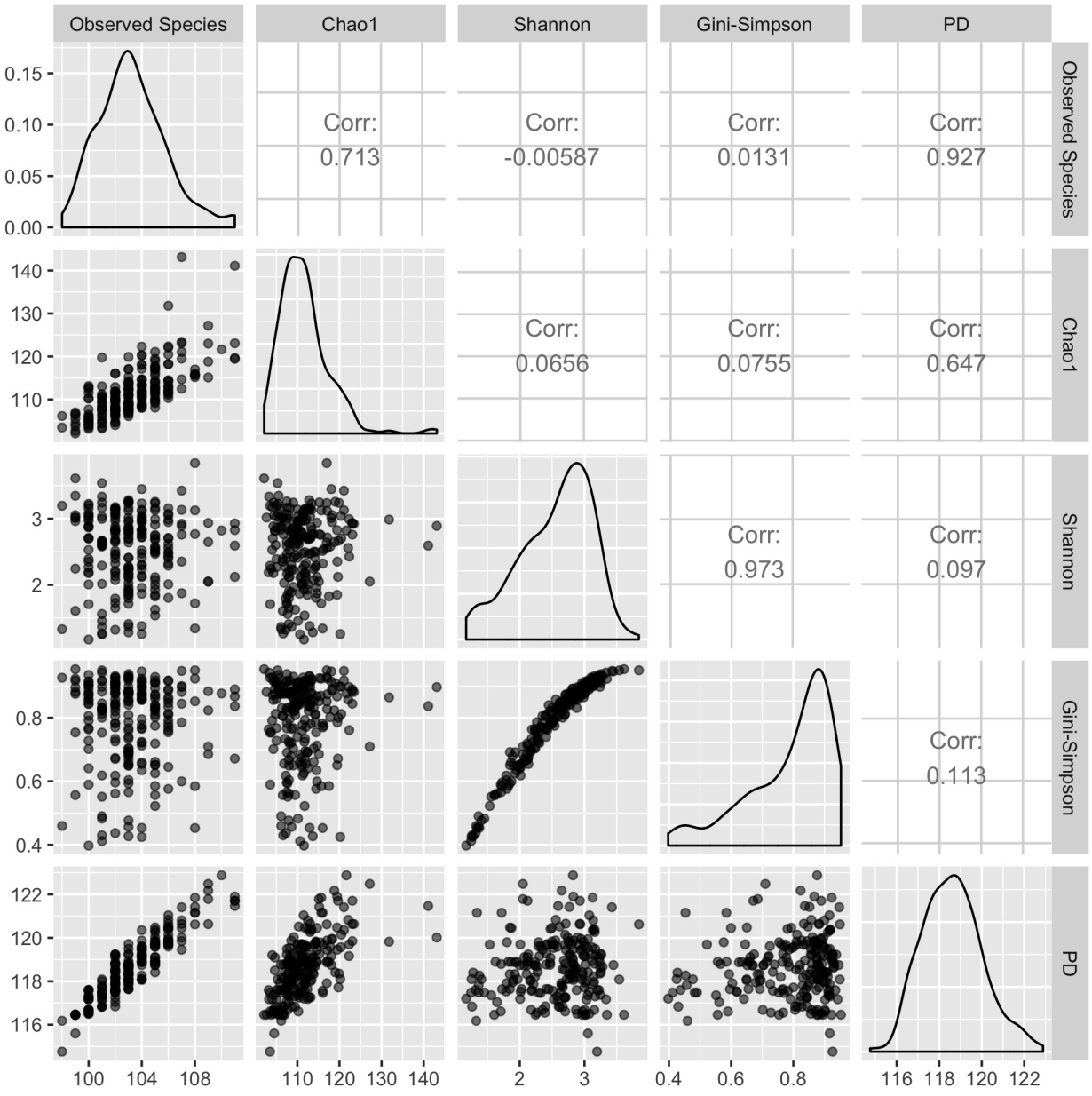


Figure 2: Different alpha diversity measures using the diet swap data at the genus level.

taxa, and 1 corresponds to the case when the two samples have complete different observed absolute abundances. Thus larger value correspond to more diversity between samples. On the other hand, Jaccard distance is a dissimilarity measure for presence or absence of taxa

Table 4: Formulas for calculating beta diversities.

Category	**Metric**	**Formula**
Non-phylogeny	Bray-Curtis dissimilarity	$BC_{AB} = \frac{\sum_{i=1}^{m} \lvert O_{iA} - O_{iB} \rvert}{\sum_{i=1}^{m} (O_{iA} + O_{iB})}$
	Jaccard distance	$J_{AB} = 1 - \frac{\sum_{i=1}^{m} I(O_{iA}>0) I(O_{iB}>0)}{\sum_{i=1}^{m} I(O_{iA}+O_{iB}>0)}$
Phylogeny	Unweighted UniFrac	$UU_{AB} = \frac{\sum_{b=1}^{B} l_b \lvert I(A_b>0) - I(B_b>0) \rvert}{\sum_{b=1}^{B} l_b}$
	Original Weighted UniFrac	$WU_{AB} = \sum_{b=1}^{B} l_b \lvert \frac{A_b}{A_T} - \frac{B_b}{B_T} \rvert$
	Normalized Weighted UniFrac	$WU_{AB} = \sum_{b=1}^{B} \frac{l_b \lvert \frac{A_b}{A_T} - \frac{B_b}{B_T} \rvert}{l_b (\frac{A_b}{A_T} + \frac{B_b}{B_T})}$

Note that:

(1) $I(O_{ij} > 0)$ is the indicator function which equals to 0 or 1 as taxon i is absent or present in sample j,

(2) B = the total number of branches in the phylogenetic tree,

(3) l_b = the length of branch b,

(4) j_b = total number of descendants of branch b from sample j,

(5) $I(j_b > 0)$ = indicator equal to 0 or 1 as descendants of node b absent or present in samples j.

without taking into account the abundance information. It ranges from 0 to 1, with 0 implies that the two samples share exact the same taxa, and 1 implies there is no common taxa.

The larger value the more diverse the data.

Unweighted UniFrac and Weighted UniFrac are two popular phylogeny based diversity measures which are calculated using sequence distances in the phylogenetic tree. They are based on the fraction of branch length that is shared between two samples or unique to one or the other sample. Unweighted UniFrac is purely based on sequence distances so it does not include abundance information, while Weighted UniFrac includes both sequence and abundance information by weighting branch lengths using relative abundances. Unweighted UniFrac and (Normalized) Weighted UniFrac range from 0 to 1, the larger value corresponds to larger diversity.

1.1.3 Analysis of Diversity (ANODIV)

While the alpha and beta diversities are well studied in the microbiome literature, Rao's Quadratic Entropy (Rao, 1984; Nayak, 1986; Ricotta and Marignani, 2007; Rao, 2010; Chen et al., 2018) and the resulting Analysis of Diversity (ANODIV) has not been well discussed in the microbiome literature. The ANODIV resembles the classical ANOVA but is based on diversity measures. Although a variety of diversity measures may be used in defining quadratic entropy, for simplicity of exposition, in this paper we shall use Gini-Simpson index when defining Rao's quadratic entropy. In practice one may consider more informative measures depending on the available information and scientific question. Analogous to ANOVA, the ANODIV provides a general framework for analyzing complex designs including multifactorial studies with covariate adjustments (Nayak, 1986), (Rao, 2010). As demonstrated in (Rao, 2010) the total diversity (SST) can be partitioned into various components such as within group (SSW) and between group (SSB), which can be further decomposed into other components depending upon the study design. Based on the asymptotic theory developed in (Nayak, 1986), one can formally test the null hypothesis that the compositions of two or more ecosystems are same. Thus, the classical machinery of ANOVA or ANCOVA can be easily imported into ANODIV. Although, mathematically, the asymptotic theory developed by Nayak (Nayak, 1986) is designed for the case when the sample sizes are larger than the number of taxa, there is an opportunity to extend those results for high-dimensional mi-

crobiome data. If the analysis are carried out at higher order of the phylogeny, say at the phylum or class or perhaps even order level, where the sample sizes might be larger than the number of taxa (i.e. phylum, class or order) then ANODIV can be applied directly.

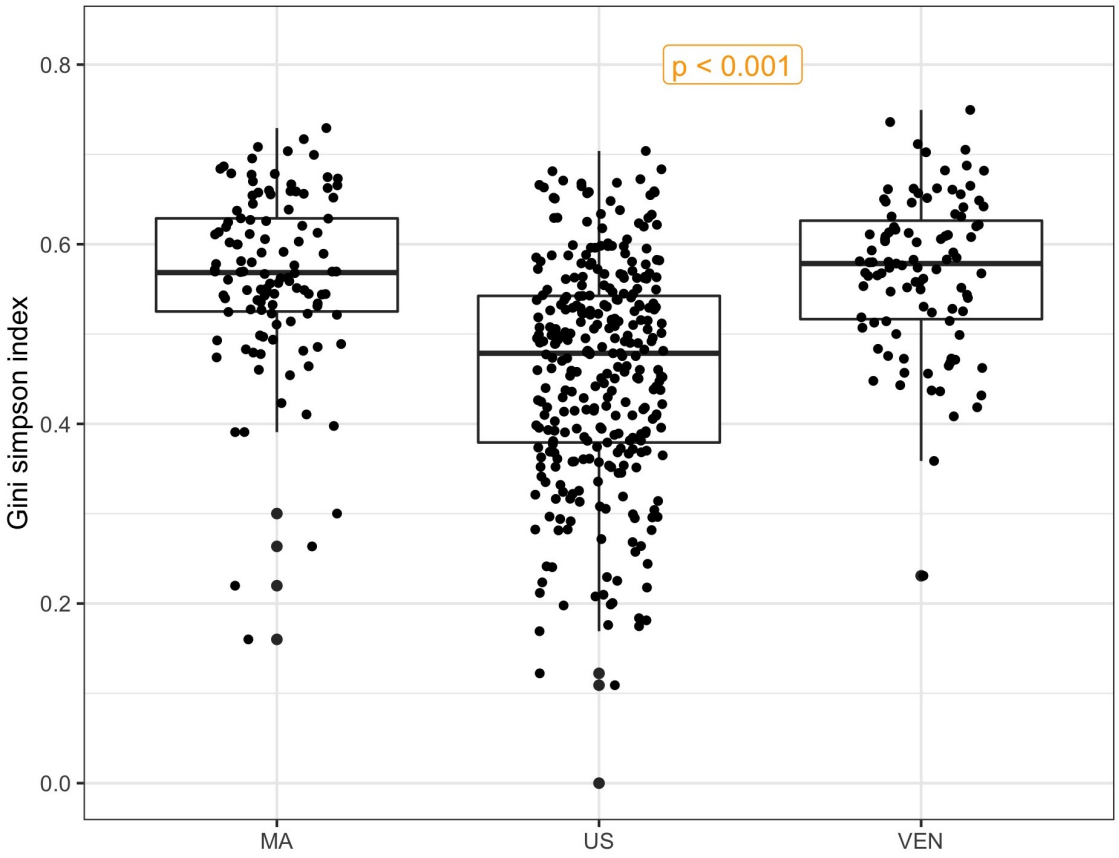


Figure 3: Box plot of Rao's quadratic entropy.

Let $\hat{P}_g = (\hat{P}_{g1}, \hat{P}_{g2}, \ldots, \hat{P}_{gm})^T$ denote the sample proportions of m taxa in the g^{th} group, which are estimated using n_g observations in the g^{th} group, $g = 1, 2, \ldots, G$. Let $\hat{P}_.$ denote the weighted average of $\hat{P}_1, \hat{P}_2, \ldots, \hat{P}_G$, weights being the sample sizes $n_1, n_2, \ldots, n_G$. For a vector of proportions P, define $H(P) = P^T \Delta P$, where Δ is a suitably chosen $m \times m$ matrix. For example, in the case of Gini-Simpson index, which is used in this paper, $\Delta = J - I$, J is a matrix where all elements are 1 and I is the identity matrix. Then the total diversity is defined as $SST = H(\hat{P}_.)$ and within-sample diversity is $SSW = \sum_{g=1}^{G} \frac{n_g}{n} H(\hat{P}_g)$ and the between sample diversity is $SSB = SST - SSW = -\sum_{g=1}^{G} \frac{n_g}{n} (\hat{P}_g - \hat{P}_.)^T \Delta (\hat{P}_g - \hat{P}_.)$. Note that since $H(P)$ defined above is a concave function, therefore SSB, SSW and SST are all

non-negative. Under the null hypothesis that all populations have same relative abundance of all taxa, the asymptotic distribution of $(g-1)(n-1)SSB/SST$ is central $\chi^2_{(g-1)(m-1)}$.

To illustrate ANODIV, we use the global gut microbiota data (Yatsunenko et al., 2012) and analyze it at the phylum level. Using data on 39 phyla with subjects from Malawi (MA, $n_1 = 114$), USA (US, $n_2 = 317$), and Venezuela (VEN, $n_3 = 99$), we found the total diversity (SST) in the three samples to be 0.574, the within sample diversity SSW = 0.567, and the between sample diversity SSB = 0.007, with a p-value of 4.30×10^{-18}, we reject the null hypothesis that the phyla compositions are same among the three countries. The box plot in Figure 3 seems to confirm this finding. Although the variation within each box is very large, which seems to be consistent with very large SSW, and numerically SSB is small, three box plots appear to be significantly different, and the statistical test is sensitive to find a statistically significant p-value. Thus the ANODIV is a useful method to test hypotheses regarding the equality of microbial compositions in two or more groups.

1.2 Differential abundance analysis

1.2.1 Normalization methods

As we described intuitively in the introduction, a main obstacle for performing DA analysis is the unknown sampling fractions. Therefore, normalization is critical to enable meaningful comparison of absolute abundances from different experimental conditions by eliminating artificial biases caused by the variability of sampling fractions. The primary objective of normalization is to transform the observed absolute abundances in samples so that expected differences in the mean absolute abundances between two ecosystems is not confounded by the differences in the sampling fractions. Thus correcting for the bias induced by differential sampling fractions should be an important objective of a normalization procedure. Failure to do so will result in a systematic bias that increases the false discovery rate (FDR) and loss of power.

Rarefaction or subsampling, is a widely used normalization method in the microbiome

studies. It was first recommended for microbiome data in order to moderate differences in the presence of rare taxa (Lozupone et al., 2011). Rarefaction curves represent the diversity as a function of library sizes. If the lines in the plot appear to "level out" (i.e., approach a slope of zero) at some library size along the x-axis, that suggests that the diversity of the samples has been fully observed or sequenced. Otherwise, increasing the minimum library size would be likely to result in the observation of additional features. Originally, the diversity metrics used in rarefaction curves was alpha diversity (Gotelli and Colwell, 2001; Brewer and Williamson, 1994). However, recent years have seen studies using beta diversities (Horner-Devine et al., 2004; Jernvall and Wright, 1998) as well. Although rarefaction has been criticized for potential loss of statistical power when a relatively large proportion of data is removed, some studies (Weiss et al., 2017) have demonstrated that it remains to be a promising technique for ordination/clustering and that control of false positive rate due to rarefaction outweighs any loss in power.

Scaling the data is another popular method of normalization of microbiome data. The basic idea is to multiply each element in the feature table by a "normalization factor" to eliminate biases resulting from unequal sampling fractions. Some commonly used normalization methods include Cumulative-Sum Scaling (CSS) implemented in metagenomeSeq (Paulson et al., 2013), Median (MED) in DESeq2 (Love et al., 2014), Upper Quartile (UQ) (Bullard et al., 2010) and Trimmed Mean of M-values (TMM) (Robinson and Oshlack, 2010) in edgeR (Robinson et al., 2010), Wrench (Kumar et al., 2018), and Total-Sum Scaling (TSS) that simply transforms the abundance table (feature table) into relative abundance table, i.e. scale by each sample's library size. Note that as stated in the user manual of edgeR (Chen et al., 2014), the author suggests that to address the "RNA composition" effect, one should multiply the normalization factors with the corresponding library size to account for "effective library size". Hence, we also considered modified versions of UQ and TMM, denoted as "ELib-UQ" (Effective library size using UQ) and "ELib-TMM" (Effective library size using TMM) in this paper. Since the literature is not often very explicit regarding the mathematical formulas used by various methods, we provide some useful formulas in Table 5.

TSS is known to have bias in differential abundance estimates (Bullard et al., 2010; Dillies

Table 5: Summary of different normalization methods.

Method	**Sampling Fraction Estimate**
ANCOM-BC	$\log(\hat{c}_j^{\text{ANCOM-BC}}) = \frac{1}{m}\sum_{i=1}^{m}(y_{ij} - x_j^T\hat{\beta}_i)$
CSS	$\hat{c}_j^{\text{CSS}} = \frac{s_j^{\hat{l}}+1}{N}$
MED	$\hat{c}_j^{\text{MED}} = \underset{i:O_i^R\neq 0}{\text{median}}\frac{O_{ij}}{O_i^R}$
UQ	$\hat{c}_j^{\text{UQ}} = \underset{i:O_{ij}>0}{\text{UQ}}(\frac{O_{ij}}{O_{\cdot j}})$
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	$\hat{c}_j^{\text{Elib-UQ}} = O_{\cdot j}\hat{c}_j^{\text{UQ}}$
Elib-TMM	$\hat{c}_j^{\text{Elib-TMM}} = O_{\cdot j}\hat{c}_j^{\text{TMM}}$
Wrench	$\hat{c}_j^{\text{Wrench}} = \frac{1}{m}\sum_{i=1}^{m} b_{ij}\frac{r_{ij}}{\bar{r}_{i\cdot}}$
TSS	$\hat{c}_j^{\text{TSS}} = O_{\cdot j}$

Where

(1) $\hat{\beta}_i$ is obtained from ANCOM-BC algorithm,

(2) N = an approximately choose normalization constant,

(3) $s_j^{\hat{l}} = \sum_{i:O_{ij}\leq q_j^{\hat{l}}} O_{ij}$,

(4) $q_j^{\hat{l}} = \hat{l}^{th}$ quantile of sample k,

(5) $O_i^R = (\prod_{j=1}^{n} O_{ij})^{\frac{1}{n}}$,

(6) UQ(X) denotes the upper quartile of X,

(7) $M_{ij} = \log_2(\frac{O_{ij}}{O_{\cdot j}}) - \log_2(\frac{O_{ij'}}{O_{\cdot j'}})$, where j' is the reference sample,

(8) $w_{ij} = \frac{O_{\cdot j}-O_{ij}}{O_{\cdot j}O_{ij}} + \frac{O_{\cdot j'}-O_{ij'}}{O_{\cdot j'}O_{ij'}}$, where j' is the reference sample,

(9) $G*$ represents a set of taxa that were not considered as extreme data for fold-change (M values) and average intensity (A values).

(10) b_{ij} represents the taxon-specific weight.

et al., 2013) since a few preferentially sampled 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. The Cumulative-Sum Scaling (CSS) (Paulson et al., 2013) in metagenomeSeq modifies Total-Sum Scaling (TSS) in a sample-specific manner to reduce biases resulting from preferentially sampled taxa. CSS assumes that read counts of all samples should be roughly independent and identically distributed up to a specific quantile. The Median normalization (MED) method used in DESeq2 (Love et al., 2014) assumes that a large proportion of taxa are not differentially abundant. While this may be a reasonable assumption in gene expression studies where most genes are not differentially expressed, but in the case of microbiome data it is not a reasonable assumption. Depending upon the application, a very large proportion of taxa may be differentially abundant between two conditions. The Upper Quartile normalization (UQ) and the Trimmed Mean of M-values (TMM) used in edgeR have similar issues as MED in DESeq2. UQ assumes that the upper quartile can capture the invariant segment of the count distribution, however, choosing the most effective quantile is nontrivial (Paulson et al., 2013; Robinson et al., 2010; Bullard et al., 2010; Dillies et al., 2013; Anders and Huber, 2010; Agresti and Hitchcock, 2005). Similarly, 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 absolute 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 (absolute abundance level). Wrench (Kumar et al., 2018) assumes the observed absolute 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 most taxa do not change across conditions, or the effect sizes of differentially abundant taxa are not too large. One must exercise caution when using scaling methods as well. Most importantly, a scaling method is likely to overestimate or underestimate the fractions of zero counts depending on the corresponding library size of each sample (Friedman and Alm, 2012; Agresti and Hitchcock, 2005). This problem becomes more obvious for microbiome data since its feature table is typically sparse.

In this dissertation, we proposed a novel method called Analysis of Compositions of Microbiome with Bias Correction (ANCOM-BC), which will be discussed in detail in the next two chapters, it assumes that the observed sample is an unknown fraction of a unit volume of the ecosystem, and the sampling fraction varies 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 data. The offset term serves as the bias correction.

Finally, Aitchison's log-ratio transformation (Aitchison, 1982) implemented in in methods such as ALDEx2 (Fernandes et al., 2014), ANCOM (Mandal et al., 2015), and DR (Morton et al., 2019), is another alternative normalization method for compositional data. By taking log-ratios on observed absolute abundances or relative abundances within each sample, one is eliminating the effect of sampling fraction inherent to a given sample. There are three obvious choices for the log-ratio transformation, described below:

Definition 1.2.1 (additive log-ratio transformation (alr) (Aitchison, 1982), $\mathbb{S}^m \to \mathbb{R}^{m-1}$)**.**

$$alr(O_j) = [\log(\frac{O_{1j}}{O_{i'j}}), \ldots, \log(\frac{O_{mj}}{O_{i'j}})], \tag{1.2}$$

where

(1) O_j is the observed absolute abundances for sample j,

(2) i' is taken to be the reference taxon.

Since alr projects the observed absolute abundances, which originally reside in a m dimensional simplex, into $m-1$ dimensional Euclidean space, standard calculus of Euclidean geometry becomes valid. Note that alr transformation is an isomorphism, but not isometry, meaning that distances on transformed values will not be equivalent to distances on the original compositions in the simplex. One apparent drawback with alr is the choice of reference taxon (Morton et al., 2019). For different reference taxa, one gets different interpretations of the data.

The ambiguity of the chosen of reference taxon can be reduced by selecting the center-of-mass as the reference, allowing a one-to-one transformation of all taxa. This can be achieved by the so-called centered log-ratio transformation (clr):

Definition 1.2.2 (centered log-ratio transformation (clr) (Aitchison, 1982), $\mathbb{S}^m \to \mathbb{U}^m$)**.**

$$clr(O_j) = [\log(\frac{O_{1j}}{g(O_{ij})}), \ldots, \log(\frac{O_{mj}}{g(O_{ij})}], \tag{1.3}$$

where

(1) O_j is the observed absolute abundances for sample j,

(2) $g(x)$ is the geometric mean of x,

(3) $U^m = \{[u_1, \ldots, u_m] : u_1 + \ldots + u_m = 0\}$ is a hyperplane of $\mathbb{R}^m$.

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.

Neither alr nor clr transformation can be directly linked to an orthogonal coordinate system in the simplex. The isometric log-ratio transformation (ilr) (Egozcue et al., 2003) transformation (also known as balance), which is an isometry between $\mathbb{S}^m$ and $\mathbb{R}^{m-1}$, provides a solution to this problem.

Definition 1.2.3 (isometric log-ratio transformation (ilr), $\mathbb{S}^m \to \mathbb{R}^{m-1}$)**.**

$$ilr(O_j) = clr(O_j)\Psi^T, \tag{1.4}$$

where

(1) O_j is the observed absolute abundances for sample j,

(2) Ψ is a $(m-1, m)$-matrix of 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, where each element in the ilr transformed data is of the following form:

$$b_j = \sqrt{\frac{|j_L||j_R|}{|j_L| + |j_R|}} \log[\frac{g(j_L)}{g(j_R)}], \tag{1.5}$$

where

(1) b_j is the balance at internal node j,

(2) j_L is the set of relative abundances contained in the left subtree at internal node j,

(3) j_R is the set of relative abundances contained in the right subtree at internal node j ,

(4) $|j_L|$ is the number of taxa contained in j_L,

(5) $|j_R|$ is the the number of taxa contained in j_R,

(6) $g(x)$ is the geometric mean of x.

One caveat of applying log-ratio transformation is the choice of pseudo count. Because of the nature of log transformation, the addition of a pseudo count is necessary to handle sampling zeros. Studies have shown that differential abundance or clustering results could be sensitive to the choice of pseudo count (Costea et al., 2014; Paulson et al., 2014). Although different values of pseudo count have also been exhaustively discussed (Egozcue et al., 2003; Costea et al., 2014; Paulson et al., 2014; Greenacre, 2011), to our best knowledge, there is no consensus on how to choose the optimal value.

1.2.2 Methods of differential abundance analysis

One of the objectives in this dissertation is to identify taxa that are differentially abundant between two or more groups, and determine the biological functions and processes associated with such taxa. 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 test for two groups; the Kruskal-Wallis test for multiple groups) after rarefying the feature table. Unfortunately, these standard nonparametric tests do not take into account the compositional structure of microbiome data.

As alternatives to standard nonparametric tests, many parametric models have been proposed in the literature based on transcriptomics data such as RNA-Seq data for testing differences across experimental groups. Among them, DESeq2 (Love et al., 2014) and edgeR (Robinson et al., 2010) are two most popular methods. They both model the count data using negative binomial distributions to allow for extra variation, and use shrinkage estimation for dispersions to improve stability and reliability of estimates.

1.2.2.1 RNA-seq based methods: DESeq2 and edgeR Both DESeq2 and edgeR model the observed absolute abundances using the negative binomial distribution. While both methods are in general 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 deferentially abundant. This assumption is not valid for microbiome data. As a consequence, the test statistics used by these methods are biased under the null hypothesis. As demonstrated in several previous studies (Mandal et al., 2015; Weiss et al., 2017), the bias in the test statistic results in inflated FDRs for these methods. What is worse, as the sample size increases, the FDR increases for these methods.

1.2.2.2 MetanegomeSeq Instead of using a negative binomial model, metagenomeSeq (Paulson et al., 2013) used a zero-inflated Gaussian mixture (ZIG) model, with the zero mass tackling excess zeros due to insufficient sequencing depth or biological nature, and the Gaussian distribution modeling the non-zero counts. However, as shown in simulation studies (Weiss et al., 2017), metagenomeSeq was the only method, among all parametric models, that increased FDR when using rarefied data. This might due to its zero-inflated model which requires the raw library size to capture the zero proportion. Even with its own normalization method (CSS), metagonomeSeq still has a highly inflated FDR, and it gets worse when sample size or the fold change increases (Mandal et al., 2015; Weiss et al., 2017).

The authors of metagenomeSeq, modified their procedure and recommend using zero-inflated Log-Gaussian (ZILG) mixture model instead of zero-inflated Gaussian (ZIG) mixture model for each feature. Although this improves the FDR control, the procedure becomes extremely conservative, with FDR close to zero and substantial loss in power.

1.2.2.3 ALDEx2 Inherited from the original version of ANOVA-Like Differential Expression (ALDEx) analysis (Fernandes et al., 2013), ALDEx2 was proposed as a compositional data analysis tool that is applicable to three different types of data: RNA-Seq, ChIP-Seq and 16S rRNA sequencing (Fernandes et al., 2014). By acknowledging these high-throughput sequencing datasets are fundamentally compositional, the methodology of

ALDEx2 can be summarized as follows:

(1) The observed absolute 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 absolute abundance O_{ij} is represented by a vector of MC samples of relative abundances $(r_{ij}^{(1)}, \ldots, r_{ij}^{(K)})^T$,

(2) Within each sample j and each MC Dirichlet realization $k, k = 1, \ldots, K$, the relative abundances $(r_{1j}^{(k)}, \ldots, r_{mj}^{(k)})^T$ is clr transformed giving a vector of transformed values,

(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-H[71995Benjamini and Hochberg] 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) between two or more sample groups, relative to the *geometric mean absolute abundance*. Thus, the parameter of interest in ALDEx2 is different from the parameter of interest in DA analysis. Throughout this dissertation, a differentially abundant taxon is the one whose mean *absolute abundance* in the *ecosystem* is significantly different with regard to the covariate of interest. As a result, ALDEx2 not only generally exceeds the nominal level of FDR (5%), but also has substantially smaller power as compared to competing DA methods (Morton et al., 2019).

1.2.2.4 Analysis of composition of microbiomes (ANCOM) is an Aitchison's log-ratio based methodology (Aitchison, 1982), which accounts for the compositional structure of microbiome data. Suppose there are a total of m taxa, ANCOM relies on two mild assumptions as follows. Under these assumptions, the authors proved that one can test the null hypotheses regarding absolute abundance in a unit volume using relative abundances.

Assumption 1.2.1. *The mean abundance (in log scale) of at most $m-2$ taxa are different. Thus, some two taxa are assumed to be not differentially abundant.*

Assumption 1.2.2. *The mean abundance (in log scale) of all m taxa do not differ by the same amount between the two study groups.*

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

$$\log \frac{r_{ij}^{(g)}}{r_{i'j}^{(g)}} = \alpha_{ii'} + \beta_{ii'}^{(g)} + \sum_k x_{jk}\beta_{ii'k} + \epsilon_{ii'j}^{(g)}, \tag{1.6}$$

where i' is the reference taxon, $i' \neq i = 1, 2, \ldots, m$, $g = 1, 2, \ldots, G$ is the number of groups, $\alpha_{ii'}$ is the overall common mean, $\beta_{ii'}^{(g)}$ is factor of interest at the g^{th} level, x_{jk} are adjusting covariates indexed by k. $\epsilon_{ii'j}^{(g)} \sim N(0, \sigma_{ii'}^2)$.

By virtue of Assumption 1.2.1 and Assumption 1.2.2, to test whether a taxon i is differentially abundant according to a factor of interest with G levels, it is equivalent to test:

$$H_{0(ii')} : \beta_{ii'}^{(1)} = \ldots = \beta_{ii'}^{(G)} = 0,$$
$$H_{1(ii')} : \text{Not all } \beta_{ii'}^{(g)} \text{ equals to } 0,$$

for every $i \neq i'$.

Altogether $\frac{m(m-1)}{2}$ distinct null hypotheses $H_{0(ii')}$, $i \neq i'$ are tested using a multiple testing correction such as the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995). For each taxon, the number of rejections, denoted as W_i, is counted, and ANCOM makes use of the empirical distribution of $\{W_1, W_2, \ldots, W_m\}$ to determine the cut-off value of significant taxon. The rule of thumb is when the value of W_i is larger, it is more likely that taxon i tends to be differentially abundant. The author recommends using 70 percentile of the W distribution as the empirical cut-off value.

