Statistical methods for microbiome data and antimicrobial resistance analysis

by

Chaohui Yuan

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Chong Wang, Co-major Professor
Peng Liu, Co-major Professor
Dan Nettleton
Daniel J. Nordman
Annette M. O'Connor

Iowa State University

Ames, Iowa

2018

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Page

ACKNOWLEDGMENTS

First and foremost I would like to express my sincerest gratitude to my advisors, Dr. Chong Wang and Dr. Peng Liu, for their generous support, insightful guidance, and constant encouragement throughout my graduate study. They were always available to listen and to give advice whenever I encountered obstacles in my research. This dissertation would not have been possible without their help and guidance.

I would also like to take this opportunity to thank my committee members for all of their time and suggestions on my work: Dr. Dan Nettleton, Dr. Dan Nordman, and Dr. Annette O'Connor.

Additionally, I would especially like to thank Dr. Annette O'Connor for funding me as a research assistant, and providing me the great opportunity to work on diverse and exciting projects. I have been fortunate to work with her. She always inspired me with her hardworking and enthusiastic attitude.

I would like to thank all of the faculty who I have had an opportunity to learn from in courses at ISU. I also want to thank everyone in the microbiome group meeting for stimulating discussions and offering me advice on my causal mediation project. Also many thanks to the people who helped me out in proofreading my thesis draft. I am also grateful to my friends who I have met in Ames. Thank you for sharing with me your knowledge. My time in Ames was enriched a large part due to you all.

Last, but certainly not least, I would like to thank my family for their long time support and encouragement in all my pursuits.

ABSTRACT

This thesis consists of three projects motivated by biological problems: (i) detecting differentially abundant taxa in multiple metagenomic samples (chapter 2), (ii) developing a two-stage causal mediation model for identifying taxa mediating the effect of environmental conditions on an outcome in the analysis of microbiome data (chapter 3), and (iii) analyzing temporal changes of the antimicrobial susceptibility (chapter 4).

Although the emerging field of metagenomics has revolutionized our understanding of the microbial world, the analysis of metagenomic data raises some statistical challenges, including modeling high-dimensional overdispersed count data with excessive zeros. In the first project (chapter 2), we propose a hypothesis testing framework based on a Poisson Hurdle hierarchical model to address the considerable zeros issue in the metagenomic data, and a full Bayesian inference is performed to identify the differentially abundant taxa among multiple treatment groups. Simulation studies demonstrate our model outperforms the existing approaches in terms of false discovery rate control at desired level of significance and statistical power as well. In the second project (chapter 3), we develop a causal mediation model to investigate the effect of a treatment on an outcome transmitted through microbes. Considering the sparsity and high-dimensional overdispersed count natures of the metagenomic data, we propose a novel screening procedure to reduce the dimension to a moderate size. Then a Bayesian variable selection strategy with a shrink and diffuse prior is used to select the key taxa with mediation effects. The performance of the proposed method is illustrated via simulation studies.

In the third project (chapter 4), we present a hierarchical Bayesian latent class mixture model to detect the temporal changes in antibiotic resistance using minimum inhibitory concentration (MIC) values. By taking the censorship into account, our proposed approach would achieve less bias in the estimation of mean MIC. We also apply this proposed method to the data from CDC

NARMS program and show that evidence of temporal changes in mean MIC exist in spite of no changes or changes of adverse direction in the proportion of resistance.

CHAPTER 1. GENERAL INTRODUCTION

In this chapter we are going to briefly introduce the background knowledge in metagenomics, antimicrobial resistance data, and several existing statistical methods discussed in this thesis.

1.1 Metagenomics

Microbes, including bacteria, fungi, protists and so on, are small microorganisms that can only be observed through a microscope. They play an essential role in all life on earth (Council et al., 2007). The collection of microbes or microorganisms forming a "mini-ecosystem" environment is referred to as a microbiome community (Sun and Dudeja, 2018). Historically, the traditional pure laboratory cultivation approaches were used to study the functions of microbes. However, these methods have some limitations. First, only about 1% of microbes can be cultured in the laboratory (Riesenfeld et al., 2004). The majority of microbes cannot be cultured and thus cannot be sequenced. Second, these traditional approaches are not efficient to study the whole microbial community since they are only able to study a few microbes each time. Instead of studying a single microbe, Handelsman et al. (1998) first proposed culturing the entire microbial community analogous to a single genome, known as metagenomics. Metagenomics is the genomic analysis, consisting of the whole collective genome of microorganisms directly extracted from their natural living environmental samples. Previous research has studied soil root microbiome of plants (review paper (Fierer, 2017)), host-associated environmental microbiome samples (such as the human gut microbiome (Turnbaugh et al., 2007)), and so on. Unlike conventional genomics approaches, metagenomics does not require a culture and a clone of each individual microbe species (Wooley and Ye, 2010). The recent rapid development of next generation sequencing (NGS) technology provides researchers the ability to determine the genomes of organisms more efficiently and economically. It can parallel sequence millions of DNA sequences and study the whole microbiome communities in a sample at the same time.

A review article by Thomas et al. (2012) provides a detail about the metagenomics sequencing technology. Here we just briefly introduce the basic steps of metagenomics. Usually, the first step of metagenomics is to extract microbial DNA from environmental samples of interest. To determine the component of the microbiome community, two sequencing approaches are widely used. One is called whole metagenomic approach by sequencing DNA fragments from all microbial genomics in the sample (Tyson et al., 2004). In contrast, the other sequencing method targets a specific marker gene from the sample, such as 16S ribosomal RNA (rRNA) gene (Tringe and Rubin, 2005). Usually 16S rRNA is chosen as it appears almost in every microbial species and its genomes contain hyper variable regions that can be used to distinguish between different species. In short, specific PCR primers corresponding to the conserved region of the 16S rRNA gene are created to amplify the genomic region of this gene. Then those amplicons are sequenced with high-throughput sequencing technology, such as Illumina MiSeq and 454 pyrosequencing, resulting in massive of DNA fragments, known as reads, which are then counted. After certain standard pre-processing procedures, including removal redundant and low-quality sequences, then similar sequences are grouped together based on a given sequence similarity threshold (usually 97% or sometimes more stringent cutoff 99% used). Those sequences within one cluster are assumed to be identical and each cluster is referred to as an Operational Taxonomic Unit (OTU) (also called as taxon/feature). The reason of clustering sequences based on similarity into OTUs rather than species is that it is impossible to ensure sequences to be chopped at the same location all the time. Then each OTU is represented by a consensus sequence and this sequence is aligned against a reference 16S rRNA bacteria database. In other words, OTU is a group of related microbes clustered based on their DNA sequence similarity of a specific taxonomic marker gene. The result is represented in a count matrix format, known as a OTU table, with rows corresponding to different types of OTUs and columns to different biological samples collected. These OTUs counts could be further summarized into higher phylogeny level, e.g., species, genus, family, etc. Figure 1.1 shows a summary of the pipeline for 16S rRNA sequence data processing.

Metagenomics provides new insights into the diversity of microbial communities in an ecosystem and enables researchers to reveal the functional features of those microbes not only over genetic and environmental areas, but also over corresponding phenotypes, such as plant yield. In agriculture, microbes are associated with almost every plant tissue: leaves, stems, root, and so on. The most studied is the microorganisms in the plant root. For example, it has been reported that rhizobia and mycorrhizal fungi have great impact on stimulating plant growth in the production of phytohormones and providing nutrients including carbohydrates and amino acids (Moe, 2013). Additionally, the plant microbiome also influences the plant secondary metabolites (Berg et al., 2015; Weston and Mathesius, 2013). Several studies have also shown that the plant microbiome is also involved in plant survival in different condition such as drought condition and nutrient-poor soils (Yang et al., 2009; Rolli et al., 2015). For example, Pseudomonas is the most frequent microbe under humid conditions (Mendes et al., 2013) and Bacillus is abundant under arid conditions (Köberl et al., 2011). In turn, the plant root also recruit certain microorganisms to assemble around root so that these microorganisms would provide better support to the growth of plant. For instance, Lundberg et al. (2012) reported that only a specific set of bacterial in the soil is colonized around the plant roots of Arabidopsis thaliana. Therefore, in order to unravel the influence of the microbes on the plant science, great effects have been made to develop new technologies to study the microbial communities.

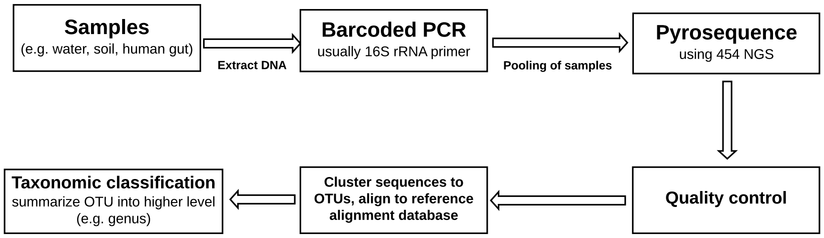

Samples
(e.g. water, soil, human gut)
Extract DNA
Barcoded PCR
usually 16S rRNA primer
Pooling of samples
Pyrosequence
using 454 NGS
Quality control
Cluster sequences to OTUs, align to reference alignment database
Taxonomic classification
summarize OTU into higher level (e.g. genus)

Figure 1.1: Illustration of the pipeline for 16S rRNA microbiome data.

From the statistical analysis prospective, there are three central themes following in the metagenomic analysis. First, describe the relative taxonomic component in the given samples, from species, genus, family to class (Clemente et al., 2011; Jiang et al., 2012). Second, identify the taxa that are over or under abundant under different certain environmental conditions (Parks and Beiko, 2010; Paulson et al., 2013). This is often referred to as differential abundance analysis. Third, assess the association (even causality if possible) between microbiome abundance and biological (or clinical) outcomes. The goal is to allow researchers have a better understanding of the microbiome community in order to potentially shape the component of the community to benefit desired biological phenotype or better clinical outcomes (Dini-Andreote and Raaijmakers, 2018; Huttenhower et al., 2012).

1.1.1 Statistical challenges for microbiome data analysis

In this thesis, we will focus on developing statistical methods specifically for 16S metagenomic data at taxonomic level (i.e. OTUs). However, there are several unique challenges for such metagenomic data.

First, metagenomic data are usually high-dimensional data, i.e. consisting of hundreds or even thousands of OTUs but with only a few biological samples as it is still expensive to conduct biological experiments. Therefore classical statistical models cannot be directly applied to such kinds of data because number of parameters needed to be estimated is much larger than the sample size. Furthermore, it is often difficult to derive the asymptotic distribution for the statistics used for inference and usually small sample size would violate the asymptotic property as well.

In addition, it has been widely reported that most OTU tables contain excessive zero counts compared to other genomics data (such as RNA sequence data), say as high as 90% (Paulson et al., 2013). This could be due to two reasons: one is that some zero values arise by random chance due to the random sampling or because of true absence of microbes in the real environment. The second reason might be due to sequencing detection errors or the amount of sequences are too small to de detected (McMurdie and Holmes, 2014; Hamady and Knight, 2009).

Third, the OTUs counts are overdispersed, i.e. the observed variance of the counts is much larger than what would be expected from a Poisson distribution, which is a common distribution used for count data. Therefore, methods should allow for overdispersion to model the OTU count data.

To illustrate those issues, Figure 1.2 exhibits a OTU count table from a study of comparison of root microbes of Sorghum under low and full nitrogen conditions from our collaborators. There are 5581 OTUs (at the 97% similarity cutoff) and 80 samples. As we can see from the OTU table that the dimension of OTU counts are extremely larger than the sample size, and contains 70.5% zeros. Figure 1.3 shows that only a few OTUs are present across samples, and the majority of OTUs are only seen in a few samples. Therefore, the distribution of OTUs are skewed. This might be one potential explanation for the excessive zeros observed in OTU table (Figure 1.2). Figure 1.4 displays the scatter plot of variances of the OTU abundances vs means of OTU abundances (both on log scale) and demonstrates that the overdispersion issue in this microbiome dataset as $R^=$0.95. Consequently, it is critical to adequately model the excessive zero counts and overdispersed for the high-dimensional microbiome data.

1.1.2 Zero-inflated regression methods

There are two common used methodologies for analyzing count data with far more zeros than expected under different distributions: the zero-inflated regression model and the hurdle regression model. There are two reasons that result in excessive zeros: one is the corresponding event is impossible or irrelevant for some case, which is referred to as structure zero (Ridout et al., 1998). For example, in metagenomic data, it is possible that some taxa are not supposed to occur under certain condition. The other reason is due to randomness that a zero event is observed by chance, which is often referred to as sampling zero. For instance, it is possible that the corresponding sequence reads are too low to be detected due to machine limitation for some taxa (McMurdie and Holmes, 2014).

