Simulation studies to assess the power of set testing methods for microbiome data

by

Lauren McKeen

A Creative Component submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Statistics

Program of Study Committee:
Chong Wang
Peng Liu
Max Morris

Iowa State University

Ames, Iowa

2020

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

With advances in sequencing methods, the study of the microbiome has greatly increased. Microbiome data, in the form of an OTU or ASV count table, can be used to identify specific ASVs that function differently across treatment conditions. Such analysis is deemed differential abundance analysis. ASVs are grouped by their taxonomic rank, and ASVs sharing the same rank have similar biological traits. By studying groups or sets of ASVs, and identifying if the set is differentially abundant, the biological interpretation of a microbiome study is enhanced. We review current approaches in set testing methods and apply them to a microbiome data set from a 2017 study. We propose a new set testing method based on an existing Poisson hurdle model, and compare performance across all methods through a model based simulation study. We find that under certain conditions, our proposed model outperforms existing approaches. We discuss the limitations of our model and conclude that more simulation studies, specifically non-parametric simulation studies, are needed to better compare across possible methods.

CHAPTER 1. INTRODUCTION

A microbiome is a collection or community of micro-organisms, such as bacteria, fungi, and viruses, that populate an environment, and create what can be referred to as a “mini-ecosystem” [1]. Such organisms are vital to the health and functioning of both humans and plants and are found in the human gut as well as the soil of a plant. With recent advances in DNA sequencing, the study of these complex communities is increasing. As more and more sequencing data become available, and more tools for analysis are developed, the understanding of these communities so too increases.

Currently, one of the main approaches in sequencing is amplicon sequencing. This approach involves amplifying and then sequencing a specific gene, often, the 16S rRNA gene. This gene has been found to have many important features that make it the target of such sequencing, including being present in all bacteria. Once a sample is collected, the DNA is extracted and sequenced, and is then further classified. Commonly, this process involves classification into Operational Taxanomic Units (OTUs), in which sequences are grouped based on their similarity to OTUs, or other sequences in the community [2]. Another approach in sequencing is whole-metagenome shotgun sequencing, which does not target a specific gene. After a process called marker gene analysis, sequences are classified into Amplicon Sequence Variants (ASVs) [3].

To distinguish between the above two approaches of classification is not central to this project, as both result in the same data structure. Typically, thousands of ASVs/OTUs (which we will refer to as ASVs from now on) are observed in every sample, and the data are arranged in a count table with rows and columns representing ASVs and samples, respectively. These data often have a large number of zeros, along with relatively few samples compared to the number of ASVs. Table 1.1 below shows an example of an ASV count table, which exhibits the “large p, small n” phenomenon;

i.e., there are very few biological replicates (samples), n, compared to the number of ASVs, p. Here, there are 2 treatment conditions with two replicates each, and 1000 total ASVs.

ASV	Sample (Treatment)			
	1 (Low Nitrogen)	2 (Low Nitrogen)	3 (High Nitrogen)	4 (High Nitrogen)
1	0	0	0	10
2	0	2	2	6
3	15	8	7	2
⋮	⋮	⋮	⋮	⋮
1000	24	4	0	2

Table 1.1: ASV count table example

Microbiome data in the form of an ASV count table, like the one shown above, can be used to pinpoint the ASVs that function differently across treatment conditions, which in turn provides important information about the "mini-ecosystem" and how it responds to treatments. Such analysis is called differential abundance. There are several methods available to identify differentially abundant (DA) ASVs, many of which are similar to differential expression (DE) methods for genes; examples of these methods will be discussed later. Since ASVs differ from genes in that the data have excess zeros, there have also been several methods proposed that account for this.

Recently, there have been methods proposed to identify DE gene *categories*. Identifying DE gene categories requires prior knowledge to group genes into said categories. This strengthens the understanding of how a treatment works, and how different genes, and sets, respond to treatment [4]. While there has been much work done in the field of gene set testing, there is not much existing literature on set testing for ASVs. ASVs have an inherent structure to them in their taxanomic rank (domain, kingdom, phylum, etc.); for example, ASVs belonging to the same phylum share similar traits. Due to this biological structure, ASVs can easily be grouped into categories. Thus, the goal of this project is to identify categories, or sets, of DA ASVs. We briefly review methods of identifying DA/DE ASVs/genes, review and analyze current methods in gene set testing, and provide suggestions to determine a method that can be used to identify DA sets of ASVs.

CHAPTER 2. EXISTING METHODS

2.1 Differential Abundance

Identifying DA ASVs across treatment conditions is similar to differential expression analysis for RNA-seq data; i.e., for genes. Like genes, ASVs are considered DA if their mean count is significantly different across two or more treatment conditions. Thus, the statistical tools that were originally developed for differential expression analysis of RNA-seq data, such as edgeR and DESeq, have been suggested to identify DA ASVs [2].

Briefly, both edgeR and DESeq require proper normalization of the count data. The aim of normalization is to account for any "systematic technical differences among samples occurring in the data" [2], so that technical bias is removed. Once the data have been appropriately normalized, both methods use a negative binomial model and focus on estimation of an overdispersion parameter. Additionally, for experimental designs with only two treatment conditions, edgeR implements an exact test, similar to Fisher's Exact Test, to detect DE genes [5, 6].

While edgeR and DESeq are commonly used to identify DA ASVs, such methods do not account for the greater abundance of zeros typically found in ASV count data. A method called metagenomeSeq was designed specifically for microbiome/ASV data and is based on a zero-inflated Gaussian model [7]. It is important to note that this project will be based on a model proposed as an alternative to the three previously mentioned methods, and is further explained in section 3.