As shown by simulation studies (Mandal et al., 2015; Weiss et al., 2017), ANCOM successfully controls the FDR under the nominal level (5%) while maintaining adequate power. However, ANCOM can be computationally intensive especially if the number of taxa is large. In addition, the statistical decision made by ANCOM depends on the quantile of its test statistic rather than quantitative measures such p-values, which some biologists find it difficult to interpret.

1.2.2.5 Differential Ranking (DR) exploits the fact that the ranks of relative differentials are identical to the ranks of absolute differentials. They estimate relative differentials using a multinomial regression where OTUs/SVs are the explanatory variables in the model. The regression coefficients corresponding to different taxa are ranked in order to determine the most important to least important taxa.

The multinomial regression model is formulated using additive log-ratio (ALR) transformation:

$$\begin{aligned} \beta_{ik} &\sim N(0, \mu_\beta) \\ \eta_j &= \text{alr}^{-1}(\beta_i^T x_j) \\ O_{\cdot j} &\sim \text{Multinomial}(\eta_j), \end{aligned} \tag{1.7}$$

The model parameters are estimated using a maximum a posteriori priori (MAP) estimation by stochastic gradient descent. Since the regression parameters are estimated under the constraint that they sum to 0, this method does not require to pre-specify the reference taxon and hence is robust to the choice of reference taxon. Secondly, it does take into account the compositional structure of microbiome data.

Note that, unlike other methods which use a p-value, this method makes decisions solely on the magnitude of β_{ik} and the ranks of taxa derived from there.

1.2.2.6 Gneiss (Morton et al., 2017) is different from all above DA methods in the sense that it aims to move away from identifying differential abundance properties of individual taxon; Instead, in its implementations, it explore the concept of balances (Egozcue and Pawlowsky-Glahn, 2005; Pawlowsky-Glahn and Egozcue, 2011) to infer meaningful properties of sub-communities. It is important to note that gneiss is not designed to infer changes of absolute abundance for each individual taxon, but it can limit the number of possible scenarios with regards to the absolute changes of a group of taxa.

2.0 Analysis of Compositions of Microbiomes with Bias Correction (ANCOM-BC)

2.1 Introduction

As introduced in the last chapter, a number of procedures have been proposed and used in the literature for identifying deferentially abundant taxa between two or more ecosystems. A detailed survey of some of the existing methods and their performance has been discussed in (Weiss et al., 2017). As noted in a list of studies (Mandal et al., 2015; Gloor and Reid, 2016; Gloor et al., 2016, 2017; Morton et al., 2019), the observed microbiome data are relative abundances, hence they are compositional. Standard statistical methods are not appropriate for analyzing compositional data (Aitchison, 1982). Methods such as ANOVA, Kruskal-Wallis test do not appropriately take into consideration the compositional feature of microbiome data when performing differential abundance (DA) analysis. As demonstrated in literatures (Weiss et al., 2017; Mandal et al., 2015), these methods are subject to inflated false discovery rates (FDR). Although metagenomeSeq (Paulson et al., 2013) was specifically developed for microbiome data, it too is subject to inflated FDR under the Gaussian mixture model (Weiss et al., 2017; Mandal et al., 2015).

Aitchison's methodology converts relative abundances, which are points in a simplex (i.e. compositional), into points in a lower dimensional Euclidean space by taking suitable log-ratios of each taxon with respect to a pre-specified reference taxon or the geometric means of all taxa. However, there are two caveats to keep in mind when using this class of methods (Morton et al., 2019). Firstly, the results and interpretation of data depend on the reference frame. Secondly, some of these methods are appropriate for relative abundance and not the absolute abundance(Morton et al., 2019). Although ANCOM (Mandal et al., 2015) uses Aitchison's framework, it is important to remind that unlike other compositional methods, in its implementation, the ANCOM algorithm does not fix one single taxon (or the geometric mean) as a reference, but uses all taxa and pools results from all such analyses when declaring differential abundance of a particular taxon. Thus, the reference taxon issue

(Morton et al., 2019) does not apply to ANCOM. By assuming that the mean absolute abundance of two taxa are not different between two ecosystems, ANCOM (Mandal et al., 2015) developed a strategy that enables researchers to infer about mean absolute abundances by testing hypotheses regarding mean relative abundances. Thus, unless the absolute mean abundance of almost all taxa changed, ANCOM should perform well in terms of FDR and power. One of the deficiencies of ANCOM is that it does not provide p-value for individual taxon, nor can it provide standard errors or confidence intervals of differential abundance for each taxon, and it can be computationally intensive. According to an extensive simulation study (Weiss et al., 2017), among the available methods for DA analysis, only ANCOM performs well in controlling FDR at the desired level while maintaining high power, as long as the sample size is not too small (e.g. $n = 5$ per group).

The Differential Ranking (DR) methodology (Morton et al., 2019) reformulates the problem as a multinomial regression problem. By imposing the constraint that sum of the regression coefficients is zero, the DR methodology accounts for compositionality in the relative abundance of microbiome data. Thus, unlike ALDEX2 (Gloor, 2015), they do not require the pre-specification of a reference frame. This makes their method more flexible than ALDEX2. Also, as demonstrated in their paper (Morton et al., 2019), the ranks of relative differentials perfectly correlate with ranks of absolute differentials. This result is consistent with the analytical results obtained by ANCOM, provided 2 taxa are not differentially abundant (in mean absolute abundance). Similar to ANCOM, the DR procedure does not provide explicit p-values or confidence intervals to declare statistical significance.

Since not all samples have the same sampling fraction, all DA methodologies require the counts to be properly normalized to account for differences in sampling fractions across samples. Sampling fraction is determined by two components, namely, the microbial load in a unit volume of the ecosystem (e.g. gut) and the library size of the corresponding sample (e.g. total species abundance sequenced from a subject's stool sample). Therefore it is not sufficient to normalize the library size across samples as one needs to take into consideration the differences in the microbial loads. As shown in Figure 1, if a normalization method is based only on the library size and ignores the sampling fraction, then the two samples (A and B) would be considered as normalized, which leads to a false negative conclusion. Thus,

normalizing data on the basis of sampling fractions gives a better description of the truth than normalization methods that rely purely on the library sizes.

Ideally, under the null hypothesis, the test statistic for DA analysis should be (at least approximately) centered at zero (i.e. unbiased). However, for many DA methods, this is not always true for at least one of the following reasons: (1) The test statistic may not be designed for testing hypothesis regarding the actual parameter of interest. For example, the statistic is designed to test hypotheses regarding relative abundance but the null hypothesis is regarding the absolute abundance; (2) Data are not properly normalized. For example, data are normalized to correct for differences in library sizes only but not account for differences in the sampling fractions; (3) Underlying structure, such as compositionality, is ignored; (4) The methodology imposes strong parametric assumption on the data, which could lead to a potential model misspecification problem. For instance, although DESeq2(Love et al., 2014) and edgeR(Robinson et al., 2010) have been widely used for DA analysis, studies on RNA-Seq data have shown that they could yield high FDR as the negative binomial model does not fit the data well when there are many zeros (Weiss et al., 2017). Applying non-parametric tests, such as Wilcoxon rank-sum test, to the OTU table directly, not only neglects the compositional structure of the absolute abundance data, but also implicitly assumes equivalent sampling fractions for all samples.

Motivated by the above reasons, in this chapter we propose a novel methodology called Analysis of Compositions of Microbiomes with Bias Correction (ANCOM-BC) that 1) explicitly tests for differential absolute abundance, 2) normalizes the OTU table for differences in sampling fractions among samples, 3) account for the compositional structure of the OTU table properly, and 4) does not make strong parametric assumptions on the data. As in ANCOM and DR, ANCOM-BC assumes that the observed sample is an unknown fraction of a unit volume of the ecosystem, and the sampling fraction varies 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 data. The offset term serves as the bias correction, and the linear regression framework in log scale is analogous to log-ratio transformation to deal with the compositionality of microbiome data. The case of zero counts is also discussed in Methods section. This methodology has some conceptual similar-

ities with DR, but is fundamentally different. With ANCOM-BC, one can perform standard statistical tests and construct confidence intervals for differential abundance. Moreover, as demonstrated in benchmark simulation studies, ANCOM-BC (a) controls the FDR very well while maintaining adequate power compared to other popular methods, and (b) it is substantially faster the ANCOM. The CPU time is 0.28 mins vs 63 mins when the number of taxa is 500. The CPU time for ANCOM increases dramatically as the number of taxa increases to 1,000. In this case, the CPU times for ANCOM-BC and ANCOM are 0.51 mins and 211 mins, respectively. In addition to results based on synthetic data, we also illustrate ANCOM-BC using the well-known global gut microbiota dataset (Yatsunenko et al., 2012).

2.2 Methods

2.2.1 Model assumptions

Assumption 2.2.1.

$$\begin{aligned} E(O_{ij}|A_{ij}) &= c_j A_{ij}, \\ Var(O_{ij}|A_{ij}) &= \sigma^2_{w,ij}, \end{aligned} \tag{2.1}$$

where $\sigma^2_{w,ij}$ = variability between specimens *within* the j^{th} sample. Therefore, $\sigma^2_{w,ij}$ characterizes the within-sample variability. Typically, researchers do not obtain more than one specimen at a given time in most microbiome studies. Consequently, variability between specimens within sample is usually not estimated. Throughout this paper, we use "sample" and "specimen" exchangeably.

According to Assumption 2.2.1, in expectation the absolute abundance of a taxon in a random sample is in constant proportion to the absolute abundance in the ecosystem of the sample. In other words, the expected relative abundance of each taxon in a random sample is equal to the relative abundance of the taxon in the ecosystem of the sample.

Assumption 2.2.2. *For each taxon i, $A_{ij}, j = 1, \ldots, n$, are independently distributed with*

$$\begin{aligned} E(A_{ij}|b_i, x_j) &= b_i^T x_j, \\ Var(A_{ij}|b_i, x_j) &= \sigma^2_{b,ij}, \end{aligned} \tag{2.2}$$

where

(1) $x_j = (x_{j1}, x_{j2}, \ldots, x_{jp})^T$ *are the covariates of interest for the* j^{th} *sample,*

(2) $b_i = (b_{i1}, b_{i2}, \ldots, b_{ip})$ *are the corresponding coefficients for* x_j*,*

(3) $\sigma_{b,ij}$ = *between sample variation for the* i^{th} *taxon.*

The Assumption 2.2.2 states that for a given taxon, all samples are independent.

2.2.2 ANCOM-BC for fixed effects models

2.2.2.1 Regression framework From Assumptions 2.2.1 & 2.2.2, we have:

$$\begin{aligned} &E(O_{ij}|b_i, x_j) = c_j b_i^T x_j, \\ &Var(O_{ij}|b_i, x_j) = f(\sigma^2_{w,ij}, \sigma^2_{b,ij}) := \sigma^2_{t,ij}. \end{aligned} \tag{2.3}$$

Motivated by the above set-up, we introduce the following linear model framework for log-transformed absolute abundances:

$$y_{ij} = d_j + \beta_i^T x_j + \epsilon_{ij}, \tag{2.4}$$

with

$$\begin{aligned} &E(\epsilon_{ij}) = 0, \\ &E(y_{ij}) = d_j + \beta_i^T x_j, \\ &Var(y_{ijk}) = Var(\epsilon_{ijk}) := \sigma^2_{ij}. \end{aligned} \tag{2.5}$$

Note that the above log-transformation of data is inspired by the Box-Cox family of transformations (Box and Cox, 1964) which are routinely used in data analysis.

Rewrite the model (2.4) in the vector form, we have

$$y_i = d + X\beta_i + \epsilon_i, \tag{2.6}$$

with

$$
\begin{aligned}
&E(\epsilon_i) = (0, \ldots, 0)^T, \\
&E(y_i) = d + X\beta_i, \\
&Cov(y_i) = Cov(\epsilon_i) = \begin{bmatrix} \sigma_{i1}^2 & 0 & \ldots & 0 \\ 0 & \sigma_{i2}^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_{in}^2 \end{bmatrix}.
\end{aligned} \tag{2.7}
$$

where

(1) $y_i = (y_{i1}, y_{i2}, \ldots, y_{in})^T$,

(2) $d = (d_1, d_2, \ldots, d_n)^T$,

(3) $\beta_i = (\beta_{i1}, \beta_{i2}, \ldots, \beta_{ip})^T$,

(4) $\epsilon_i = (\epsilon_{i1}, \epsilon_{i2}, \ldots, \epsilon_{in})^T$,

(5) $X = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}$.

It is important to note that within each subject j, for taxa $i \neq i'$, ϵ_{ij} and $\epsilon_{i'j}$ are not independent. Thus the column vectors y_i and $y_{i'}$ are not independent random vectors.

For ease of exposition, define the adjusted log absolute abundance $y_i^{adj} = y_i - d$, then by (2.6)

$$y_i^{adj} = X\beta_i + \epsilon_i. \tag{2.8}$$

From the above model, the ordinary least squares (OLS) estimators of d and β_i can be obtained by iteratively solving the following system of equations.

Suppose on convergence, $d \leftarrow d^*, y_i^{adj} \leftarrow y_i^{adj^*}, \beta_i \leftarrow \beta_i^*$, we have

$$
\begin{aligned}
d^* &= \frac{1}{m}\sum_{i=1}^{m}(y_i - X\beta_i^*), \\
y_i^{adj^*} &= y_i - d^*, \\
\beta_i^* &= (X^TX)^{-1}X^T y_i^{adj^*}.
\end{aligned} \tag{2.9}
$$

Algorithm 1 Iterative least square regression

1: **Initialize:**

For $i = 1, \ldots, m$

$d \leftarrow 0$

$y_i^{adj} \leftarrow y_i - d = y_i$

$\beta_i \leftarrow (X^T X)^{-1} X^T y_i^{adj} = (X^T X)^{-1} X^T y_i$

2: **while** not converge **do**

3: $d \leftarrow \frac{1}{m} \sum_{i=1}^{m} (y_i - X\beta_i)$

4: $y_i^{adj} \leftarrow y_i - d$

5: $\beta_i \leftarrow (X^T X)^{-1} X^T y_i^{adj}$

6: **end while**

Therefore

$$\begin{aligned} d^* &= \frac{1}{m} \sum_{i=1}^{m} (y_i - X\beta_i^*) = \frac{1}{m} \sum_{i=1}^{m} (y_i - P y_i^{adj*}) \\ &= \frac{1}{m} \sum_{i=1}^{m} (y_i - P y_i + P d^*) = \frac{1}{m} \sum_{i=1}^{m} (y_i^{adj} + d - P(y_i^{adj} + d) + P d^*) \\ &= (I - P)d + P d^* + \frac{1}{m} \sum_{i=1}^{m} (I - P) y_i^{adj} \\ &= (I - P)d + P d^* + \frac{1}{m} \sum_{i=1}^{m} e_i, \end{aligned} \tag{2.10}$$

where

(1) $P = X(X^T X)^{-1} X$ is the projection matrix,

(2) $e_i = (I - P) y_i^{adj}$ with $E(e_i) = 0$.

By (2.10), it is easy to see that

$$\begin{aligned} (I - P)d^* &= (I - P)d + \frac{1}{m} \sum_{i=1}^{m} e_i \\ \iff (I - P)[E(d^*) - d] &= 0. \end{aligned} \tag{2.11}$$

As $P \subset \mathcal{C}(X)$, the equation (2.11) holds as long as either of the following is valid:

(1) $E(d^*) - d = 0$,

(2) $E(d^*) - d \in \mathcal{C}(X)$.

Suppose there exists a vector $\delta \in \mathbb{R}^p$, such that

$$E(d^*) = d - X\delta. \tag{2.12}$$

Clearly, a zero vector of δ corresponds to condition (1) stated above; While a nonzero vector of δ corresponds to condition (2). Therefore, by (2.9),

$$E(\beta_i^*) = \delta + \beta_i. \tag{2.13}$$

We shall denote d^* and β_i^* obtained from the above iterative algorithm as preliminary estimators of d and β_i. Without loss of generality, throughout this paper we assume X^TX is a full rank matrix. If it is not a full rank matrix, then we shall use any generalized inverse of X^TX. Since $X\beta_i^*$ in (2.10) is invariant of the choice of generalized inverse $(X^TX)^g$ used in $\beta_i^* = (X^TX)^gX^Tz_i$. Thus the preliminary estimator d^* provided above is invariant of the choice of generalized inverse used in deriving β_i^*. Furthermore, throughout this paper we are interested in testing hypothesis regarding the parameter $A\beta_i$ where we implicitly assume that $\mathcal{C}(A^T) \subset \mathcal{C}(X^T)$. Consequently, $A\beta_i$ is estimable and $A\beta_i^*$ is invariant of the generalized inverse used in the calculation of β_i^* when X^TX is not full rank. If X^TX is full rank then $\mathcal{C}(A^T) \subset \mathcal{C}(X^T)$ is trivially satisfied. Hence throughout this text we shall assume X^TX is full rank.

For each taxon $i = 1, 2, \ldots, m$, by (2.13), β_i^* is a biased estimator if $\delta \neq (0, \ldots, 0)^T$. Suppose we wish test the hypothesis

$$\begin{aligned} H_0 &: A\beta_i = A\beta_i^{(0)}, \\ H_1 &: A\beta_i \neq A\beta_i^{(0)}. \end{aligned} \tag{2.14}$$

Under the null hypothesis, $E(A\beta_i^*) - A\beta_i^{(0)} = A\delta \neq 0$ and hence biased. The goal of ANCOM-BC is to estimate this bias and accordingly modify the estimator $A\beta_i^*$ so that the resulting estimator is asymptotically centered at $A\beta_i^{(0)}$ under the null hypothesis and hence the test statistic is asymptotically centered at zero.

First we make the following observations. Since $E(\beta_i^*) = \delta + \beta_i$, from Cramér Wold Theorem, we note that as $n \to \infty$

$$\Sigma_i^{-\frac{1}{2}}(\beta_i^* - (\delta + \beta_i)) \to_d N_p(0, I), \tag{2.15}$$

where

$$\Sigma_i = \lim_{n\to\infty}(X^TX)^{-1}(\sum_{j=1}^{n}\sigma_{ij}^2 x_j x_j^T)(X^TX)^{-1}. \tag{2.16}$$

Since

$$E(d^* + X\beta_i^*) = d - X\delta + X(\delta + \beta_i) = d + X\beta_i, \tag{2.17}$$

i.e. $d^* + X\beta_i^*$ is an unbiased estimator of $d + X\beta_i$, hence we could obtain the empirical estimator for Σ_i as

$$\hat{\Sigma}_i = (X^TX)^{-1}(\sum_{j=1}^{n}(y_{ij} - d_j^* - \beta_i^{*T}x_j)^2 x_j x_j^T)(X^TX)^{-1}. \tag{2.18}$$

Under some mild regularity conditions(Peddada and Smith, 1997), we have the following consistency result

$$n(\hat{\Sigma}_i - \Sigma_i) \to_p 0, \text{ as } n \to \infty. \tag{2.19}$$

Therefore, replacing Σ_i with $\hat{\Sigma}_i$ in (2.15) and appealing to Slutsky's theorem, we have:

$$\hat{\Sigma}_i^{-\frac{1}{2}}(\beta_i^* - (\delta + \beta_i)) \to_d N_p(0, 1), \text{ as } n \to \infty. \tag{2.20}$$

By (2.16) and (2.19), under some mild regularity conditions(Peddada and Smith, 1997), we obtain

$$\hat{\Sigma}_i \to_p 0, \text{ as } n \to \infty. \tag{2.21}$$

Consequently,

$$\beta_i^* \to_p \delta + \beta_i, \text{ as } n \to \infty. \tag{2.22}$$

The above observation regarding the convergence of β_i^* plays a critical role in the following. Since the sampling fraction is constant for all taxa within a sample, we attempt to pool information across taxa within each sample when estimating δ. We model each taxa abundance using the following Gaussian mixture model. For the i^{th} taxon and the k^{th} covariate (note that for a categorical covariate of s levels, it results in s coefficients,

e.g. $\beta_{i1}, \ldots, \beta_{is}$, and we will fit the Gaussian mixture model for these s coefficients separately), let C_0 denote the set of taxa that are not differentially abundant with regard to x_{ik}, i.e. $C_0 = \{i \in (1, 2, \ldots, m) : \beta_{ik} = 0\}$, C_1 denote the set of taxa whose absolute abundance decreases as the increase of x_{ik}, i.e. $C_1 = \{i \in (1, 2, \ldots, m) : \beta_{ik} < 0\}$, and let C_2 denote the set of taxa whose absolute abundance increases as the increase of x_{ik}, i.e. $C_2 = \{i \in (1, 2, \ldots, m) : \beta_{ik} > 0\}$, Let π_r denote the probability that a taxon belongs to set C_r, $r = 0, 1, 2$. For simplicity of estimation of parameters, similar to GEE, we shall assume that $\beta_{ik}, i = 1, 2, \ldots, m$, are independently distributed. Thus, we ignore the underlying correlation structure when estimating δ. This is similar to what is often done in other omics studies. Thus, we model the distribution of β^*_{ik} by Gaussian mixture model as follows:

$$f(\beta^*_{ik}) = \pi_0 \phi(\frac{\beta^*_{ik} - \delta_k}{\nu_{i0}}) + \pi_1 \phi(\frac{\beta^*_{ik} - (\delta_k + l_1)}{\nu_{i1}}) + \pi_2 \phi(\frac{\beta^*_{ik} - (\delta_k + l_2)}{\nu_{i2}}), \tag{2.23}$$

where

(1) ϕ is the normal density function,

(2) $\delta_k, \delta_k + l_1$, and $\delta_k + l_2$ are means for $\beta^*_{ik}|C_0, \beta^*_{ik}|C_1$, and $\beta^*_{ik}|C_2$, respectively. $l_1 < 0$, $l_2 > 0$,

(3) ν_{i0}, ν_{i1}, and ν_{i2} are variances of $\beta^*_{ik}|C_0, \beta^*_{ik}|C_1$, and $\beta^*_{ik}|C_2$, respectively.

Note that instead of fitting a multivariate Gaussian mixture model for all covariates together, we choose to fit a univariate Gaussian mixture model repeatedly for each single covariate. This repetition is simply because the sets of of taxa $\{C_0, C_1, C_2\}$ are not necessarily the same for different covariates.

For computational simplicity, we assume that $\nu_{i1} > \nu_{i0}, \nu_{i2} > \nu_{i0}$. Thus, Without loss of generality for $\kappa_1, \kappa_2 > 0$, let $\nu_{i1} = \nu_{i0} + \kappa_1$ and $\nu_{i2} = \nu_{i0} + \kappa_2$. While this assumption is not a requirement for our method, it is reasonable to assume that variability among differentially abundant taxa is larger than that among the null taxa. By making this assumption, we speed-up the computation time.

Assuming samples are independent, we begin by first estimating $\nu^2_{i0} = Var(\beta^*_{ik})$. Note that ν^2_{i0} is the function of heteroscedastic variances, the consistent estimator of ν^2_{i0}, which we refer to as $\hat{\nu}^2_{i0}$, is the k^{th} diagonal element of $\hat{\Sigma}_i$ stated in (2.18). In all future calculations, we plug in $\hat{\nu}^2_{i0}$ for ν^2_{i0}. This is similar in spirit to many statistical procedures involving nuisance parameters. The following lemma (McLachlan and Krishnan, 2007) is useful in the sequel.

Lemma 2.2.1. $\frac{\partial}{\partial\theta}\log f(x) = E_{f(z|x)}[\frac{\partial}{\partial\theta}\log f(z) + \frac{\partial}{\partial\theta}\log f(x|z)]$.

Let $\Theta = (\delta_k, \pi_1, \pi_2, \pi_3, l_1, l_2, \kappa_1, \kappa_2)^T$ denote the set of unknown parameters, then for each taxon the log-likelihood can be reformulated using Lemma 2.2.1, as follows:

$$\Theta \leftarrow \arg\max_{\Theta} \sum_{i=1}^{m}\sum_{r=0}^{2} p_{r,i}[\log Pr(i \in C_r) + \log f(\beta_{ik}|i \in C_r)]. \tag{2.24}$$

Then the E-M algorithm is described as follows:

- E-step: Compute conditional probabilities of the latent variable. Define $p_{r,i} = Pr(i \in C_r|\beta_{ik}) = \frac{\pi_r \phi(\frac{\beta_{ik}-(\delta+l_r)}{\nu_{ir}})}{\sum_r \pi_r \phi(\frac{\beta_{ik}-(\delta+l_r)}{\nu_{ir}})}, r = 0, 1, 2; i = 1, \ldots, m$, which are conditional probabilities representing the probability that an observed value follows each distribution. Note that $l_0 = 0$.
- M-step: Maximize the likelihood function with respect to the parameters, given the conditional probabilities.

We shall denote the resulting estimator of δ_k by $\hat{\delta}_k^{EM}$.

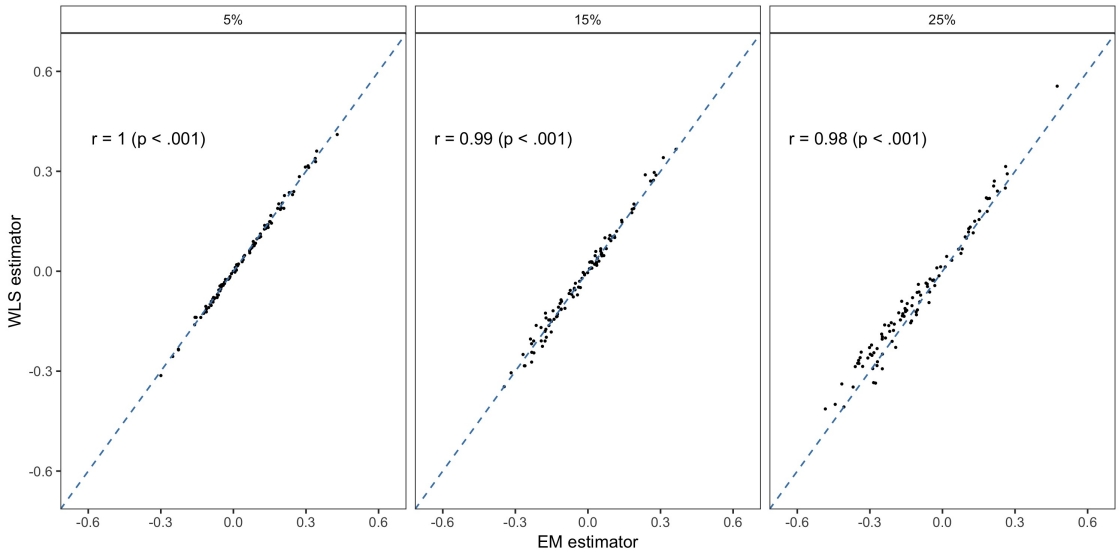


Figure 4: EM and WLS estimators of the bias term are highly correlated.

Next we estimate $Var(\hat{\delta}_{EM})$. Since the likelihood function is not a regular likelihood and hence it is not feasible to derive the Fisher information. Consequently, we take a simpler and

a pragmatic approach to derive an approximate estimator of $Var(\hat{\delta}_{EM})$ using the variance of weighted least square (WLS) estimator, $Var(\hat{\delta}_{WLS})$, which is defined below. Extensive simulation studies suggest that $\hat{\delta}_{EM}$ and $\hat{\delta}_{WLS}$ are highly correlated (Figure 4) and it appears to be reasonable to approximate $Var(\hat{\delta}_{EM})$ by $Var(\hat{\delta}_{WLS})$.