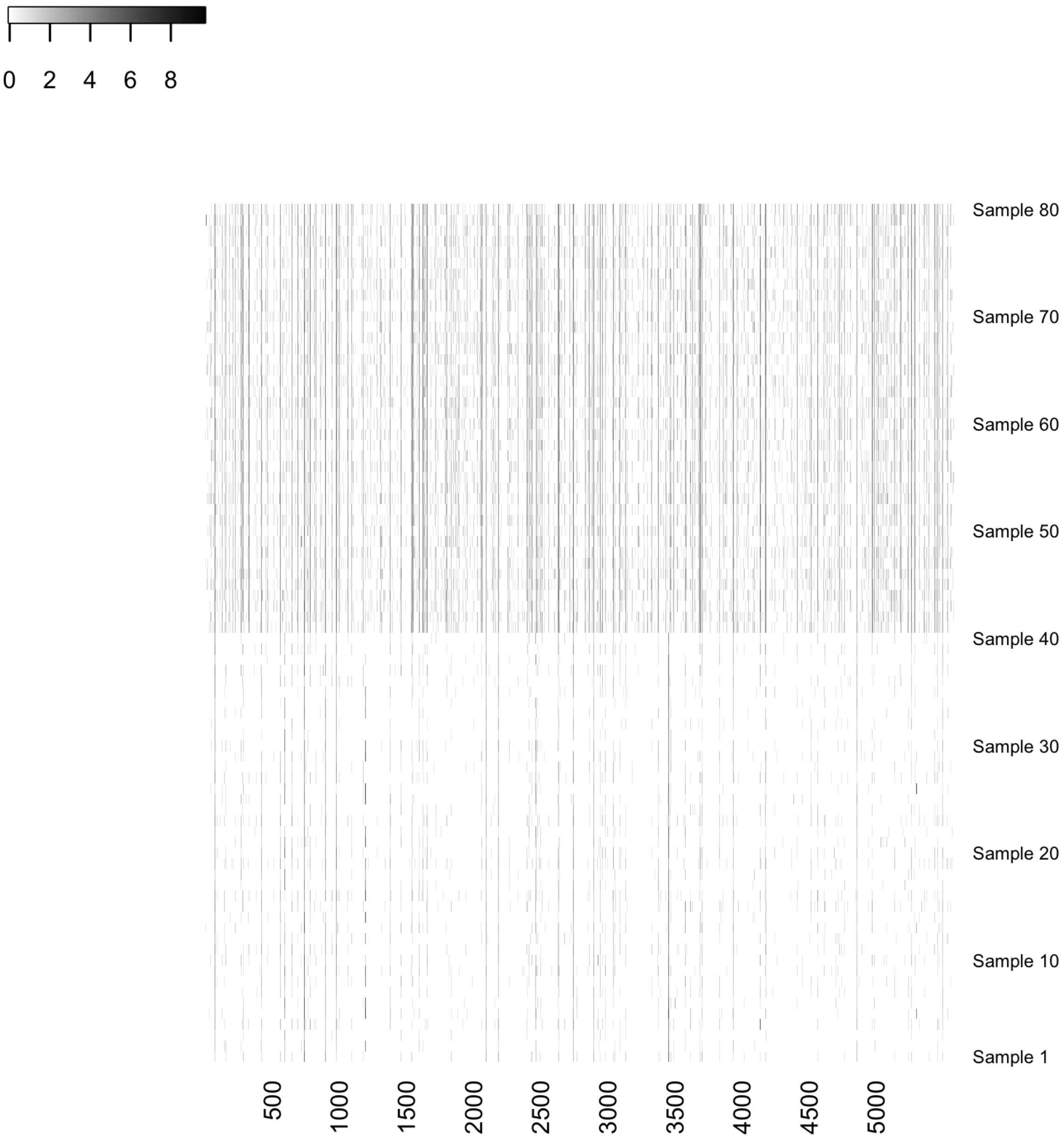


Figure 1.2: The heatmap shows the OTU counts on log scale. It takes value zero if OTU count is zero. Each row represents sample and each column denotes OTU count on log scale. There are 80 samples and 5581 OTUs.

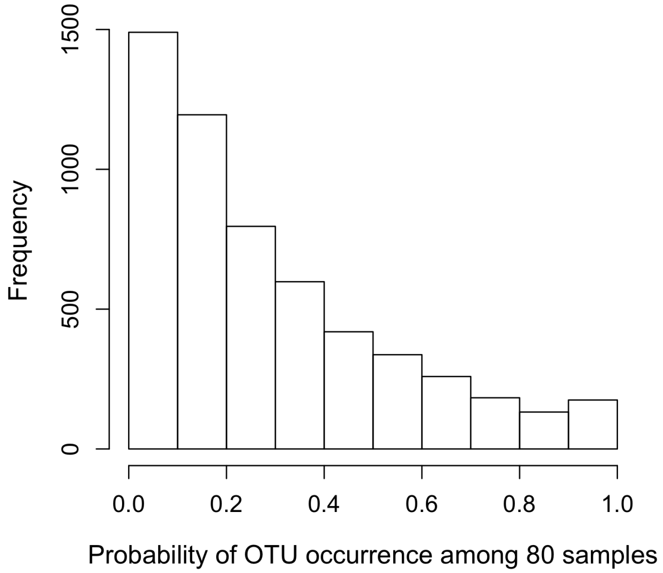


Figure 1.3: The histogram of probability of OTU occurrence across 80 samples.

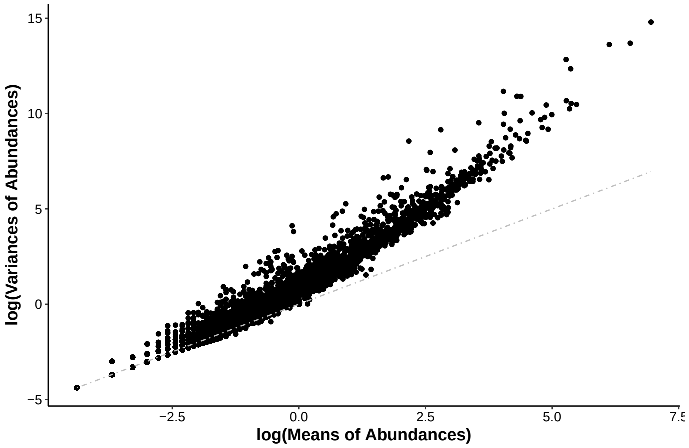


Figure 1.4: Scatter plot between variances of abundances on log scale against the means of abundances of abundances on log scale using 5581 OTUs from 80 samples. The simple linear model results the $R^2 = 0.95$.

The zero-inflated model models the "extra zeros" using a binomial process and the remaining realizations are modeled using an appropriate discrete distribution. The zero-inflated model allows that the observed high occurrence of zeros comes from in both the binomial process and the discrete distribution as well. On the other hand, the hurdle model utilizes a binomial process to model all the zero counts and a truncated-at-zero type discrete distribution is used for the positive realizations. Both the zero-inflated model and the hurdle model use the two-stage modeling processes to account for "excess zeros". In practice, the main differences in these two approaches are in the computational complexity. Besides, it is possible in reality that fewer zero counts than expected would be observed in some circumstance. When zero-deflation exists, zero-inflated type models would fail but hurdle model could still work.

In this thesis, we consider Poisson distribution to model excess zeros, but similar methodology could be extended to other discrete distributions, such as Negative binomial distribution if there is evidence in the data for over-dispersion.

1.1.2.1 Poisson Hurdle (PH) regression model

Hurdle model, also known as two-part model, models zero count and nonzero counts separately. It was proposed by Heilbron (1989). Let Y be the response variable, and $y_i, i = 1, 2, \cdots, n$ be the n independent realizations of Y. Under the assumption that the non-zero count is modeled as a Poisson distribution, the marginal probability mass function (pmf) for Y can be written as:

$$Pr(Y = y_i) = \begin{cases} 1 - p_i, & \text{if } y_i = 0, \\ p_i \frac{\mu_i^{y_i}}{[\exp(\mu_i) - 1] y_i!}, & \text{if } y_i = 1, 2, 3, \cdots \end{cases} \tag{1.1}$$

where p_i denotes the proportion of non-zero counts and μ_i represents the mean of Poisson distribution with scaling. Under the generalized linear regression framework, it is common to relate the parameter p_i to covariates of interests using logit link function and μ_i with log link, i.e.,

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = x_i^T \beta_1, \quad \log(\mu_i) = z_i^T \beta_2, \tag{1.2}$$

where covariates x_i and z_i can be the same, different, or overlapping with each other. Then the log-likelihood can be written as:

$$
\begin{aligned}
lnL_{PH} =& \sum_{i=1}^{n} \left[x_i^T\beta_1 - 2\log\left(1+\exp(x_i^T\beta_1)\right)\right] \\
&+ \sum_{i=1}^{n} I(y_i>0)\left[\log\left(\exp(\exp(z_i^T\beta_2))-1\right) - y_i z_i^T\beta_2 - \log(y_i!)\right],
\end{aligned}
\tag{1.3}
$$

where $I(\cdot)$ is the indicator function defined as:

$$
I(y>0) = \begin{cases} 1, & \text{if } y>0 \\ 0, & \text{if } y=0 \end{cases} .
\tag{1.4}
$$

Notice that in Equation (1.3), the log-likelihood function can be decomposed into two independent parts. The first part is for estimating parameters β_1 related to proportion and the second part is only related parameters β_2 in the zero-truncated Poisson mean.

1.1.2.2 Zero-Inflated Poisson (ZIP) regression model

The Zero-Inflated model models the zeros via a mixture of point mass at zero and an appropriate distribution for count event. A zero-inflated Poisson model was first proposed by Lambert (1992) with an application to detect defects in manufacturing. The marginal distribution of the ZIP model can be written as:

$$
Pr(Y=y_i) = \begin{cases} 1-\pi_i+\pi_i\exp(-\mu_i), & \text{if } y_i=0, \\ \pi_i\frac{\mu_i^{y_i}\exp(-\mu_i)}{y_i!}, & \text{if } y_i = 1,2,3\cdots, \end{cases}
\tag{1.5}
$$

Similar to the Hurdle model, under the same generalized linear regression framework, the log-likelihood can be written as:

$$
\begin{aligned}
lnL_{ZIP} =& \sum_{i=1}^{n} I(y_i=0)\left\{\log\left((1-p_i)+p_i\exp(-\mu_i)\right)\right\} \\
&+ \sum_{i=1}^{n} I(y_i>0)\left\{\log(p_i)+y_i\log(\mu_i)-\mu_i-\log(y_i!)\right\},
\end{aligned}
\tag{1.6}
$$

where $p_i = \exp(x_i^T\beta_1)/(1+\exp(x_i^T\beta_1))$ and $\mu_i = \exp(z_i^T\beta_2)$. All the parameters β_1 and β_2 are not separable in the log-likelihood function for Zero-Inflated Poisson in Equation (1.6), which increases the computational complexity compared to the separable property in the log-likelihood function in Equation (1.3).

1.1.3 Bayesian false discovery rate

In order to determine whether an observed outcome occurs by random chance alone, a single hypothesis or multiple hypotheses are usually conducted for statistical inference. A discovery is obtained by reject the null hypothesis and thus it is in favor of the alternative. Usually there are two error criteria used as guidelines in hypothesis testing. One is the Type I error when the null hypothesis is true but was falsely rejected based on the decision. The second is the Type II error occurring if a false null hypothesis is accepted. In practice, it is common that several hypotheses are tested simultaneously. For instance in genomic data analysis, usually researchers conduct a hypothesis testing for each gene at the same time. Each of the individual test would potentially have Type I and Type II errors. Therefore it is necessary to find a threshold such that the associated testing procedure is able to identify as many true discoveries as possible (i.e. maximum power), while maintaining a relatively lower number of false discoveries (i.e. a reasonable error rate bound). This is the basic rationale of False Discovery Rate (FDR), which refers to as the proportion of false discoveries relative to the total number of discoveries. False discovery rate has been widely used in high-dimensional genomics data analysis.

The concept of FDR was first introduced by Benjamini and Hochberg and they provided a procedure to calculate the adjusted p values. This procedure is often called BH procedure (Benjamini and Hochberg, 1995). Keeping the same notation of Benjamini and Hochberg (1995) in this section, we consider testing m independent null hypotheses simultaneously, of which m_0 are true. The testing result is summarized in Table 1.1. R is the observed total number of hypotheses being rejected, which is a random variable. T is the number of Type II errors, where hypotheses actually are from alternative but are classified as not significant. V is the number of Type I errors, where hypotheses are classified as significant when they are actually from the null hypotheses. All V, U, S and T are unobservable random variables.

Table 1.1: Result of m null hypotheses tests

	Declared non-significant	Declared significant	Total
True null hypotheses	U	V	m_0
Non-true null hypotheses	T	S	$m - m_0$
Total	$m - R$	R	m

The FDR is defined as the expectation of the false rejection rate expressed in Equation (1.7)

$$\text{FDR} = E\left(\frac{V}{R}|R > 0\right) Pr(R > 0). \tag{1.7}$$

The BH procedure works as follows:

1. Compute the p values $p_1, p_2, \cdots, p_m$ for each hypothesis test.

2. Order the p values increasingly $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$.

3. Find value k such that $k = \max\left\{k : p_{(k)} \leq \alpha\frac{k}{m}\right\}$.

4. If k exists, then reject all null hypotheses $H_{0i}, i = 1, 2, \cdots, k$; otherwise no hypotheses are rejected.

Benjamini and Hochberg (1995) proved the above BH procedure would control the FDR at a more stringent level $m_0\alpha/m$ for independent tests.