2.2 Set (or Category) Testing

As previously emphasized, methods used to identify DE genes can also be used to identify DA ASVs; this can also be extended to methods in set testing. There are two main approaches to set testing found in the literature. The first is a test of enrichment, where categories are compared to each other and a given category is tested for over-representation compared to the remaining categories. The second is not a test of enrichment; instead, each category is tested individually. This section focuses on some key methods in gene set testing that may be useful for ASV set testing, along with one method that has been developed for microbiome data.

2.2.1 GOseq

GOseq, a method proposed for RNA-seq data, advances previous methods of Gene Ontology (GO) analysis. GO methods test for over-representation of a category and assume that every gene has equal probability of being detected as DE. Unlike typical GO methods, where all genes are assumed independent, GOseq incorporates selection bias; i.e., GOseq accounts for categories with a greater number of "longer" genes (genes with a more reads), or categories with highly expressed genes, as both situations lead to a higher chance of a category being found as over-represented. GOseq proposes a three step methodology.

First, genes that are significantly DE between treatment conditions are identified. This can be done using techniques such as edgeR or DESeq, and p-values for differential expression need to be provided. Then, a probability weighting function (PWF) is estimated from the data. This function captures the relationship between the probability of differential expression and transcript length (or any other type of bias that may be of interest, such as read count), by dichotomizing the p-values into the following: a zero if the gene is not DE, and 1 otherwise. Lastly, a p-value for each category is computed through resampling. A random set of genes is selected, the same size as the set of DE genes, and the number of genes associated with a given category are counted, weighted by the PWF. The resampling is then repeated many times and creates a sampling distribution. The sampling distribution allows for a calculation of a p-value for whether each category is over-represented in the set of DE genes [8].

For the purposes of this project, it is important to note that GOseq's null hypothesis is based on the enrichment of a category compared to other categories. For each of C categories, $c = 1, \dots, C$, the null hypothesis is:

$$H_{0c} : \text{Category } c \text{ is not over represented, or enriched, compared to all other categories.}$$

A major drawback to GOseq is in dichotomizing the p-values of DE genes, the rank order is lost. Further, to dichotomize the p-values depends on an arbitrary threshold for significance. Several other methods exist that do not have such a drawback.

2.2.2 SAFE

Significance Analysis of Function and Expression (SAFE) was also proposed for microarray data. SAFE is a two-stage permutation-based approach to set testing. First, gene-specific statistics are calculated. These statistics are deemed "local statistics" and measure the association between the expression profile of a gene and the response variable of interest. Local statistics may include t-statistics if appropriate, or a ranking statistic based on p-values from differential expression testing. Next, a larger-scale "global statistic" is computed as a function of the local statistics. This global statistic aims to assess how the distribution of local statistics within a category differs from local statistics outside of the category. Common global statistics include the Wilcoxon rank sum and Kolmogorov-Smirnov statistics. The significance of the global statistic for each category is then calculated through the use of permutations by randomly swapping treatment labels. [9].

Overall, SAFE is a very flexible approach to the identification of over-represented gene categories, as users are able to choose their own local and global statistics. It also has the ability to incorporate the dichotomization discussed earlier, by viewing local statistics as the presence or absence of a gene on a list and the global statistic as a sum of local statistics. By allowing for other statistics, however, the power to identify DE categories can be improved. Again, it is important to note that the null hypothesis of SAFE is the same of that of GOseq.

2.2.3 MRPP

Several authors have noted that information is lost when continuous gene specific measures of differential expression (i.e., p-values) are dichotomized to produce a list of DE genes and non-DE genes (this is the approach taken by GOseq) [9, 10, 4]. The rank order of the genes in terms of differential expression is lost. Results of analysis can also be sensitive to the threshold of significance.

Categories may go undetected and fail to be identified as over-represented if many changes on the individual fail to reach a certain significance level. Further, methods similar to SAFE may fail to detect treatment effects on the multivariate expression distribution of genes in a category.

A multi-response permutation procedure (MRPP) is an approach that can overcome the issues with both GOseq and SAFE. MRPP is not a test of enrichment, unlike GOseq and SAFE, and is thus not a comparison among categories. Instead of comparing groups of genes against each other based on gene-specific analysis of differential expression, MRPP identifies gene categories whose "joint expression distribution differs across treatments" [4].

The null hypothesis in this case is that the multivariate distribution of genes in a given category is the same for all treatment conditions. When this null hypothesis is false, the category is said to be DE. MRPP is based on distances. Euclidean distances is commonly used, but not required. Like previously mentioned methods, MRPP uses a permutation based p-value by swapping treatment labels to provide a p-value for a gene category of interest.

2.2.4 PERMANOVA

Similar to MRPP, PERMANOVA is a multivariate test. However, PERMANOVA was developed for microbiome data, and essentially compares collections of objects by examining the centroids and dispersion of said collection. More explicitly, PERMANOVA is a nonparametric procedure for testing the hypothesis of no difference in centroids and dispersion between two or more treatment conditions (i.e., samples). A distance metric is used to capture the multivariate relationship between each pair of samples. Again, Euclidean distance can be used as a distance metric, but other forms of distances can also be used. The flexibility in choice of distance allows for some microbiome-specific distance measures to be used. Although we will not discuss such distance metrics here, some of the most common ones include alpha and beta diversity, and have biological interpretations for microbiome data. Again, a permutation p-value is computed by permuting treatment labels [11].