Define

$$\begin{aligned}\hat{\delta}_{WLS} &= \frac{\sum_{i\in C_0}\frac{\beta^*_{ik}}{\hat{\nu}^2_{i0}} + \sum_{i\in C_1}\frac{\beta^*_{ik}-\hat{l}_1}{\hat{\nu}^2_{i1}} + \sum_{i\in C_2}\frac{\beta^*_{ik}-\hat{l}_2}{\hat{\nu}^2_{i2}}}{\sum_{i\in C_0}\frac{1}{\hat{\nu}^2_{i0}} + \sum_{i\in C_1}\frac{1}{\hat{\nu}^2_{i1}} + \sum_{i\in C_2}\frac{1}{\hat{\nu}^2_{i2}}} \\ &= \frac{\sum_{i\in C_0}\frac{\beta^*_{ik}}{\nu^2_{i0}} + \sum_{i\in C_1}\frac{\beta^*_{ik}-l_1}{\nu^2_{i1}} + \sum_{i\in C_2}\frac{\beta^*_{ik}-l_2}{\nu^2_{i2}}}{\sum_{i\in C_0}\frac{1}{\nu^2_{i0}} + \sum_{i\in C_1}\frac{1}{\nu^2_{i1}} + \sum_{i\in C_2}\frac{1}{\nu^2_{i2}}} + o_p(1).\end{aligned} \tag{2.25}$$

The above expression is of the form

$$\frac{a_1^T x_1 + a_2^T x_2 + a_3^T x_3}{a_1^T\mathbf{1} + a_2^T\mathbf{1} + a_3^T\mathbf{1}} \equiv \frac{\alpha^T u}{\alpha^T\mathbf{1}}, \tag{2.26}$$

where

(1) $\mathbf{1} = (1,\ldots,1)^T$,

(2) $a_r = (a_{r1}, a_{r2}, \ldots, a_{rm_r})^T := (\frac{1}{\nu^2_{ir}})^T,\ i \in C_r,\ r = 0,1,2,$

(3) $x_r = (x_{r1}, x_{r2}, \ldots, x_{rm_r})^T := (\beta^*_{ik} - l_i)^T,\ i \in C_r,\ r = 0,1,2.$ Note that $l_0 = 0,$

(4) $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)^T \equiv (a_1^T, a_2^T, a_3^T)^T,$

(5) $u = (u_1, u_2, \ldots, u_m)^T \equiv (x_1^T, x_2^T, x_3^T)^T.$

For the simplicity of notation we relabel a and x by α and u, respectively. Denote $Cov(x) = Cov(u)$ by Ω, and let $\omega_{ii'}$ denotes the (i,i') element of Ω. We make the following assumption

Assumption 2.2.3.

$$\frac{\sum_{i\neq i'}^m \omega_{ii'}}{m^2} = o(1). \tag{2.27}$$

Using the above expressions, we compute the variance as follows:

$$\begin{aligned}Var(\hat{\delta}_{WLS}) &= Var(\frac{\alpha^T u}{\alpha^T\mathbf{1}}) \\ &= \frac{\sum_{i=1}^m \alpha_i^2\omega_{ii}}{(\sum_{i=1}^m \alpha_i)^2} + \frac{\sum_{i\neq i'}^m \alpha_i\alpha_{i'}\omega_{ii'}}{(\sum_{i=1}^m \alpha_i)^2}.\end{aligned} \tag{2.28}$$

Recall that (a) for $i \in C_0$, $\omega_{ii} = Var(\beta^*_{ik}) = \nu^2_{i0} = O(n^{-1})$, (b) for $i \in C_1$, $\omega_{ii} = Var(\beta^*_{ik}) = \nu^2_{i1} = \nu^2_{i0} + \kappa_1 = O(1)$, and (c) for $i \in C_2$, $\omega_{ii} = Var(\beta^*_{ik}) = \nu^2_{i2} = \nu^2_{i0} + \kappa_2 = O(1)$. Note that $\alpha_i = \frac{1}{Var(\beta^*_{ik})} = \frac{1}{\omega_{ii}}$, thus we have:

$$\begin{aligned} Var(\frac{\alpha^T u}{\alpha^T \mathbf{1}}) &= \frac{\sum_{i=1}^m \alpha_i^2 \omega_{ii}}{(\sum_{i=1}^m \alpha_i)^2} + \frac{\sum_{i \neq i'}^m \alpha_i \alpha_{i'} \omega_{ii'}}{(\sum_{i=1}^m \alpha_i)^2} \\ &= \frac{1}{\sum_{i=1}^m \alpha_i} + \frac{\sum_{i \neq i'}^m \alpha_i \alpha_{i'} \omega_{ii'}}{(\sum_{i=1}^m \alpha_i)^2}. \end{aligned} \tag{2.29}$$

Since $\nu^2_{i0} = O(n^{-1})$, $\nu^2_{i1} = O(1)$, and $\nu^2_{i2} = O(1)$, consequently, $a_{1i} = O(n)$, $a_{2i} = a_{3i} = O(1)$, and

$$\begin{aligned} \sum_{i=1}^m \alpha_i &= \mathbf{1}^T a_1 + \mathbf{1}^T a_2 + \mathbf{1}^T a_3 \\ &= \sum_{i \in C_0} O(n) + \sum_{i \in C_1} O(1) + \sum_{i \in C_2} O(1) \\ &= O(m_0 n) + O(m_1) + O(m_2) \\ &= O(m_0 n) \quad \text{if } m_0 n \geq \max\{m_1, m_2\}. \end{aligned} \tag{2.30}$$

Using these facts and Assumption 2.2.3 in (2.29), we get

$$\begin{aligned} Var(\frac{\alpha^T u}{\alpha^T \mathbf{1}}) &= O(m_0^{-1} n^{-1}) + \frac{\sum_{i \neq i'}^m \{n^{-1} m^{-1} \alpha_i\}\{n^{-1} m^{-1} \alpha_{i'}\} \omega_{ii'}}{n^{-2} m^{-2} (\sum_{i=1}^m \alpha_i)^2} \\ &= O(m_0^{-1} n^{-1}) + \frac{1}{m^2} \frac{\sum_{i \neq i'}^m \{n^{-1} \alpha_i\}\{n^{-1} \alpha_{i'}\} \omega_{ii'}}{(\sum_{i=1}^m n^{-1} m^{-1} \alpha_i)^2} \\ &= O(m_0^{-1} n^{-1}) + \frac{1}{m^2} \frac{O(1) o(m^2)}{O(1)} \\ &= O(m_0^{-1} n^{-1}). \end{aligned} \tag{2.31}$$

Thus, under Assumption 2.2.3 regarding $\omega_{ii'}$, the contribution of the covariance terms in the above variance expression is negligible as long as m is very large compared to n, which is usually the case. Hence

$$Var(\hat{\delta}_{WLS}) = Var(\frac{\alpha^T u}{\alpha^T \mathbf{1}}) = O(m_0^{-1} n^{-1}). \tag{2.32}$$

Furthermore, appealing to Cauchy-Schwartz inequality we get

$$\begin{aligned} Cov(\hat{\mu}_{i1} - \hat{\mu}_{i2}, \hat{\delta}_{WLS}) &\leq \sqrt{Var(\hat{\mu}_{i1} - \hat{\mu}_{i2}) Var(\hat{\delta}_{WLS})} \\ &\leq O(n^{-1/2}) O(m_0^{-1/2} n^{-1/2}) \\ &= O(n^{-1} m_0^{-1/2}). \end{aligned} \tag{2.33}$$

Hence, as long as m_0 is large, the contribution made by $Var(\hat{\delta}_{WLS})$ and $Cov(\hat{\mu}_{i1} - \hat{\mu}_{i2}, \hat{\delta}_{WLS})$ relative to $Var(\hat{\mu}_{i1} - \hat{\mu}_{i2})$ is negligible. Otherwise, we replace $\hat{\delta}_k^{EM}$ with

$$\hat{\delta}_k^{WLS} = \frac{\sum_{i \in C_0} \frac{\beta_{ik}^*}{\hat{\nu}_{i0}^2} + \sum_{i \in C_1} \frac{\beta_{ik}^* - \hat{l}_1}{\hat{\nu}_{i1}^2} + \sum_{i \in C_2} \frac{\beta_{ik}^* - \hat{l}_2}{\hat{\nu}_{i2}^2}}{\sum_{i \in C_0} \frac{1}{\hat{\nu}_{i0}^2} + \sum_{i \in C_1} \frac{1}{\hat{\nu}_{i1}^2} + \sum_{i \in C_2} \frac{1}{\hat{\nu}_{i2}^2}}, \tag{2.34}$$

consequently, $Var(\beta_{ik}^*)$ and $Cov(\beta_{ik}^*, \beta_{ik'}^*)$ in Σ_i are replaced by the upper bound of $Var(\beta_{ik}^* - \hat{\delta}_k^{WLS})$ and $Cov(\beta_{ik}^* - \hat{\delta}_k^{WLS}, \beta_{ik'}^* - \hat{\delta}_{k'}^{WLS})$, respectively.

The above procedure is applied to every $\beta_{ik}, k = 1, \ldots, p$, eventually, we obtain the estimator of δ as

$$\hat{\delta}^{EM} = (\hat{\delta}_1^{EM}, \hat{\delta}_2^{EM}, \ldots, \hat{\delta}_p^{EM})^T. \tag{2.35}$$

Therefore, the final estimator of β is defined as

$$\hat{\beta}_i = \beta_i^* - \hat{\delta}^{EM}, \tag{2.36}$$

with

$$\hat{\beta}_i \to_p \beta_i, \text{ as } n \to \infty, \tag{2.37}$$

given that $\hat{\delta}^{EM}$ is a good approximation of δ.

We provide the following algorithm for summarizing the above estimating procedure

For taxon i, we test the following hypothesis

$$H_0 : A\beta_i = A\beta_i^{(0)},$$
$$H_1 : A\beta_i \neq A\beta_i^{(0)}.$$

Algorithm 2 Bias correction

Input:

$\beta_i^*, \Sigma_i, i = 1, \ldots, m$

procedure E-M(β_i^*, Σ_i)

    return $\hat{\delta}_k^{EM}, k = 1, \ldots, p$

end procedure

for $k = 1, \ldots, p$ **do**

    $\hat{\beta}_{ik} \leftarrow \beta_{ik}^* - \hat{\delta}_k^{EM}$

end for

From Slutsky's theorem, as $n \to \infty$, the following test statistic is approximately central chi-square distributed under the null hypothesis

$$\begin{aligned} W_i &= (A\hat{\beta}_i - A\beta_i)^T (A\hat{\Sigma}_i A^T)^{-1} (A\hat{\beta}_i - A\beta_i) \\ &= (A\beta_i^* - A\hat{\delta}^{EM} - A\beta_i)^T (A\hat{\Sigma}_i A^T)^{-1} (A\beta_i^* - A\hat{\delta}^{EM} - A\beta_i) \\ &\to_d \chi_q^2, \end{aligned} \tag{2.38}$$

where $q = rank(A)$.

To control the FDR due to multiple comparisons, We recommend applying Holm method (Holm, 1979) instead of Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995) in the algorithm to adjust raw p-values, since the Holm method does not require any dependence structure in the underlying p-values, and research has showed that it is a more appropriate method to control the FDR when p-values are not accurate (Lim et al., 2013).

2.2.2.2 Sampling fraction estimation After obtaining $\hat{\delta}^{EM}$, the estimator of sampling fraction d is defined as follows:

$$\hat{d} = \frac{1}{m} \sum_{i=1}^{m} (y_i - X\hat{\beta}_i). \tag{2.39}$$

Let Σ_j denote an $m \times m$ covariance matrix of $\epsilon_j = (\epsilon_{1j}, \epsilon_{2j}, \ldots, \epsilon_{mj})^T$, where $\sigma_{ii'j}$ is the $(i, i')^{th}$ element of Σ_j and σ_{ij}^2 is the i^{th} diagonal element of Σ_j. Furthermore, suppose

Assumption 2.2.4.

$$\begin{aligned} &\sigma_{ij}^2 < \sigma_0^2 < \infty, \\ &\frac{\sum_{i \neq i'}^{m} \sigma_{ii'j}}{m^2} = o(1). \end{aligned} \tag{2.40}$$

Denote $\mathbf{1} = (1, 1, \ldots, 1)^T$, based on Assumption 2.2.4, we have

$$0 \leq \mathbf{1}^T \Sigma \mathbf{1} = \sum_{i=1}^{m} \sum_{i'=1}^{m} \sigma_{ii'j} = \sum_{i=1}^{m} \sigma_{ij}^2 + \sum_{i \neq i'}^{m} \sigma_{ii'j} \leq m\sigma_0^2 + \sum_{i \neq i'}^{m} \sigma_{ii'j}. \tag{2.41}$$

Hence

$$0 \leq \frac{\mathbf{1}^T \Sigma \mathbf{1}}{m^2} \leq \frac{\sigma_0^2}{m} + \frac{\sum_{i \neq i'}^{m} \sigma_{ii'j}}{m^2} = o(1). \tag{2.42}$$

Thus, for each taxon $i = 1, 2, \ldots, m$, we have

$$\frac{1}{m} \sum_{i=1}^{m} (y_i - (d + X\beta_i)) \to_p 0, \quad \text{as } m \to \infty. \tag{2.43}$$

Therefore, according to (2.37) and (2.43), as $m, n \to \infty$,

$$\hat{d} \to d. \tag{2.44}$$

2.2.3 ANCOM-BC for mixed effects models

Similar to the fixed effects model, for each taxon $i, i = 1, \ldots, m$, and each subject $j, j = 1, \ldots, n$, the offset-based mixed effects log-linear model is set up as

$$y_{ij} = d_j + X_j \beta_i + Z_j \alpha_j + \epsilon_{ij}, \tag{2.45}$$

where d_j is the n_j-vector sampling fractions, X_j is the $n_j \times p$ design matrix for fixed effects, β_i is the p-vector of fixed effects regression coefficients to be estimated, Z_j is the $n_j \times q$ design matrix for the random effects, α_j is the q-vector random effects, ϵ_{ij} is the n_j-vector residuals. Note that $\sum_j n_j = n$. The following distributional assumptions are made

$$\begin{aligned} &\alpha_j \sim N(0, D), \\ &\epsilon_{ij} \sim N(0, \sigma_j^2 I_{n_j}), \\ &\epsilon_{i1}, \ldots, \epsilon_{in}, \alpha_1, \ldots, \alpha_n \text{ independent.} \end{aligned} \tag{2.46}$$

Thus, for each taxon $i, i = 1, \ldots, m$, and each subject $j, j = 1, \ldots, n$, we have

$$y_{ij} \sim N(d_j + X_j\beta_i, H_j(\theta)), \tag{2.47}$$

where $H_j(\theta) = Z_j D Z_j^T + \sigma_j^2 I_{n_j}$ denotes a general covariance matrix parametrized by θ.

To stack up all subject's data

$$y_i = d + X\beta_i + Z\alpha + \epsilon_i, \tag{2.48}$$

where

$$y_i = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad d = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_n \end{bmatrix}, \quad X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix},$$

$$Z = \begin{bmatrix} Z_1 & 0 & \ldots & 0 \\ 0 & Z_2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & Z_n \end{bmatrix}, \quad \alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix}, \quad \epsilon_i = \begin{bmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{in} \end{bmatrix}.$$

That is,

$$y_i \sim N\left(d + X\beta_i, H(\theta) = \begin{bmatrix} H_1(\theta) & 0 & \ldots & 0 \\ 0 & H_2(\theta) & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & H_n(\theta) \end{bmatrix}\right), \tag{2.49}$$

where $H(\theta)$ (or H for short) is a block diagonal matrix.

Similarly, we run the iterative least square algorithm to obtain preliminary estimators for d and β_i. As compared to fixed effects model, the algorithm is modified slightly

Algorithm 3 Iterative least square regression

1: **Initialize:**

For $i = 1, \ldots, m$

$d \leftarrow 0$

$y_i^{adj} \leftarrow y_i - d = y_i$

$\beta_i \leftarrow ReML(y_i^{adj}) = ReML(y_i)$

2: **while** not converge **do**

3: $d \leftarrow \frac{1}{m} \sum_{i=1}^{m} (y_i - X\beta_i)$

4: $y_i^{adj} \leftarrow y_i - d$

5: $\beta_i \leftarrow ReML(y_i^{adj})$

6: **end while**

where in the algorithm, we maximize the restricted log-likelihood function (dropping constant terms) w.r.t. variance components θ to obtain the restricted maximum likelihood (ReML) (Patterson and Thompson, 1971; Harville, 1974) estimator for the covariance matrix $H_i(\theta)$, and the corresponding estimator for regression coefficients β_i,

$$\begin{aligned} \mathcal{L}(\theta|y) = &- \sum_{j=1}^{n} \log |H_j| - \sum_{j=1}^{n} \log |X_j^T H_j^{-1} X_j| - \\ &\sum_{j=1}^{n} (y_{ij} - X_j\beta_i)^T H_j^{-1} (y_{ij} - X_j\beta_i), \end{aligned} \tag{2.50}$$

where $\beta_i \leftarrow (X^T H^{-1} X)^{-1} X^T H^{-1} y_i$. As close-form solutions of (2.50) do not exist, Newton-Raphson method is usually employed (Lindstrom and Bates, 1988).

Suppose on convergence, $d \leftarrow d^*, y_i^{adj} \leftarrow y_i^{adj^*}, H \leftarrow H^*, \beta_i \leftarrow \beta_i^*$, we have

$$\begin{aligned} d^* &= \frac{1}{m} \sum_{i=1}^{m} (y_i - X\beta_i^*), \\ y_i^{adj^*} &= y_i - d^*, \\ \beta_i^* &= (X^T H^{*-1} X)^{-1} X^T H^{*-1} y_i^{adj^*}. \end{aligned} \tag{2.51}$$

It is easy to show that there exists a vector $\delta \in \mathbb{R}^p$, such that

$$\begin{aligned} E(d^*) &= d - X\delta, \\ E(\beta_i^*) &= \delta + \beta_i. \end{aligned} \tag{2.52}$$

i.e., $\hat{\beta}_i$ is a biased estimator for β_i.

Similar to the case of fixed effects model, we fit the Gaussian mixture model to each $\beta_{ik}, k = 1, \ldots, p$ separately, to correct this bias δ, and final estimators for β_i and d are given by

$$\begin{aligned} \hat{\beta}_i &= \beta_i^* - \hat{\delta}^{EM}, \\ \hat{d} &= \frac{1}{m} \sum_{i=1}^{m} (y_i - X\hat{\beta}_i). \end{aligned} \tag{2.53}$$

Thus, for taxon i, the Wald statistic for the following hypothesis

$$\begin{aligned} H_0 &: A\beta_i = A\beta_i^{(0)}, \\ H_1 &: A\beta_i \neq A\beta_i^{(0)}, \end{aligned}$$

is given by

$$\begin{aligned} W_i &= (A\hat{\beta}_i - A\beta_i)^T (A\hat{\Sigma}_i A^T)^{-1} (A\hat{\beta}_i - A\beta_i) \\ &= (A\beta_i^* - A\hat{\delta}^{EM} - A\beta_i)^T (A\hat{\Sigma}_i A^T)^{-1} (A\beta_i^* - A\hat{\delta}^{EM} - A\beta_i) \\ &\rightarrow_d \chi_q^2, \end{aligned} \tag{2.54}$$

where

(1) $\hat{\Sigma}_i = (X^T H^{*-1} X)^{-1}$,

(2) $q = rank(A)$.

2.3 Simulation study

2.3.1 Normalization

Using simulated data, we illustrate how the existing normalization methods fail to eliminate the bias introduced by differences in sampling fractions across samples, whereas the normalization method introduced in ANCOM-BC performs well. Specifically, we compare our proposed method with Cumulative-Sum Scaling (CSS) implemented in metagenomeSeq (Paulson et al., 2013), Median (MED) in DESeq2 (Love et al., 2014), Upper Quartile (UQ) and Trimmed Mean of M-values (TMM), and Total-Sum Scaling (TSS)). Additionally, we

also considered modified versions of UQ and TMM. These are obtained by multiplying the normalization factors with the corresponding library size to account for "effective library size" (Chen et al., 2014) and implemented in edgeR (Robinson et al., 2010). We shall refer to these methods as ELib-UQ and ELib-TMM (Appendix A).

We considered a simulation study where there are unbalanced microbial loads in two experimental groups and balanced library sizes for samples. This results in a large variability in sampling fractions. Thus, we simulated data where sampling fraction in Group 1 is systematically different from sampling fraction in Group 2. Consequently, observed absolute abundances in the samples in the two groups were systematically different even though the actual absolute abundances in the ecosystems are same. To evaluate the performance of each normalization method, we introduced a residual measure (Appendix B) that estimates the deviation between the estimated sampling fraction and the true sampling fraction. For simplicity of exposition, we plotted the centered residuals, by subtracting the group average of the residuals. If a normalization method is effective then it should eliminate the bias due to the differences in the sampling fractions so that samples from the two groups (circles and triangles) should intermix and not cluster by the group labels.

From Figure 5, we notice that the samples normalized by ANCOM-BC are nicely intermixed and do not cluster by the group labels. This is not case with most of the remaining methods where residuals cluster by group labels, thus indicating that they are unable to eliminate the underlying differences in sampling fractions between the two groups. Thus, under the null hypothesis of no difference in the absolute abundance of a taxon in two groups, their test statistics are not centered at zero. This results in inflated FDR (Appendix A). We also note from Figure 5, that not only ANCOM-BC does well in estimating the bias due to differences in sampling fraction, the variability in the estimates of the sampling fractions is very small as seen from the height of the box plot for ANCOM-BC. This is an important observation because it suggests that the variability in the estimator of bias due to sampling fraction is potentially negligible in the test statistic described in Methods section. Clearly, as seen in Figure 6, the normalization of data has a major effect on the FDR and power of various methods.

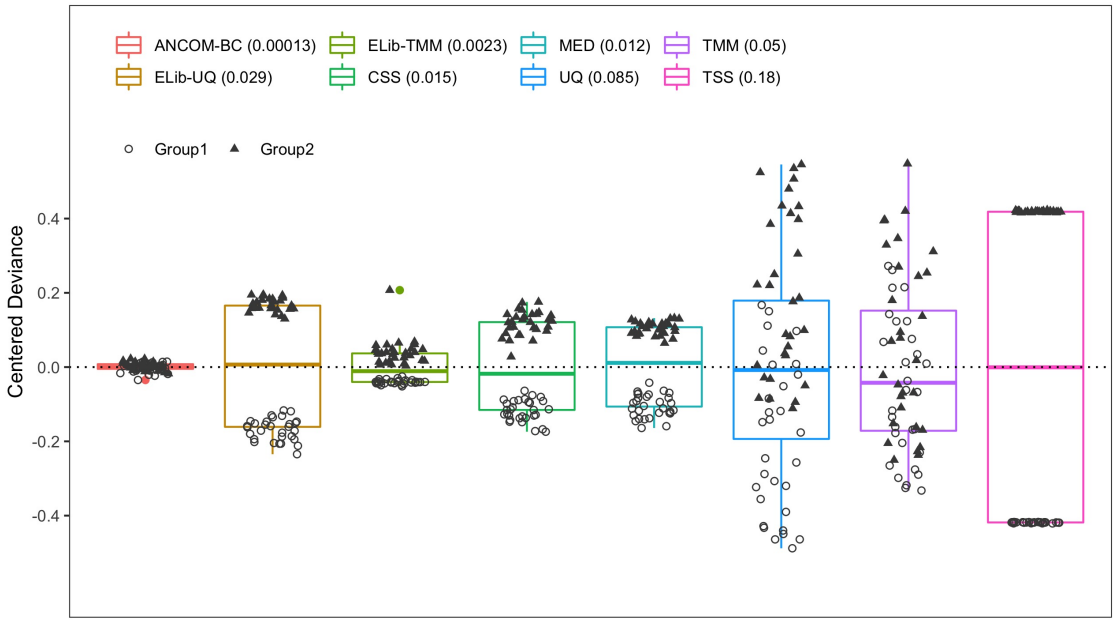


Figure 5: Box plot of residuals between true sampling fraction and its estimate.

2.3.2 Differential abundance analyses

Simulating data from Poisson-Gamma distributions, we evaluated the performance of various methods in terms of FDR and power. Since there is no hard threshold available for DR to declare whether a taxon is differentially abundant or not, it was not included in this simulation study.

Not surprisingly, standard Wilcoxon rank-sum test applied to relative abundance data leads to highly inflated FDR (Figure 6) in all simulation scenarios. This is primarily because such standard tests ignore the compositional structure of the data, and seen from Figure 5, TSS does not successfully normalize the data. Simply applying non-parametric tests without any normalization can also be problematic when the sampling fractions are different across experimental groups (Figure 6a). The two widely used count-based methods in RNA-Seq literature, edgeR (implemented using ELib-TMM (Chen et al., 2014) by default) and DESeq2, generally exceed the 5% nominal FDR level when there are differences in sampling fractions

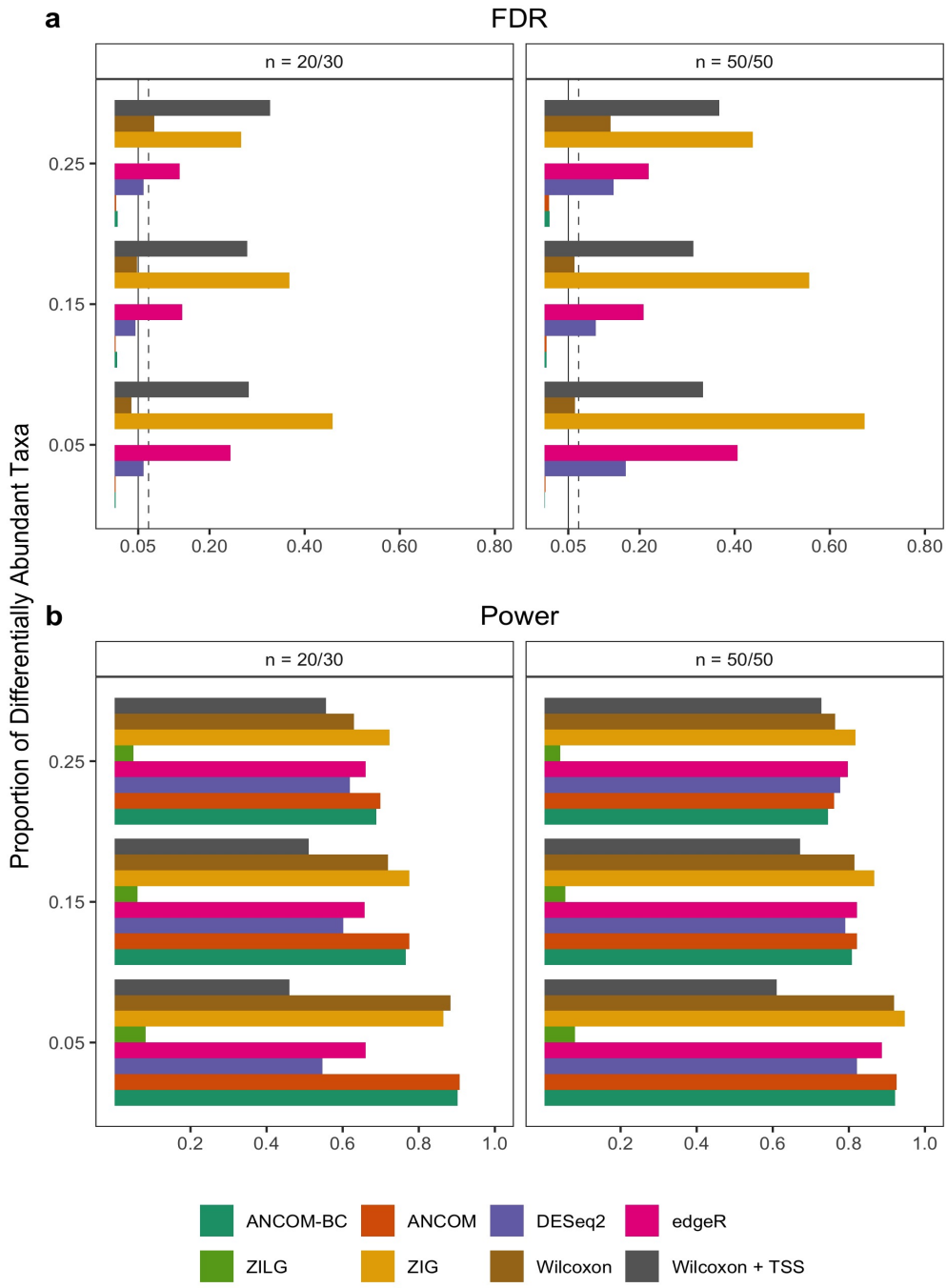


Figure 6: FDR and power comparisons using synthetic data.

(Figure 6a). For instance, edgeR has FDR as large as 40% (Figure 6a), meaning that 40% of findings could be potentially false positives. The zero-inflated Gaussian mixture model

used in metagenomeSeq (ZIG) consistently has the largest FDR when sampling fractions are not constant (Figure 6a). In some cases, the FDR could be as much as 70%, which perhaps is partly due to the Gaussian distribution assumption on log abundance data. Although metagenomeSeq using zero-inflated Log-Gaussian mixture model (ZILG) successfully controls the FDR under 5% in all simulations, it suffers a severe loss of power (Figure 6b). The power of detecting differentially abundant taxa could be lower than 10%.

Similar to ANCOM, ANCOM-BC not only controls the FDR at the nominal level (5%) but also maintains adequate power in all simulation settings considered here. An important observation to be made regarding all methods, other than ANCOM and ANCOM-BC, is that as the sample size within each group increases, so does the FDR. This is perhaps a consequence of the fact that the test statistics are not centered at the true null parameter but are shifted due to differences in the sampling fraction. Hence asymptotically, these tests fail to control the false positive as well as FDR (Appendix A).

In addition to the above Poisson-Gamma model, we performed simulations using the real global patterns data (Caporaso et al., 2011), to get a broader perspective on the performance of the various methods. In this case again, ANCOM and ANCOM-BC controlled the FDR and competed well in terms of power with all other methods. The estimated FDR of DESeq2 and edgeR increased further in this simulation set-up (Figure 7) compared to the simulation using Poisson-Gamma distribution. Note that DESeq2 and edgeR were designed for Poisson-Gamma distribution, and hence it is not surprising that these methods performed poorly in this new set-up.

2.4 Illustration using gut microbiota data

We illustrate ANCOM-BC by analyzing the US, Malawi and Venezuela gut microbiota data (Yatsunenko et al., 2012). This dataset consists of 11,905 OTUs obtained from subjects in the USA (n = 317), Malawi (n = 114), and Venezuela (n = 99). We first assessed the performance of different normalization methods mentioned above. One heuristic approach to gain insights on the impact of normalization is to examine how well the normalized samples

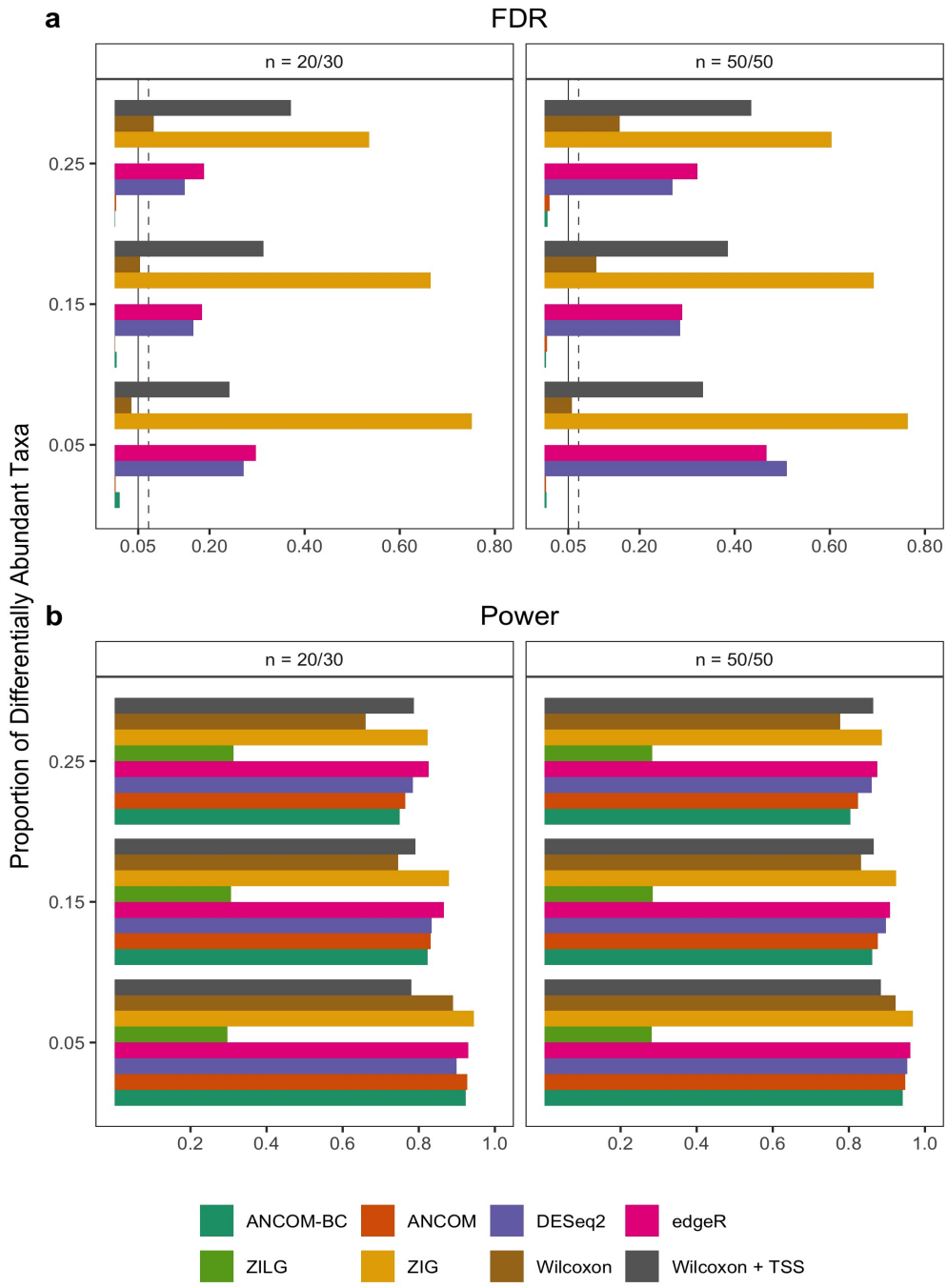


Figure 7: FDR and power comparisons using global pattern data.

separate with each other according to their phenotypes in a non-metric multidimensional scaling (NMDS) plot. We provide the results for Malawi and Venezuela populations in

Figure 8.