Newton et al. (2004) also proposed a Bayesian FDR approach (BFDR), which utilizes the posterior sampling distributions of the null hypothesis and the alternative hypothesis statistics. It has been widely used in the Bayesian inference to assess corresponding statistical significance. The BFDR is evaluated at α level through

$$\widehat{\text{BFDR}}(\alpha) = \frac{\sum_{i=1}^{m}(1 - v_i)\delta_i}{\sum_{i=1}^{m}\delta_i}, \tag{1.8}$$

where v_i denotes the posterior probability of the ith alternative hypothesis test. Then $1 - v_i$ represents the posterior probability of the ith null hypothesis test. $\delta_i = I(v_i > c)$ is the decision rule, i.e., $\delta_i = 1$ means the ith hypothesis is rejected and it takes value zero for the ith null hypothesis fail to be rejected at cutoff c. Thus $D = \sum_{i=1}^{m}\delta_i$ is the total number of rejections. Therefore, the

BFDR could be interpreted as the posterior proportion of false positives among the list that are classified into alternative groups. Thus it is straightforward to assess the error rate based on the posterior distribution of statistics for the null hypothesis (or for the alternative hypothesis) in the Bayesian framework. This brings the popularity for the BFDR, especially Bayesian methods are getting more and more attractive to deal with the complex data structure nowadays.

1.1.4 Causal mediation analysis

Establishing a causal relationship is one of the central aims in scientific research. In metagenomics analysis, a causal inference is desired for identifying mediating factors (i.e. the microbes) accounting for treatment effects on outcome (Waldron, 2018; Li, 2015; Xia and Sun, 2017). Causal mediation analysis has been widely used to assess the effect of the treatment on the outcome mediated through some possible intermediate pathway. The identification of mediators is important because this will, to some extent, contribute to improve the treatment effect by focusing on relatively important mediators. Therefore, mediation analysis in the metagenomics studies is an ongoing area and will continue to grow.

Most research focuses on establishing or testing how variable T, representing treatment variable, impact outcome variable Y. For simplicity, we only consider outcome variable is an univariate case in this thesis. A mediator, also known as a intervening variable or an intermediary variable, is conceptualized as the mechanism on the pathway between the treatment T and the outcome Y (i.e. the pathway $T \to M \to Y$ in Figure 1.5). In other words, the treatment T influences the mediator M, which in turn causes the variation in the outcome Y. However, this does not indicate that the entire pathway of treatment effect T on outcome Y has to go through mediator M, as part of the effect could be direct from treatment T to outcome variable Y, such as the pathway $T \to Y$ in Figure 1.5. Typically, researchers are interested in how much the treatment effect on outcome through the mediator M. Other than just one single mediator, there might be multiple mediators between the treatment and the outcome, or a sequence of mediators (VanderWeele and Vansteelandt, 2014).

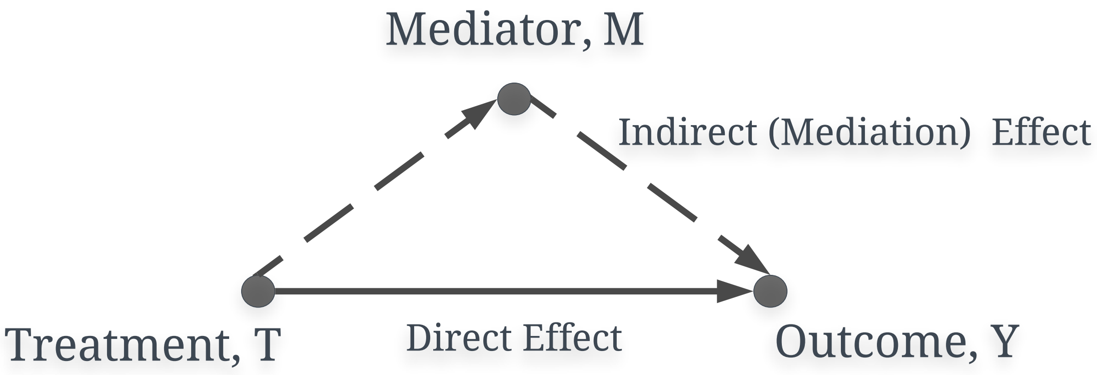


Figure 1.5: A simple illustration of the three variables path diagram of the standard causal mediation model framework. The independent variable T presumably affects the dependent variable Y, by having a direct effect ($T \rightarrow Y$) and an indirect effect mediated by mediator M ($T \rightarrow M \rightarrow Y$). All three variables, T, Y, and M are scalars.

Mediation analysis was first proposed by Baron and Kenny (1986). Under the regression framework, the basic schematic used for a mediation model can be written as:

$$Y = cT + \epsilon_1, \tag{1.9}$$

$$M = aT + \epsilon_2, \tag{1.10}$$

$$Y = bM + c^{'}T + \epsilon_3, \tag{1.11}$$

where $E(\epsilon_1) = E(\epsilon_2) = E(\epsilon_3) = 0$. The parameter c in Equation (1.9) denotes the total effect on treatment T on Y. The parameter a denotes the association of T on M in Equation (1.10). Under this case, it is easy to show that:

$$ab + c' = c, \tag{1.12}$$

holds, meaning the total effect can be decomposed into direct effect and indirect effect. Thus, under the assumption that the above mediation model is correctly specified, then the parameters can be estimated by Ordinary Least Squares (OLS). Baron and Kenny (1986) and Judd and Kenny (1981) proposed three steps using the multiple regression framework to establish causal mediation effect:

1. Show that the treatment is significantly related to the outcome, i.e., estimate and test parameter c in Equation (1.9). This would suggest that there is an effect might be mediated.

2. Show that the treatment is significantly associated with the mediator, i.e., estimate and test parameter a in Equation (1.10).

3. The mediator is significantly related to the outcome, controlling for treatment, i.e., estimate and test parameter b in Equation (1.11). This is different from just testing correlation between mediator M and outcome Y. This is because that the mediator M and outcome Y might be correlated due to the fact that they are both related to treatment T.

Then the mediation effect could be assessed by testing whether the product ab (i.e. indirect effect) is zero.

Later on, there are a lot of methodological developments about the mediation analysis due to its ubiquitous nature of causal mechanisms interpretation (Krull and MacKinnon, 1999; MacKinnon et al., 2007b,a; VanderWeele and Vansteelandt, 2014; Judd and Kenny, 1981). One of the most important developments is based on the potential outcomes framework (Imai et al., 2010a; Robins and Greenland, 1992).

1.1.5 Bayesian variable selection

With the increasing availability of massive and high-dimensional datasets, extracting valuable knowledge from such challenging data has become a fundamental research concern. This leads to the intensive focus on variable selection methodology, selecting a parsimony subset of observed covariates (also known as features) which provides insights on the contributes to the observed phenomenon. The linear regression model is perhaps the most well-known statistical tool used to evaluate the relationship between the independent variables $\boldsymbol{X}$ and the continuous outcome of interest Y. A typical multiple linear regression model for the ith sample unit has the form as following:

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad i = 1, 2, \cdots, n, \tag{1.13}$$

where y_i is the ith observed response, $\boldsymbol{x}_i$ is the covariates vector associated of the ith sample with size p, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ is the unknown (and fixed) regression coefficient vector, and $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)^T$ is a random error with the assumption that uncorrelated error terms (i.e. $\mathrm{Cov}(\epsilon_j, \epsilon_k) = 0\ \forall j \neq k$) with $E(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2 > 0$ for all observations. Often, the error terms are also assumed to be normal distributed in order to make statistical inference. When $n > p$, i.e., the total number of observations is larger than the total number of parameters needed to be estimated, ordinary least squares (OLS) is a common approach of estimating $\boldsymbol{\beta}$ by minimizing the residual sums of squares (RSS) (i.e. $\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2$). However, when p is larger than the number of observations n, or some of the covariates are highly correlated, the OLS estimates are unsatisfactory with high variance and inflated standard errors. However, it is very common to have the $p > n$ case in genomics analysis. Thus variable selection is necessary to identify those important covariates that should be included in the final model and then estimate the corresponding effects efficiently as well. To address this concern, a lot of methods have been proposed, including least absolute shrinkage and selection operator (LASSO) regression (Tibshirani, 1996), the elastic net (Zou and Hastie, 2005), smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), adaptive LASSO (Zou, 2006), the group Lasso (Yuan and Lin, 2006), etc. To assess the uncertainty of the regression coefficients and to obtain the corresponding confidence interval or the p-value for the hypothesis testing, such as bootstrap (Efron, 1982), valid post-selection inference (Berk et al., 2013), or covariance test statistics (Lockhart et al., 2014).

Unlike frequentist methods aforementioned looking for a single optimal model, a Bayesian approach is capable of not only obtaining the point estimations of coefficients, but the entire joint posterior probability of all parameters under consideration. This makes the Bayesian inference simple and nature. For the variable selection, one common used Bayesian prior is known as the spike and slab priors. The basic underlying idea of these priors is to specific two mixing components such that the first component shrinks the coefficient towards zero with certain probability and the other component allows for nonzero values. Thus the sparsity of the model would be determined by this prior. The original spike and slab prior formulation was proposed by Lempers Lempers (1971), but the term was proposed by Mitchell and Beauchamp (1988). These priors include a point-mass mixture prior, i.e. a point mass at zero and a continuous prior elsewhere for each regression coefficient β_j in Equation 1.13, written as:

$$\beta_j \sim (1 - Z_j)\delta_0 + Z_j f_j, \quad j = 1, 2, \cdots, p, \tag{1.14}$$

where Z_j is a binary random variable implying whether β_j is zero or not, δ_0 is a Dirac delta measure with mass on zero (i.e. the "spike" part), and f_j is a continuous density (i.e. the "slab" part). Although the intuitive idea of point mass prior on the spike component is clear, it causes computational difficulties to get the MCMC chains converge. Recently, Narisetty et al. (2014) adapted the spike and slab Gaussian priors with prior variances depend on sample size, and proved the strong selection consistency of their proposed priors under certain regularity conditions. More specifically, the so-called shrinking and diffusing priors can be expressed as follows:

$$\begin{aligned}
&\beta_j|\sigma^2, Z_j = 0 \sim N(0, \sigma^2\tau_{0,n}^2),\\
&\beta_j|\sigma^2, Z_j = 1 \sim N(0, \sigma^2\tau_{1,n}^2),\\
&P(Z_j = 1) = 1 - P(Z_j = 0) = q_n,\\
&\sigma^2 \sim \text{IG}(a, b),
\end{aligned} \tag{1.15}$$

where a, b are constants, and $\tau_{0,n}^2, \tau_{1,n}^2, q_n$ are constants depend on sample size n. $\text{IG}(a, b)$ represents the inverse-gamma distribution with shape parameter a and rate parameter b.

1.2 Antimicrobial Resistance

In the past few decades, it has been reported that the number of antimicrobials that are effective in treating infections are decreasing (Spellberg et al., 2008). Therefore, the emerging crisis of antibiotic resistance is a serious threat for public health. In the United States, it is estimated that more than \$20 billion is spent on antimicrobial resistance (AMR) each year (Marston et al., 2016). *Salmonella* causes about 1.2 million illnesses, 23,000 hospitalizations, and 450 deaths in the United States every year (Scallan et al., 2011). Food of animal origin, such as beef and poultry, is the main source for *Salmonella* infections (Medalla et al., 2017).

1.2.1 Antimicrobial resistance data

Typically, a dilution test is a common approach used to collect AMR data (Caprioli et al., 2000), such as *Salmonella* spp. . This test results in a Minimum Inhibitory Concentration (MIC) value (milligram per milliliter (mg/ml)). The minimum inhibitory concentration is defined as the lowest concentration of an antimicrobial that prevent the visible growth of a microorganism. A MIC value depends on both the microorganism and the antibiotic of interest. The minimum inhibitory concentration of a particular antibiotic is reported between the concentration of the last well where no bacteria grew and the next lower concentration where bacterial was still observed. For example, the same amount of bacteria doses are cultured in the arrays of wells consisting of concentrations of the antimicrobial agent. Each well increases the percent concentration of antimicrobial by 2 (i.e. 2 mg/ml, 4 mg/ml, 8 mg/ml, 16 mg/ml, and 32 mg/ml), such that the smallest concentration in this experiment is 2 mg/ml and the largest one is 32 mg/ml. If no visual bacteria growth is in the first well (2 mg/ml), then the reported MIC value would be ≤ 2 mg/ml, which will be considered as a left-censored event in our analysis. If some bacteria are observed at the concentrations of 2 and 4 mg/ml, but are inhibited of growth at 8 mg/ml. Then the reported MIC value in this case would equal 8 mg/ml. We will denote this case as an interval-censored event since the underlying true inhibition of growth actually occurs at the concentration between 4 mg/ml and 8 mg/ml. Similarly, if bacteria is still observed at 32 mg/ml, then the true MIC value will treated as right-censored in our analysis, and the MIC is reported as > 32 mg/ml.

1.2.2 Censored data

Statistically, such serial dilution experiment can generate censored data (i.e. left, right or interval censored). That is, the left censored data means the event of interest has already occurred before recording. For the right censored data, we only know that the event of interest occurs after the censoring time. For the interval-censored, the event occurs between two time points.