CHAPTER 3. PROPOSED METHOD

There exist additional challenges for microbiome data that were not addressed by any of the methods reviewed in section 2. Specifically, none of the methods account for the large number of zeros that are present in microbiome data. To suggest a method that is suitable for microbiome data, we build upon an existing model that incorporates zero inflation.

We present our proposed method based on a zero-inflated model for identifying DA ASVs. Chaohui Yuan et. al. proposed a hierearchical Poisson hurdle model to identify DA ASVs as an alternative to methods such as edgeR, DESeq, and metagenomeSeq. The model is defined as follows.

Let Y_{gij} be the ASV count for the gth ($g = 1, \ldots, G$) ASV of the jth $j = 1, \ldots, n_i$) replicate in the ith (i = 1,2) treatment group. Let Z_{gij} be a Bernoulli random variable that takes the value one if Y_{gij} is positive. Then,

$$Z_{gij}|p_{gi} \sim Ber(p_{gi}),$$

$$\mathrm{Y}_{gij}|Z_{gij}, \mu_{gij} \sim \begin{cases} 0, & \text{if } Z_{gij} = 0, \\ \mathrm{TP}(\mu_{gij}), & \text{if } Z_{gij} = 1, \end{cases} \tag{3.1}$$

where TP(μ_{gij}) represents a zero truncated Poisson distribution and p_{gi} is the probability of the ASV count being positive. Further,

$$\mathrm{logit}(p_{gi}) = \alpha_{gi}, \tag{3.2}$$

$$\log(\mu_{gij}) = log(s_{ij}) + \beta_{gi} \tag{3.3}$$

where s_{ij} represents a normalization factor [6].

Differential abundance across treatment conditions is affected by both the probability of having a positive count, and the count given the count is positive. This accounts for the large proportion of zeros that is typical of microbiome data. Thus, for each ASV g, the following hypotheses are of interest:

$$H_0^g : \alpha_{g1} = \alpha_{g2} \text{ and } \beta_{g1} = \beta_{g2} \tag{3.4}$$

Let $\eta_{\alpha g}$ be an indicator variable such that $\eta_{\alpha g} = 0$ represents $\alpha_{g1} = \alpha_{g2}$, i.e., the probability of having a positive count is not affected by the treatment. Similarly, define $\eta_{\beta g}$ which represents the probability of differential abundance given the count is positive. If either $\eta_{\alpha g} = 1$ or $\eta_{\beta g} = 1$, then we reject the null hypothesis and conclude there is a treatment effect on ASV g.

The indicator variables are further modeled through use of a mixture model. First, the zero component is modeled using a Bernoulli distribution; i.e., $\eta_{\alpha g}|\pi_{\alpha 0} \sim \text{Ber}(1 - \pi_{\alpha 0})$, with $\pi_{\alpha 0}$ equal to the chance that $\alpha_{g1} = \alpha_{g2}$. Then,

$$(\alpha_{g1}, \alpha_{g2})' = \begin{cases} \tilde{\alpha}_{g0}(1,1)', & \text{if } \eta_{\alpha g} = 0, \\ (\tilde{\alpha}_{g1}, \tilde{\alpha}_{g2})', & \text{if } \eta_{\alpha g} = 1, \end{cases} \tag{3.5}$$

and

$$\tilde{\alpha}_{g0}|\phi_0, \tau_0 \sim N(\phi_0, \tau_0^2),$$
$$\tilde{\alpha}_{gi}|\phi_i, \tau_i \sim N(\phi_i, \tau_i^2).$$

Similarly, the count component is modeled using a Bernoulli distribution; i.e., $\eta_{\beta g}|\pi_{\beta 0} \sim \text{Ber}(1-\pi_{\beta 0})$, with $\pi_{\beta 0}$ equal to the chance that $\beta_{g1} = \beta_{g2}$. Then,

$$(\beta_{g1}, \beta_{g2})' = \begin{cases} \tilde{\beta}_{g0}(1,1)', & \text{if } \eta_{\beta g} = 0, \\ (\tilde{\beta}_{g1}, \tilde{\beta}_{g2})', & \text{if } \eta_{\beta g} = 1, \end{cases} \tag{3.6}$$

and

$$\tilde{\beta}_{g0}|\theta_0, \sigma_0 \sim N(\theta_0, \sigma_0^2),$$

$$\tilde{\beta}_{gi}|\theta_i, \sigma_i \sim N(\theta_i, \sigma_i^2).$$

The prior distributions are defined as follows:

$$\phi_i \sim \mathrm{N}(0, 10^4)$$

$$\phi_0 \sim \mathrm{N}(0, 10^4)$$

$$\theta_i \sim \mathrm{N}(0, 10^4)$$

$$\theta_0 \sim \mathrm{N}(0, 10^4)$$

$$\tau_i^2 \sim \mathrm{IG}(0.001, .001)$$

$$\tau_0^2 \sim \mathrm{IG}(0.001, .001)$$

$$\sigma_i^2 \sim \mathrm{IG}(0.001, .001)$$

$$\sigma_0^2 \sim \mathrm{IG}(0.001, .001)$$

$$\pi_{\alpha 0} \sim \mathrm{Unif}(0, 1)$$

$$\pi_{\beta 0} \sim \mathrm{Unif}(0, 1)$$

To estimate these parameters, a fully Bayesian approach is used, and is implemented through JAGS.