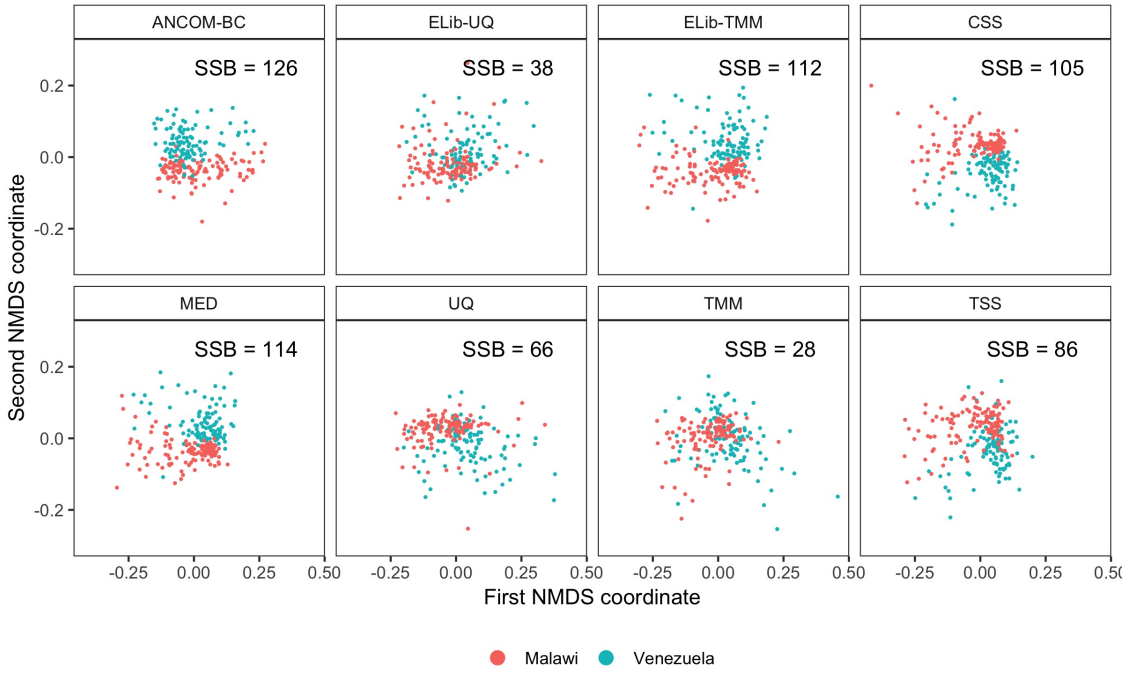


Figure 8: Non-metric multidimensional scaling (NMDS) visualizations of normalized data.

As seen from this figure, ANCOM-BC appears to perform very well visually in separating samples from the two populations and has the largest between-group sum of squares (SSB). SSB measures how well clusters are separated. Larger the SSB value the better a method is in clustering objects according to group labels. ELib-TMM, CSS and MED also performed well. Consistent with the bias correction and FDR/Power simulations reported in Figure 5 and Figure 6, where ELib-UQ, UQ, TMM and TSS perform poorly in correcting biases and have poor FDR control, they also have poor performances in distinguishing samples based on their nationalities.

We also report results of pairwise DA analyses at phylum level among the above three countries using ANCOM-BC. It is well-known that the infant gut microbiota evolve with their age (Lozupone et al., 2013b) due to changes in the feeding patterns, diet, and other exposures. Hence, for illustration purposes, we performed a stratified analysis by considering two age groups, infants below 2 years (labeled as "infants") and adults between 18 to 40 (labeled as "adults"). Results of all pairwise comparisons are provided in Figure 9a. Note that ANCOM-

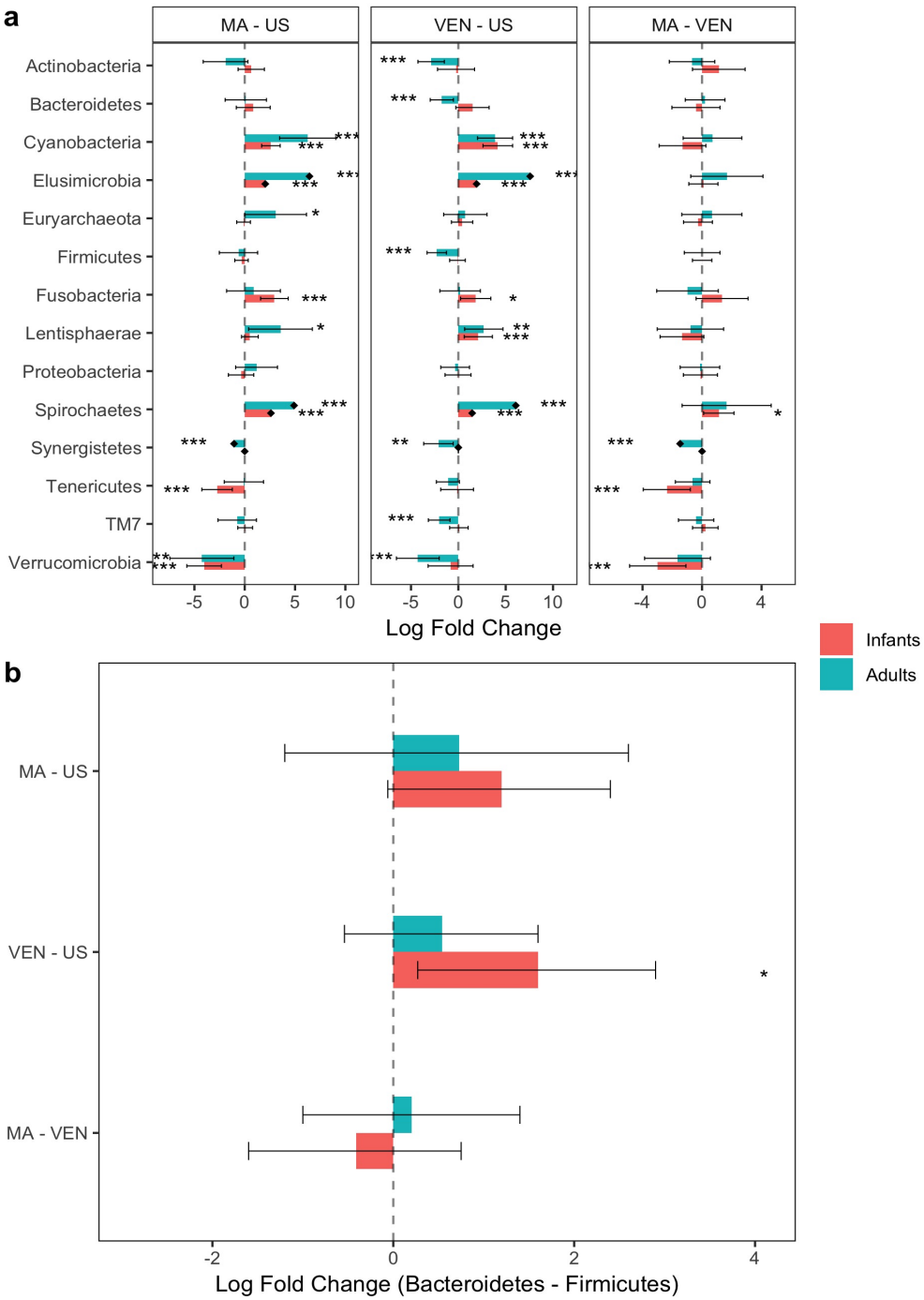


Figure 9: Analysis of the global gut microbiota data in phylum level.

BC is the first method in the literature that can not only identify differentially abundant taxa while controlling the FDR for multiple testing, it also provides 95% simultaneous confidence intervals for the mean differential abundance of each taxon in two experimental groups. These confidence intervals are adjusted for multiplicity using Bonferroni method. Thus, a researcher can evaluate the effect size associated with each taxon when comparing two experimental groups. This is particularly important in the present climate when researchers are increasingly skeptical about making decisions based on p-values (alone) (Amrhein et al., 2019).

Interestingly, it turns out that phyla such as Cyanobacteria, Elusimicrobia, Euryarcheota, and Spirochaetes, which are known to be associated with rural environment and hygiene (Codd, 1995; Herlemann et al., 2007; Obregon-Tito et al., 2015; Halperin, 2010), are significantly more abundant among Malawi than the US infants and adults. We discover an interesting trend in the absolute abundance of phylum Verrucomicrobia, whose absolute abundance is known to increase with antibiotics usage to protect against pathogens and other opportunistic bacteria (Dubourg et al., 2013). Consistent with the high usage of antibiotics in the western world among infants as well as adults, we discover a significant increase in the absolute abundance of Verrucomicrobia in US relative to Malawi adults and infants, and relative to Venezuelan adults (Figure 9a). Similarly, there is a significant increase in its absolute abundance among Venezuelan infants compared to Malawi (Figure 9a).

It is well-documented in the literature that BMI is linked to the ratio of Bacteroidetes to Firmicutes (Castaner et al., 2018). In our sample, the US infants, as well as adults, had higher BMI than their counterparts in Malawi; The US infants also had higher BMI than Venezuela infants (Table 6). Interestingly the ratio of Bacteroidetes to Firmicutes was larger among Malawi infants than the US infants (Figure 9b, Table 6). Similarly, the ratio was significantly larger among Venezuela infants than the US infants (Figure 9b, Table 6). Although the differences of the ratio of Bacteroidetes to Firmicutes between US and non-US adults were not significant, the effect sizes showed a similar trend as infants indicating that US adults had smaller ratio of Bacteroidetes to Firmicutes. We did not find any significant differences between Malawi and Venezuelan infants as well as adults. These results are in

line with with our findings that there were no differences in the mean absolute abundances of Firmicutes as well as Bacteroidetes among Malawi and Venezuelan infants as well as adults (Figure 9a).

2.5 Discussion

Identifying the microbial taxa that differentiate obtained samples is a challenging problem (Gloor et al., 2017; Morton et al., 2019), in part due to inaccessibility of data necessary for drawing inferences on differential abundance in two or more ecosystems. An important unobservable parameter that impacts DA analysis is the sampling fraction of a sample drawn from a unit volume of ecosystem. As noted in previous studies (Gloor et al., 2017; Morton et al., 2019), the bias correction due to sampling fraction is a major hurdle. While, ANCOM as well as DR procedures find ways to get around the problem from different perspectives, there is room for improvements.

ANCOM-BC enjoys several important unique characteristics. Firstly, it is the only method available in the literature that estimates the sampling fraction and performs DA analysis by correcting bias due to differential sampling fractions across samples. It is the only procedure that provides valid p-values and confidence intervals for each taxon. Secondly, unlike ANCOM, it simplifies DA analysis by recasting the problem as a linear regression problem with an off-set. The off-set is due to the sampling fraction. By virtue of linear regression formulation, ANCOM-BC can be applied to a broad collection of study designs, including longitudinal data, repeated measurements design, covariance adjusted analysis, and so on. Using a broad range of simulations studies, we demonstrate that ANCOM-BC, like ANCOM, controls the FDR very well, while almost all other methods investigated in this paper fail.

The ANCOM-BC methodology may not perform well when the sample sizes are very small, such as $n = 5$ per group. The FDR is not controlled by ANCOM-BC in such cases (Figure 10). However, when the sample size increases to 10, our simulation results indicate that ANCOM-BC controls FDR with adequate power (Figure 10). We also evaluated the performance of ANCOM-BC when the number of taxa is small, as when researchers perform

DA analysis at the phylum or class levels. Even in such instances, ANCOM-BC controls the FDR very well while maintaining high power (Table 7). ANCOM-BC performs best in terms of FDR control when the proportion of differentially abundant taxa (denoted by "Prop. DA") is not too large (e.g. less than 75%). Otherwise, it may have slightly elevated FDR (Figure 11). However, none of the other methods control the FDR either, in fact, they have larger FDRs than ANCOM-BC.

In summary, the proposed ANCOM-BC methodology (1) explicitly tests hypothesis regarding differential absolute abundance of individual taxon and provides valid confidence intervals; (2) provides an approach to correct the bias induced by (unobservable) differential sampling fractions across samples; (3) takes into account the compositionality of the OTU table, and (4) does not rely on strong parametric assumptions. With the linear regression framework adopted in ANCOM-BC, it allows researchers to derive p-value associated with each taxon as well as confidence interval estimation for differential absolute abundance. These are unique to ANCOM-BC, to our best knowledge. Last but not the least, because of the regression framework adopted in ANCOM-BC, it can be extended to more general settings involving multi-group comparisons, adjusting covariates as well as applying to longitudinal/repeated measurements data.

Table 6: Summary of the global gut microbiota data.

Infants (age $\leq$ 2, n = 133)			
	MA (n = 56)	**US (n = 49)**	**VEN (n = 28)**
Age			
Min	0.033	0.083	0.250
Max	2.0	2.0	2.0
Mean (SD)	0.99 (0.63)	0.55 (0.42)	1.1 (0.58)
BMI			
Min	11	14	14
Max	22	24	19
Mean (SD)	16 (1.9)	18 (3.4)	16 (1.4)
Gender (%)			
F	26 (46)	26 (53)	10 (36)
M	30 (54)	23 (47)	15 (54)
NA	0 (0)	0 (0)	3 (11)
Adults (18 $\leq$ age $\leq$ 40, n = 83)			
	MA (n = 21)	**US (n = 41)**	**VEN (n = 21)**
Age			
Min	20	23	18
Max	38	40	40
Mean (SD)	27 (4.9)	29 (5.3)	29 (7.4)
BMI			
Min	20	18	21
Max	26	66	41
Mean (SD)	22 (2.0)	27 (11)	30 (5.2)
Gender (%)			
F	21 (100)	39 (95)	20 (95)
M	0 (0)	2 (5)	1 (5)
NA	0 (0)	0 (0)	0 (0)

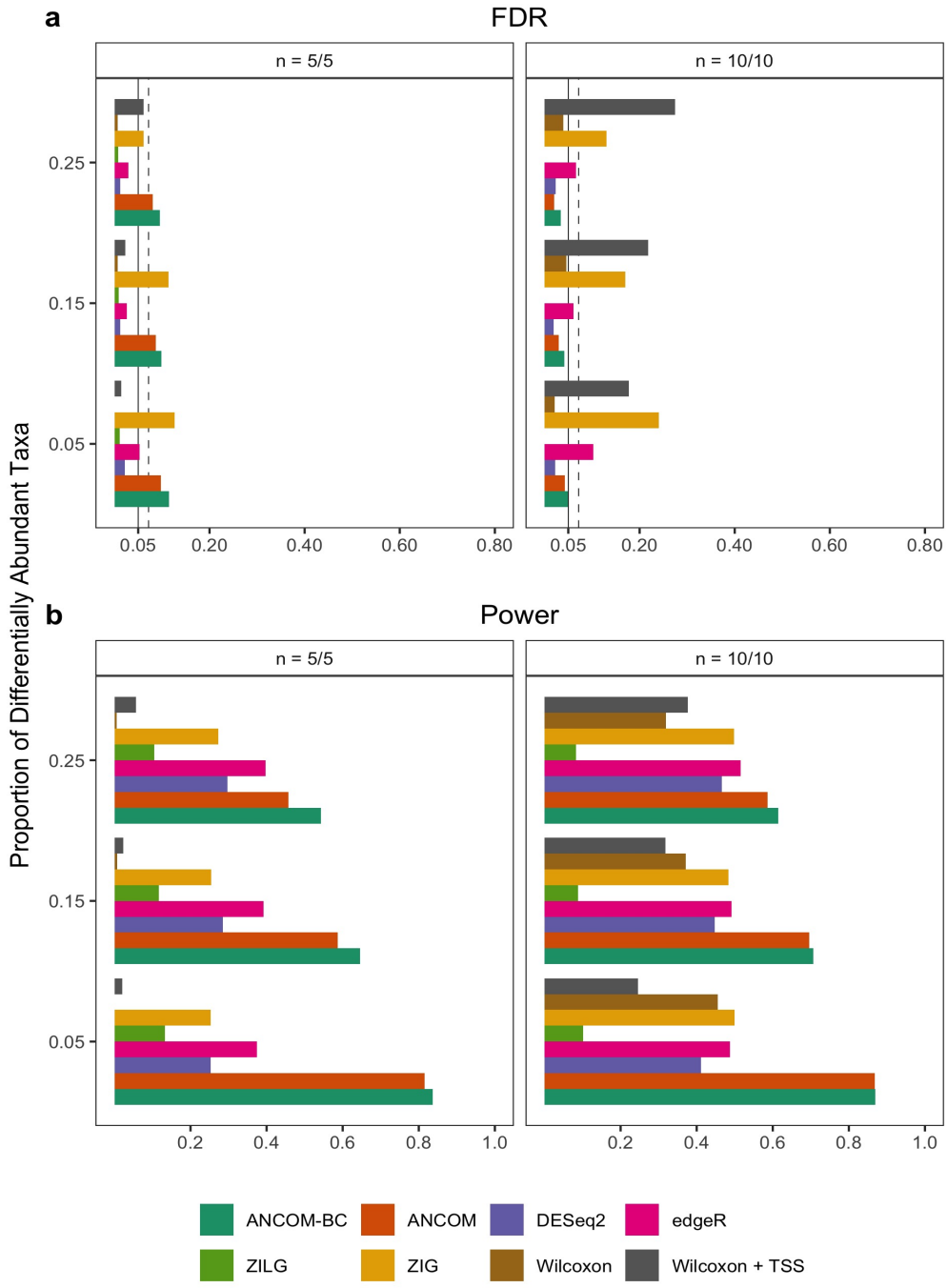


Figure 10: FDR and power comparisons with small sample size.

Table 7: FDR and power of ANCOM-BC when the number of taxa is small.

# Taxa	Sample Size	Prop. DA (%)	FDR	FDRSD	Power	PowerSD
10	20/30	25	0	0	0.96	0.14
10	50/50	25	0.0073	0.07	0.96	0.13
50	20/30	25	0.012	0.037	0.79	0.15
50	50/50	25	0.012	0.047	0.84	0.13

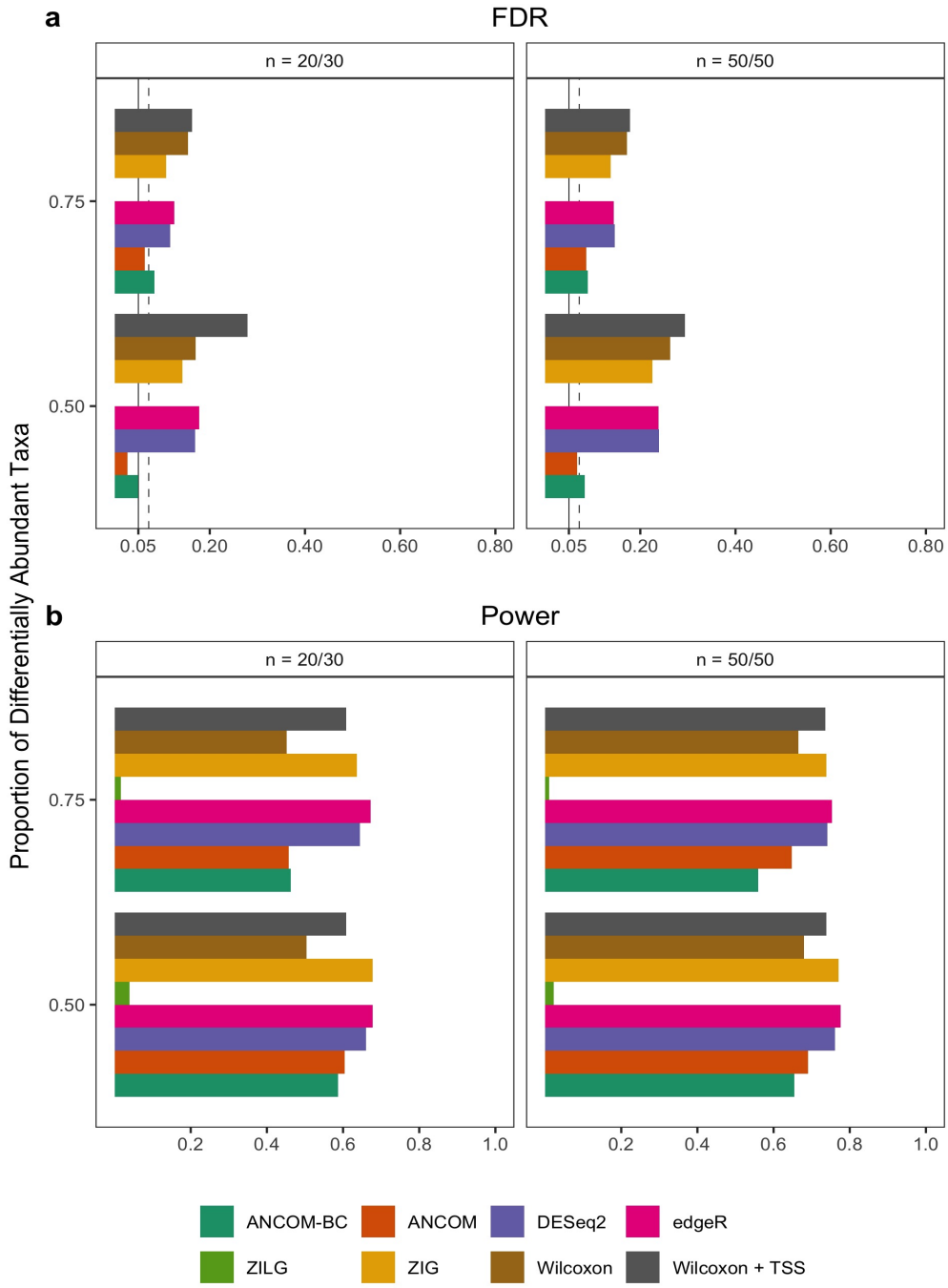


Figure 11: FDR and power comparisons with large Prop. DA.

3.0 Multi-group Analysis of Compositions of Microbiomes with Bias Correction

3.1 Introduction

In some applications, for a given taxon, researchers are interested in drawing inferences regarding differential abundance in more than two groups. For example, for a given taxon, researchers may want to test whether there exists at least one experimental group that is significantly different from others. We shall refer to this type of test as the global test. However, in some situations, researchers may be interested in testing whether the absolute abundance of the taxon increased or decreased across two or more pre-specified groups, and this type of test is known as the directional test. A special case of directional testing would be then one that tests if the abundance of a taxon in one or more groups is larger or smaller than its abundance in the control group, which is analogous to the one-sided version of the alternative hypothesis in Dunnett's test (Dunnett, 1955; Dunnett and Tamhane, 1991, 1992). In this setting, the researcher is not interested in comparing abundance among the remaining groups, we shall refer to this kind of test as the test against control. Lastly, when the experimental groups are ordered naturally, such as doses of exposure or duration of exposure or stages of a disease etc., for a given taxon, researchers may be interested in testing whether the abundance of the taxon is changing with the ordered experimental groups according to some specific pattern. Some examples of patterns include *monotonically increasing, umbrella shaped, loop shape* etc (Figure 12).

When performing the directional test or the test against control, two decisions are made for each taxon under this framework. Firstly, one needs to decide whether the taxon is differentially abundant between two groups. A statistical test at this stage can potentially lead to a Type I error. Once the taxon is declared to be differentially abundant, The second decision will be made to decide whether the abundance is increasing or decreasing in the second group as compared to the first one. Due to the underlying variability in the data, one can potentially make a wrong decision regarding the direction. For instance, a taxon

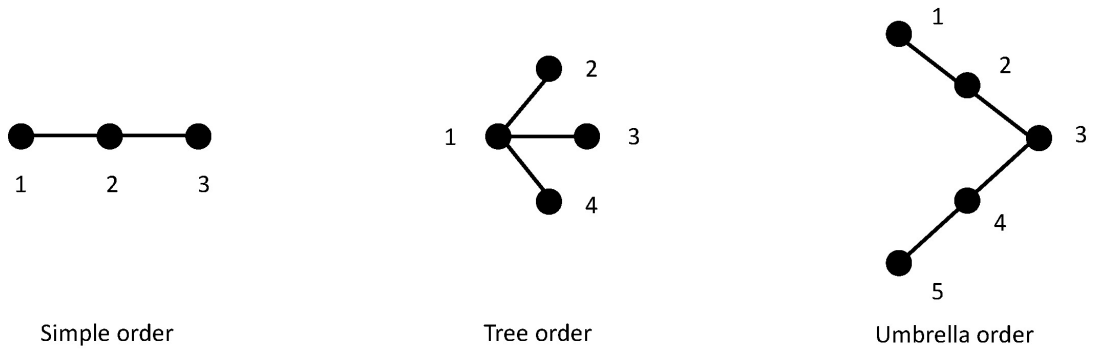


Figure 12: Graphs for ordered restrictions.

whose abundance is increasing, is falsely declared to have a decreasing direction. Thus, the total error one may want to control contains both Type I error and the directional error. In the context of multiple testing, involving thousands of taxa, one may want to control the false discovery rate together with the total directional error rate. This total error rate is known as mixed directional FDR (mdFDR), that accounts for all pairwise tests of interest are performed. Besides the general FDR controlling procedure such as B-H procedure (Benjamini and Hochberg, 1995), the mdFDR controlling procedure (Grandhi et al., 2016; Guo et al., 2010) should also be considered.

There are several patterns of common interest when dealing with tests for patterns. A common pattern is the *simple order* or *monotonic order*, which takes the form:

$$\mu_1 \leq \mu_2 \leq \ldots \leq \mu_g,$$

where g denotes the total number of experimental groups. Similarly, if the study aims to compares some new treatments to the standard treatment, it is straightforward to assume a *tree order*, which is defined as,

$$\mu_1 \leq \mu_i, \ i = 2, \ldots, g.$$

In some cases, researchers may also be interested in identifying taxa that have more complicated patterns such as *umbrella* order, where, for example, the mean absolute abundance of

taxa have the pattern such as the following. Typically, the location of the peak is also an unknown parameter.

$$\mu_1 \leq \mu_2 \leq \mu_{i-1} \leq \mu_i \geq \mu_{i+1} \geq \mu_{g-1} \geq \mu_g.$$

3.2 Methods

3.2.1 Global test

For simplicity of exposition, we split the covariates X into two parts, where X_1 stands for the assignment of group, and X_2 denotes the remaining covariates, i.e., in the log-linear regression model

$$y_i = d + X_1\beta_i + X_2\gamma_i + \epsilon_i, \tag{3.1}$$

where

(1) β_i is the vector of group effects of the order $g \times 1$,

(1) X_1 is the design matrix of the order $n \times g$ consisting of 0s and 1s,

(2) X_2 is the known matrix of other covariates of the order $n \times (p-g)$ with the corresponding regression parameter vector γ_i of the order $(p-g) \times 1$.

The global test intends to test

$$H_{0,i} : \cap_{k \neq k' \in \{1,\ldots,g\}} \beta_{ik} = \beta_{ik'},$$
$$H_{1,i} : \cup_{k \neq k' \in \{1,\ldots,g\}} \beta_{ik} \neq \beta_{ik'},$$

which can be reformulated as

$$H_{0,i} : A\beta_i = 0,$$
$$H_{1,i} : A\beta_i \neq 0.$$

where

$$A = \begin{bmatrix} 1 & -1 & 0 & \ldots & 0 \\ 1 & 0 & -1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & \ldots & 0 & -1 \end{bmatrix}$$

with the test statistic

$$W_i = (A\hat{\beta}_i - A\beta_i)^T (A\hat{\Sigma}_i A^T)^{-1} (A\hat{\beta}_i - A\beta_i) \to_d \chi^2_q. \tag{3.2}$$

3.2.2 Directional test

If we are interested in knowing whether the absolute abundance increased or decreased between all pairs of group, this is equal to test

$$H_{0,i,k,k'} : \beta_{ik} = \beta_{ik'}$$
$$H_{1,i,k,k'} : \{\beta_{ik} < \beta_{ik'}\} \cup \{\beta_{ik} > \beta_{ik'}\},$$

where $k \neq k' \in \{1, \ \ldots, \ g\}$.

Denote the test statistic for pairwise comparison as

$$W_{i,kk'} = \frac{\hat{\beta}_{ik} - \hat{\beta}_{ik'}}{\sqrt{\widehat{Var}(\hat{\beta}_{ik}) + \widehat{Var}(\hat{\beta}_{ik'})}} \to_d N(0,1), \quad \text{as } n \to \infty, \tag{3.3}$$

where $\widehat{Var}(\hat{\beta}_{ik})$, $\widehat{Var}(\hat{\beta}_{ik'})$ are the k^{th} and k'^{th} diagonal elements of $\hat{\Sigma}_i$ defined in (2.18), respectively.

Thus, the raw p-value for comparing group k and group k' is defined as:

$$p_{i,kk'} = 2[1 - \phi(|W_{i,kk'}|)]. \tag{3.4}$$

Note that the null and alternative hypothesis for global test are denoted as $H_{0,i}$ and $H_{1,i}$, a Type I error might occur due to wrongly rejecting $H_{0,i}$ or correctly rejecting $H_{0,i}$ but wrongly rejecting $H_{0,i,k,k'}$ for some $i = 1, \ldots, m$; A directional error might occur due to correctly rejecting $H_{0,i}$ but wrong assignment of the direction between β_{ik} and $\beta_{ik'}$ while correctly rejecting $H_{0,i,k,k'}$. In this case, we need to control the error rate combining both Type I and the directional errors in the FDR framework, which is referred to as mixed directional FDR (mdFDR) (Guo et al., 2010; Grandhi et al., 2016).