Censored data leads to information loss since we do not know the exact value of the event occurring. Take the right-censored MIC value aforementioned as an example. We only know the

MIC value should be larger than 32 mg/ml for that particular bacteria and antimicrobial agent, but we do not know the exact value. Therefore, one of the easiest approaches would be to impute the censored data, such as delete the reported censored observation and substituted with a reasonable value, assuming the observations are independent with each other. Then further statistical analysis can be adopted using this "imputed dataset". However, this deletion and substitution method leads to the biased parameters estimation and therefore are not recommended (Gilbert, 1987). Alternatively, a model based approach using likelihood function is often used to analyze censored data, although it requires an assumption of the distribution for the response. In this approach, it is assumed that the censoring mechanism is independent with the underlying data generating mechanism. Let y_i denotes the observation with probability distribution function (or probability mass function) $f(y_i;\theta)$ for $i = 1, 2, \cdots, n$, and the corresponding cumulative distribution function is $F(\cdot)$. Following the common notations in the censoring, for each subject i, a pair of (y_i, δ_i) is observed, where y_i is the observed value and δ_i is used to indicate whether the observation is censored or not. Then the likelihoods for different type of censoring can be expressed as follows:

1. Left censoring at value c: $L(\theta; y) = \prod_{i=1}^{n} f(y;\theta)^{\delta_i} F(c)^{1-\delta_i}$, where δ_i takes value 1 if y_i is observed otherwise it takes value 0.

2. Right censoring at c: $L(\theta; y) = \prod_{i=1}^{n} f(y;\theta)^{\delta_i} (1 - F(c))^{1-\delta_i}$, where δ_i takes value 1 if y_i is observed otherwise it takes value 0.

3. Interval censoring between c_1 and c_2 $L(\theta; y) = \prod_{i=1}^{n} f(y;\theta)^{\delta_i} (F(c_2) - F(c_1))^{1-\delta_i}$, where δ_i takes value 1 if y_i is observed otherwise it takes value 0.

1.3 Dissertation Organization

The rest of the dissertation is organized as follows. Chapter 2 presents a statistical model to detect the differentially abundant taxa among different metagenomic samples. Chapter 3 develops a two-stage causal mediation model for identifying taxa mediating the treatment effect on the outcome. Chapter 4 provides a mixture model to monitor the trend of antimicrobial susceptibility using MIC values.

CHAPTER 2. A BAYESIAN HIERARCHICAL POISSON HURDLE MODEL FOR DIFFERENTIAL ABUNDANCE ANALYSIS OF MICROBIOME DATA

A microbiome refers to the collection of all microorganisms in an environment, and the recent advent of high-throughput sequencing technologies have dramatically advanced our ability to study microbiomes with unprecedented resolution. One commonly used technology is called amplicon sequencing, which results in sequencing fragments that are clustered into Operational Taxonomic Units (OTUs). A fundamental step in the analysis of OTU count data is differential abundance analysis, i.e., to detect the OTUs whose abundances change across treatments and conditions. There are several challenges in differential abundance analysis of microbiome data. First, there are often considerable number of zeros in each sample and in most OTUs. Second, microbiome data is one of those "small n, large p" cases, where the dimension of variables is high while the number of replicates is small. Besides, the distribution of OTU counts can be highly skewed due to a few extremely abundant OTUs in some samples. In this manuscript, we propose use of Poisson Hurdle model to deal with excessive zeros and we do so in a hierarchical model framework to borrow information across OTUs. We develop a fully Bayesian approach for differential abundance analysis while controlling false discovery rate. Comprehensive simulation studies demonstrate that our proposed method outperforms other existing methods in terms of statistical power and false discovery rate control. We also apply our method to two plant root microbiome studies.

2.1 Introduction

Microbes, which are small microorganisms that can only be observed through a microscope, play an essential role in all life on earth. The collection of microbes or microorganisms forming a "mini-ecosystem" environment is referred to as a microbiome (or microbial community) (Sun and

Dudeja, 2018). The advance of sequencing technologies has greatly facilitated studies of microbiomes in the past decade. Nowadays, different microbes from a given sample can be measured simultaneously using next-generation sequencing technologies by either whole-metagenome shotgun (WMS) sequencing or amplicon sequencing. With WMS sequencing, millions of random DNA fragments from the whole microbial community are sequenced in parallel while the amplicon sequencing technology targets specific microbial amplicons, predominantly the bacterial 16S ribosomal RNA (16S rRNA) gene because it appears almost in every bacterial species and there are hyper variable regions that can be used to differentiate different species (Wooley and Ye, 2010; Morgan and Huttenhower, 2012; Li, 2015). With both WMS sequencing and amplicon sequencing, we can obtain taxonomic profiling of the microbiota under study. Taking amplicon sequencing as an example, the sequenced DNA fragments (i.e. reads) are clustered into Operational Taxonomic Units (OTUs) based on a given sequence similarity cutoff (usually 95%, 97% or 99%). Each OTU is represented by a consensus sequence and this sequence is aligned against a reference 16S rRNA bacterial database. Thousands of OTUs are typically observed in each sample, and data from all samples can be organized in a count data table. Marker gene analysis of WMS sequencing data results in similar count data table. In order to understand how microbes work across treatments/conditions, one fundamental step of data analysis is to identify those OTUs whose abundance levels differ across treatments/conditions. Such analysis is usually referred to as differential abundance analysis in the field of metagenomics (Paulson et al., 2013; Li, 2015).

There are several challenges in differential abundance analysis of microbiome data. First, there are often considerable number of zeros in each sample. The observed zero counts might be due to non-existence of certain microbes in a given sample or as a result of detection limit because not all microbes existing in a sample are guaranteed to be observed using the current technology (Wooley and Ye, 2010; McMurdie and Holmes, 2014). Second, microbiome data is one of those "small n, large p" cases, where n refers to the number of biological replicates and p refers to the number of OTUs measured in each sample. Due to the high cost of metagenomics experiments, only a few biological replicates can be afforded while thousands of OTUs are measured simultaneously for each sample. Using a small number of biological replicates tends to result in unstable parameter estimation, and statistical methods that rely on asymptotic properties may not work well. Besides, distribution of OTU counts could be highly skewed due to a few extremely abundant OTUs in some samples. Taking all these challenges into consideration, robust statistical methods are needed to draw reliable inferences from differential abundance analysis.

Several statistical methods have been developed for differential abundance analysis of microbiome data. Rodriguez-Brito et al. (2006) proposed a permutation-based approach that estimates the median difference between two samples and obtains the associated p-value through bootstrapping. Other non-parametric methods, such as Wilcoxon-Mann-Whitney test for assessing differential abundance between two groups or Kruskal-Wallis test for multiple-group comparison, were also applied to the microbiome data analysis (White et al., 2009; Parks and Beiko, 2010; Segata et al., 2011; Parks et al., 2014). However, those non-parametric or permutation-based methods have a number of limitations. One major drawback is that small sample size results in low power and granular distribution of p-values, and the latter introduces problems for the control of multiple testing error. Another problem is that they do not take into account the excessive zeros in microbiome data. Methods within the generalized linear model framework without considering excessive zeros were also applied for functional comparison of metagenomes, such as ShotgunFunctionalizeR (Kristiansson et al., 2009). More recently, zero-inflated model based methods have been proposed to deal with the challenge of excessive zeros in microbiome data. MetagenomeSeq is an R package that applies a zero-inflated Gaussian method to address the issue of extra zeros (Paulson et al., 2013). Sohn et al. (2015) considered a ratio-based approach for identifying differential abundant (RAIDA) taxa, which was based on a modified zero-inflated log-normal model. Those studies, however, suffer from the fact that they proposed continuous distributions to model count data, which are discrete in nature. These methods thus require data transformation, say log-transformation on the count data in OTU table, where zero counts need to be replaced by arbitrary values such as ones to avoid logarithm of zero. Such arbitrary modifications do not match the natural variation in data. Furthermore, the RAIDA method might not be appropriate for comparison of environmentally different samples since the majority of taxa were assumed not differentially abundant (Sohn et al., 2015). For metagenomeSeq, McMurdie and Holmes (2014) argued that it tends to produce high false positive rate. Chen and Li (2016) provided a two-part mixed-effect Beta regression model by transforming read counts into compositional data. A potential problem for compositional data, which is obtained by normalizing the abundances of all microbes in the sample by dividing by the total read counts, is that the read sequence count is not statistically stable for skewed data. Instead of working with transformed data, more recent attention has focused on studying the OTU count data directly. Zhang et al. (2017) used a negative binomial mixed model to account for the correlation among samples for each OTU without considering the extra zero nature of the microbiome data. Chen et al. (2017) developed a zero-inflated negative binomial model for each OTU separately and considered covariates of interest in modeling both negative binomial mean and zero-inflation part. Existing studies considering excessive zeros have only focused on fitting a zero-inflated or two-part model for each OTU. However, there are also considerable number of OTUs that have non-zero counts across all samples, and forcing a zero-component for such OTUs results in unstable numerical results since the parameter related to the proportion of zero counts lies on the boundary of the parameter space.

We use a real dataset to illustrate the characteristic of excessive zeros in microbiome data. One of the studies presented by Lundberg et al. (2012) investigates the relationship between the root rhizosphere compartment microbiota and plant genotypes. There are a total of 44 samples and 18774 OTUs for this study. Figure 2.1 presents a histogram of proportion of zero counts in each OTU. In Figure 2.1, apparently numerous number of OTUs contain excessive zeros (more than 75% OTUs have zero counts in at least 90% of 44 samples). Additionally, there is great variation in the proportion of zero counts among OTUs, with about 1% OTUs having no zero counts at all. Figure 2.2 presents count frequencies of four specific OTUs with zero-inflation detected to be significant based on likelihood ratio tests between zero-inflated and non-inflated models. Besides zero-inflation, Figure 2.2 shows that the proportions of zeros vary between different OTUs and also vary among different genotypes for the same OTU.

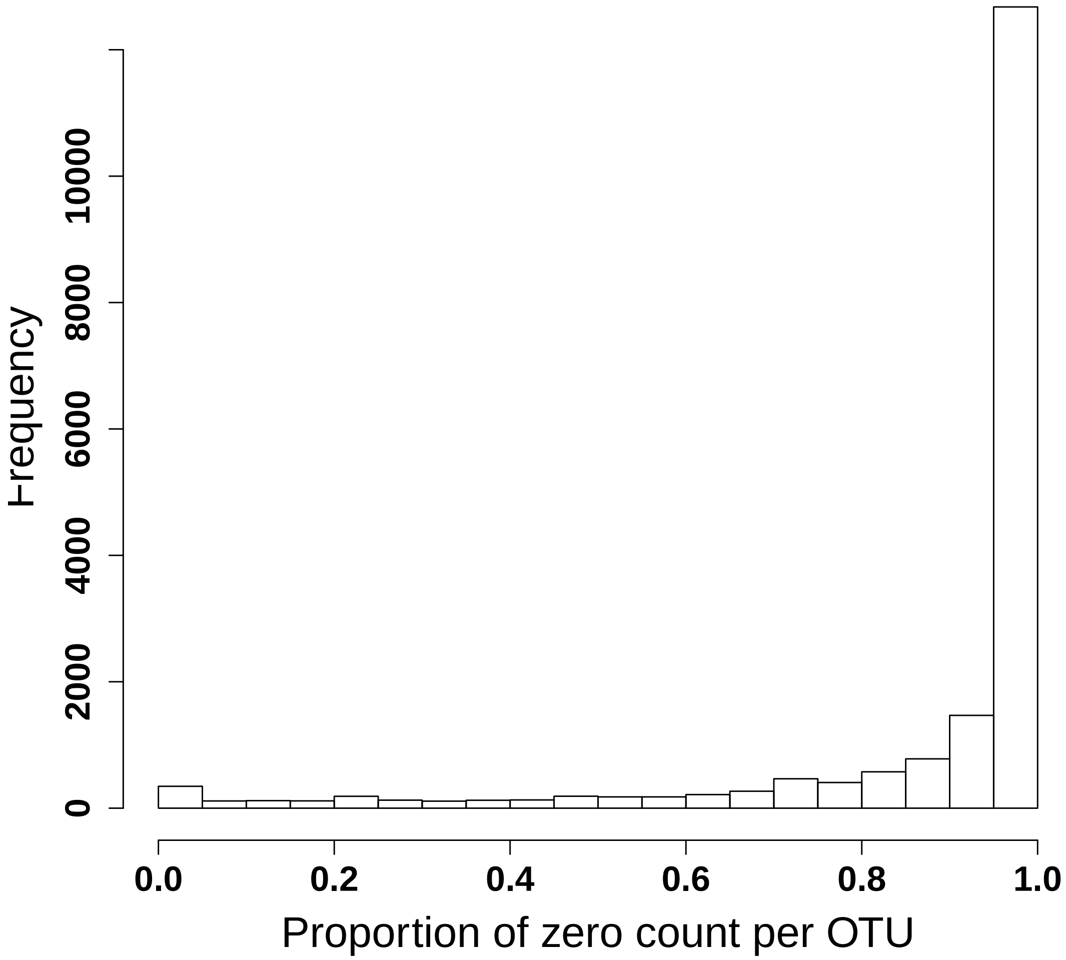


Figure 2.1: Histogram of the proportion of zero counts in each OTU in the *Arabidopsis* root rhizosphere microbiome dataset.