Since this model involves heavy computation, this project uses the model presented above to first identify DA ASVs and then defines a category or set of ASVs as DA if any individual ASV belonging to that category is DA. A posterior probability for each category is calculated. For the rest of this paper, we will refer to this model as PHSeq [12].

CHAPTER 4. REAL DATA ANALYSIS

In this section, we apply each of the existing methods for set testing to a real data set.

4.1 Data

A microbiome data set from a 2017 study consists of samples taken from the soil, root, and rhizosphere (sample type) of sorghum plants, across a variety of treatment conditions, including nitrogen levels and genotype. The experiment was repeated on different days and in different locations. This large microbiome data set contains over one hundred thousand ASVs and over 3300 samples. Additionally, the annotations of the ASVs are provided; i.e., the taxonomic classifications for each ASV are given.

A subset of this data set was chosen for the purposes of this project. To select a subset, first one location, Central City, was chosen. Then for every combination of date, sample type, and genotype, the total number of ASVs present (after some filtering criteria) were identified, along with the total number of DA ASVs based on edgeR. The combination of conditions that maximized the total number of ASVs present and the total number of DA ASVs was chosen as the subset of data to be used for real data analysis. The resulting data set contains 16 samples from two treatment conditions, high and low nitrogen. (Samples were taken on July 18th, are from the rhizosphere of the plant, and have genotype Chinese Amber.)

Based on this subset of conditions, there are 142,111 total ASVs. However, many of these ASVs do not appear in any of the 16 samples, across either treatment condition. Figure 4.1 shows that a large number of ASVs are present in no samples; that is, there are a large number of ASVs that have entirely zero counts in all 16 samples. After filtering to consider only the ASVs that are present in at least one-third of all samples (at least 6 total samples), there are 660 ASVs. The process of filtering ASVs with low abundance described here has been performed in many papers.

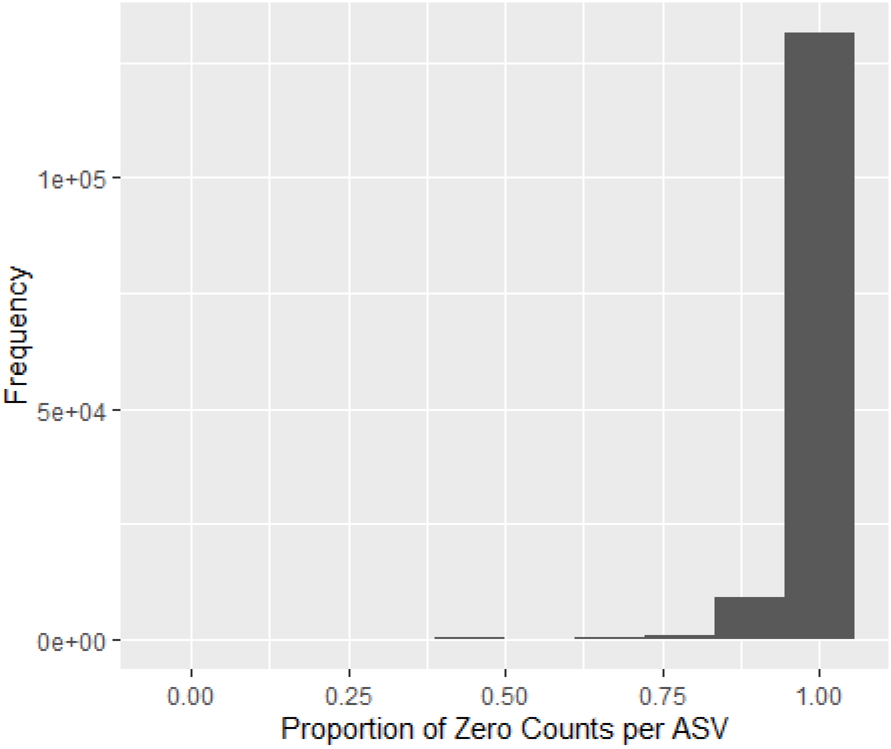


Figure 4.1: Proportion of zero counts per ASV

The final data set consists of 660 ASVs across 16 samples (8 samples from one treatment condition - high nitrogen, and 8 samples from another treatment condition - low nitrogen). The taxanomic classifications for each ASV is also given, at the following levels: domain, phylum, class, order, family, genus, and species. All ASVs belong to the bacteria domain; a breakdown of the number of ASVs belonging to each phylum is given in Table 4.1.

	Phylum	n
1	Proteobacteria	294
2	Actinobacteria	119
3	Acidobacteria	70
4	Bacteroidetes	56
5	Chloroflexi	30
6	Verrucomicrobia	23
7	Firmicutes	18
8	Gemmatimonadetes	18
9	Planctomycetes	11
10	Entotheonellaeota	6
11	Nitrospirae	4
12	Rokubacteria	4
13	Armatimonadetes	3
14	Cyanobacteria	2
15	Latescibacteria	2

Table 4.1: Breakdown of ASVs belonging to each phylum

The phylum level is chosen to represent categories in the remainder of the real data analysis.

4.2 Methods

Each of the existing methods described in section 2 were tested on our real data set, in order to identify sets of DA ASVs. Default settings were used for each. To implement GOseq, edgeR was first used to identify DA ASVs (after proper normalization) and no correction for selection bias was applied. To implement SAFE, the local statistic was chosen as a t-statistic, and the global statistic was chosen as the Wilcoxon Rank Sum statistic. For both MRPP and PERMANOVA, Euclidean distances were used.

4.3 Results

The results below do not account for multiple testing error (i.e., p-values for categories/sets are not adjusted), although all methods do provide means by which to account for this in their paper and respective code. Table 4.2 provides p-values for each category for each of the existing methods described in section 2.