Definition 3.2.1 (mdFDR)**.** Let $V(i)$ denote the indicator function of at least one Type I error or directional error committed, i.e.

$$V(i) = \begin{cases} 1 & \text{if Type I or directional error occurs,} \\ 0 & \text{otherwise.} \end{cases} \tag{3.5}$$

Then, mdFDR is defined as the expected proportion of Type I and directional errors among all discovered taxa.

$$mdFDR = E(\frac{\sum_{i=1}^{m} V(i)}{\max(R, 1)}). \tag{3.6}$$

To control the mdFDR for all pairwise tests, we adopt the general mdFDR controlling procedure (Grandhi et al., 2016), and do the following:

(1) Apply global test method stated above to obtain the p-value (p_i) for each taxon. Apply Bonferroni correction (default) or BH procedure to identify taxa that are differentially abundant in at least one pairwise comparison. Let R denote the number of taxa being discovered.

(2) For each taxon discovered in step 1, apply the Bonferroni correction to the pairwise p-values ($p_{i,kk'}$) at level $R\alpha/m$.

(3) For a given taxon discovered in step 1, if a pairwise hypothesis is rejected in step 2, then we declare $\beta_{ik} < \beta_{ik'}$ or $\beta_{ik} > \beta_{ik'}$ according to $W_{i,kk'} < 0$ or > 0.

It has been proved that under assumption of independence of p-values obtained from the global test, i.e., $p_i, i = 1, \ldots, m$, the mdFDR of the above procedure is strongly controlled at level α (Grandhi et al., 2016).

3.2.3 Test against a specific group

Often researchers are interested in knowing whether the absolute abundance increased or decreased in an ecosystem relative a pre-specified group, say the control group. Suppose group 1 is the control group, it aims to test

$$H_{0,i,k} : \beta_{ik} = \beta_{i1},$$
$$H_{1,i,k} : \{\beta_{ik} < \beta_{i1}\} \cup \{\beta_{ik} > \beta_{i1}\},$$

where $k \in \{2, \ldots, g\}$.

Similarly, the pairwise test statistic is defined as

$$W_{i,k} = \frac{\hat{\beta}_{ik} - \hat{\beta}_{i1}}{\sqrt{\widehat{Var}(\hat{\beta}_{ik}) + \widehat{Var}(\hat{\beta}_{i1})}} \to_d N(0,1), \quad \text{as } n, \to \infty \tag{3.7}$$

where $\widehat{Var}(\hat{\beta}_{ik})$, $\widehat{Var}(\hat{\beta}_{i1})$ are the k^{th} and 1^{th} diagonal elements of $\hat{\Sigma}_i$ defined in (2.18), respectively.

Thus, the raw p-value for comparing group k and group 1 is defined as:

$$p_{i,k} = 2[1 - \phi(|W_{i,k}|)] \tag{3.8}$$

Likewise, we apply the mdFDR controlling procedure for all pairwise tests.

3.2.4 Test for patterns

3.2.4.1 Simple order The most common pattern is called simple order or monotonic order. We intend to test

$$H_{0,i} : \beta_{i1} = \beta_{i2} = \ldots = \beta_{ig},$$

$$H_{1,i} : \beta_{i1} \leq \beta_{i2} \leq \ldots \leq \beta_{ig} \text{ with at least one strict inequality,}$$

$$\text{or } H_{1,i} : \beta_{i1} \geq \beta_{i2} \geq \ldots \geq \beta_{ig} \text{ with at least one strict inequality.}$$

WLOG, suppose the alternative hypothesis is $H_{1,i} : \beta_{i1} \leq \beta_{i2} \leq \ldots \leq \beta_{ig}$ with at least one strict inequality.

Obtaining the estimate for β_i under constraint is asymptotically equivalent to the following optimization problem

$$\begin{aligned} \hat{\beta}_i^{opt} &= \underset{\beta_i \in \mathbb{R}^g}{\arg\min} (\hat{\beta}_i - \beta_i)^T (\widehat{Var}(\hat{\beta}_i))^{-1} (\hat{\beta}_i - \beta_i), \\ &\text{s.t. } \beta_{i1} \leq \beta_{i2} \leq \ldots \leq \beta_{ig} \text{ with at least one strict inequality,} \end{aligned} \tag{3.9}$$

where $\hat{\beta}_i$ is the first g elements in (2.36), and $\widehat{Var}(\hat{\beta}_i)$ is upper left $g \times g$ block of $\hat{\Sigma}_i$ defined in (2.18).

The solution to (3.9) can be numerically obtained through some convex optimization algorithms, such as CVRX (Fu et al., 2017).

To test the hypothesis stated above, we use William's type test statistic

$$W_i = \frac{|\hat{\beta}_{ig}^{opt} - \hat{\beta}_{i1}^{opt}|}{\sqrt{\widehat{Var}(\hat{\beta}_{ig}) + \widehat{Var}(\hat{\beta}_{i1})}}. \tag{3.10}$$

The null distribution of the test statistic is constructed by simulation. Under null, the expectations for $\hat{\beta}_{ik}, k = 1, \ldots, g$ are the same, thus, we can then construct the null distribution of W_i by doing the following:

(1) Generate $\hat{\beta}_{ik} \sim \sqrt{\widehat{Var}(\hat{\beta}_{ik})}N(0,1), l = 1, \ldots, g$,

(2) Get constrained regression estimators for $\hat{\beta}_{ik}^{opt,(b)}$,

(3) Compute $W_i^{(b)} = \frac{|\hat{\beta}_{ig}^{opt,(b)} - \hat{\beta}_{i1}^{opt,(b)}|}{\sqrt{\widehat{Var}(\hat{\beta}_{ig}) + \widehat{Var}(\hat{\beta}_{i1})}}$,

(4) Repeat the above steps B times, we get the null distribution of W_i by $(W_i^{(1)}, \ldots, W_i^{(B)})^T$.

The raw p-value is calculated as

$$p_i = \frac{1}{B}\sum_{b=1}^{B} I(W_i^{(b)} > W_i). \tag{3.11}$$

Then we apply the Holm method or BH procedure on raw p-values to control the FDR.

3.2.4.2 Tree order If the study aims to compares some new treatments to the standard treatment, it is straightforward to assume a tree order, which aims to test

$$H_{0,i} : \beta_{i1} = \beta_{i2} = \ldots = \beta_{ig},$$
$$H_{1,i} : \beta_{i1} \leq \beta_{il}, l = 2, \ldots, g \text{ with at least one strict inequality.}$$

The testing procedure is similar to the case of simple order:

(1) Obtain constrained regression estimators $\hat{\beta}_{i1}^{opt}$ and $\hat{\beta}_{il}^{opt}$,

(2) Use William's type of test statistic $W_i = \max_{l=2,\ldots,g} \frac{|\hat{\beta}_{il}^{opt} - \hat{\beta}_{i1}^{opt}|}{\sqrt{\widehat{Var}(\hat{\beta}_{il}) + \widehat{Var}(\hat{\beta}_{i1})}}$

(3) Construct the null distribution of test statistic by simulation,

(4) Compute raw p-values,

(5) Apply the Holm method or BH procedure on raw p-values to control the FDR.

3.2.4.3 Umbrella order In some cases, researchers may also be interested in identifying taxa that have more complicated patterns such as umbrella order, where the effect of covariate of interest has the pattern like

$$H_{0,i} : \beta_{il} = \beta_{i2} = \ldots = \beta_{ig},$$

$$H_{1,i} : \beta_{i1} \leq \ldots \leq \beta_{i,k-1} \leq \beta_{ik} \geq \beta_{i,k+1} \geq \ldots \geq \beta_{ig}, \text{ with at least one strict inequality.}$$

The testing procedure is similar to the case of simple order:

(1) Obtain constrained regression estimators $\hat{\beta}_{i1}^{opt}$, $\hat{\beta}_{ik}^{opt}$, and $\hat{\beta}_{ig}^{opt}$,

(2) Use William's type of test statistic $W_i = \max(\frac{|\hat{\beta}_{ik}^{opt} - \hat{\beta}_{i1}^{opt}|}{\sqrt{\widehat{Var}(\hat{\beta}_{ik}) + \widehat{Var}(\hat{\beta}_{i1})}}, \frac{|\hat{\beta}_{ik}^{opt} - \hat{\beta}_{ig}^{opt}|}{\sqrt{\widehat{Var}(\hat{\beta}_{ik}) + \widehat{Var}(\hat{\beta}_{ig})}})$

(3) Construct the null distribution of test statistic by simulation,

(4) Compute raw p-values,

(5) Apply the Holm method or BH procedure on raw p-values to control the FDR.

3.2.5 ANCOM-BC for mixed effects models

Global test and directional test under mixed effects model are similar to those of fixed effects model, therefore, we skip the derivations here.

To draw statistical inference under inequality constraints in linear mixed effects model, we use William's type test statistic as specified in the last section, and adopt Constrainted Linear Mixed Effects (CLME) framework (Jelsema et al., 2016) into ANCOM-BC. The procedure is summarized as follows:

(1) Obtain $\hat{\beta}_i$, the estimate of β_i under the null hypothesis,

(2) Compute the observed values of random effects and residuals.
$\hat{\epsilon}_i = (I - X(X^T \hat{H}^{-1} X)^{-1} X^T \hat{H}^{-1}) y_i^*$, and $\hat{\alpha} = \hat{D} Z^T \hat{H}^{-1} \hat{\epsilon}_i$,

(3) Standardize the observed values of random effects and residuals. Define $\hat{a} = SE(\hat{\alpha})^{-1} \hat{\alpha}$, and $\hat{e} = SE(\hat{\epsilon})^{-1} \hat{\epsilon}$, where $SE(\cdot)$ denotes the standard error,

(4) Obtain bootstrap samples. Let a^* and e^* denote the bootstrap samples of a and e, respectively. Then define $\hat{\alpha}^* = SE(\hat{\alpha}) a^*$ and $\hat{\epsilon}^* = SE(\hat{\epsilon}) e^*$. Finally construct the final bootstrap sample as: $y^{(b)} = \hat{d} + X\hat{\beta}_i + Z\hat{\alpha}^* + \hat{\epsilon}^*$,

(5) Repeat B times ($b = 1, \ldots, B$), we construct the null distribution for the test statistic.

(6) Compute raw p-values,

(7) Apply the Holm method (default) or BH procedure on raw p-values to control the FDR.

3.3 Simulation study

We now evaluate the performance of ANCOM-BC using three different simulation studies: The first study is designed for global test, and the second focuses on pairwise directional tests, and the last study to detect patterns across different groups. Taxa abundance in the ecosystem are generated using the Poisson-Gamma model.

Each simulation data consists of 500 taxa and 4 experimental groups. We considered two patterns of sample sizes. The pattern corresponding to small sample size consists of 15 samples in each group, summing up to a total of 60 samples in the four groups. The pattern corresponding to large sample size consists of 25 samples within each group, resulting in a total of 100 samples. The magnitude and the distribution of absolute abundance are the same as two-group settings. We considered five different patterns of Prop. DA's, ranging from 10% to 50%, with a 10% increments. The log fold change for differentially abundant taxa were either uniformly distributed from 1 to 10, representing increase in abundance relative to the control group, or uniformly distributed from 0.1 to 1, representing decrease in abundance relative to the control group. Furthermore, 20% of taxa were structural zeros in the control group. All simulation experiments were repeated 100 times to estimate the FDR.

The results of global test are summarized in Figure 13. As seen from this figure, under all configurations, ANCOM-BC has estimated FDR below the nominal level of 0.05 (in fact less than 0.02), while enjoying high powers (greater than 0.70). Furthermore, as one would desire, the power of this test increases with sample size. The FDR trend with respect to Prop. DA is consistent with the previously described ANCOM-BC methodology. The FDR decreases with the Prop. DA (from 10% to 40 %), and then increases (at 50%) due to small samples sizes in the built-in EM algorithm used in ANCOM-BC methodology. Interestingly, the power decreases with an increase in Prop. DA. This is due to the simulation settings

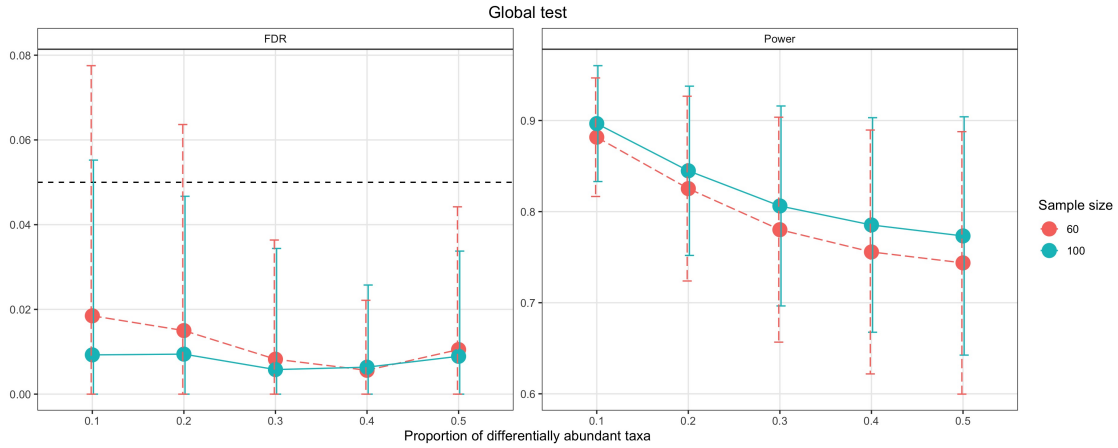


Figure 13: FDR and power comparisons for global test.

in relation to structural zeros. When the Prop. DA is small, ANCOM-BC benefits from primarily detecting structural zeros and hence has large powers.

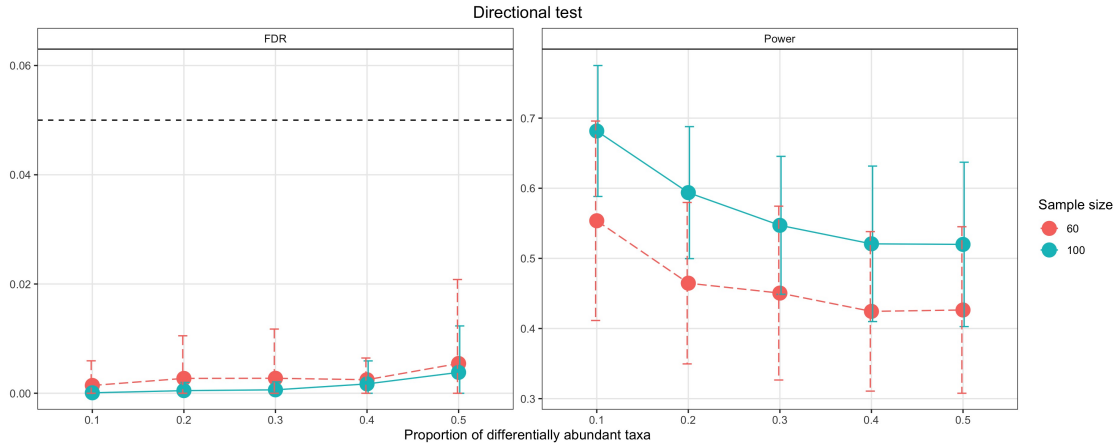


Figure 14: FDR and power comparisons for directional test.

In the case of directional test, the FDR and power comparison results are also reported in Figure 14. Based on results of simulations, ANCOM-BC has estimated FDRs well below the nominal level (even less than 0.01), and powers around 0.50. As compared to the global test, one may notice a reduced trend in both FDR and power decreases in the equivalent setting. This is because the mdFDR controlling procedure is more conservative than general

FDR controlling procedures, say BH procedure which can inflate the FDR since it does not account for all possible pairwise comparisons.

We also evaluated the performance of ANCOM-BC for three commonly encountered patterned alternative hypotheses, namely, *simple order*, *tree order*, and *umbrella order*. To generate data accordingly, we considered a total of 3 patterns in this simulation study: *Monotonically increasing* in absolute abundance from the first to the fourth group, which corresponds to data with *simple order*, and the null and the alternative hypotheses are given by:

$$H_0 : \mu_{i1} = \mu_{i2} = \mu_{i3} = \mu_{i4},$$
$$H_1 : \{\mu_{i1} \leq \mu_{i2} \leq \mu_{i3} \leq \mu_{i4}\} \cup \{\mu_{i1} \geq \mu_{i2} \geq \mu_{i3} \geq \mu_{i4}\}.$$

To generate data with *tree order*, with the control group having the smallest mean absolute abundance compared to other groups. Thus, the alternative hypothesis is given by H_1 below.

$$H_0 : \mu_{i1} = \mu_{i2} = \mu_{i3} = \mu_{i4},$$
$$H_1 : \{\mu_{i1} \leq \mu_{ik}, k \neq 1\} \cup \{\mu_{i1} \geq \mu_{ik}, k \neq 1\}.$$

In the case of *umbrella order*, we generated data so that the mean absolute abundance was maximum abundance in group 2. More precisely, the mean absolute abundance increases from group 1 to group 2 and then decreases as in H_1 described below. Thus the null and the alternative hypotheses are given by:

$$H_0 : \mu_{i1} = \mu_{i2} = \mu_{i3} = \mu_{i4},$$
$$H_1 : \{\mu_{i1} \leq \mu_{i2} \leq \mu_{i3} \leq \mu_{i4}\} \cup \{\mu_{i1} \leq \mu_{i2} \leq \mu_{i3} \geq \mu_{i4}\}$$
$$\{\mu_{i1} \leq \mu_{i2} \geq \mu_{i3} \geq \mu_{i4}\} \cup \{\mu_{i1} \geq \mu_{i2} \geq \mu_{i3} \geq \mu_{i4}\}$$
$$\{\mu_{i1} \geq \mu_{i2} \geq \mu_{i3} \leq \mu_{i4}\} \cup \{\mu_{i1} \geq \mu_{i2} \leq \mu_{i3} \leq \mu_{i4}\}.$$

We first estimated FDR and power without without paying attention to the pattern i.e., a rejection of null hypothesis given there truly exists a pattern regardless which pattern it is will be treated as true positive; While a false rejection of null hypothesis will be considered as false positive. Based on Figure 15, as expected, the power of detecting underlying patterns increases with larger sample size for all simulated patterns, and has good power, usually greater than 0.70. The FDR is often well controlled under the nominal level except the cases

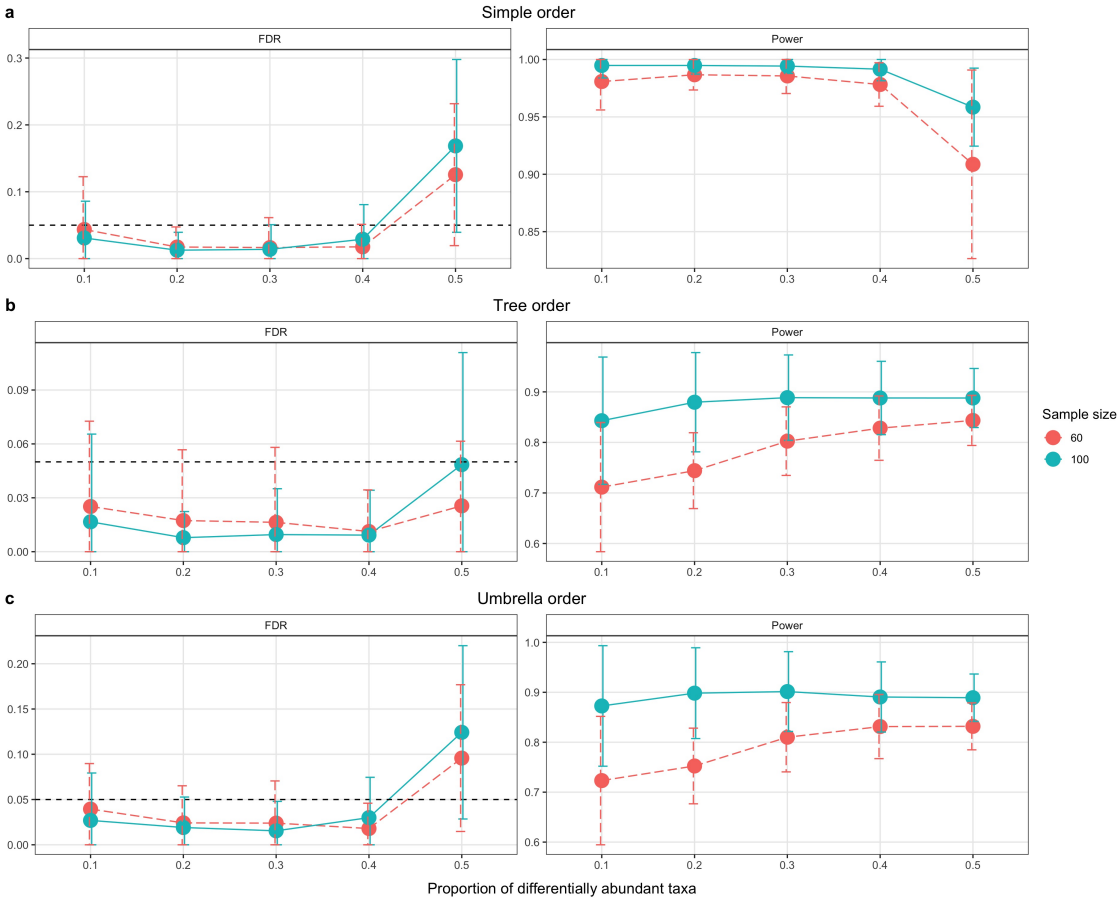


Figure 15: FDR and power comparisons for testing patterns (pattern matching ignored).

of simple order and umbrella order, and Prop. DA reaches 0.5. The decreasing trend of FDR vs. Prop. DA maintains when the Prop. DA is relatively small (from 10% to 40%), while one may expect an elevated FDR when Prop. DA is getting larger, as we can see from the case where Prop. DA is 50%.

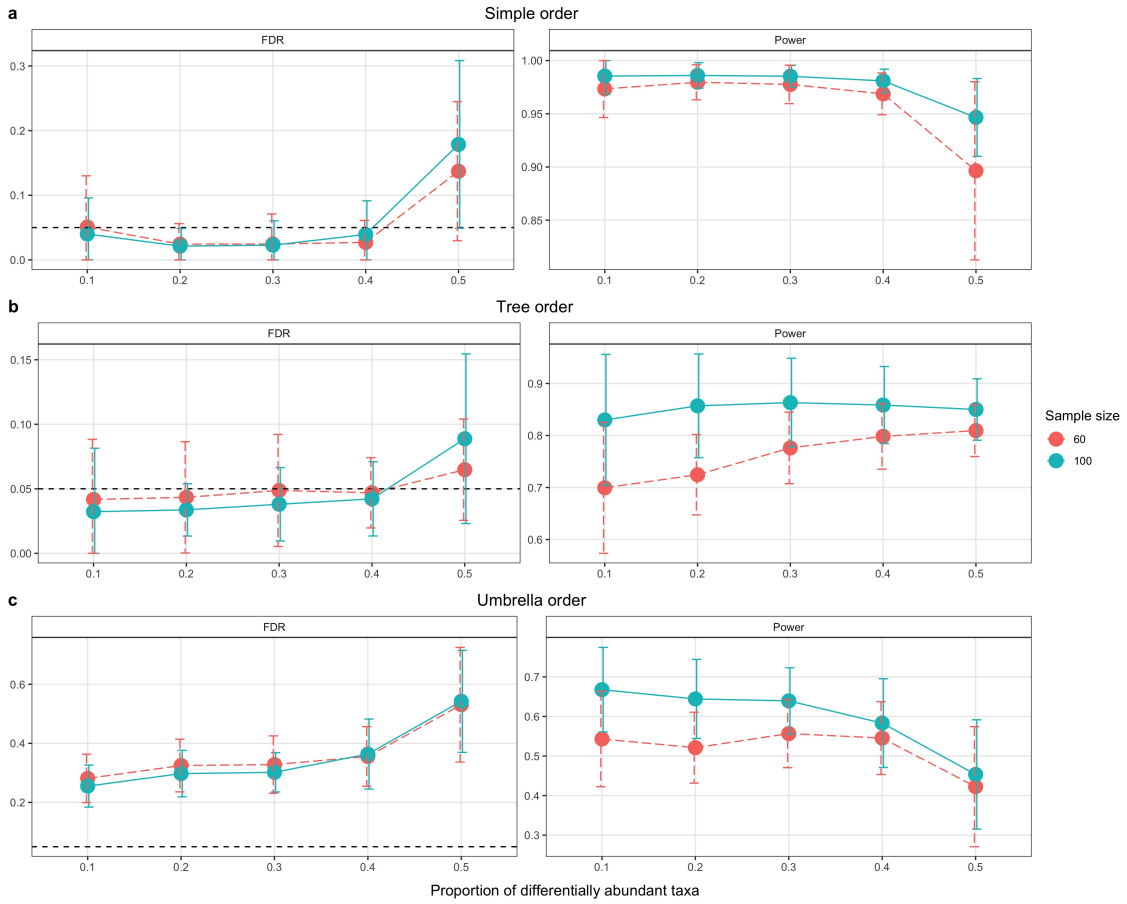


Figure 16: FDR and power comparisons for testing pattern (pattern matching considered).

We also estimated the FDR and power in case when the goal is to identify the exact pattern. FDR and power when the exact pattern matching is taken into account, i.e. a true positive shall be represented by rejecting the null hypothesis with the correct alternative; Otherwise, the rejection of the null will be regarded as false positive. As we can see from Figure 16, whether or not taking the exact pattern matching into account, it barely affects the FDR and power when testing patterns of *simple order* or *tree order*; However, in the case of *umbrella order*, there exist much larger number of alternatives, it is inevitable to assign

incorrect umbrella pattern. (from generally greater than 0.70 to smaller than 0.70), and an inflated FDR (greater than 0.30). One caveat is, in this case, a larger Prop. DA would lead to a even highly inflated FDR (around 0.50).

3.4 Analysis of global human gut microbiome data

We illustrate ANCOM-BC using a publicly available global human gut microbiome data (Yatsunenko et al., 2012). We subdivide the data into two age categories "$\leq$ 2 years" and "$>$ 2 years." This stratification is done since it is well-known that microbial composition of infants changes drastically when they switch over from breast milk (or formula milk) to solid food (Lozupone et al., 2013a). The sample sizes in the two strata ($\leq$ 2 years, $>$ 2 years) for Malawi (MA), US, and Venezuela (VEN) samples were (47, 36), (50, 260) and (27, 70), respectively. After several pre-processing steps of the raw OTU data we obtain for each age group and each location, a matrix of observations with $m = 11,905$ OTUs and inherent structural zeros. In agreement with the two-group comparison, we aggregate the OTU table and perform the DA analysis at phylum level. The total number of phyla is 39.

It is well-known that during early stages of human life gut microbial composition and diversity changes dramatically due to increased exposure to baby's environment, such as parents, siblings, diet and so on. In particular, diet plays an important role in changing the gut microbiome composition as babies begin to rely less on breast/formula milk and start eating other foods. Interestingly, our findings reported in the left panel of Figure 17, are consistent with this. As expected the abundance of the phylum Bacteroidetes significantly increased with age, while Actinobacteria decreased with age. This is directly related to change in feeding habits, less dependence on breast/formula milk. However, it is well-known that after 2 years, at the phylum level, there is very little difference in the microbial composition of a two year old and the mother. This is because in most cultures, the diet and environmental exposures of a two-year old is similar to that of the mother. Thus, we do not expect much temporal change in the gut microbiome composition after 2 years. This is reflected in our analysis, summarized in the right panel of Figure 17. As expected the

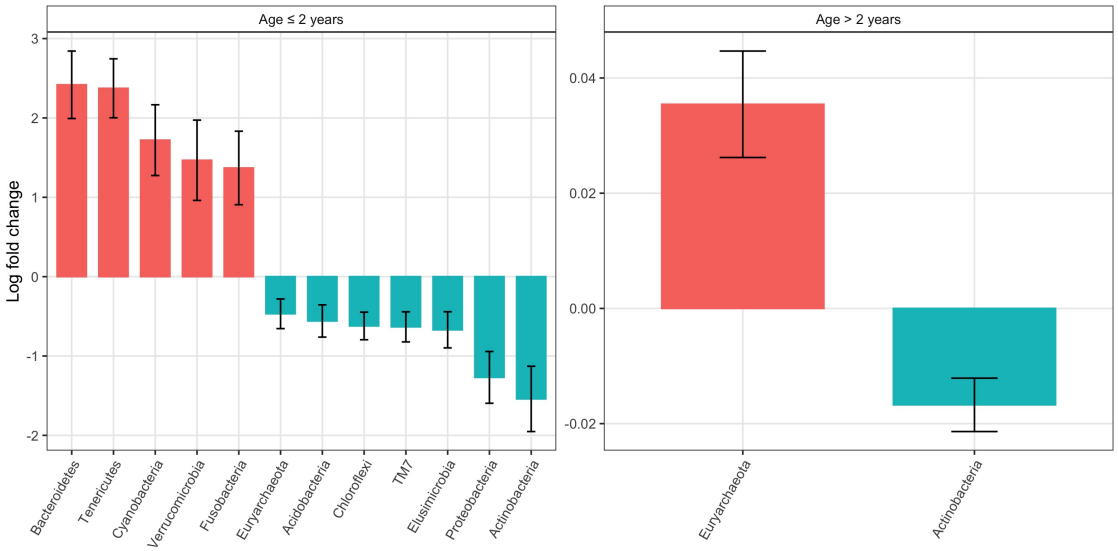


Figure 17: Age effect on microbial absolute abundance of the global gut microbiota data.

abundance of Actinobacteria decreases with age.