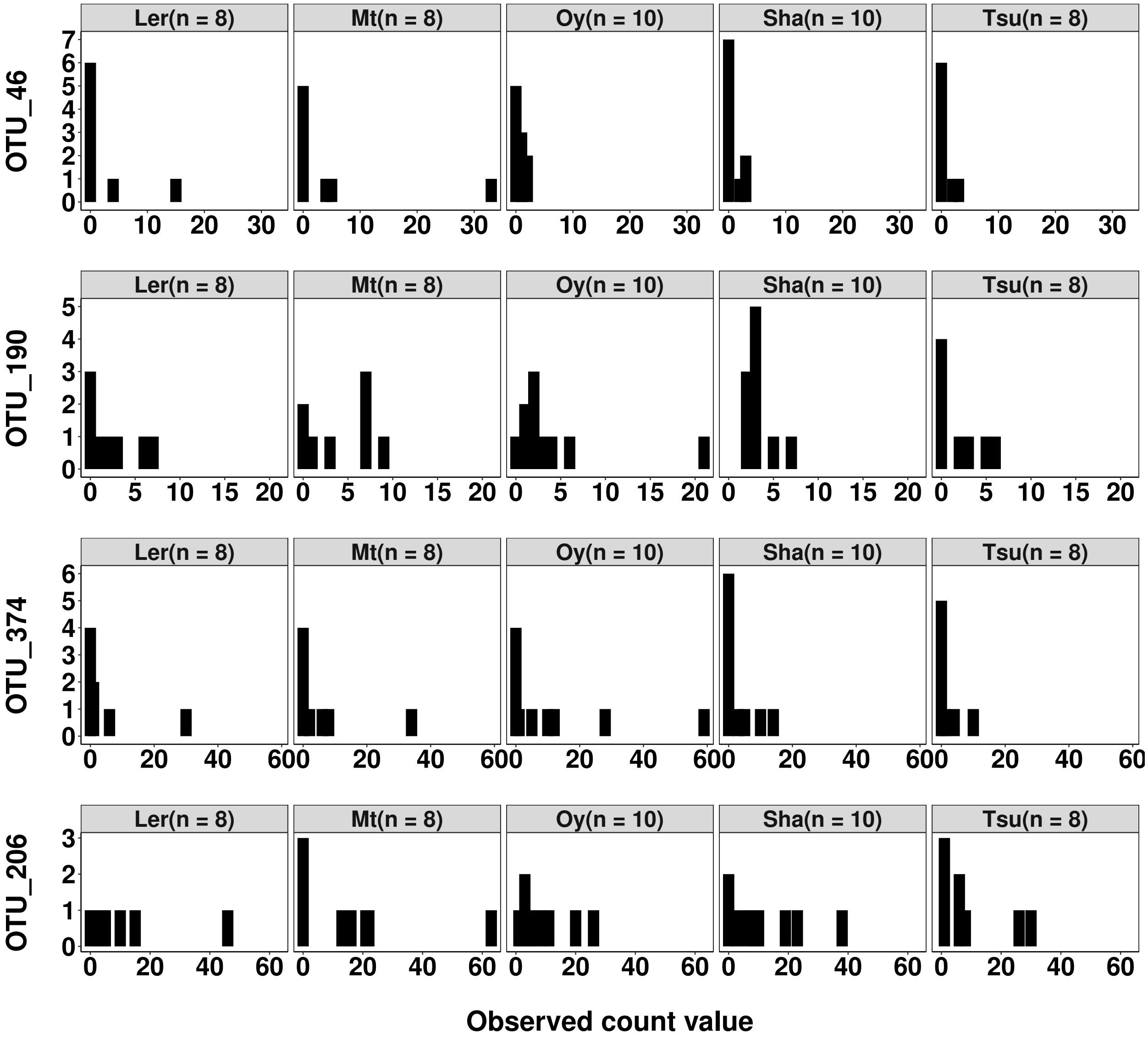


Figure 2.2: Count frequency distribution of four specific OTUs in the *Arabidopsis* root rhizosphere microbiome dataset. Each row corresponds to an OTU and each column corresponds to a genotype (Ler, Mt, Oy, Sha, and Tsu). The value n in each figure title is the number of biological replicates in the corresponding genotype group. The horizontal axis corresponds to the observed count and the vertical axis corresponds to frequency.

Two types of statistical models have commonly been applied to deal with count data with extra zeros: zero-inflated models and hurdle models (also known as the two part models) (Hilbe, 2011). Mathematically it can be shown that zero-inflated models are special cases of hurdle models; yet hurdle models are more general as they may be used to specify zero-inflation as well as zero-deflation. In addition, it has been shown that estimates based on hurdle models tend to be more computationally stable especially in the case of small amount of zeros (Xu et al., 2015). Hence, in this paper, we propose to use Poisson Hurdle models to deal with excessive zeros in the framework of hierarchical models. Hierarchical models have been widely applied for "-omics" data analysis such as transcriptomic data generated by RNA-sequencing experiments (Newton et al., 2004; Do et al., 2005). Such models allow for borrowing information across variables such as genes or OTUs and result in improved statistical inference. Ji and Liu (2010) demonstrated that hierarchical models are efficient when analyzing high-throughput data. We develop a fully Bayesian approach for differential abundance analysis while controlling false discovery rate (FDR). Our method offers the flexibility of not only examining the mean abundance difference across treatments but also other biological relevant null hypotheses in terms of model parameters. For example, we can also test the treatment effect on the proportion of zeros.

The remainder of this manuscript is organized as follows. Section 2.2 describes our proposed method. In Section 2.3, we show results from several simulation studies that compare our proposed method with existing methods. In Section 2.4, we analyze two real datasets using our proposed method. Section 2.5 provides some discussions.

2.2 Model

2.2.1 Hierarchical poisson hurdle model

In this section, we describe our proposed hierarchical Poisson Hurdle model for analyzing microbiome data. A Poisson Hurdle model is a mixture model of two components. One component is a degenerate distribution with a point mass at zero that models the zero counts, and the other component is a truncated Poisson distribution that models the non-zero counts. Let Y_{gij} be the read count for the gth $(g = 1, 2, \cdots, G)$ OTU of the jth $(j = 1, 2, \cdots, n_i)$ biological replicate in the ith $(i = 1, 2, \cdots, I)$ treatment group. Let Z_{gij} be a Bernoulli random variable indicating which of the two components in the mixture Y_{gij} comes from. Z_{gij} takes value one if and only if the observed count Y_{gij} is positive, i.e., it comes from the zero truncated Poisson component. Thus, our model of the indicator Z_{gij} and the read count Y_{gij} is

$$\begin{aligned} Z_{gij} \mid p_{gi} \sim & \ \mathrm{Ber}(p_{gi}), \\ Y_{gij} \mid Z_{gij}, \mu_{gij} \sim & \begin{cases} 0, & \text{if } Z_{gij} = 0, \\ \mathrm{TP}\left(\mu_{gij}\right), & \text{if } Z_{gij} = 1, \end{cases} \end{aligned} \tag{2.1}$$

where $\mathrm{TP}(\mu)$ represents a zero truncated Poisson distribution with probability mass function (pmf) expressed as $f_{TP}(y|\mu) = \frac{f_p(y|\mu)}{1-f_p(0|\mu)}$, and $f_p(\cdot|\mu)$ is the pmf of a Poisson distribution with mean μ. The parameter p_{gi} is the probability of the non-zero component and the parameter μ_{gij} characterizes the abundance given Y_{gij} is positive. Following common practice in the generalized linear model framework, we model p_{gi} and μ_{gij} by α_{gi} and β_{gi} through logit and log link functions respectively as follows:

$$\begin{aligned} \mathrm{logit}(p_{gi}) = \log\left(\frac{p_{gi}}{1 - p_{gi}}\right) = \alpha_{gi}, \\ \log(\mu_{gij}) = \log(s_{ij}) + \beta_{gi}, \end{aligned} \tag{2.2}$$

where s_{ij} represents the normalization factor of replicate j in treatment group i and is used to adjust for potential systematic variations between samples due to technical issues such as sequencing depth. We estimate the normalization factor using the method in DESeq (Anders and Huber, 2010). In brief, we calculate the ratio of each observed count versus the geometric mean count across all samples for the corresponding OTU, and the normalization factor for a sample is calculated as the median of these ratios across all OTUs in that sample. We prefer this method over the total sum normalization, which sums up the total number of read counts in each sample. The reason is that it has been shown that total sum normalization would lead to biased estimation for RNA-seq data and result in potential bias for microbiome data (Paulson et al., 2013).

The treatments could affect an OTU by affecting the probability of having a positive count, or the abundance given that the count is positive, or both. For each OTU g $(g = 1, 2, \cdots, G)$, we are interested in testing the following hypothesis: $H_0^g : \alpha_{g1} = \cdots = \alpha_{gI}$ and $\beta_{g1} = \cdots = \beta_{gI}$ versus H_1^g : not H_0^g.

As the above null hypothesis involves two conditions, we use two indicators to study these two conditions respectively. First, let $\eta_{\alpha g}$ be an indicator related to condition $\alpha_{g1} = \cdots = \alpha_{gI}$ in the null hypothesis. If $\eta_{\alpha g} = 0$ then the probability of having a positive count is not affected by the treatment, i.e., $\alpha_{g1} = \cdots = \alpha_{gI}$. We model this indicator using a Bernoulli distribution $\eta_{\alpha g} \mid \pi_{\alpha 0} \sim \text{Ber}(1 - \pi_{\alpha 0})$ with $\pi_{\alpha 0}$ equal to the chance of condition $\alpha_{g1} = \cdots = \alpha_{gI}$. Hence, a mixture model is utilized for parameters $(\alpha_{g1}, \cdots, \alpha_{gI})^{'}$ conditional on the indicator $\eta_{\alpha g}$:

$$\begin{aligned}
(\alpha_{g1}, \cdots, \alpha_{gI})^{'} &= \begin{cases} \tilde{\alpha}_{g0}(1, 1, \cdots, 1)^{'}, & \text{if } \eta_{\alpha g} = 0, \\ (\tilde{\alpha}_{g1}, \cdots, \tilde{\alpha}_{gI})^{'}, & \text{if } \eta_{\alpha g} = 1, \end{cases} \\
\tilde{\alpha}_{g0} \mid \phi_0, \tau_0 &\sim N(\phi_0, \tau_0^2), \\
\tilde{\alpha}_{gi} \mid \phi_i, \tau_i &\sim N(\phi_i, \tau_i^2),
\end{aligned} \tag{2.3}$$

where $g = 1, 2, \cdots, G$, $i = 1, 2, \cdots, I$ and $\phi_0, \tau_0, \phi_i, \tau_i$ are hyper-parameters. Similarly, we define indicator $\eta_{\beta g}$ so that $\eta_{\beta g} = 0$ if $\beta_{g1} = \cdots = \beta_{gI}$. Indicator $\eta_{\beta g}$ is assumed to have a Bernoulli distribution $\eta_{\beta g} \mid \pi_{\beta 0} \sim \text{Ber}(1 - \pi_{\beta 0})$ with $\pi_{\beta 0}$ as the chance of condition $\beta_{g1} = \cdots = \beta_{gI}$. The parameters $(\beta_{g1}, \cdots, \beta_{gI})^{'}$ are modeled using a mixture distribution:

$$\begin{aligned}
(\beta_{g1}, \cdots, \beta_{gI})^{'} &= \begin{cases} \tilde{\beta}_{g0}(1, 1, \cdots, 1)^{'}, & \text{if } \eta_{\beta g} = 0, \\ \left(\tilde{\beta}_{g1}, \cdots, \tilde{\beta}_{gI}\right)^{'}, & \text{if } \eta_{\beta g} = 1, \end{cases} \\
\tilde{\beta}_{g0} \mid \theta_0, \sigma_0 &\sim N(\theta_0, \sigma_0^2), \\
\tilde{\beta}_{gi} \mid \theta_i, \sigma_i &\sim N(\theta_i, \sigma_i^2),
\end{aligned} \tag{2.4}$$

where $\theta_0, \sigma_0, \theta_i, \sigma_i$ are hyper-parameters.

In our model, $\eta_{\alpha g}$ and $\eta_{\beta g}$ can be interpreted as latent random variables indicating whether the treatment affects the probability of having positive count and the abundance given that the count is positive, respectively. Thus if either $\eta_{\alpha g} = 1$ or $\eta_{\beta g} = 1$ we reject the null hypothesis and conclude there is a treatment effect on the OTU g.