	Category	Category Size	GOseq	SAFE	MRPP	PERMANOVA
1	Proteobacteria	294	0.248	0.553	0.689	0.537
2	Actinobacteria	119	0.910	0.804	0.616	0.631
3	Acidobacteria	70	1.000	0.598	0.962	0.940
4	Bacteroidetes	56	0.658	0.034	0.466	0.528
5	Chloroflexi	30	1.000	0.949	0.853	0.948
6	Verrucomicrobia	23	0.349	0.150	0.137	0.126
7	Firmicutes	18	0.284	0.082	0.196	0.126
8	Gemmatimonadetes	18	1.000	0.532	0.588	0.574
9	Planctomycetes	11	1.000	0.948	0.813	0.809
10	Entotheonellaeota	6	1.000	0.144	0.080	0.061
11	Nitrospirae	4	1.000	0.901	0.940	0.830
12	Rokubacteria	4	1.000	0.091	0.324	0.333
13	Armatimonadetes	3	1.000	0.589	0.356	0.348
14	Cyanobacteria	2	0.036	0.076	0.008	0.003
15	Latescibacteria	2	1.000	0.332	0.144	0.528

Table 4.2: P-values for differential abundance of all categories (at the phylum level) for GOseq, SAFE, MRPP, and PERMANOVA

Based on just Table 4.2, it is difficult to compare the performance of these methods. Table 4.3 shows the top DA categories by method, where a (*) indicates significance at $\alpha = .05$. Clearly, not all methods agree in terms of which category is the most DA and even in whether or not a category is DA. Consequently, the results of the real data analysis are hard to interpret, as the true status of the categories remain unknown even after analysis. This motivates a simulation study to assess, and compare, all methods.

GoSeq	SAFE	MRPP	PERMANOVA
Cyanobacteria* (2)	Bacteriodetes* (56)	Cyanobacteria* (2)	Cyanobacteria* (2)
Proteobacteria (294)	Firmicutes (18)	Entotheonellaeota (6)	Entotheonellaeota (6)
Firmicutes (18)	Cyanobacteria (2)	Verrucomicrobia (23)	Verrucomicrobia (23)
Verrucomicrobia (23)	Rokubacteria (4)	Latescibacteria (2)	Firmicutes (18)
Bacteriodetes (56)	Entotheonellaeota (6)	Firmicutes (18)	Rokubacteria (4)

Table 4.3: Top 5 DA categories (size) by method, where a (*) indicates significance at $\alpha = .05$

CHAPTER 5. MODEL BASED SIMULATION

While the results in section 4 show that current methods, whether they be for genes or ASVs, can be applied to microbiome data, they do not prove that any of the methods are superior to another, since the true status of each category is unknown. In this section, we examine the performance of current set testing methods along with our proposed method based on PHSeq. Through a model based simulation study, we assess the power and type one error of all methods.

5.1 Model Fit

We base our simulation study on the fit of a real data set. Fitting the PHSeq model to the data set described in section 4 results in estimates of the model parameters as shown in Table 5.1.

	statistics.Mean	statistics.SD	statistics.Naive.SE	statistics.Time.series.SE
$\pi_{\alpha 0}$	0.27369	0.02250	0.00038	0.00152
$\pi_{\beta 0}$	0.52483	0.02620	0.00044	0.00093
σ_1	0.85073	0.03870	0.00065	0.00225
σ_2	0.76651	0.03389	0.00057	0.00150
σ_0	0.64258	0.03442	0.00058	0.00619
τ_1	0.06819	0.03800	0.00064	0.00466
τ_2	0.07922	0.04864	0.00082	0.00712
τ_0	0.10197	0.06466	0.00109	0.00968
θ_1	3.84183	0.05737	0.00097	0.00238
θ_2	3.83193	0.05073	0.00086	0.00278
θ_0	3.31246	0.03896	0.00066	0.00171
ϕ_1	-0.16021	0.03990	0.00067	0.00406
ϕ_2	0.25221	0.03778	0.00064	0.00335
ϕ_0	1.77872	0.07585	0.00128	0.00981

Table 5.1: Posterior estimates of fit of PHSeq model to real data set

5.2 Simulation Procedure

The following procedure was used to generate each of $n = 100$ data sets.

1. Generate 4000 ASVs based on the fit of the Poisson Hurdle Model, such that there are 2080 DA ASVs and 1920 non DA ASVs. Randomly select 1000 non DA ASVs. Note that we set $\pi_{\alpha_0} = .6$ and $\pi_{\beta_0} = .8$, and s_{ij} is fixed at 1; the remaining parameters are set as the estimates from the fit of the model to the real data set, shown in Table 5.1.

2. Let n_{g_1} be the size of the DA category of interest. Let $n_{g_2} = 1000 - n_{g_1}$ be the number of non DA ASVs.

3. Split the 1000 non DA ASVs into two groups of size n_{g_1} and n_{g_2} and call these two groups g_1 and g_2 respectively. Replace some proportion (p) of ASVs in g_1 with DA ASVs from step 1. g_1 is now the category of DA ASVs to be tested.

4. Test g_1 (against g_2 where appropriate) using each of GOseq, SAFE, MRPP, and PERMANOVA, and the proposed PHSeq method.

5. Repeat steps 3-4 for values of $p \in [0, 1]$.

6. Repeat steps 2-5 for various category sizes of DA ASVs ($n_{g_1} = 500, 300, 200, 100, 50, 20, 10, 5$). Category sizes are chosen based on the real data.