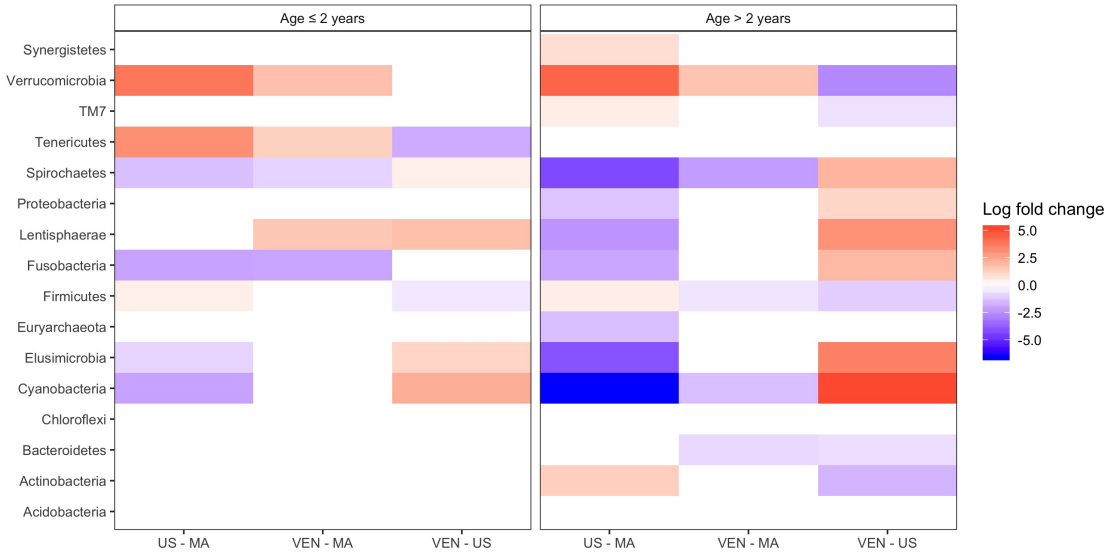


Figure 18: Pairwise differential abundance analyses on locations using ANCOM-BC.

Instead of performing a stratified analysis by different age groups, in this section, we look into the location effect by performing the DA analysis adjusted for age directly and

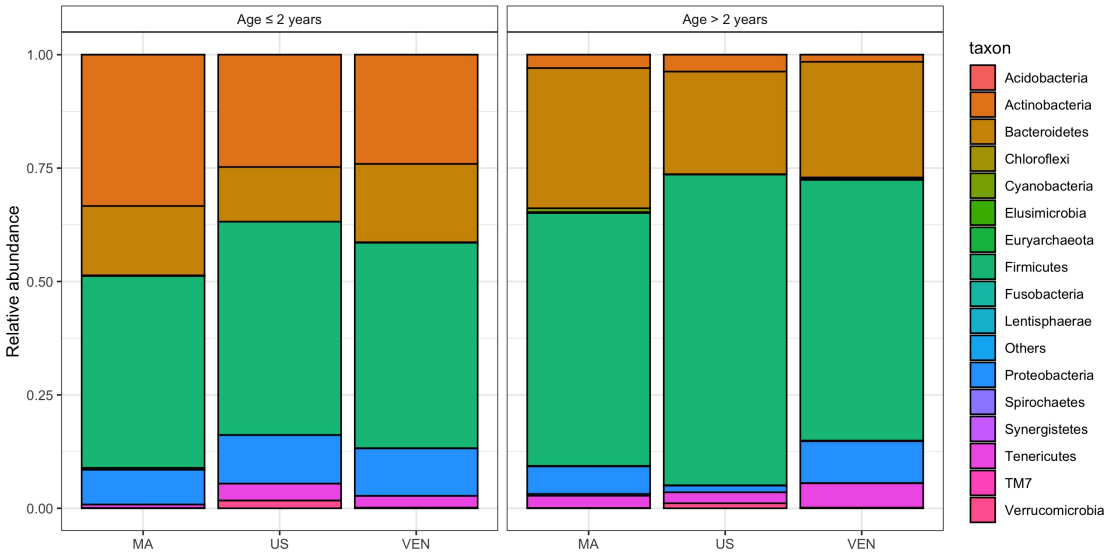


Figure 19: Relative abundance by location in phylum level.

controlling the mdFDR. Based on Figure 18, in line with what we found earlier, phyla such as Cyanobacteria, Elusimicrobia, and Spirochaetes, are significantly more abundant among Malawi than the US regardless of the age group. On the other hand, the absolute abundance of Verrucomicrobia shows a similar trend as we discovered previously that it is more abundant in US relative to Malawi adults and infants, and relative to Venezuelan adults. The relative abundance plot (Figure 19) could also be used as a reference to verify the results stated above.

In addition to pairwise DA analyses, we also apply ANCOM-BC for pattern discovery and see if there exist any patterns between absolute abundance with regards to the location effect while adjusting for the age effect. The results are summarized in Figure 20.

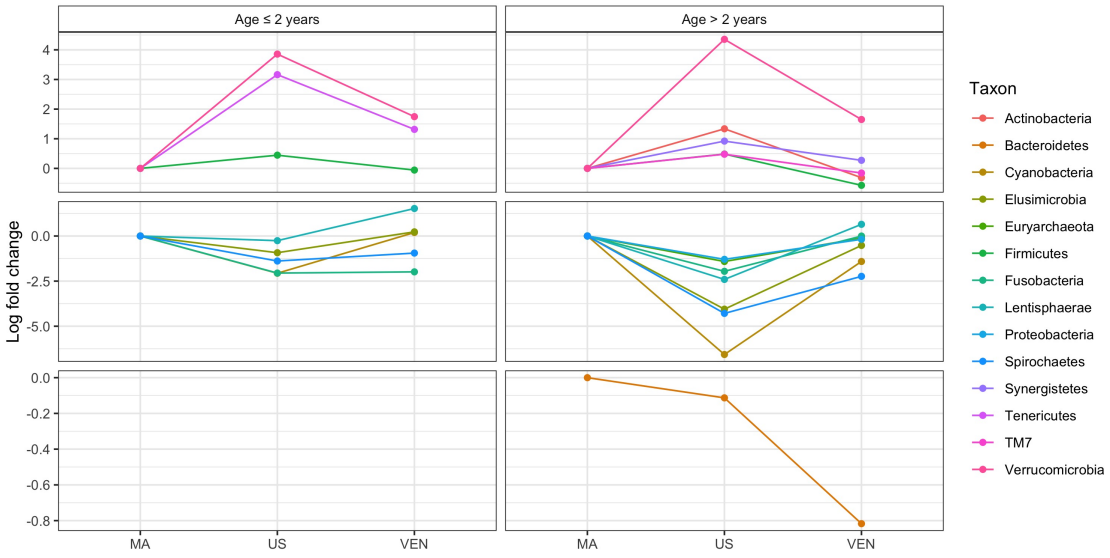


Figure 20: Testing for patterns with respect to location effect.

4.0 Distance Correlation for Microbiome (DICOM)

4.1 Introduction

In some applications, researchers are interested in elucidating complex inter-relationships among taxa within an ecosystem or across systems (e.g., gut and oral environments), and in identifying clusters of taxa, which might be biologically meaningful for certain pathways. Thus, a common goal of analyzing microbiome data is to identify the dependence between taxa, and the primary strategy of achieving this goal is to perform the correlation analysis.

Recall that the OTU/SV table is inherently compositional and only provides information in the form of relative abundances regardless of different microbial loads and library sizes. Due to the compositional structure, standard methods for computing either Pearson correlation or Spearman correlation are invalid and might introduce negative correlations between taxa regardless of the true underlying correlations (Mosimann, 1962; Aitchison, 2003; Pawlowsky-Glahn and Buccianti, 2011). SparCC (Friedman and Alm, 2012) was developed to overcome the limitation of standard methods, which introduce spurious correlations (Mosimann, 1962; Mandal et al., 2015), for estimating the Pearson correlation coefficients. Like most compositional data analysis techniques, SparCC exploits the log-ratio transformation of the data, while relying on two mild assumptions: 1) the number of taxa is large; 2) most taxa are not strongly correlated with each other. Simulations and real data applications show that SparCC could achieve high accuracy when inferring the Pearson correlations.

However, not all taxa are expected to be linearly related; More importantly, dependence is generally not equivalent to correlations. Therefore, distance correlation (Székely et al., 2007) seems to be a better mathematical tool as compared to Pearson or Spearman correlations, since it is a measure of statistical dependence between two paired random vectors.

Definition 4.1.1 (Distance correlation)**.** Suppose two random vectors, $X = (x_1, \ldots, x_n)^T$ and $Y = (y_1, \ldots, y_n)^T$. The distance covariance and correlation are defined as

(1) $dCov^2(X, Y) = \frac{1}{\pi^2} \iint \frac{\|f_{X,Y}(t,s) - f_X(t) f_Y(s)\|^2}{t^2 s^2} dt ds,$

(2) $dVar^2(X) = \frac{1}{\pi^2}\iint \frac{\|f_{X,X}(t,s)-f_X(t)f_X(s)\|^2}{t^2 s^2} dtds$,

(3) $dCor(X,Y) = \frac{dCov(X,Y)}{\sqrt{dVar(X)dVar(Y)}}$.

where $\|\cdot\|$ denotes Euclidean norm.

Define pairwise distances as $d_{ij} = |x_i - x_j|, e_{ij} = |y_i - y_j|$, and doubly centered distances as $D_{ij} = d_{ij} - \bar{d}_{i\cdot} - \bar{d}_{\cdot j} + \bar{d}_{\cdot\cdot}, E_{ij} = e_{ij} - \bar{e}_{i\cdot} - \bar{e}_{\cdot j} + \bar{e}_{\cdot\cdot}, i \neq j$, the empirical estimates of distance covariance and correlation can be obtained by

(1) $dCov_n^2(X,Y) = \frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n D_{ij}E_{ij}$,

(2) $dVar_n^2(X,Y) = \frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n D_{ij}^2$.

Note that

(1) $0 \leq dCor(X,Y) \leq 1$,

(2) $dCor(X,Y) = 0$ if and only if X and Y are independent,

(3) $dCor(X,Y) = 1$ implies that dimensions of the linear subspaces spanned by X and Y samples respectively are almost surely equal.

To illustrate the difference between distance correlation and Pearson/Spearman correlations, consider the toy example shown in Figure 21. As expected, when two random variables are linearly related all three measures (Pearson correlation, Spearman correlation, and distance correlation) have values equal to 1, indicating that they all detect the linear dependence between x and y; on the contrary, when the relationship between x and y is nonlinear, $y = (x-5)^2$ in this case, both Pearson and Spearman correlation lose track of the nonlinear association. In contrast, the distance correlation is the only one declaring that x and y are not independent.

To further illustrate the importance of considering distance correlation in analyzing microbiome data, we generated a synthetic dataset using the Poisson-Gamma model, with the number of taxa equals to 10 (T1 to T10), and the number of samples equals to 60 (S1 to S60). The first five taxa are designed to be linearly or nonlinearly associated with each other, as shown in Figure 22. We calculated pairwise correlations among ten taxa using different measures and visualized the result using a network graph (Figure 23). As we can see from the network, while the linear relationship between T1 and T2 is well detected by all methods, none of the correlation measures, except the distance correlation, successfully

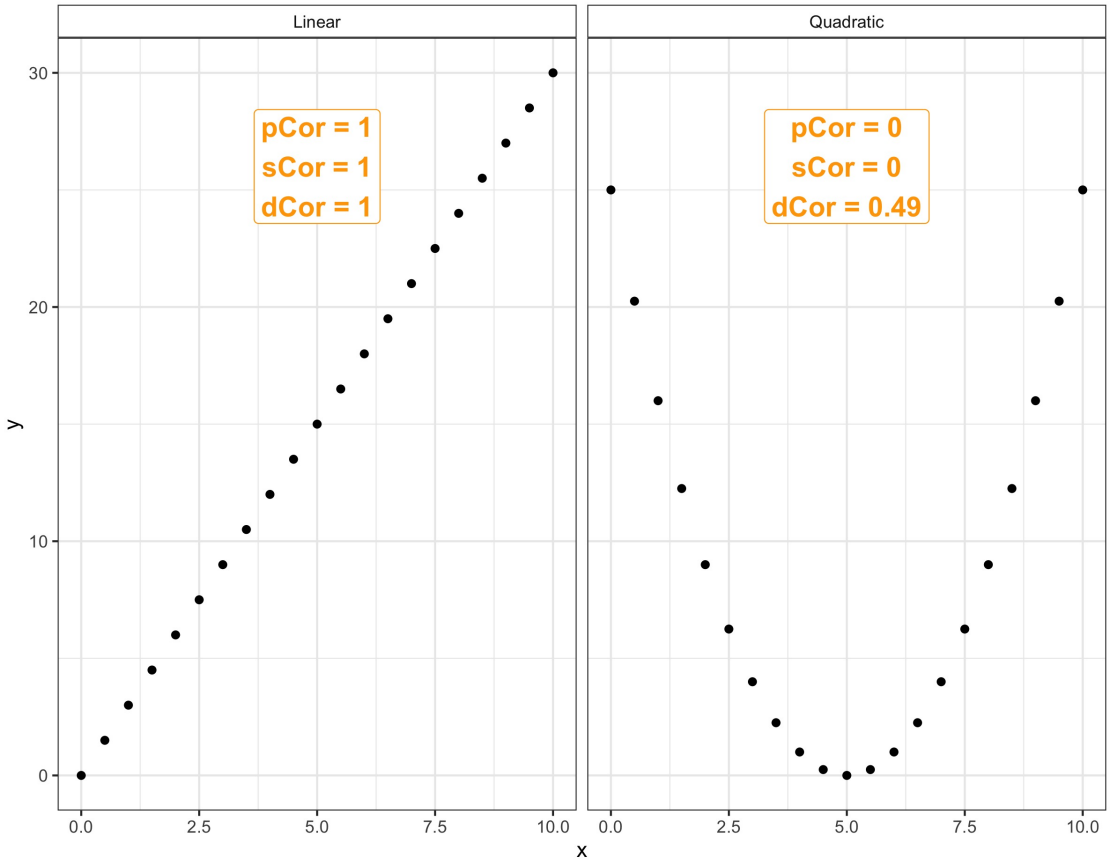


Figure 21: Pearson correlation vs. Spearman correlation vs. Distance correlation.

captured the nonlinear relationship among T1 to T5. Thus, this toy example illustrates that even after transforming simplex into Euclidean space, the standard methods, such as Pearson or Spearman correlations, which are designed for linear associations, fail to detect important nonlinear relationships that exist among microbial communities. On the other hand, the distance correlation based method considered in this chapter is very encouraging for detecting any type of dependence, whether linear or nonlinear. Therefore, in this chapter, we develop a tool called Distance Correlations for Microbiome (DICOM), based on distance correlations to describe dependence (linear or nonlinear) among microbes.

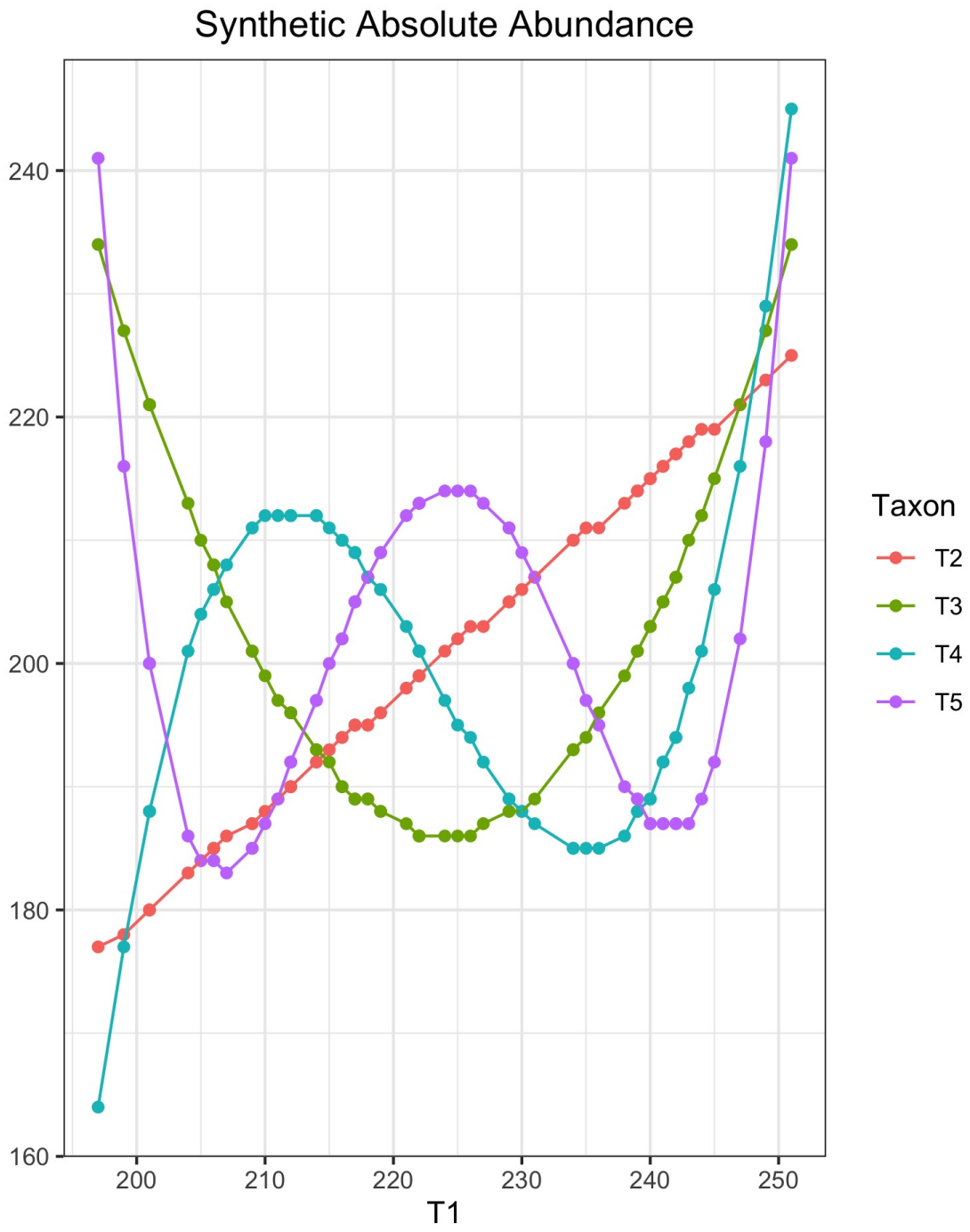


Figure 22: Various kinds of associations between taxon 1 (T1) and Taxa 2 to 5 (T2 to T5).

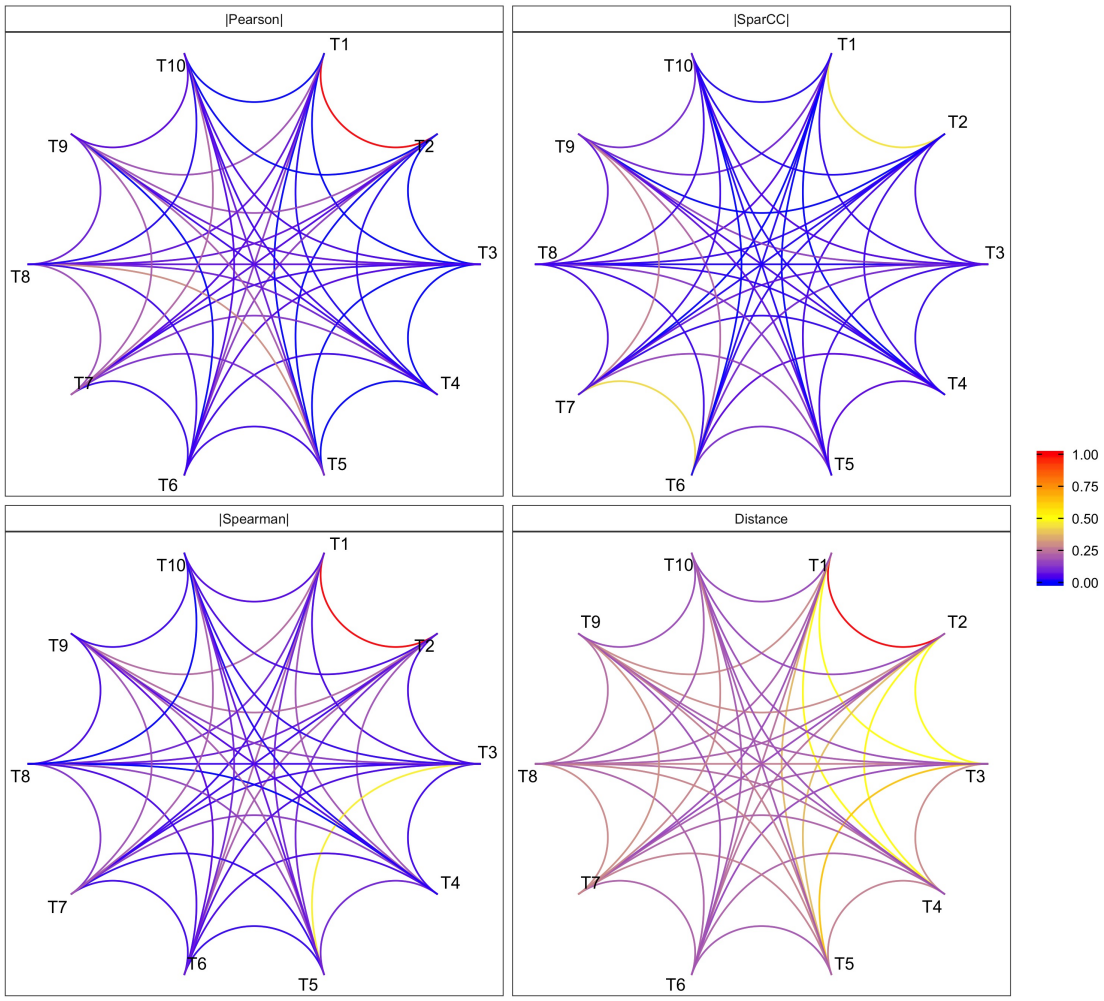


Figure 23: Network visualizations for different correlation measures.

4.2 Distance Correlations for Microbiome (DICOM)

4.2.1 The relative abundances in the sample are reasonable estimates of the relative abundances in the ecosystem

The main hurdle in performing the correlation analysis on microbiome data is the lack of access to the absolute abundance data in the ecosystem level. Since we do not observe the entire ecosystem, any concept of correlations among the observed taxa does not trans-

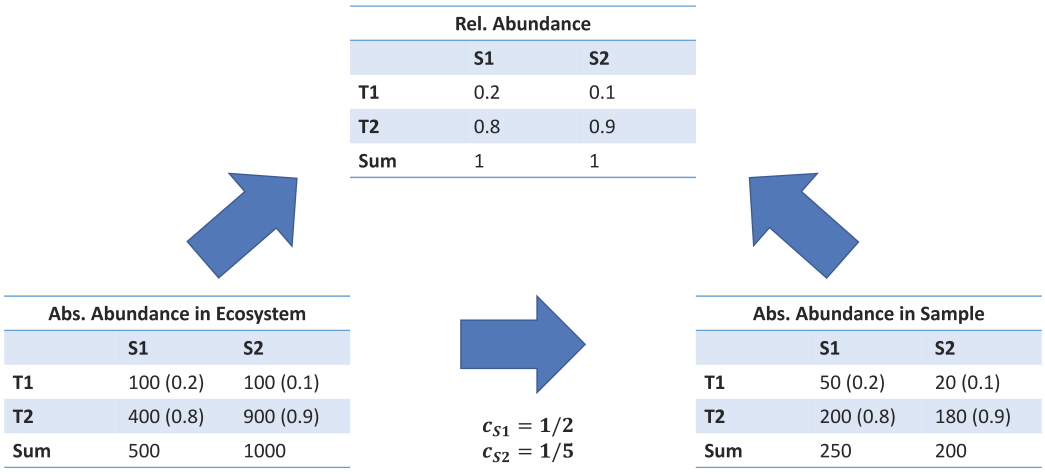


Figure 24: The relative abundances is the same between the ecosystem and sample.

late easily to correlations among the absolute abundances in the ecosystem. The unknown sampling fractions play a critical role. The sample level data (OTU/SV table) only contains information on relative abundances (see Figure 24). As discussed in previous chapters, the observed absolute abundances are not in the same proportions as in the ecosystems due to the differential sampling fractions across samples. In the example shown in Figure 24, the absolute abundances in sample 1 (S1) are half of those in its ecosystem, while sample 2 (S2) is $\frac{1}{5}$ of its ecosystem. This makes the correlation analysis at the sample level difficult. However, the relative abundances in the sample are reasonable estimates of the relative abundances in the ecosystem. This observation is well-known to classical statisticians working in the area of survey sampling. This fact becomes useful for us as we develop our DICOM methodology.

4.2.2 Dependence structure is carried by ranks

We first note that dependence in the rank transformed data relates to dependence in the original data (Kendall, 1938; Sukhatme et al., 1970). We show this using a simple quadratic

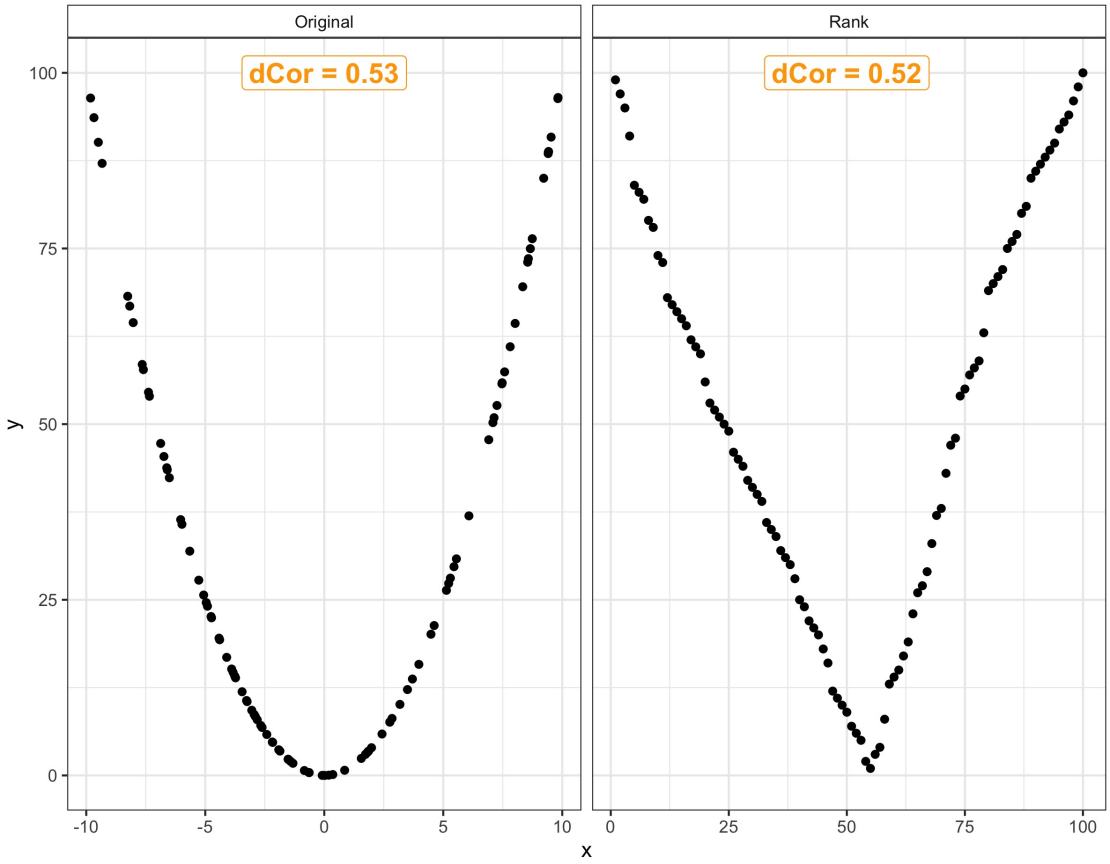


Figure 25: Distance correlations calculated by the original data vs. by ranks.

example (Figure 25), here $y = x^2$. Even though the shape of (X, Y) is not perfectly kept using $(rank(X), rank(Y))$, it turns out that $dCor(X, Y)$ and $dCor(rank(X), rank(Y))$ are close to each other, meaning that the dependence structure is well preserved. Given that the absolute abundances are usually not observable in microbiome studies, ranks would be more accessible and robust.

4.2.3 The relationship between absolute ratios and relative ratios

The changes between two ecosystems (j and j'), where j' is the reference sample, with respect to absolute abundances of m taxa, $A_j = (A_{1j}, \ldots, A_{mj})^T$ and $A_{j'} = (A_{1j'}, \ldots, A_{mj'})^T$,

can be computed as follows:

$$\frac{A_j}{A_{j'}} = (\frac{A_{1j}}{A_{1j'}}, \ldots, \frac{A_{mj}}{A_{mj'}})^T. \tag{4.1}$$

If we are only able to measure relative abundances, the change of relative abundances are related to the change of absolute abundance in the sense that:

$$\frac{A_j}{A_{j'}} = (\frac{A_{\cdot j} \times \gamma_{1j}}{A_{\cdot j'} \times \gamma_{1j'}}, \ldots, \frac{A_{\cdot j} \times \gamma_{mj}}{A_{\cdot j'} \times \gamma_{mj'}})^T = \frac{\gamma_j}{\gamma_{j'}} \times \frac{A_{\cdot j}}{A_{\cdot j'}} := \frac{\gamma_j}{\gamma_{j'}} \times \Lambda_{jj'}, \tag{4.2}$$

which implies the ratio between relative abundances is in proportional to the ratio of absolute abundance. This proportion $\Lambda_{jj'}$, defined as biomass bias, is the ratio of microbial loads $\frac{A_{\cdot j}}{A_{\cdot j'}}$.

Assumption 4.2.1. *There exist some taxa that are non-differentially abundant.*

If the absolute abundances of taxon i in sample j is more abundant than its correspondence in sample j', then obviously, $\frac{A_{ij}}{A_{ij'}} > 1$; on the contrary, if taxon i is less abundant in sample j, then $\frac{A_{ij}}{A_{ij'}} < 1$; lastly, if this is a non-differentially abundant taxon, then one should expect that $\frac{A_{ij}}{A_{ij'}} = 1$. The Assumption 4.2.1 indicates that not all taxa are differentially abundant; Therefore, it is reasonable to assume the ratio between absolute abundances (absolute ratio) has the property:

$$\text{median}\, \frac{A_j}{A_{j'}} \approx 1. \tag{4.3}$$

Based on equation (4.2), the median of ratios between relative abundances (relative ratios) will be:

$$\text{median}\, \frac{\gamma_j}{\gamma_{j'}} \approx \Lambda_{jj'}^{-1}. \tag{4.4}$$

Thus, we propose to estimate biomass bias $\Lambda_{jj'}$ by

$$\hat{\Lambda}_{jj'} = (\text{median}\, \frac{\gamma_j}{\gamma_{j'}})^{-1}. \tag{4.5}$$

With $\hat{\Lambda}_{jj'}$, although the data of absolute ratios is unattainable due to the unobservable ecosystem level data, one can still make inference on it using the relative ratio data after adjusting the biomass bias.