2.2.2 Parameter estimation

There are a total of $4(I+1)+2$ unknown parameters in our hierarchical mixture model:

$$\mathbf{\Theta} = \{\phi_0, \phi_1, \cdots, \phi_I, \tau_0, \tau_1, \cdots, \tau_I, \theta_0, \theta_1, \cdots, \theta_I, \sigma_0, \sigma_1, \cdots, \sigma_I, \pi_{\alpha 0}, \pi_{\beta 0}\}. \tag{2.5}$$

We propose a fully Bayesian approach for the parameter estimation. For all unknown hyper parameters, we assume independent non-informative priors as follows:

$$\begin{aligned} \phi_k \sim N(0, 10^4), &\quad \theta_k \sim N(0, 10^4), \\ \tau_k \sim \mathrm{IG}(0.001, 0.001), &\quad \sigma_k \sim \mathrm{IG}(0.001, 0.001), \\ \pi_{\alpha 0} \sim \mathrm{Unif}(0,1), &\quad \pi_{\beta 0} \sim \mathrm{Unif}(0,1), \end{aligned} \tag{2.6}$$

where $k = 0, 1, \cdots, I$.

Markov Chain Monte Carlo (MCMC) Gibbs sampling is implemented to make posterior inference using rjags (Plummer et al., 2003). The posterior estimates of parameters, including all the latent indicators $\eta_{\alpha g}$ and $\eta_{\beta g}$, were obtained from posterior samples. More specifically, we run two independent MCMC chains with different initial values, and each chain ran 70,000 iterations. After 20,000 burn-in iterations, we draw samples at every 25 iterations and obtain a sample of size 2,000 for posterior inference. Convergence was checked via Gelman-Rubin criteria (Gelman and Rubin, 1992). Details about MCMC sampling procedure are available at 2.2.2.1.

2.2.2.1 Markov chain monte carlo implementation

In this section, we provide a detailed description of a Markov Chain Monte Carlo (MCMC) algorithm to implement our proposed model in Section 2.1. More specifically, we use an overall Gibbs sampling algorithm. In each Gibbs step, parameters are updated either by using conjugacies from the model structure or by applying the univariate stepping-out slice sampling proposed by Neal (2003).

For our proposd model, the likelihood function with observed data $\mathbf{Y}$ can be expressed in the Equation (2.7):

$$\begin{aligned} L(\boldsymbol{\Theta}|\boldsymbol{Y}) = & \prod_{g=1}^{G}\prod_{i=1}^{I}\prod_{j=1}^{n_i}\left\{[(1-p_{gi}(\alpha_{gi}))]^{(1-Z_{gij})}\left[\frac{p_{gi}(\alpha_{gi})}{1-\exp(\mu_{gij}(\beta_{gi}))}f_p(y_{gij};\mu_{gij}(\beta_{gi}))\right]^{Z_{gij}}\right\} \\ & \times\prod_{g=1}^{G}\left\{\left[\pi_{\alpha 0}f_N(\tilde{\alpha}_{g0};\phi_0,\tau_0^2)\right]^{(1-\eta_{\alpha g})}\left[(1-\pi_{\alpha 0})\prod_{i=1}^{I}f_N(\tilde{\alpha}_{gi};\phi_i,\tau_i^2)\right]^{\eta_{\alpha g}}\right\} \\ & \times\prod_{g=1}^{G}\left\{\left[\pi_{\beta 0}f_N(\tilde{\beta}_{g0};\theta_0,\sigma_0^2)\right]^{(1-\eta_{\beta g})}\left[(1-\pi_{\beta 0})\prod_{i=1}^{I}f_N(\tilde{\beta}_{gi};\theta_i,\sigma_i^2)\right]^{\eta_{\beta g}}\right\}. \end{aligned} \tag{2.7}$$

To draw Bayesian inference, we not only sample the parameters in the set

$$\boldsymbol{\Theta} = \{\phi_0,\phi_1,\cdots,\phi_I,\tau_0,\tau_1,\cdots,\tau_I,\theta_0,\theta_1,\cdots,\theta_I,\sigma_0,\sigma_1,\cdots,\sigma_I,\pi_{\alpha 0},\pi_{\beta 0}\},$$

but also update the following parameters for each OTU g ($g = 1, 2, \cdots, G$)

$$\{\eta_{\beta g},\eta_{\alpha g},\alpha_{g0},\alpha_{g1},\cdots,\alpha_{gI},\beta_{g0},\beta_{g1},\cdots,\beta_{gI}\}.$$

In this section, without loss of generality, we use the notation $f(\theta|\ \cdot)$ to denote the generic notation of density function for any parameter θ of interest conditioning on everything else. Each of the following full conditional distributions is sampled in one iteration of the Gibbs sampling:

1. For parameter θ_0, we consider a non-informative Normal distribution as prior, $\theta_0 \sim N(a_0, b_0^2)$ namely $f_N(\theta_0; a_0, b_0) \propto \frac{1}{\sqrt{2\pi b_0^2}}\exp\left(-\frac{1}{2b_0^2}(\theta_0 - a_0)^2\right)$, where $a_0 = 0$, $b_0^2 = 10^4$. Then the full conditional distribution can be written as:

$$\begin{aligned} f(\theta_0|\ \cdot) &\propto \left\{\prod_{g=1}^{G} f_N(\tilde{\beta}_{g0}|\eta_{\beta g}=0,\theta_0,\sigma_0)\right\}\{f_N(\theta_0;a_0,b_0^2)\} \\ &\propto \exp\left\{-\sum_{g=1}^{G} I(\eta_{\beta g}=0)\frac{(\tilde{\beta}_{g0}-\theta_0)^2}{2\sigma_0^2}\right\}\exp\left\{-\frac{(\theta_0-a_0)^2}{2\times b_0^2}\right\}, \end{aligned}$$

i.e., $\quad \theta_0|\cdot \sim N(m, v),$

where

$$v = \left[\frac{\sum_{g=1}^{G} I(\eta_{\beta g}=0)}{\sigma_0^2} + \frac{1}{b_0^2}\right]^{-1}, m = v\left[\frac{\sum_{g=1}^{G} I(\eta_{\beta g}=0)\tilde{\beta}_{g0}}{\sigma_0^2} + \frac{a_0}{b_0^2}\right].$$

2. For parameters $(\theta_1, \theta_2, \cdots, \theta_I)^{'}$, we assume an independent non-informative normal distribution as prior, i.e.,

$$(\theta_1, \theta_2, \cdots, \theta_I)^{'} \sim \prod_{i=1}^{I} N(a_i, b_i^2),$$

and $f_N(\theta_i; a_i, b_i) \propto \frac{1}{\sqrt{2\pi b_i^2}} \exp\left(-\frac{1}{2b_i^2}(\theta_i - a_i)^2\right)$, where $a_i = 0, b_i^2 = 10^4$. Then the full conditional distribution can be written as:

$$\begin{aligned} f(\theta| \cdot) &\propto \left\{\prod_{g=1}^{G}\prod_{i=1}^{I} f_N(\beta_{gi}|\eta_{\beta g} = 1, \theta_i, \sigma_i)\right\}\{f_N(\theta_i; a_0, b_0^2)\} \\ &\propto \prod_{i=1}^{I} \exp\left\{-\sum_{g=1}^{G} I(\eta_{\beta g} = 1)\frac{(\tilde{\beta}_{gi} - \theta_i)^2}{2\sigma_i^2}\right\} \exp\left\{-\frac{(\theta_i - a_0)^2}{2 \times b_0^2}\right\}, \end{aligned}$$

$$\text{i.e.,} \qquad \theta_i| \cdot \sim N\left(m, v\right),$$

where

$$v = \left[\frac{\sum_{g=1}^{G} I(\eta_{\beta g} = 1)}{\sigma_i^2} + \frac{1}{b_0^2}\right]^{-1}, m = v\left[\frac{\sum_{g=1}^{G} I(\eta_{\beta g} = 1)\tilde{\beta}_{gi}}{\sigma_i^2} + \frac{a_0}{b_0^2}\right], \quad i = 1, 2, \cdots, I.$$

3. Assuming a conjugate prior $\sigma_0^2 \sim \text{IG}(a_0, b_0)$ for parameter σ_0 yields a full conditional distribution of the form:

$$\begin{aligned} f(\sigma_0^2| \cdot) &\propto \left\{\prod_{g=1}^{G} f_N(\tilde{\beta}_{g0}|\eta_{\beta g} = 0, \theta_0, \sigma_0^2)\right\}\{f_{IG}(\sigma_0^2; a_0, b_0)\} \\ &\propto (\sigma_0^2)^{-\sum_{g=1}^{G} I(\eta_{\beta g}=0)/2} \exp\left\{-\sum_{g=1}^{G} I(\eta_{\beta g} = 0)\frac{(\tilde{\beta}_{g0} - \theta_0)^2}{2\sigma_0^2}\right\}(\sigma_0^2)^{-a_0-1}\exp\left(-\frac{b_0}{\sigma_0^2}\right), \end{aligned}$$

$$\text{i.e.,} \qquad \sigma_0^2| \cdot \sim \text{IG}\left(a_0 + \frac{1}{2}\sum_{g=1}^{G} I(\eta_{\beta g} = 0), b_0 + \frac{1}{2}\sum_{g=1}^{G} I(\eta_{\beta g} = 0)(\tilde{\beta}_{g0} - \theta_0)^2\right),$$

where $f_{IG}(\sigma_0^2; a_0, b_0) \propto (\sigma_0^2)^{-a_0-1}\exp(-\frac{b_0}{\sigma_0^2})$, and $a_0 = b_0 = 0.001$.

4. For parameters $(\sigma_1^2, \sigma_2^2, \cdots, \sigma_I^2)^{'}$, we consider the independent conjugate prior $(\sigma_1^2, \sigma_2^2, \cdots, \sigma_I^2)^{'} \sim \prod_{i=1}^{I} \text{IG}(a_i, b_i)$, and $f_{IG}(\sigma_i^2; a_i, b_i) \propto (\sigma_i^2)^{-a_i-1}\exp(-\frac{b_i}{\sigma_i^2})$, where $a_i = b_i = 0.001$. Then the full conditional distribution is:

$$\begin{aligned} f(\sigma^2| \cdot) &\propto \left\{\prod_{g=1}^{G}\prod_{i=1}^{I} f_N(\tilde{\beta}_{gi}|\eta_{\beta g} = 1, \theta_i, \sigma_i^2)\right\}\{f_{IG}(\sigma_i^2; a_0, b_0)\} \\ &\propto \prod_{i=1}^{I}(\sigma_i^2)^{-\sum_{g=1}^{G} I(\eta_{\beta g}=1)/2} \exp\left\{-\sum_{g=1}^{G} I(\eta_{\beta g} = 1)\frac{(\tilde{\beta}_{gi} - \theta_i)^2}{2\sigma_i^2}\right\}(\sigma_i^2)^{-a_0-1}\exp\left(-\frac{b_0}{\sigma_i^2}\right), \end{aligned}$$

$$\text{i.e.,} \quad \sigma_i^2| \cdot \sim \text{IG}\left(a_i + \frac{1}{2}\sum_{g=1}^{G} I(\eta_{\beta g} = 1), b_i + \frac{1}{2}\sum_{g=1}^{G} I(\eta_{\beta g} = 1)(\tilde{\beta}_{gi} - \theta_i)^2\right), \quad i = 1, 2, \cdots, I.$$

5. We consider a non-informative Normal distribution prior for the parameter ϕ_0, i.e., $\phi_0 \sim N(a_0, b_0^2)$ namely $f_N(\phi_0; a_0, b_0) \propto \frac{1}{\sqrt{2\pi b_0^2}} \exp\left(-\frac{1}{2b_0^2}(\phi_0 - a_0)^2\right)$, where $a_0 = 0, b_0^2 = 10^4$. Then the full conditional distribution is:

$$\begin{aligned} f(\phi_0| \cdot) &\propto \left\{\prod_{g=1}^{G} f_N(\tilde{\alpha}_{g0}|\eta_{\alpha g} = 0, \phi_0, \tau_0)\right\} \{f_N(\phi_0; a_0, b_0^2)\} \\ &\propto \exp\left\{-\sum_{g=1}^{G} I(\eta_{\alpha g} = 0)\frac{(\tilde{\alpha}_{g0} - \phi_0)^2}{2\tau_0^2}\right\} \exp\left\{-\frac{(\phi_0 - a_0)^2}{2 \times b_0^2}\right\}, \end{aligned}$$

i.e., $$\phi_0| \cdot \sim N(m, v),$$

where

$$v = \left[\frac{\sum_{g=1}^{G} I(\eta_{\alpha g} = 0)}{\tau_0^2} + \frac{1}{b_0^2}\right]^{-1}, m = v\left[\frac{\sum_{g=1}^{G} I(\eta_{\alpha g} = 0)}{\tau_0^2} + \frac{a_0}{b_0^2}\right].$$