By varying category size, and the proportion of DA ASVs within a category, the methods are assessed under a variety of conditions. When the proportion of DA ASVs in a category is 0, i.e., $p = 0$, the type one error rate of each method is assessed. The remaining values of p assess the methods in terms of the power.

5.3 Results

Results for each category size are shown in Figure 5.2, based on 100 simulated data sets. Figure 5.1 shows results for 200 simulated data sets for a category size of 5. Each point represents the total number of simulated data sets (out of either 100 or 200) that reject the null hypothesis of the category (p-value $<$ .05) being non DA. When all ASVs in a category are DA ($p = 1$), the category should be DA, and when no ASVs in a category are DA, the category should be non DA. The plos can then be interpreted in terms of power for $p > 0$, and error when $p = 0$.

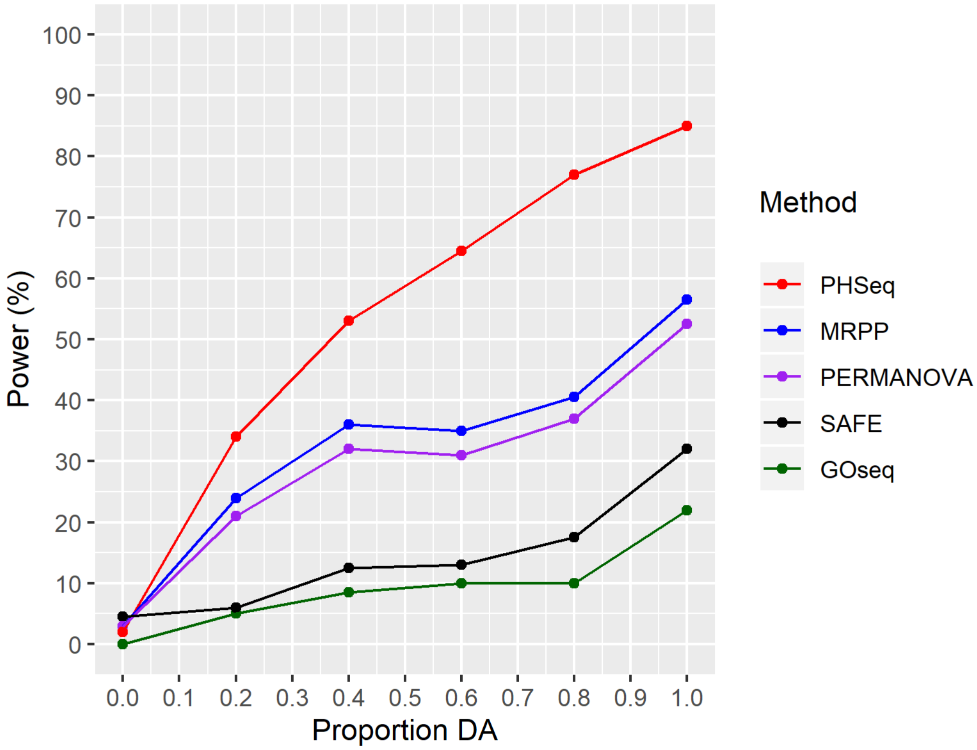
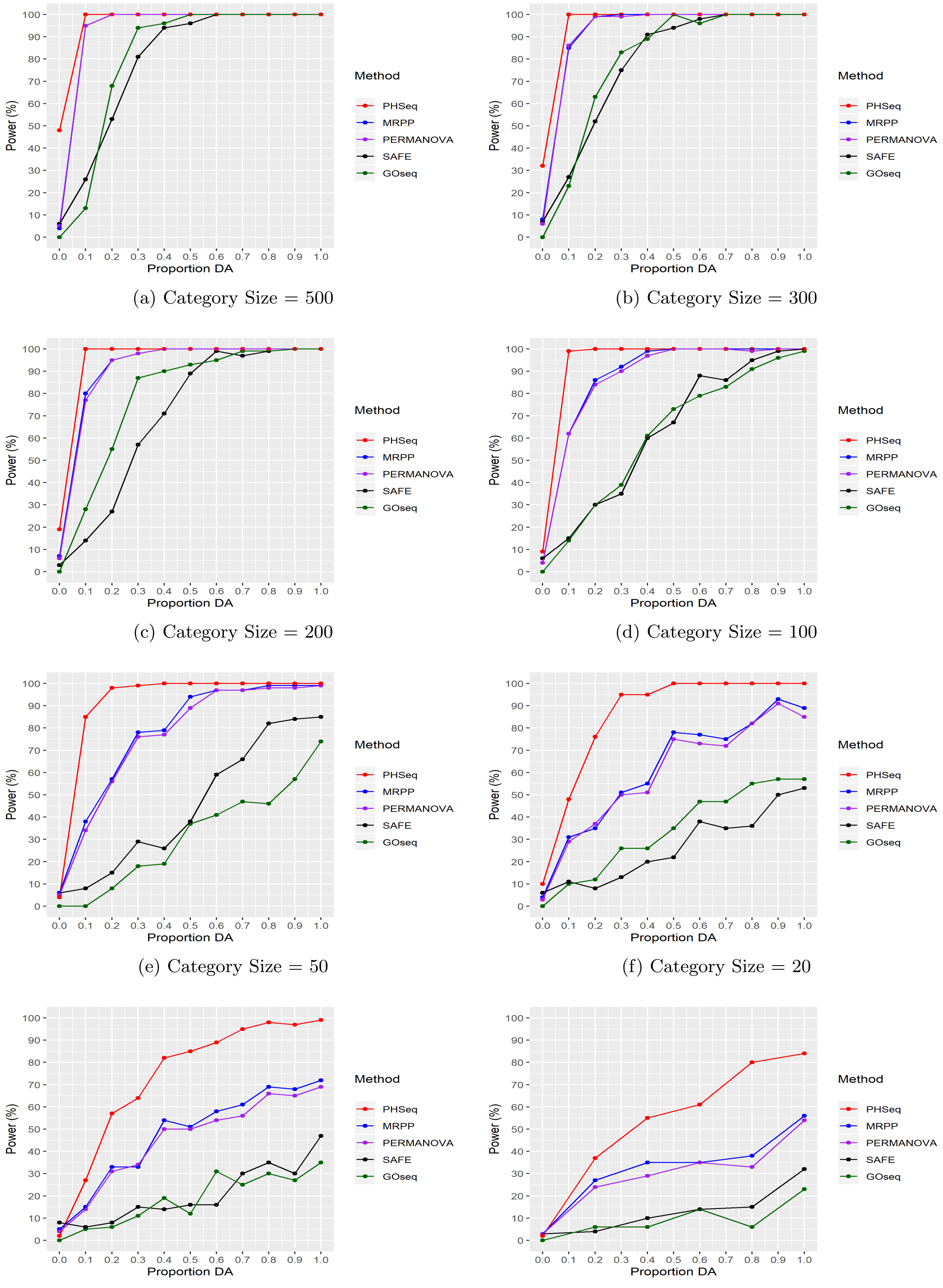