Table 8: Summary of synthetic absolute abundance table.

	S1	S2	S3	$\frac{S1}{S3}$	$\frac{S2}{S3}$	$r(\frac{S1}{S3})\dagger$	$r(\frac{S2}{S3})$
T1	5	10	20	0.25	0.5	1	2
T2	10	10	10	1	1	1	1
T3	10	10	10	1	1	1	1
T4	10	10	10	1	1	1	1
T5	65	10	50	1.3	0.2	2	1
Sum	100	50	100				

$\dagger$ $r(X)$ denotes the rank of X.

Table 9: Summary of synthetic relative abundance table (unadjusted).

	S1	S2	S3	$\frac{S1}{S3}$	$\frac{S2}{S3}$	$r(\frac{S1}{S3})\dagger$	$r(\frac{S2}{S3})$
T1	0.05	0.2	0.2	0.25	1	1	2
T2	0.1	0.2	0.1	1	2	1	2
T3	0.1	0.2	0.1	1	2	1	2
T4	0.1	0.2	0.1	1	2	1	2
T5	0.65	0.2	0.5	1.3	0.4	2	1

$\dagger$ $r(X)$ denotes the rank of X.

4.2.4 Retrieve the rank within each taxon

We now show how to retrieve the rank within each taxon by utilizing the ranks of relative ratios.

To illustrate the procedure, we simulated the absolute abundance table (at the ecosystem level) using the Poisson-Gamma model. The number of taxa is set to be 5 (T1 to T5), and the number of samples is 3 (S1, S2, and S3). The absolute abundance table and the corresponding unadjusted relative abundance table are summarized in Table 8 and Table 9.

Table 10: Summary of synthetic relative abundance table (adjusted).

	S1	S2	S3	$\frac{S1}{S3} \times \hat{\Lambda}_{1G}$	$\frac{S2}{S3} \times \hat{\Lambda}_{2G}$	$r(\frac{S1}{S3} \times \hat{\Lambda}_{1G})\dagger$	$r(\frac{S2}{\mu_G} \times \hat{\Lambda}_{2G})$
T1	0.05	0.2	0.2	0.25	0.5	1	2
T2	0.1	0.2	0.1	1	1	1	1
T3	0.1	0.2	0.1	1	1	1	1
T4	0.1	0.2	0.1	1	1	1	1
T5	0.65	0.2	0.5	1.3	0.2	2	1

$\dagger$ $r(X)$ denotes the rank of X.

Let $r(X)$ denotes the rank of X. In Table 8, suppose S3 is chosen as the reference sample. Then, because $\frac{S1}{S3} = \frac{S2}{S3} = 1$, we therefore have $r(\frac{S1}{S3}) = r(\frac{S2}{S3}) = 1$. Given the absolute ratios have a common reference (S3) in the denominator, we conclude that $r(S1) = r(S2) = 1$ for T2, which is consistent with ranks derived from absolute abundances ($A(S1) = A(S2) = 10$). Similarly, we also have $r(S1) = r(S3)$ and $r(S2) = r(S3)$ by setting S2 and S1 as the reference, respectively.

However, if we calculate ranks within T2, in which S3 is chosen as the reference again, using the relative abundance table (Table 9), it leads to $r(S1) = 1 < r(S2) = 2$ since $\frac{S1}{S3} = 1 < \frac{S2}{S3} = 2$. As the underlying truth is $r(S1) = r(S2) = 1$ based on the absolute abundance table, this erroneous conclusion results from the bias due to unequal biomass bias among relative ratios because $\Lambda_{13} = 1$ and $\Lambda_{23} = 0.5$.

This misleading result based on the raw relative abundance table, which is referred to as the unadjusted relative abundance table, can be avoided by adjusting relative ratios with estimates of biomass biases. According to equation (4.5), it is easy to see that $\hat{\Lambda}_{13} = 1$ and $\hat{\Lambda}_{23} = 2$ in this toy example. The adjusted relative ratios are shown in Table 10. As we can see, with biomass bias taking into account, $r(\frac{S1}{S3} \times \hat{\Lambda}_{1G}) = r(\frac{S2}{S3} \times \hat{\Lambda}_{2G}) = 1$, thus $r(S1) = r(S2) = 1$ within T2.

4.2.5 Implementation of DICOM

The methodology of DICOM can be summarized as follows:

(1) Obtain the estimates of relative abundance table from the observed OTU/SV table.

(2) Calculate the relative ratios for every sample $j, j = 1, \ldots, n$, with respect to a reference $j', j' = 1, \ldots, n$. This results in n matrices of relative ratios.

(3) Adjust the relative ratios using the estimates of biomass bias.

(4) For each taxon $i, i = 1, \ldots, m$, obtain the ranks across samples using the adjusted relative ratios obtained from step (3).

(5) **Raw distance correlations:** Calculate distance correlations between each pair of taxa using ranks given the reference sample j'.

(6) **Permutation**: Permute the samples in the relative ratio table in (2), repeat (3) – (4), and obtain the "background" distance correlations.

(7) **Denoise**: Any values in (4) less than the corresponding distance correlation in (5) will be assigned as 0.

(8) **Final estimate**: Get the final estimate of distance correlation by averaging values across n matrices (each matrix corresponds to a chosen reference sample j').

We first show the implementation of DICOM using the synthetic dataset generated from the Poisson-Gamma model, where the number of taxa is 10, the number of groups equals 2, and the sample size is 50 per group. Taxon 1 (T1) is designed to be linearly correlated with taxon 2 (T2) and nonlinearly correlated with taxon 3 (T3). Taxon 5 (T5) is more abundant in group 1, while Taxon 7 (T7) is more abundant in group 2. Because of the differential abundances, T5 and T5 are also correlated. We summarized the estimated distance correlations in Figure 26. The x-axis denotes the reference sample, ranging from 1 to 100; the y-axis represents the estimate of distance correlation. The estimate will be colored in red if two taxa are genuinely dependent on each other (i.e., T1 vs. T2, T1 vs. T3, T2 vs. T3, and T5 vs. T7); otherwise, it will be colored in green. As we can see from Figure 26, there is a clear separation between distances correlations from dependent pairs and those from independent pairs, as the estimated values by DICOM tend to be larger when two taxa are indeed dependent.

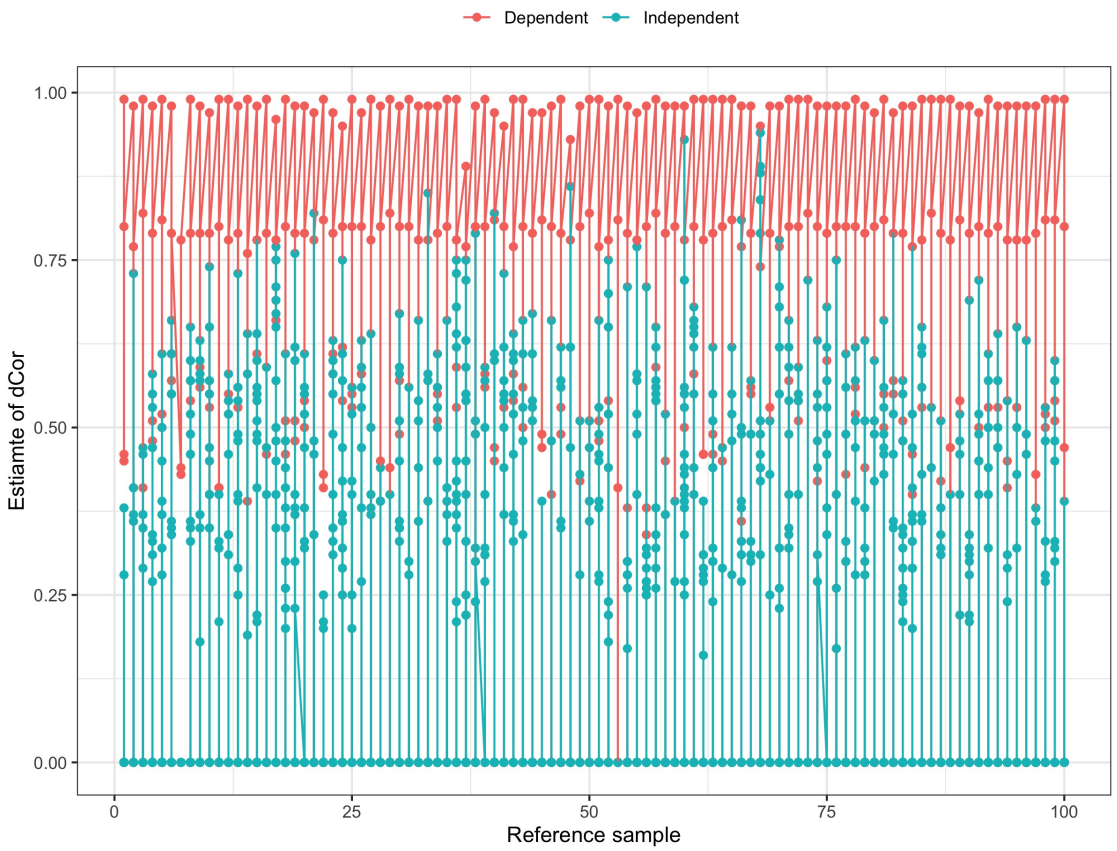


Figure 26: Estimated distance correlations using DICOM.

For ease of comparison, we also visualized the simulation result in Figure 27. It shows that DICOM (right panel) can infer dependence among taxa (left panel) with high accuracy in this simulation study. Note that the DICOM network tends to have more purely blue edges, where distance correlation equals to 0, than the true network. This is due to the denoising step implemented in the DICOM procedure as it forces the distance correlations below the threshold to be exactly 0s.

Lastly, to illustrate a potential application of DICOM, we infer a rich ecological network connecting ten interacting phyla across samples from the global gut data (Yatsunenko et al., 2012). Figure 28 shows that as compared to more mature subjects (age $>$ 2 years old), in the gut environments of infants (age $\leq$ 2 years old), phyla tend to be more strongly

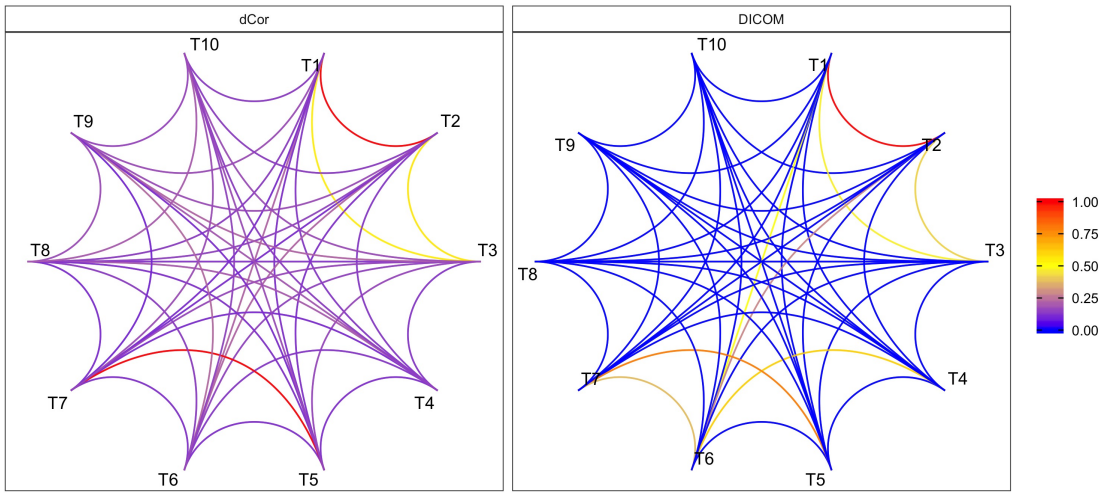


Figure 27: Network visualization of DICOM using synthetic data.

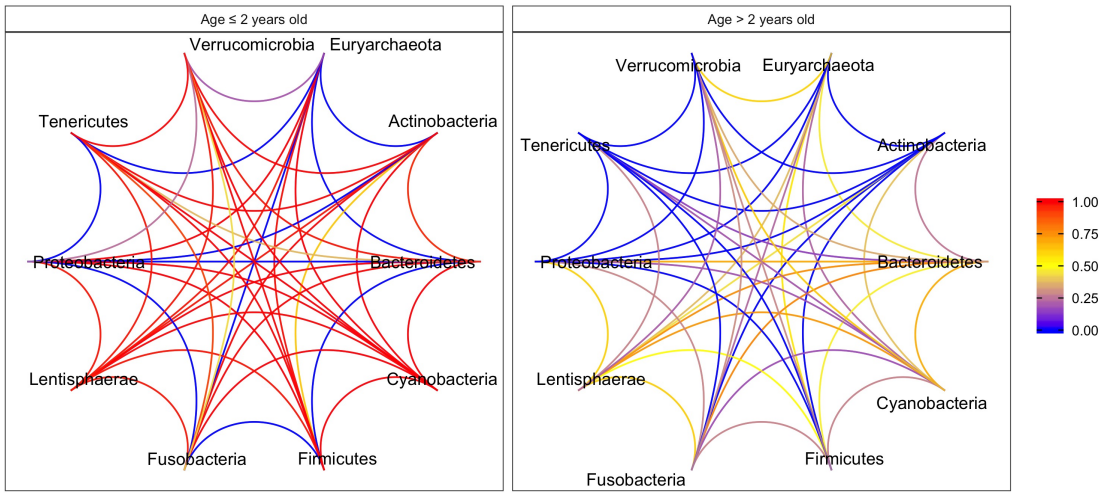


Figure 28: Implementation of DICOM using global gut microbiota data.

dependent on each other. This is consistent with previous studies that there is a strong temporal development of the infant gut microbiome.

5.0 Discussion and Future Work

Increasingly researchers are conducting microbiome studies to ask a wide range of questions of scientific interest. However, as we know from previous studies (Mandal et al., 2015; Morton et al., 2017, 2019), the question of "who are all there?" is still an important basic question before we can answer questions such as "What are they doing?" and "How are they doing?" etc.

As the observed microbiome data are relative abundances (compositional) with lots of zeros (Gloor and Reid, 2016; Gloor et al., 2017; Mandal et al., 2015; Morton et al., 2017, 2019), the method of differential abundance (DA) analysis is not necessarily routine. Numerous methods are available in the literature to analyze these data. There have been misunderstandings and controversies, in part because there is a lack of clarity on what parameters are to be tested and what hypotheses a given method/statistic is really testing. The only exception being (Mandal et al., 2015) who clearly described the various parameters associated with microbiome studies and proposed statistically rigorous method called ANCOM for performing DA analysis. A major hurdle in performing DA analyses is the bias introduced by differences in the sampling fractions, which is determined by the library size of each sample and its corresponding microbial load of the ecosystem of interest, across sample. In this dissertation we propose a novel method called Analysis of Compositions of Microbiomes with Bias Correction (ANCOM-BC), which estimates the unknown sampling fractions and corrects the bias induced by it. The resulting sample abundance data are modeled using a linear regression framework. This formulation makes a fundamental advancement in the field because, unlike any of the existing methods, it (a) provides statistically valid test with appropriate p-values, (b) controls the False Discovery Rate (FDR), (c) maintains adequate power, (d) provides biologically meaningful results, (e) is computationally simple to implement, and (f) is straightforward to apply to more complex multi-group comparison scenarios. Besides establishing methodology for DA analysis, in this dissertation, we also develop a general framework for describing dependence among various taxa. Note that for a biological system to function effectively, the various taxa must work in harmony.

This implies that taxa are potentially dependent on each other, where the relationships can be either linear or nonlinear. Standard correlation based methods seek to explore linear relationships, which ignores complex nonlinear relationships. In this dissertation we propose a novel methodology called Distance Correlations for Microbiome (DICOM), which exploits the concept of distance correlation and properly addresses the compositional structure of the microbiome data, to elicit dependence among taxa.

This dissertation research opens new models, thoughts and directions for future methodological research for analyzing microbiome data. Future work can be done on the top of the development of this dissertation. For example, ANCOM-BC algorithm can be improved by combing the sampling zeros into model. Sampling zero arises due to insufficient sequencing depth, and currently is simply addressed by adding a pseudo-count (e.g. 1) to the observed absolute abundance table. For future modifications of ANCOM-BC, in addition to the offset-based log linear model, sampling zeros can be more properly taken care of by implementing the hurdle model

$$y_i j = \begin{cases} 0 & \text{with prob. } \pi_{ij} \\ d_j + \beta_i^T x_j + \epsilon_i j & \text{with prob. } (1 - \pi_{ij}) \end{cases} \tag{5.1}$$

where π_{ij} is the subject-specific and taxon-specific probability, and can be further modeled by logistic regression, such as

$$\pi_{ij} = \frac{\exp(\eta_i^T x_j)}{\exp(\eta_i^T x_j) + 1} \tag{5.2}$$

Such a formulation might improve the performance of ANCOM-BC when there are (a) very large number of differentially abundant taxa, or (b) when the sample sizes are small.

The DICOM methodology is not fully developed in this dissertation, for example, a multinomial logistic regression model can be adopted to come up with better estimates of relative abundances from the observed OTU/SV table. Let $\Delta_{ij} = I(y_i j = 0)$

$$p_{ij} = E(r_{ij}|\Delta_{ij}) = \frac{(1 - \Delta_{ij}) \exp(\alpha_i^T x_j)}{\sum_{i=1}^{m-1} (1 - \Delta_{ij}) \exp(\alpha_i^T x_j) + (1 - \Delta_{ij})} \tag{5.3}$$

A more efficient algorithm can be developed to prevent false positives. We hope to explore the area further, especially to describe temporal changes in the dependence among microbiome. For example, a researcher may be interested in understanding the changes in absolute abundances as well as changes in the dependence among taxa as the disease of a

patient progresses. Such investigations will be useful for deriving suitable treatments and interventions.

Appendix A

Inflated false positive rates of some standard methods

Two potential reasons why some differential abundance (DA) analysis methods for microbiome data result in inflated positive rates, and hence inflated false discovery rates (FDR), are as follows:

(1) The test statistic may not be designed for testing the hypothesis of interest. For example, the test statistic may be designed for testing hypothesis regarding relative abundance but is used for testing absolute abundance.

(2) Data are not properly normalized to account for bias due to variability in sampling fractions.

In the following we discuss some commonly used methods in the literature, namely, Wilcoxon rank sum test (with and without TSS)(Mann and Whitney, 1947), DESeq2 (Love et al., 2014), edgeR (Robinson et al., 2010), metagenomeSeq (Paulson et al., 2013). We begin with the following simple lemma.

Lemma A.0.1. *Suppose, for a taxon i, $E(\hat{\beta}_i) = \beta_i + \delta_i$, and $\widehat{SE}(\hat{\beta}_i)$ is $O_p(n^{-1/2})$. Further assume that, under H_0, $\beta_i = 0$, $\delta_i \neq 0$ and*

$$T_{\beta_i} = \frac{\hat{\beta}_i - \beta_i}{\widehat{SE}(\hat{\beta}_i)} \rightarrow_d N(\frac{\delta_i}{SE(\hat{\beta}_i)}, 1).$$

Suppose $z_{1-\alpha/2}$ is $(1-\alpha/2)\times 100$ percentile of standard normal distribution then the probability of Type I error associated with the critical region:

$$|T_{\beta_i=0}| \geq z_{1-\alpha/2}$$

increases with sample size. Equivalently, the p-value based on $|T_{\beta_i=0}|$ stochastically decreases with n.

Proof. Note that under the null hypothesis we have T_{β_i} is centered at δ_i. Since $\delta_i \neq 0$, and $\widehat{SE}(\hat{\beta}_i)$ grows at the rate of $\sqrt{n}$, therefore $|T_{\beta_i}|$ stochastically increases with n, and p-value decreases stochastically. This results in inflated Type I error. □

With slight abuse of notation, we use i to denote taxon, j to denote group, and k to denote sample. In the following sections, suppose taxon i is not differentially abundant between two groups. For simplicity of exposition, we assume there are only two experimental groups, and the sample sizes are equal between the two groups.

A.1 Wilcoxon rank-sum test with no normalization

Suppose for $k = 1, \ldots, n, O_{i1k} \sim_{iid} F_{i1}$ and $O_{i2k} \sim_{iid} F_{i2}$. Under no normalization, the Wilcoxon rank-sum test aim to test the following hypotheses

$$H_0 : F_{i1} = F_{i2}$$
$$H_1 : F_{i1} \neq F_{i2}$$

The test statistic is given by:

$$U = \frac{1}{n^2} \sum_{k=1}^{n} \sum_{k'=1}^{n} I(O_{i1k} \leq O_{i2k'}). \tag{A.1}$$

Asymptotically, under the null hypothesis we know that:

$$U \sim AN(\frac{1}{2}, \frac{1}{6n}). \tag{A.2}$$

The basic assumption made by the above U statistic is that under the null hypothesis O_{i1k} is equally likely to be small or large compared to O_{i2k}. Thus the indicator random variable $I(O_{i1k} \leq O_{i2k})$ has the same distribution as $I(O_{i2k} \leq O_{i1k})$. Note that in the existing implementation of these tests the samples are not normalized for unequal sampling fractions. Therefore under the null hypothesis, the U statistic is not centered at $\frac{1}{2}$. Hence the Type I error is not controlled at α according to Lemma A.0.1.

A.2 Wilcoxon rank-sum test with TSS

TSS normalization transforms the absolute abundance table into the relative abundance table. Using these relative abundance data, for $k = 1, \ldots, n, r_{i1k} \sim_{iid} G_{i1}$ and $r_{i2k} \sim_{iid} G_{i2}$, the Wilcoxon Rank-Sum test is used for testing the following hypotheses:

$$H_0 : G_{i1} = G_{i2}$$
$$H_1 : G_{i1} \neq G_{i2}.$$

Under the above normalization, even if the expected absolute abundance of a taxon is same between two ecosystems, its relative abundances may not be same. Thus, testing the null hypothesis of equality of relative abundance of a taxon between two ecosystems is not equivalent to the null hypothesis that the absolute abundances are equal. Furthermore, the Wilcoxon rank-sum test applied directly to the relative abundance data ignores the compositional structure. Consequently, asymptotically the Type I error will not be controlled as indicated in Lemma A.0.1.

A.3 DESeq2

DESeq2 assumes a negative-binomial model for absolute abundances. Thus, the observed count data and the corresponding parameters are modeled as follows:

$$\begin{aligned} &O_{ijk} \sim NB(s_{jk}q_{ij}, \phi_i) \\ &s_{jk} = \underset{i:O_i^R \neq 0}{\text{median}} \frac{O_{ijk}}{O_i^R} \\ &\log q_{ij} = \beta_{i0} + \beta_{i1} I(j = 1), \quad j = 1, 2 \\ &\hat{\beta}_{i1} = \arg\max_{\beta_{i1}} (\sum_{j=1}^{2} \sum_{k=1}^{n} \log f_{NB}(O_{ijk}; s_{jk}q_{ij}, \phi_i) + \Lambda(\beta)) \end{aligned} \tag{A.3}$$

where

(1) $O_i^R = (\prod_{j=1}^{2} \prod_{k=1}^{n} O_{ijk})^{\frac{1}{2n}}$,

(2) $\Lambda(\beta) = -(\frac{\beta_{i0}^2}{2\sigma_0^2} + \frac{\beta_{i1}^2}{2\sigma_1^2})$,

(3) σ_0^2, σ_1^2 are prior variances for β_{i0}, β_{i1}, respectively.

DESeq2 first scales the OTU table by the normalization factor s_{jk}, and then tests for differential abundance, consequently it does not take into account the uncertainty associated with s_{jk}.

Recall from the regression framework of ANCOM-BC that:

$$\begin{aligned} E(O_{ijk}) &= c_{jk}\theta_{ij} \\ E(y_{ijk}) &= d_{jk} + \mu_{ij} \end{aligned} \tag{A.4}$$

Compared to (A.3), DESeq2 estimates the sampling fraction c_{jk} by s_{jk}, i.e. $\hat{c}_{jk}^{\text{MED}} = s_{jk}$ and therefore $\hat{d}_{jk}^{\text{MED}} = \log s_{jk}$. Thus, we have

$$\begin{aligned} \hat{d}_{jk}^{\text{MED}} &= \underset{i:O_i^R \neq 0}{\text{median}}(\log O_{ijk} - \frac{1}{2n}\sum_{j=1}^{2}\sum_{k=1}^{n}\log O_{ijk}) \\ &= \underset{i:O_i^R \neq 0}{\text{median}}(y_{ijk} - \frac{1}{2n}\sum_{j=1}^{2}\sum_{k=1}^{n} y_{ijk}) \\ &= \underset{i:O_i^R \neq 0}{\text{median}}(d_{jk} + \mu_{ij} + \epsilon_{ijk} - \frac{1}{2n}\sum_{j=1}^{2}\sum_{k=1}^{n} y_{ijk}) \\ &= \underset{i:O_i^R \neq 0}{\text{median}}(d_{jk} - \bar{d}_{..} + \mu_{ij} - \bar{\mu}_{i.} + \epsilon_{ijk} - \bar{\epsilon}_{i..}) \\ &= d_{jk} - \bar{d}_{..} + \underset{i:O_i^R \neq 0}{\text{median}}(\mu_{ij} - \bar{\mu}_{i.} + \epsilon_{ijk} - \bar{\epsilon}_{i..}) \\ &:= d_{jk} - \bar{d}_{..} + \mu_{a_{jk}j} - \bar{\mu}_{a_{jk}\cdot} + \epsilon_{a_{jk}jk} - \bar{\epsilon}_{a_{jk}\cdot\cdot}. \end{aligned} \tag{A.5}$$

In the expressions a_{jk} denotes the index that corresponds to the taxon for which $\underset{i:O_i^R \neq 0}{\text{median}}(\mu_{ij} - \bar{\mu}_{i.} + \epsilon_{ijk} - \bar{\epsilon}_{i..}) = \mu_{a_{jk}j} - \bar{\mu}_{a_{jk}\cdot} + \epsilon_{a_{jk}jk} - \bar{\epsilon}_{a_{jk}\cdot\cdot}$. Averaging over all samples $k = 1, 2, \ldots, n$ in group j, we get

$$\bar{\hat{d}}_{j\cdot}^{\text{MED}} = \bar{d}_{j\cdot} - \bar{d}_{..} + \tilde{\mu}_{\cdot(j)j} - \tilde{\mu}_{\cdot(j)\cdot} + \tilde{\epsilon}_{\cdot(j)j\cdot} - \tilde{\epsilon}_{\cdot(j)\cdot\cdot}. \tag{A.6}$$

Since each subject k in group j, may potentially have a different taxon that yields the median value $\mu_{a_{jk}j} - \bar{\mu}_{a_{jk}\cdot} + \epsilon_{a_{jk}jk} - \bar{\epsilon}_{a_{jk}\cdot\cdot}$, in the above expression $\tilde{x}$ represents the mean of variable x taken over the suitable subset of taxa. Secondly, the notation $x_{\cdot(j)}$ represents the mean taken within group j.

The test statistic for DESeq2 is of the form:

$$W_i^{\text{DESeq2}} = \frac{\hat{\mu}_{i1} - \hat{\mu}_{i2} - \hat{\delta}^{\text{MED}}}{\text{SE}(\hat{\mu}_{i1} - \hat{\mu}_{i2} - \hat{\delta}^{\text{MED}})} \tag{A.7}$$

The MED estimator of the bias term is:

$$\begin{aligned}
\hat{\delta}^{\text{MED}} &:= \bar{\hat{d}}_{1\cdot}^{\text{MED}} - \bar{\hat{d}}_{2\cdot}^{\text{MED}} \\
&= \bar{d}_{1\cdot} - \bar{d}_{2\cdot} + \{\tilde{\mu}_{\cdot(1)1} - \tilde{\mu}_{\cdot(1)\cdot} + \tilde{\epsilon}_{\cdot(1)1\cdot} - \tilde{\epsilon}_{\cdot(1)\cdot\cdot}\} - \{\tilde{\mu}_{\cdot(2)2} - \tilde{\mu}_{\cdot(2)\cdot} + \tilde{\epsilon}_{\cdot(2)2\cdot} - \tilde{\epsilon}_{\cdot(2)\cdot\cdot}\} \\
&= \delta + \{\tilde{\mu}_{\cdot(1)1} - \tilde{\mu}_{\cdot(1)\cdot} + \tilde{\epsilon}_{\cdot(1)1\cdot} - \tilde{\epsilon}_{\cdot(1)\cdot\cdot}\} - \{\tilde{\mu}_{\cdot(2)2\cdot} - \tilde{\mu}_{\cdot(2)\cdot} + \tilde{\epsilon}_{\cdot(2)2\cdot} - \tilde{\epsilon}_{\cdot(2)\cdot\cdot}\}
\end{aligned} \tag{A.8}$$

Note that $E(\tilde{\epsilon}_{\cdot(j)j\cdot} - \tilde{\epsilon}_{\cdot(j)\cdot\cdot}) = 0$. However, unless $E_{\mathcal{S}}(\tilde{\mu}_{\cdot(1)1} - \tilde{\mu}_{\cdot(1)\cdot}) = 0$ and $E_{\mathcal{S}}(\tilde{\mu}_{\cdot(2)2} - \tilde{\mu}_{\cdot(2)\cdot}) = 0$, where the subscript $\mathcal{S}$ denotes the collection of all suitable subsets of taxa $\{1, 2, \ldots, m\}$, the MED estimator does not estimate the bias term in the null hypothesis unbiasedly, i.e.

$$E(\hat{\delta}^{\text{MED}}) \neq \delta. \tag{A.9}$$

Thus, under the null hypothesis

$$E(\hat{\mu}_{i1} - \hat{\mu}_{i2} - \hat{\delta}^{\text{MED}}) \neq 0 \tag{A.10}$$

As seen from the figures presented in Chapter 1, the normalization method used in DESeq2 fails to eliminate the bias due to variability in the sampling fraction (Figure 5). Consequently, the test statistic used in DESeq2 intrinsically tests a biased hypothesis and hence from Lemma A.0.1, it can potentially inflate the false positive rate.