6. The full conditional distribution of parameters $(\phi_1, \phi_2, \cdots, \phi_I)'$ is factorized into the product of independent terms, since we consider the independent non-informative normal distribution as prior,

$$(\phi_1, \phi_2, \cdots, \phi_I)' \sim \prod_{i=1}^{I} N(a_i, b_i^2),$$

and $f_N(\theta_i; a_i, b_i) \propto \frac{1}{\sqrt{2\pi b_i^2}} \exp\left(-\frac{1}{2b_i^2}(\phi_i - a_i)^2\right)$, where $a_i = 0, b_i^2 = 10^4$. The full conditional distribution is:

$$\begin{aligned} f(\phi| \cdot) &\propto \left\{\prod_{g=1}^{G}\prod_{i=1}^{I} f_N(\tilde{\alpha}_{gi}|\eta_{\alpha g} = 1, \phi_i, \tau_i)\right\} \{f_N(\phi_i; a_0, b_0^2)\} \\ &\propto \prod_{i=1}^{I} \exp\left\{-\sum_{g=1}^{G} I(\eta_{\alpha g} = 1)\frac{(\tilde{\alpha}_{gi} - \phi_i)^2}{2\tau_i^2}\right\} \exp\left\{-\frac{(\phi_i - a_i)^2}{2 \times b_i^2}\right\}, \end{aligned}$$

i.e., $$\phi_i| \cdot \sim N(m, v),$$

where

$$v = \left[\frac{\sum_{g=1}^{G} I(\eta_{\alpha g} = 1)}{\tau_i^2} + \frac{1}{b_i^2}\right]^{-1}, m = v\left[\frac{\sum_{g=1}^{G} I(\eta_{\alpha g} = 1)}{\tau_i^2} + \frac{a_i}{b_i^2}\right], \quad i = 1, 2, \cdots, I.$$

7. For parameter τ_0, we consider a conjugate prior $\tau_0^2 \sim \text{IG}(a_0, b_0)$. Then the full conditional distribution is:

$$
\begin{aligned}
f(\tau_0^2|\cdot) &\propto \left\{\prod_{g=1}^{G} f_N(\tilde{\alpha}_{g0}|\eta_{\alpha g}=0, \phi_0, \tau_0^2)\right\}\left\{f_{IG}(\tau_0^2; a_0, b_0)\right\} \\
&\propto (\tau_0^2)^{-\sum_{g=1}^{G} I(\eta_{\alpha g}=0)/2} \exp\left\{-\sum_{g=1}^{G} I(\eta_{\alpha g}=0)\frac{(\tilde{\alpha}_{g0}-\phi_0)^2}{2\tau_0^2}\right\}(\tau_0^2)^{-a_0-1}\exp\left(-\frac{b_0}{\tau_0^2}\right),
\end{aligned}
$$

$$
\text{i.e.,}\quad \tau_0^2|\cdot \sim \text{IG}\left(a_0+\frac{1}{2}\sum_{g=1}^{G} I(\eta_{\alpha g}=0), b_0+\frac{1}{2}\sum_{g=1}^{G} I(\eta_{\alpha g}=0)(\tilde{\alpha}_{g0}-\phi_0)^2\right).
$$

8. For parameters $(\tau_1^2, \tau_2^2, \cdots, \tau_I^2)'$, we consider the independent conjugate prior $(\tau_1^2, \tau_2^2, \cdots, \tau_I^2)' \sim \prod_{i=1}^{I} IG(a_i, b_i)$, and $f_{IG}(\tau_i^2; a_i, b_i) \propto (\tau_i^2)^{-a_i-1}\exp(-\frac{b_i}{\sigma_i^2})$, where $a_i = b_i = 0.001$. Then the full conditional distribution is:

$$
\begin{aligned}
f(\tau^2|\cdot) &\propto \left\{\prod_{g=1}^{G}\prod_{i=1}^{I} f_N(\tilde{\alpha}_{gi}|\eta_{\alpha g}=1, \phi_i, \tau_i^2)\right\}\left\{f_{IG}(\tau_i^2; a_0, b_0)\right\} \\
&\propto \prod_{i=1}^{I}(\tau_i^2)^{-\sum_{g=1}^{G} I(\eta_{\alpha g}=1)/2} \exp\left\{-\sum_{g=1}^{G} I(\eta_{\alpha g}=1)\frac{(\tilde{\alpha}_{gi}-\phi_i)^2}{2\tau_i^2}\right\}(\tau_i^2)^{-a_i-1}\exp\left(-\frac{b_i}{\tau_i^2}\right),
\end{aligned}
$$

$$
\text{i.e.,}\quad \tau_i^2|\cdot \sim \text{IG}\left(a_i+\frac{1}{2}\sum_{g=1}^{G} I(\eta_{\alpha g}=1), b_i+\frac{1}{2}\sum_{g=1}^{G} I(\eta_{\alpha g}=1)(\tilde{\alpha}_{gi}-\phi_i)^2\right), \quad i = 1, 2, \cdots, I.
$$

9. The full conditional distribution for the parameters $\eta_{\beta g}$ $(g = 1, 2, \cdots, G)$ can be obtained by:

$$
\begin{aligned}
f(\eta_{\beta g}=1|\cdot) &\propto \prod_{i=1}^{I} f_N(\tilde{\beta}_{gi}|\eta_{\beta g}=1, \theta_i, \sigma_i^2) f(\eta_{\beta g}=1|\pi_{\beta 0}) \\
&\propto (1-\pi_{\beta 0})\prod_{i=1}^{I} f_N(\tilde{\beta}_{gi}|\eta_{\beta g}=1, \theta_i, \sigma_i^2), \\
f(\eta_{\beta g}=0|\cdot) &\propto f_N(\tilde{\beta}_{g0}|\eta_{\beta g}=0, \theta_0, \sigma_0^2) f(\eta_{\beta g}=0|\pi_{\beta 0}) \\
&\propto \pi_{\beta 0} f_N(\tilde{\beta}_{g0}|\eta_{\beta g}=0, \theta_0, \sigma_0^2).
\end{aligned}
$$

Therefore, each $\eta_{\beta g}$ can be sampled from a Bernoulli distribution given by:

$$
\eta_{\beta g} \sim \text{Ber}\left(\frac{(1-\pi_{\beta 0})\prod_{i=1}^{I} f_N(\tilde{\beta}_{gi}|\eta_{\beta g}=1, \theta_i, \sigma_i^2)}{(1-\pi_{\beta 0})\prod_{i=1}^{I} f_N(\tilde{\beta}_{gi}|\eta_{\beta g}=1, \theta_i, \sigma_i^2) + \pi_{\beta 0} f_N(\tilde{\beta}_{g0}|\eta_{\beta g}=0, \theta_0, \sigma_0^2)}\right).
$$

10. Similarly, the full conditional distribution of parameter $\eta_{\alpha g}$ $(g = 1, 2, \cdots, G)$ can be obtained by:

$$\begin{aligned} f(\eta_{\alpha g} = 1| \cdot) &\propto \prod_{i=1}^{I} f_N(\tilde{\alpha}_{gi}|\eta_{\alpha g} = 1, \phi_i, \tau_i^2) f(\eta_{\alpha g} = 1|\pi_{\alpha 0}) \\ &\propto (1 - \pi_{\alpha 0}) \prod_{i=1}^{I} f_N(\tilde{\alpha}_{gi}|\eta_{\alpha g} = 1, \phi_i, \tau_i^2), \\ f(\eta_{\alpha g} = 0| \cdot) &\propto f_N(\tilde{\alpha}_{g0}|\eta_{\alpha g} = 0, \xi_0, \tau_0^2) f(\eta_{\alpha g} = 0|\pi_{\alpha 0}) \\ &\propto \pi_{\alpha 0} f_N(\tilde{\alpha}_{g0}|\eta_{\alpha g} = 0, \phi_0, \tau_0^2). \end{aligned}$$

Hence, each $\eta_{\alpha g}$ can be sampled from the following Bernoulli distribution:

$$\eta_{\alpha g} \sim \text{Ber}\left(\frac{(1 - \pi_{\alpha 0}) \prod_{i=1}^{I} f_N(\tilde{\alpha}_{gi}|\eta_{\alpha g} = 1, \phi_i, \tau_i^2)}{(1 - \pi_{\alpha 0}) \prod_{i=1}^{I} f_N(\tilde{\alpha}_{gi}|\eta_{\alpha g} = 1, \phi_i, \tau_i^2) + \pi_{\alpha 0} f_N(\tilde{\alpha}_{g0}|\eta_{\alpha g} = 0, \phi_0, \tau_0^2)} \right).$$

11. The full conditional distribution of the paramters β_{gi} $(g = 1, 2, \cdots, G,\ i = 0, 1, 2, \cdots, I)$ can be written as:

$$\begin{aligned} & f(\tilde{\beta}_{g0}|\eta_{\beta g} = 0, \cdot) \propto \prod_{i=1}^{I} \prod_{j=1}^{n_i} f(y_{gij}|\alpha_g, \beta_{g0}) f(\tilde{\beta}_{g0}|\eta_{\beta g} = 0, \theta_0, \sigma_0^2) \\ & \propto \left\{ \prod_{i=1}^{I} \prod_{j=1}^{n_i} \frac{1}{1 - \exp[-\exp(\log(s_{ij}) + \tilde{\beta}_{g0})]} \frac{(\exp[\log(s_{ij}) + \tilde{\beta}_{g0}])^{y_{gij}}}{y_{gij}!} \exp[-\exp(\log(s_{ij}) + \tilde{\beta}_{g0})] \right\}^{Z_{gij}} \\ & \times \left\{ \exp\left(-\frac{(\tilde{\beta}_{g0} - \theta_0)^2}{2\sigma_0^2} \right) \right\}, \end{aligned}$$

and

$$\begin{aligned} & f(\beta_{gi}|\eta_{\beta g} = 1, \cdot) \propto \prod_{i=1}^{I} \prod_{j=1}^{n_i} f(y_{gij}|\alpha_g, \beta_{gi}) f(\beta_{gi}|\eta_{\beta g} = 1, \theta_i, \sigma_i^2) \\ & \propto \left\{ \prod_{i=1}^{I} \prod_{j=1}^{n_i} \frac{1}{1 - \exp[-\exp(\log(s_{ij}) + \beta_{gi})]} \frac{(\exp[\log(s_{ij}) + \beta_{gi}])^{y_{gij}}}{y_{gij}!} \exp[-\exp(\log(s_{ij}) + \beta_{gi})] \right\}^{Z_{gij}} \\ & \times \left\{ \exp\left(-\frac{(\beta_i - \theta_i)^2}{2\sigma_i^2} \right) \right\}. \end{aligned}$$

Since the full conditional distribution does not have a standard form, we use a slice sampler to update β_{gi} or $\tilde{\beta}_{g0}$.

12. The full conditional distribution of the paramters α_{gi} $(g = 1, 2, \cdots, G,\ i = 0, 1, 2, \cdots, I)$ can be written as:

$$
\begin{aligned}
& f(\tilde{\alpha}_{g0}|\ \eta_{\alpha g} = 0, \phi_0, \tau_0) \\
& \propto \prod_{i=1}^{I} \prod_{j=1}^{n_i} \left\{ \frac{\exp(\tilde{\alpha}_{g0})}{1 + \exp(\tilde{\alpha}_{g0})} \right\}^{1-Z_{gij}} \left\{ \frac{1}{1 + \exp(\tilde{\alpha}_{g0})} \right\}^{Z_{gij}} f_N(\alpha_{g0}|\eta_{\alpha g} = 0, \phi_0, \tau_0) \\
& \propto \prod_{i=1}^{I} \prod_{j=1}^{n_i} \left\{ \frac{\exp(\tilde{\alpha}_{g0})}{1 + \exp(\tilde{\alpha}_{g0})} \right\}^{1-Z_{gij}} \left\{ \frac{1}{1 + \exp(\tilde{\alpha}_{g0})} \right\}^{Z_{gij}} \times \left\{ \exp\left(-\frac{(\tilde{\alpha}_{g0} - \xi_0)^2}{2\tau_0^2} \right) \right\},
\end{aligned}
$$

and

$$
\begin{aligned}
f(\alpha_{gi}|\eta_{\alpha g} = 1, \phi_i, \tau_i) & \propto \prod_{j=1}^{n_i} \left\{ \frac{\exp(\alpha_{gi})}{1 + \exp(\alpha_{gi})} \right\}^{1-Z_{gij}} \left\{ \frac{1}{1 + \exp(\alpha_{gi})} \right\}^{Z_{gij}} f_N(\alpha_{gi}|\eta_{\alpha g} = 1, \phi_0, \tau_0) \quad . \\
\propto \prod_{j=1}^{n_i} \left\{ \frac{\exp(\alpha_{gi})}{1 + \exp(\alpha_{gi})} \right\}^{1-Z_{gij}} & \left\{ \frac{1}{1 + \exp(\alpha_{gi})} \right\}^{Z_{gij}} \times \left\{ \exp\left(-\frac{(\alpha_{gi} - \phi_i)^2}{2\tau_i^2} \right) \right\}.
\end{aligned}
$$

Similarly, since the full conditional distribution does not have a standard form, we use a slice sampler to update α_{gi} or $\tilde{\alpha}_{g0}$.