Figure 5.1: Power (%) of detecting a DA ASV category based on 5 set testing methods and 200 simulated data sets. (Category Size=5)



(a) Category Size = 500

(b) Category Size = 300

(c) Category Size = 200

(d) Category Size = 100

(e) Category Size = 50

(f) Category Size = 20

(g) Category Size = 10

(h) Category Size = 5

Figure 5.2: Power (%) of detecting categories of DA ASVs based on 5 set testing methods and 100 simulated data sets.

CHAPTER 6. DISCUSSION

With data sets generated from a Poisson Hurdle model, as in the simulation described in section 5, PHSeq outperforms all other methods. Note Figure 5.1, where the power of detecting a DA ASV category for PHSeq is over 80% when all ASVs in a category are DA; MRPP has the next highest power at only 56%. The advantage of PHSeq is amplified when category sizes are smaller with respect to power, but the error rate is poorly controlled with category sizes larger than 100. Unsurprisingly, MRPP and PERMANOVA perform almost identically, and SAFE and GOseq perform the worst.

6.1 Non-Parametric Simulation

While the results of the model based simulation suggest that the PHSeq method outperforms GOseq, SAFE, MRPP, and PERMANOVA, under certain conditions, the simulated data is generated by the PHSeq model described in section 3. To test the robustness of our method relative to other methods, a simulation study where data are not generated based on the PHSeq model is necessary. To conduct a simulation study that does not rely on a model to generate data, we use a non-parametric procedure for simulating non DE genes called SimSeq [13].

Using the 2017 data set as before, we select samples taken on July 18th, from the Rhizosphere, in Central City, of the sweet variety. Unlike before, we do not specify a genotype, which results in 109 total samples, 54 High Nitrogen and 55 Low Nitrogen. After filtering at $\frac{1}{3}$, we have 626 ASVs across all samples. In doing this, we have assumed that there is no effect of genotype so that the treatment conditions are independent. Now, we have a large data set with which resampling can be performed, the basis of SimSeq.

The following (non-parametric) procedure is used to generate 100 data sets with two treatment conditions (High Nitrogen, Low Nitrogen), with 8 replicates each:

1. Starting with a large data set of 109 samples (54 High Nitrogen, 55 Low Nitrogen) and 626 ASVs, we fit PHSeq in order to identify the following:

    - The top most DA ASVs. There are 210 ASVs with a posterior probability of being null of 0. Call this set of ASVs G_1.
    - The top most non-DA ASVs. There are 106 ASVs with a posterior probability of being null greater than .5. Call this set of ASVs G_0.

2. Category Size 5, Proportion p

    - Randomly select $5 * p$ ASVs from G_1 and $5 * (1 - p)$ ASVs from G_0. Call this collection of ASVs G_{da}.
    - Randomly select 95 ASVs from the set $\{G_0 \setminus G_{da}\}$. Call this collection of ASVs G_{nda}.
    - Define G, the set of all ASVs, as $G = G_{da} \cup G_{nda}$, so that we create a set of 100 ASVs.
    - Define a vector of weights $\boldsymbol{w} = (w_1, \ldots w_{100})$ as

    $$w_i = \begin{cases} 1/5, & \text{if ASV}_i \in G_{da} \\ 0, & \text{otherwise} \end{cases}$$

    - Simulate according to SimSeq using the set G of ASVs and the vector of weights $\boldsymbol{w}$ to create a data set with 2 treatment conditions, 8 replicates each, with 100 ASVs, $5 * p$ of which will be truly DA, but all 5 will be labeled as DA.

3. Implement all 5 set testing methods to identify if this small category is DA.

Figure 6.1 shows the results of the non-parametric simulation. All methods but PHSeq struggle to identify DA ASVs. This may be due to the fact that the DA ASVs, defined in step 1, belonging to set G_{da} can be DA due to a high proportion of zeros. However, PHSeq identifies the small tested category as DA even when no ASVs are DA in that category ($p = 0$). The results of this non-parametric simulation clearly suggest some issues with the PHSeq method with respect to error control. Further model issues and challenges will be discussed below.