A.4 edgeR

Similar to DESeq2, edgeR assumes a negative-binomial distribution for absolute abundance data:

$$\begin{aligned} O_{ijk} &\sim \text{NB}(O_{\cdot jk} s_{jk} p_{ij}, \phi_i) = \text{NB}(M_{jk} p_{ij}, \phi_i) \\ \log p_{ij} &= \beta_{i0} + \beta_{i1} I(j = 1), \quad j = 1, 2 \end{aligned} \tag{A.11}$$

where

(1) s_{jk} = normalization factor,

(2) M_{jk} = effective library size, which is the product of original library size and normalization factor,

(3) p_{ij} is the relative abundance of taxon j in experimental group j.

The upper-quartile (UQ) normalization used in edgeR is described as follows. Let

$$\begin{aligned} \hat{c}_{jk}^{\text{UQ}} &= s_{jk} = \underset{i:O_{ijk}>0}{\text{UQ}} \left(\frac{O_{ijk}}{O_{\cdot jk}}\right) \\ \hat{d}_{jk}^{\text{UQ}} &= \log \hat{c}_{jk}^{\text{UQ}}, \end{aligned} \tag{A.12}$$

where $\text{UQ}(X)$ is the upper quartile of X. Then

$$\begin{aligned} \hat{d}_{jk}^{\text{UQ}} &= \underset{i:O_{ijk}>0}{\text{UQ}} (\log O_{ijk} - \log O_{\cdot jk}) \\ &\quad \text{(Apply Taylor's expansion)} \\ &\approx \underset{i:O_{ijk}>0}{\text{UQ}} \left(y_{ijk} - \log c_{jk}\theta_{\cdot j} - \frac{1}{c_{jk}\theta_{\cdot j}}(O_{\cdot jk} - c_{jk}\theta_{\cdot j})\right) \\ &= \underset{i:O_{ijk}>0}{\text{UQ}} \left(d_{jk} + \mu_{ij} + \epsilon_{ijk} - d_{jk} - \log \theta_{\cdot j} - \frac{O_{\cdot jk}}{c_{jk}\theta_{\cdot j}} + 1\right) \\ &= 1 - \log \theta_{\cdot j} - \frac{O_{\cdot jk}}{c_{jk}\theta_{\cdot j}} + \underset{i:O_{ijk}>0}{\text{UQ}} (\mu_{ij} + \epsilon_{ijk}) \\ &:= 1 - \log \theta_{\cdot j} - \frac{O_{\cdot jk}}{c_{jk}\theta_{\cdot j}} + \mu_{a_{jkj}} + \epsilon_{a_{jkjk}} \end{aligned} \tag{A.13}$$

Similar to DESeq2, for the k^{th} sample in the j^{th} group, a_{jk} represents the index for the taxon such that $\underset{i:O_{ijk}>0}{\text{UQ}} (\mu_{ij} + \epsilon_{ijk}) = \mu_{a_{jkj}} + \epsilon_{a_{jkjk}}$.

Averaging over all sample $k = 1, 2, \ldots n$, we get

$$\bar{\hat{d}}_{j\cdot}^{\text{UQ}} = 1 - \log \theta_{\cdot j} - \bar{x}_{j\cdot} + \tilde{\mu}_{\cdot(j)j} + \tilde{\epsilon}_{\cdot(j)j\cdot} \tag{A.14}$$

As noted earlier, since each subject k in group j, may potentially have a different taxon that yields the upper quartile value $\mu_{a_{jkj}} + \epsilon_{a_{jkjk}}$, in the above expression $\tilde{x}$ represents the mean of variable x taken over the suitable subset of taxa. Secondly, the notation $\cdot(j)$ represents the mean taken within group j. $\bar{x}_{j\cdot}$ is the average of $\frac{O_{\cdot jk}}{c_{jk}\theta_{\cdot j}}$ over group j.

Thus, the UQ estimator of the bias term in the null hypothesis is

$$\begin{aligned}\hat{\delta}^{\text{UQ}} &:= \bar{\hat{d}}_{1\cdot}^{\text{UQ}} - \bar{\hat{d}}_{2\cdot}^{\text{UQ}} \\ &= (\log\theta_{\cdot 2} - \log\theta_{\cdot 1}) + (\bar{x}_{2\cdot} - \bar{x}_{1\cdot}) + (\tilde{\mu}_{\cdot(1)1} - \tilde{\mu}_{\cdot(2)2}) + (\tilde{\epsilon}_{\cdot(1)1\cdot} - \tilde{\epsilon}_{\cdot(2)2\cdot})\end{aligned} \tag{A.15}$$

Note that $E(\tilde{\epsilon}_{\cdot(1)1\cdot} - \tilde{\epsilon}_{\cdot(2)2\cdot}) = 0$. However, it is clear that the UQ estimator does not estimate the bias term in the null hypothesis unbiasedly, i.e. $E(\hat{\delta}^{\text{UQ}}) \neq \delta = \bar{d}_{1\cdot} - \bar{d}_{2\cdot}$.

Thus the UQ normalization method does not eliminate (even asymptotically) the bias due to variability in the sampling fraction. Consequently, the test statistic intrinsically tests a biased hypothesis and hence from Lemma A.0.1, it inflates the false positive rate.

Comparing the model used in edgeR (A.11) with regression framework of ANCOM-BC, we note that:

$$E(O_{ijk}) = M_{jk}p_{ij} \tag{A.16}$$

Therefore, it is more reasonable to define the estimated sampling fraction by the effective library size. For instance, the effective library size using UQ (ELib-UQ):

$$\begin{aligned}\hat{c}_{jk}^{\text{ELib-UQ}} &= M_{jk} = O_{\cdot jk}s_{jk} \\ \hat{d}_{jk}^{\text{ELib-UQ}} &= \log\hat{c}_{jk}^{\text{ELib-UQ}}\end{aligned} \tag{A.17}$$

Hence, we have:

$$\begin{aligned}\hat{d}_{jk}^{\text{ELib-UQ}} &= \underset{i:O_{ijk}>0}{\text{UQ}}(\log O_{ijk}) \\ &= \underset{i:O_{ijk}>0}{\text{UQ}}(y_{ijk}) \\ &= \underset{i:O_{ijk}>0}{\text{UQ}}(d_{jk} + \mu_{ij} + \epsilon_{ijk}) \\ &= d_{jk} + \mu_{a_{jkj}} + \epsilon_{a_{jkjk}}\end{aligned} \tag{A.18}$$

As before, for the k^{th} sample in the j^{th} group, a_{jk} represents the index for the taxon such that $\underset{i:O_{ijk}>0}{\text{UQ}}(d_{jk} + \mu_{ij} + \epsilon_{ijk}) = d_{jk} + \mu_{a_{jkj}} + \epsilon_{a_{jkjk}}$.

Averaging over all sample $k = 1, 2, \ldots n$, we get

$$\bar{\hat{d}}_j^{\text{ELib-UQ}} = \bar{d}_{j\cdot} + \tilde{\mu}_{\cdot(j)j} + \tilde{\epsilon}_{\cdot(j)j\cdot} \tag{A.19}$$

Since each subject k in group j may potentially have a different taxon that yields the upper quartile $\mu_{a_{jkj}} + \epsilon_{a_{jkjk}}$, in the above expression $\tilde{x}$ represents the mean of variable x taken over the suitable subset of taxa. Secondly, the notation $\cdot(j)$ represents the mean taken within group j.

Thus, the ELib-UQ estimator of the bias term in the null hypothesis is:

$$\begin{aligned}
\hat{\delta}^{\text{ELib-UQ}} &:= \bar{\hat{d}}_{1\cdot}^{\text{ELib-UQ}} - \bar{\hat{d}}_{2\cdot}^{\text{ELib-UQ}} \\
&= \bar{d}_{1\cdot} - \bar{d}_{2\cdot} + (\tilde{\mu}_{\cdot(1)1} - \tilde{\mu}_{\cdot(2)2}) + (\tilde{\epsilon}_{\cdot(1)1\cdot} - \tilde{\epsilon}_{\cdot(2)2\cdot}) \\
&= \delta + (\tilde{\mu}_{\cdot(1)1} - \tilde{\mu}_{\cdot(2)2}) + (\tilde{\epsilon}_{\cdot(1)1\cdot} - \tilde{\epsilon}_{\cdot(2)2\cdot})
\end{aligned} \tag{A.20}$$

Note that $E(\tilde{\epsilon}_{\cdot(1)1\cdot} - \tilde{\epsilon}_{\cdot(2)2\cdot}) = 0$. However, unless the average abundance of all 75^{th} percentile taxa is same between the two ecosystems, i.e. $\tilde{\mu}_{\cdot(1)1} - \tilde{\mu}_{\cdot(2)2} = 0$, the ELib-UQ estimator does not estimate the bias term in the null hypothesis unbiasedly, i.e. $E(\hat{\delta}^{\text{ELib-UQ}}) \neq \delta$.

Thus the ELib-UQ normalization method used in edgeR does not always eliminate the bias due to variability in the sampling fraction. Consequently, the test statistic used in edgeR intrinsically tests a biased hypothesis and hence from Lemma A.0.1, it inflates the false positive rate.

We skip the proofs for TMM and ELib-TMM since the arguments are similar.

A.5 metagenomeSeq

Suppose the zero-inflated Gaussian (ZIG) mixture model is used in metagenomeSeq. The framework can be summarized as

$$\begin{aligned}
&y_{ijk} = \log_2(O_{ijk} + 1) \\
&f_{\text{zig}}(y_{ijk}; O_{\cdot jk}, \mu_{ij}, \sigma_{ij}^2) = \pi_{jk}(O_{\cdot jk}) I_{\{0\}}(y_{ijk}) + (1 - \pi_{jk}(O_{\cdot jk}))\phi(y_{ijk}; \mu_{ij}, \sigma_{ij}^2) \\
&E(y_{ijk}|j = 1) = \pi_{jk} \cdot 0 + (1 - \pi_{jk}) \cdot (\beta_{i0} + \eta_i \log_2(\frac{\hat{s_{jk}^l} + 1}{N}) + \beta_{i1} I(j = 1))
\end{aligned} \tag{A.21}$$

where

(1) $N =$ an approximately choose normalization constant,

(2) $O_{\cdot jk} = \sum_{i=1}^{m} O_{ijk}$ is the library size for sample k in group j,

(3) $s_{jk}^{\hat{l}} = \sum_{i:O_{ijk} \leq q_{jk}^{\hat{l}}} O_{ijk}$,

(4) $q_{jk}^{\hat{l}} = \hat{l}^{th}$ quantile of sample k in group j.

$\hat{l}$ is determined by the smallest l that satisfies

$$\Delta_q^{l+1} - \Delta_q^l \geq 0.1\Delta_q^l \tag{A.22}$$

where

$$\begin{aligned} \Delta_q^l &= \text{median}_{jk}|q_{jk}^l - \bar{q}^l| \\ \bar{q}^l &= \text{median}_{jk} q_{jk}^l \end{aligned} \tag{A.23}$$

The null hypothesis under metagenomeSeq is as follows:

$$H_0 : \beta_{i1} = 0$$

$$H_1 : \beta_{i1} \neq 0$$

For simplicity of exposition, suppose $\pi_{jk} = 0$. Comparing the ZIG model (A.21) with the regression framework of ANCOM-BC, we define:

$$\hat{d}_{jk}^{\text{CSS}} = \log(s_{jk}^{\hat{l}} + 1) \tag{A.24}$$

Hence,

$$\begin{aligned} \hat{d}_{jk}^{\text{CSS}} &= \log(s_{jk}^{\hat{l}} + 1) \\ &\approx \log(s_{jk}^{\hat{l}}) \quad (s_{jk}^{\hat{l}} \text{ is much larger than 1}) \\ &\approx \log(E(s_{jk}^{\hat{l}})) + \frac{1}{E(s_{jk}^{\hat{l}})}(s_{jk}^{\hat{l}} - E(s_{jk}^{\hat{l}})) \quad \text{(Taylor's expansion)} \\ &= \log(\sum_{i:O_{ijk} \leq q_{jk}^{\hat{l}}} c_{jk}\theta_{ij}) + \frac{s_{jk}^{\hat{l}}}{E(s_{jk}^{\hat{l}})} - 1 \\ &= d_{jk} + \log(\sum_{i:O_{ijk} \leq q_{jk}^{\hat{l}}} \theta_{ij}) + \frac{s_{jk}^{\hat{l}}}{E(s_{jk}^{\hat{l}})} - 1 \\ &:= d_{jk} + x_{a_{jkj}} + z_{jk} - 1 \end{aligned} \tag{A.25}$$

As before, for the k^{th} sample in the j^{th} group, a_{jk} represents the index such that $\log(\sum_{i:O_{ijk} \leq q_{jk}^{\hat{l}}} \theta_{ij}) = x_{a_{jkj}}$, and $z_{jk} := \frac{s_{jk}^{\hat{l}}}{E(s_{jk}^{\hat{l}})}$.

Averaging over all sample $k = 1, 2, \ldots n$, we get

$$\bar{\hat{d}}_{j\cdot}^{\text{CSS}} = d_{j\cdot} + \tilde{x}_{\cdot(j)j} + \bar{z}_{j\cdot} - 1 \tag{A.26}$$

Since each subject k in group j, may potentially have a different total mean absolute abundance up to the $\hat{l}^{th}$ percentile, in the above expression $\tilde{x}$ represents the mean of variable x taken over the suitable subset of taxa. Secondly, the notation $\cdot(j)$ represents the mean taken within group j. $\bar{z}_{j\cdot}$ is the average of z_{jk}.

Thus, the CSS estimator of the bias term in the null hypothesis is:

$$\begin{aligned}
\hat{\delta}^{\text{CSS}} &:= \bar{\hat{d}}_{1\cdot}^{\text{CSS}} - \bar{\hat{d}}_{2\cdot}^{\text{CSS}} \\
&= \bar{d}_{1\cdot} - \bar{d}_{2\cdot} + (\tilde{x}_{\cdot(1)1} - \tilde{x}_{\cdot(2)2}) + (\bar{z}_{1\cdot} - \bar{z}_{2\cdot}) \\
&= \delta + (\tilde{x}_{\cdot(1)1} - \tilde{x}_{\cdot(2)2}) + (\bar{z}_{1\cdot} - \bar{z}_{2\cdot})
\end{aligned} \tag{A.27}$$

Note that unless $\tilde{x}_{\cdot(1)1} - \tilde{x}_{\cdot(2)2} = 0$, which means the sum up to $\hat{l}^{th}$ percentile of the mean absolute abundance is the same between two groups, the CSS estimator does not estimate the bias term in the null hypothesis unbiasedly, i.e. $E(\hat{\delta}^{\text{CSS}}) \neq \delta$.

Therefore, although metagenomeSeq directly tests for differential absolute abundance, there is a systematic bias in estimating sampling fractions. Again, according to Lemma A.0.1, it suffers from inflated FDR as well.

Appendix B

Residual analysis of normalization methods for differential sampling fractions

Although not explicitly stated, each normalization method available in the literature, such as the Cumulative-Sum Scaling (CSS) implemented in metagenomeSeq (Paulson et al., 2013), Median (MED) in DESeq2 (Love et al., 2014), Upper Quartile (UQ), Trimmed Mean of M-values (TMM), Total-Sum Scaling (TSS), as well as the modifications of UQ and TMM, denoted by ELib-UQ and ELib-TMM used in edgeR (Robinson et al., 2010), that account for "Effective Library size" (Chen et al., 2014), attempt to normalize the data for variability in sampling fractions across samples. In this section we describe a simple method to evaluate the performance of some of these available normalization methods, along with our proposed method in ANCOM-BC.

Suppose we have two experimental groups with balanced sample size, for each normalization method s, sample $k = 1, 2, \ldots, n$, in the j^{th} group, $j = 1, 2$, let the (raw) residual be denoted by

$$r_{jk}^s = \hat{d}_{jk}^s - d_{jk}. \tag{B.1}$$

Then $\bar{r}_{j\cdot}^s = \bar{\hat{d}}_{j\cdot}^s - \bar{d}_{j\cdot}$, therefore, $\bar{r}_{1\cdot}^s - \bar{r}_{2\cdot}^s = (\bar{\hat{d}}_{1\cdot}^s - \bar{\hat{d}}_{2\cdot}^s) - (\bar{d}_{1\cdot} - \bar{d}_{2\cdot})$. Since residuals generated by each normalization method will have their own center, to align the box plot of residuals at the same level, we center the raw residuals by

$$r_{jk}^{s*} = r_{jk}^s - \bar{r}_{\cdot\cdot}^s = \hat{d}_{jk}^s - d_{jk} - \bar{\hat{d}}_{\cdot\cdot}^s + \bar{d}_{\cdot\cdot}. \tag{B.2}$$

and make box plots using these (centered) residuals. Thus, if the normalization method is effective then there should be no systematic pattern among the residuals by the experimental groups. Otherwise, the normalization method is not successfully eliminating the bias due to variability in sampling fractions.

Based on our simulated data (Figure 5), as expected ANCOM-BC seems to successfully eliminate the bias induced by the differences in the sampling fractions between two experimental groups. For ANCOM-BC, the samples from the two groups (circles and triangles) are nicely intermixed with small variability of residuals. Consistent with our observations in the previous section, this is not always the case with other methods. For other methods, the group labels are not randomly distributed around zero but they are clustered by the group label (Figure 5). This suggests that the existing normalization methods do not eliminate the systematic bias introduced by the differences in the sampling fractions.

Bibliography

Agresti, A. and Hitchcock, D. B. (2005). Bayesian inference for categorical data analysis. *Statistical Methods and Applications*, 14(3):297–330.

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 139–177.

Aitchison, J. (2003). A concise guide to compositional data analysis. In *CDA Workshop, Girona*.

Amato, K. R. (2017). An introduction to microbiome analysis for human biology applications. *American Journal of Human Biology*, 29(1):e22931.

Amrhein, V., Greenland, S., and McShane, B. (2019). Scientists rise up against statistical significance.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106.

Benjamini, Y. and Hochberg, Y. (1995). 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.

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243.

Bray, J. R. and Curtis, J. T. (1957). An ordination of the upland forest communities of southern wisconsin. *Ecological monographs*, 27(4):325–349.

Brewer, A. and Williamson, M. (1994). A new relationship for rarefaction. *Biodiversity & Conservation*, 3(4):373–379.

Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC bioinformatics*, 11(1):94.

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., Fierer, N., and Knight, R. (2011). Global patterns of 16s rrna diversity at a depth of millions of sequences per sample. *Proceedings of the national academy of sciences*, 108(Supplement 1):4516–4522.

Castaner, O., Goday, A., Park, Y.-M., Lee, S.-H., Magkos, F., Shiow, S.-A. T. E., and Schröder, H. (2018). The gut microbiome profile in obesity: a systematic review. *International journal of endocrinology*, 2018.

Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of statistics*, pages 265–270.

Chen, E. Z. and Li, H. (2016). A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics*, 32(17):2611–2617.

Chen, Y., McCarthy, D., Robinson, M., and Smyth, G. K. (2014). edger: differential expression analysis of digital gene expression data user's guide.

Chen, Y., Wu, Y., and Shen, T.-J. (2018). Evaluation of the estimate bias magnitude of the rao's quadratic diversity index. *PeerJ*, 6:e5211.

Codd, G. (1995). Cyanobacterial toxins: occurrence, properties and biological significance. *Water Science and Technology*, 32(4):149–156.

Costea, P. I., Zeller, G., Sunagawa, S., and Bork, P. (2014). A fair comparison. *Nature methods*, 11(4):359.

Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., et al. (2013). A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in bioinformatics*, 14(6):671–683.

Dubourg, G., Lagier, J.-C., Armougom, F., Robert, C., Audoly, G., Papazian, L., and Raoult, D. (2013). High-level colonisation of the human gut by verrucomicrobia following broad-spectrum antibiotic treatment. *International journal of antimicrobial agents*, 41(2):149–155.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50(272):1096–1121.

Dunnett, C. W. and Tamhane, A. C. (1991). Step-down multiple tests for comparing treatments with a control in unbalanced one-way layouts. *Statistics in medicine*, 10(6):939–947.

Dunnett, C. W. and Tamhane, A. C. (1992). A step-up multiple test procedure. *Journal of the American Statistical Association*, 87(417):162–170.

Egozcue, J. J. and Pawlowsky-Glahn, V. (2005). Groups of parts and their balances in compositional data analysis. *Mathematical Geology*, 37(7):795–828.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barcelo-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300.

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biological conservation*, 61(1):1–10.

Faith, D. P. and Baker, A. M. (2006). Phylogenetic diversity (pd) and biodiversity conservation: some bioinformatics challenges. *Evolutionary bioinformatics*, 2:117693430600200007.

Fernandes, A. D., Macklaim, J. M., Linn, T. G., Reid, G., and Gloor, G. B. (2013). Anova-like differential expression (aldex) analysis for mixed population rna-seq. *PLoS One*, 8(7).

Fernandes, A. D., Reid, J. N., Macklaim, J. M., McMurrough, T. A., Edgell, D. R., and Gloor, G. B. (2014). Unifying the analysis of high-throughput sequencing datasets: characterizing rna-seq, 16s rrna gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, 2(1):15.

Friedman, J. and Alm, E. J. (2012). Inferring correlation networks from genomic survey data. *PLoS computational biology*, 8(9).

Fu, A., Narasimhan, B., and Boyd, S. (2017). Cvxr: An r package for disciplined convex optimization. *arXiv preprint arXiv:1711.07582*.

Gevers, D., Kugathasan, S., Denson, L. A., Vázquez-Baeza, Y., Van Treuren, W., Ren, B., Schwager, E., Knights, D., Song, S. J., Yassour, M., et al. (2014). The treatment-naive microbiome in new-onset crohn?s disease. *Cell host & microbe*, 15(3):382–392.

Gloor, G. (2015). Aldex2: Anova-like differential expression tool for compositional data. *ALDEX manual modular*, 20:1–11.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Frontiers in microbiology*, 8:2224.

Gloor, G. B. and Reid, G. (2016). Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. *Canadian journal of microbiology*, 62(8):692–703.

Gloor, G. B., Wu, J. R., Pawlowsky-Glahn, V., and Egozcue, J. J. (2016). It's all relative: analyzing microbiome data as compositions. *Annals of epidemiology*, 26(5):322–329.

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

Grandhi, A., Guo, W., and Peddada, S. D. (2016). A multiple testing procedure for multi-dimensional pairwise comparisons with application to gene expression studies. *BMC bioinformatics*, 17(1):104.

Greenacre, M. (2011). Measuring subcompositional incoherence. *Mathematical Geosciences*, 43(6):681–693.

Guo, W., Sarkar, S. K., and Peddada, S. D. (2010). Controlling false discoveries in multidimensional directional decisions, with applications to gene expression data on ordered categories. *Biometrics*, 66(2):485–492.

Halperin, J. J. (2010). A tale of two spirochetes: lyme disease and syphilis. *Neurologic clinics*, 28(1):277–291.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika*, 61(2):383–385.

Herlemann, D. P., Geissinger, O., and Brune, A. (2007). The termite group i phylum is highly diverse and widespread in the environment. *Appl. Environ. Microbiol.*, 73(20):6682–6685.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

Horner-Devine, M. C., Lage, M., Hughes, J. B., and Bohannan, B. J. (2004). A taxa–area relationship for bacteria. *Nature*, 432(7018):750.

Hurst, G. D. (2017). Extended genomes: symbiosis and evolution. *Interface Focus*, 7(5):20170001.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579.

Jelsema, C. M., Peddada, S. D., et al. (2016). Clme: An r package for linear mixed effects models under inequality constraints. *Journal of Statistical Software*, 75(i01).

Jernvall, J. and Wright, P. C. (1998). Diversity components of impending primate extinctions. *Proceedings of the National Academy of Sciences*, 95(19):11279–11283.

Kaul, A., Mandal, S., Davidov, O., and Peddada, S. D. (2017). Analysis of microbiome data in the presence of excess zeros. *Frontiers in microbiology*, 8:2114.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Kumar, M. S., Slud, E. V., Okrah, K., Hicks, S. C., Hannenhalli, S., and Bravo, H. C. (2018). Analysis and correction of compositional bias in sparse sequencing count data. *BMC genomics*, 19(1):799.

Lim, C., Sen, P. K., and Peddada, S. D. (2013). Robust analysis of high throughput screening (hts) assay data. *Technometrics*, 55(2):150–160.

Lindstrom, M. J. and Bates, D. M. (1988). Newton—raphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404):1014–1022.

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

Lozupone, C. and Knight, R. (2005). Unifrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.*, 71(12):8228–8235.

Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J., and Knight, R. (2011). Unifrac: an effective distance metric for microbial community comparison. *The ISME journal*, 5(2):169.

Lozupone, C. A., Hamady, M., Kelley, S. T., and Knight, R. (2007). Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Appl. Environ. Microbiol.*, 73(5):1576–1585.

Lozupone, C. A., Li, M., Campbell, T. B., Flores, S. C., Linderman, D., Gebert, M. J., Knight, R., Fontenot, A. P., and Palmer, B. E. (2013a). Alterations in the gut microbiota associated with hiv-1 infection. *Cell host & microbe*, 14(3):329–339.

Lozupone, C. A., Stombaugh, J., Gonzalez, A., Ackermann, G., Wendel, D., Vázquez-Baeza, Y., Jansson, J. K., Gordon, J. I., and Knight, R. (2013b). Meta-analyses of studies of the human microbiota. *Genome research*, 23(10):1704–1714.

Magurran, A. E. (2013). *Measuring biological diversity*. John Wiley & Sons.

Mandal, S., Van Treuren, W., White, R. A., Eggesbø, M., Knight, R., and Peddada, S. D. (2015). Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial ecology in health and disease*, 26(1):27663.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

McLachlan, G. and Krishnan, T. (2007). *The EM algorithm and extensions*, volume 382. John Wiley & Sons.

Morris, E. K., Caruso, T., Buscot, F., Fischer, M., Hancock, C., Maier, T. S., Meiners, T., Müller, C., Obermaier, E., Prati, D., et al. (2014). Choosing and using diversity indices: insights for ecological applications from the german biodiversity exploratories. *Ecology and evolution*, 4(18):3514–3524.

Morton, J. T., Marotz, C., Washburne, A., Silverman, J., Zaramela, L. S., Edlund, A., Zengler, K., and Knight, R. (2019). Establishing microbial composition measurement standards with reference frames. *Nature communications*, 10(1):2719.

Morton, J. T., Sanders, J., Quinn, R. A., McDonald, D., Gonzalez, A., Vázquez-Baeza, Y., Navas-Molina, J. A., Song, S. J., Metcalf, J. L., Hyde, E. R., et al. (2017). Balance trees reveal microbial niche differentiation. *MSystems*, 2(1):e00162–16.

Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika*, 49(1/2):65–82.

Nayak, T. K. (1986). An analysis of diversity using rao's quadratic entropy. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 315–330.

Obregon-Tito, A. J., Tito, R. Y., Metcalf, J., Sankaranarayanan, K., Clemente, J. C., Ursell, L. K., Xu, Z. Z., Van Treuren, W., Knight, R., Gaffney, P. M., et al. (2015). Subsistence strategies in traditional societies distinguish gut microbiomes. *Nature communications*, 6:6505.

O'Hara, A. M. and Shanahan, F. (2006). The gut flora as a forgotten organ. *EMBO reports*, 7(7):688–693.

O'Keefe, S. J., Li, J. V., Lahti, L., Ou, J., Carbonero, F., Mohammed, K., Posma, J. M., Kinross, J., Wahl, E., Ruder, E., et al. (2015). Fat, fibre and cancer risk in african americans and rural africans. *Nature communications*, 6:6342.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554.

Paulson, J. N., Bravo, H. C., and Pop, M. (2014). Reply to:" a fair comparison". *Nature methods*, 11(4):359.

Paulson, J. N., Stine, O. C., Bravo, H. C., and Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nature methods*, 10(12):1200.

Pawlowsky-Glahn, V. and Buccianti, A. (2011). *Compositional data analysis*. Wiley Online Library.

Pawlowsky-Glahn, V. and Egozcue, J. J. (2011). Exploring compositional data with the coda-dendrogram. *Austrian Journal of Statistics*, 40(1&2):103–113.

Peddada, S. D. and Smith, T. (1997). Consistency of a class of variance estimators in linear models under heteroscedasticity. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 1–10.

Rao, C. (2010). Quadratic entropy and analysis of diversity. *Sankhya A*, 72(1):70–80.

Rao, C. R. (1984). Convexity properties of entropy functions and analysis of diversity. *Lecture Notes-Monograph Series*, pages 68–77.

Relman, D. A. and Falkow, S. (2001). The meaning and impact of the human genome sequence for microbiology. *Trends in microbiology*, 9(5):206–208.

Ricotta, C. and Marignani, M. (2007). Computing β-diversity with rao's quadratic entropy: a change of perspective. *Diversity and Distributions*, 13(2):237–241.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, 11(3):R25.

Shah, N., Tang, H., Doak, T. G., and Ye, Y. (2011). Comparing bacterial communities inferred from 16s rrna gene sequencing and shotgun metagenomics. In *Biocomputing 2011*, pages 165–176. World Scientific.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.

Simpson, E. H. (1949). Measurement of diversity. *nature*, 163(4148):688.

Sukhatme, P. V., Sukhatme, B. V., Sukhatme, S., and Asok, C. (1970). Sampling theory of surveys with applications. ames.

Székely, G. J., Rizzo, M. L., Bakirov, N. K., et al. (2007). Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794.

Tierney, B. T., Yang, Z., Luber, J. M., Beaudin, M., Wibowo, M. C., Baek, C., Mehlenbacher, E., Patel, C. J., and Kostic, A. D. (2019). The landscape of genetic content in the gut and oral human microbiome. *Cell host & microbe*, 26(2):283–295.

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., Sogin, M. L., Jones, W. J., Roe, B. A., Affourtit, J. P., et al. (2009). A core gut microbiome in obese and lean twins. *nature*, 457(7228):480.

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J. R., Vázquez-Baeza, Y., Birmingham, A., et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5(1):27.

Whittaker, R. H. (1960). Vegetation of the siskiyou mountains, oregon and california. *Ecological monographs*, 30(3):279–338.

Whittaker, R. H. (1972). Evolution and measurement of species diversity. *Taxon*, pages 213–251.

Xia, F., Chen, J., Fung, W. K., and Li, H. (2013). A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics*, 69(4):1053–1063.

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R. N., Anokhin, A. P., et al. (2012). Human gut microbiome viewed across age and geography. *nature*, 486(7402):222.