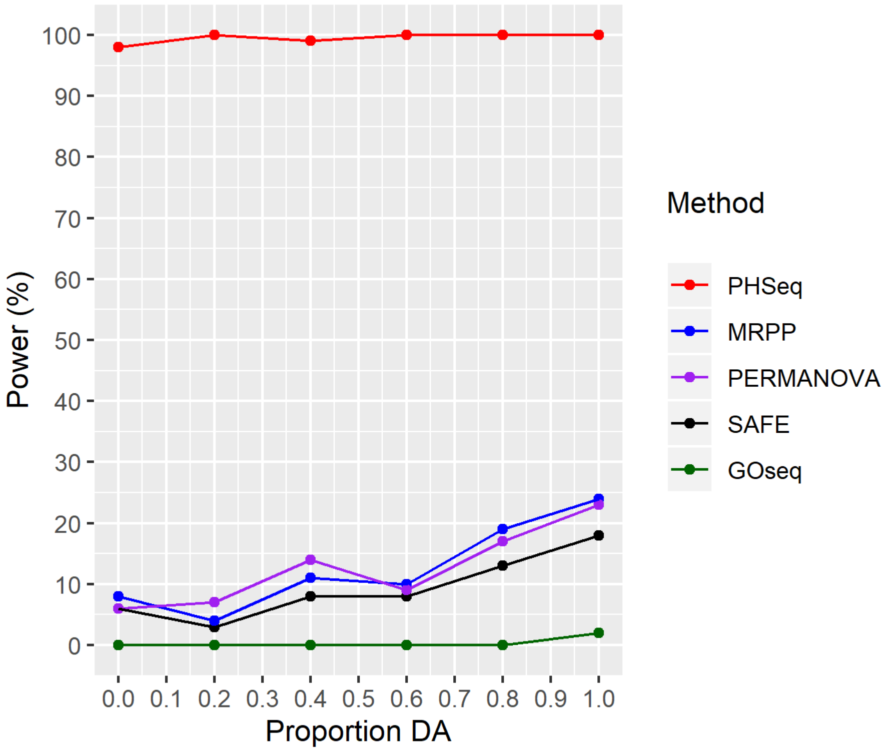


Figure 6.1: Power (%) of detecting a DA ASV category based on 5 set testing methods and 100 simulated data sets. Data are simulated based on a non-parametric approach used by SimSeq

.

6.2 Model Challenges and Future Work

The PHSeq model described in section 3 currently has poor error control. Future work must identify this issue. As of now, we have identified that the issues with the model are in the zero proportion part, so we are working with a simplified model to address such issues. Possible remedies including estimating some of the parameters from our data or changing the priors on some of our parameters. Further, our definition of a DA category is very limited; we consider a category to be DA if any ASV in a category is DA. This approach was taken due to the long computational time involved with PHSeq, so future work that decreases the time needed to run PHSeq could help to create a better model.

Additionally, different settings in existing set testing methods need to be explored. For example, GOseq can account for selection bias, but this feature was not used. SAFE has many options for local and global statistics, but only one combination was implemented. Lastly, Euclidean distances were used for both MRPP and PERMANOVA. Distances that are unique to microbiome data may be more appropriate, and more biologically meaningful.

Overall, more simulation studies are needed to compare across all methods. Specifically, more non-parametric simulation studies are needed. Such studies would allow us to directly compare PHSeq to methods in gene set testing and would provide valuable information about using existing set testing methods for microbiome data.

Bibliography

[1] Jun Sun and Pradeep K. Dudeja. *Mechanisms Underlying Host-Microbiome Interactions in Patholphysiology of Human Diseases*. Springer, 2018.

[2] Yinglin Xia, Jun Sun, and Ding-Geng Chen. *Statisical Analysis of Microbiome Data with R*. Springer, 2018.

[3] Rob Knight et. al. Best practices for analysing microbiomes. *Nature*, 16, 2018.

[4] Dan Nettleton, Justin Recknor, and James M. Reecy. Identification of differentially expressed gene categories in microarray studies using nonparametric multivariate analysis. *Bioinformatics*, 24(2):192–201, 2008.

[5] Mark D. Robinson and Davis J. McCarthy an Gordon K. Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[6] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10), 2010.

[7] Joseph N. Paulson, O. Colin Stine, Hector Corrada Bravo, and Mihai Pop. Robust methods for differential abundance analysis in marker gene surveys. *Nat. Methods*, 10(2):1200–1202, 2013.

[8] Matthew D. Young, Matthew J. Wakefield, Gordon K. Smyth, and Alicia Oshlack. Gene ontology analysis for rna-seq: accounting for selection bias. *Genome Biology*, 11, 2010.

[9] William T. Barry, Andrew B. Nobel, and Fred A. Wright. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, 21(9):1943–1949, 2005.

[10] Aravind Subramanian et. al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 102(43):15545–15550, 2005.

[11] Marti J. Anderson. A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26:32–46, 2001.

[12] Chaohui Yuan, Peng Liu, and Chong Wang. A bayesian approach based on a hierarchical poisson hurdle model for differential abundance analysis of microbiome data. *Biometrics*. Unpublished.

[13] Sam Benidt and Dan Nettleton. Simseq: a nonparametric approach to simulation of rna-sequence datasets. *Bioinformatics*, 31(13):2131–2140, 2015.