13. For parameter $\pi_{\alpha 0}$, we specify a Uniform(0,1) distribution as prior. The full conditional distribution is obtained by:

$$
\begin{aligned}
f(\pi_{\alpha 0}|\ \cdot) & \propto \prod_{g=1}^{G} f(\eta_{\alpha g}|\pi_{\alpha 0}) f(\pi_{\alpha 0}) \propto \prod_{g=1} \pi_{\alpha 0}^{1-\eta_{\alpha g}} (1 - \pi_{\alpha 0})^{\eta_{\alpha g}} \\
& \propto \pi_{\alpha 0}^{\sum_{g=1}^{G}(1-\eta_{\alpha g})} (1 - \pi_{\alpha 0})^{\sum_{g=1}^{G} \eta_{\alpha g}},
\end{aligned}
$$

$$
\text{i.e.,} \qquad \pi_{\alpha 0}|\ \cdot \sim \text{Beta}\left(\sum_{g=1}^{G} (1 - \eta_{\alpha g}) + 1, \sum_{g=1}^{G} \eta_{\alpha g} + 1 \right).
$$

14. For parameter $\pi_{\beta 0}$, we specify a Uniform(0,1) distribution as prior. The full conditional distribution is obtained by:

$$
\begin{aligned}
f(\pi_{\beta 0}|\ \cdot) & \propto \prod_{g=1}^{G} f(\eta_{\beta g}|\pi_{\beta 0}) f(\pi_{\beta 0}) \propto \prod_{g=1} \pi_{\beta}^{1-\eta_{\beta g}} (1 - \pi_{\beta 0})^{\eta_{\beta g}} \\
& \propto \pi_{\beta}^{\sum_{g=1}^{G}(1-\eta_{g\beta})} (1 - \pi_{\beta 0})^{\sum_{g=1}^{G} \eta_{\alpha g}},
\end{aligned}
$$

$$
\text{i.e.,} \qquad \pi_{\beta 0}|\ \cdot \sim \text{Beta}\left(\sum_{g=1}^{G} (1 - \eta_{\beta g}) + 1, \sum_{g=1}^{G} \eta_{\beta g} + 1 \right).
$$

We sampled each parameter from the full conditional distributions described above. Each parameter in $\mathbf{\Theta}$ was verified to achieve convergence using standard Bayesian model convergence diagnostics, including trace plots, autocorrelation statistics and Gelman-Rubin criteria.

2.2.3 Bayesian FDR

One advantage of Bayesian inference is that both hypothesis testing and estimation of quantities of interest can be easily performed by using the joint posterior distribution of model parameters. For example, our null hypothesis for each OTU g could be assessed by using the posterior probability $p_g = Pr(\eta_{\alpha g} = 0 \text{ and } \eta_{\beta g} = 0|\boldsymbol{Y})$. In practice, such a posterior probability can be used directly to make decisions of hypothesis testing while controlling a certain type of error. In high-dimensional hypothesis testing such as in genomic studies, controlling the proportion of false positives among the "significant" discoveries is of interest since a large number of hypotheses are examined simultaneously. The false discovery rate (FDR) proposed by Benjamini and Hochberg (1995) has been widely used as the error rate to control in large-scale multiple testing problems (Genovese and Wasserman, 2002; Storey, 2002). The Bayesian version of FDR could be estimated through posterior probabilities under our current framework. More specifically, for each OTU g ($g = 1, 2, \cdots, G$), the posterior probability that it is not differentially abundant is estimated via

$$\widehat{p}_g = \frac{1}{N}\sum_{n=1}^{N} I\left(\eta_{\alpha g}^{(n)} = 0 \text{ and } \eta_{\beta g}^{(n)} = 0|\boldsymbol{Y}\right), \tag{2.8}$$

where N is the number of MCMC samples, and $I(x)$ is an indicator function. Then the posterior probability that the gth OTU is differentially abundant is estimated through $1-\widehat{p}_g$. We consider the decision rule that classifies the gth OTU as being differentially abundant if the posterior probability $\widehat{p}_g < c^\star$, where $c^\star$ is the cutoff value that needs to be chosen in order to achieve a target false discovery rate, say γ, i.e.

$$c^\star = \sup\left\{c : \widehat{\mathrm{FDR}}(c) < \gamma\right\}, \tag{2.9}$$

where $\widehat{\mathrm{FDR}}(c) = \frac{\sum_{g=1}^{G}\widehat{p}_g I(\widehat{p}_g<c)}{\sum_{g=1}^{G} I(\widehat{p}_g<c)}$. Then following Newton et al. (2004), the Bayesian FDR controlled at γ level can be obtained by the following expression:

$$\widehat{\mathrm{BFDR}}(\gamma) = \frac{\sum_{g=1}^{G}\widehat{p}_g I(\widehat{p}_g < c^\star)}{\sum_{g=1}^{G} I(\widehat{p}_g < c^\star)}. \tag{2.10}$$

2.3 Simulation Studies

Simulation studies were conducted to assess the performance of our proposed approach and other methods for differential abundance analysis of microbiome data. For each simulation setting, 50 datasets were simulated based on the Poisson Hurdle model described in Section 2.2.1. As in the *Arabidopsis* dataset described in the next section, this dataset contains 1853 OTUs and 44 samples from five different genotypes. The parameters used to generate datasets were obtained from estimation of the *Arabidopsis* dataset based on our proposed model, and the values of these parameters are listed in Table 2.1. The normalization factor s_{ij} in Equation (2.2) was set to be a fixed value, i.e. $s_{ij} = 1$ for all $i = 1, 2, \cdots, I$ and $j = 1, 2, \cdots, n_i$.

Table 2.1: Parameter values estimated based on the *Arabidopsis* dataset and used in simulation studies.

Parameter	Values
$(\phi_0, \phi_1, \phi_2, \cdots, \phi_5)$	(2, -0.7, 0.2, 0.3, 0.25, -0.2)
$(\theta_0, \theta_1, \theta_2, \cdots, \theta_5)$	(-0.7, 1.5, 1.5, 1.5, 1.5, 1.5)
$(\tau_0, \tau_1, \tau_2, \cdots, \tau_5)$	(1.3, 0.1, 0.1, 0.1, 0.1, 0.1)
$(\sigma_0, \sigma_1, \sigma_2, \cdots, \sigma_5)$	(0.85, 0.8, 0.8, 0.75, 0.8, 0.8)
$n_1, n_2, \cdots, n_5$	(8,8,8,10,10,8)

Besides our proposed method that we call PHSeq, other methods under comparison in the simulation studies include edgeR (Robinson et al., 2010), DESeq (Anders and Huber, 2010), DESeq2 (Love et al., 2014), metagenomeSeq (MGSeq) (Paulson et al., 2013) and Kruskal-Wallis (KW) (Segata et al., 2011). Among these methods, edgeR, DESeq, and DESeq2 are methods widely applied for RNA-seq differential expression analysis and these methods are based on negative binomial models and do not take zero-inflation into account; metagenomeSeq is designed for microbiome data that models zero-inflation using a zero-inflated Gaussian model; and Kruskal-Wallis test is a non-parametric test for comparing multiple groups.

We evaluated the performance of different methods using the Receive Operating Characteristic (ROC) curve averaged over 50 simulated datasets and the value of the corresponding Area Under the Curve (AUC). For each simulated dataset, the OTUs were ranked by either the p-values (non-Bayesian methods) or the posterior probabilities (our proposed method). Then the true positive rates (TPRs) were calculated for a list of false positive rates (FPRs) for each dataset and then averaged across 50 simulated datasets at each given level of FPR. The higher the ROC curve, the larger the AUC value, and the better the ranking of OTUs for the corresponding method. Figure 2.4 presents the resulting average ROC curves together with the AUC values when FPR < 0.1 under four different simulation settings that differ for the combination of $\pi_{\alpha 0}$ and $\pi_{\beta 0}$. In all simulation settings, it is obvious that our proposed method performs much better than all the other methods. The AUC value for our proposed method was about 50% higher than the next best performing method. The methods edgeR, DESeq, and DESeq2 are all developed based on a negative binomial model with shrinkage estimation of the dispersion parameter, and they perform similarly and work better than the non-parametric method, Kruskal-Wallis test. The metagenomeSeq method, which is designed for handling extra zeros in metagenomic data using a log-transformation followed by a zero-inflated Gaussian mixture model (Paulson et al., 2013), had a similar performance to edgeR, DESeq, and DESeq2 for relatively large FPR level, but not for lower level of FPR.

We also examined the control of the false discovery rate (FDR) for each method because the error control is also practically important for differential abundance analysis. We applied the FDR controlling procedure described in Section 2.2.3 for our proposed method, and Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) was applied to control FDR for the other five methods under comparison. Figure 2.6 presents the plots of actual FDR versus the nominal level of FDR for all methods in all simulation settings. Overall our proposed method controlled the FDR to the nominal level, while metagenomeSeq failed to control FDR, i.e. produced a high number of false positives. This was consistent with McMurdie and Holmes's conclusion (McMurdie and Holmes, 2014) that metagenomeSeq tended to lead to higher false positive rate if there are not enough biological replicates, since this approach relies on the transformation of discrete data to approximate a continuous distribution. DESeq, DESeq2, edgeR and Kruskal-Wallis test were able to control the false discovery rate while being a bit conservative.

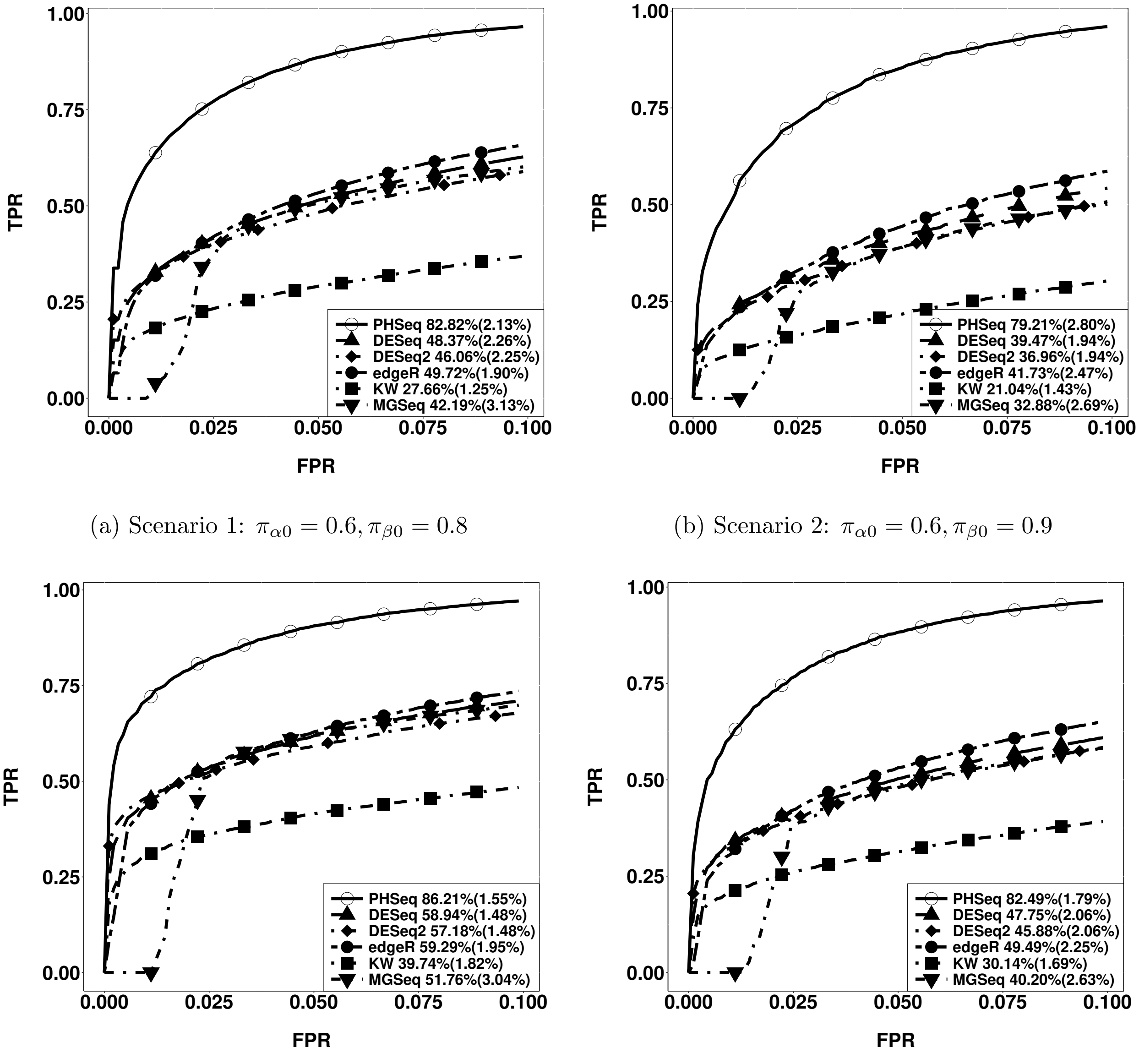
The simulation results show that our proposed Bayesian approach based on hierarchical Poisson Hurdle model outperformed all other methods in terms of both ranking differentially abundant OTUs and controlling FDR.

To test the robustness of our method, we performed another simulation study where data were not generated based on our model. In this simulation, data were generated from zero-truncated negative binomial distribution with dispersion parameter fixed as 0.05 for each OTU, and the mean of zero-truncated negative binomial were matched with the mean of zero-truncated Poisson distribution at the scenario that $\pi_{\alpha 0} = 0.6, \pi_{\beta 0} = 0.8$, while other parameters as the same as listed in Table 2.1.

The results from this simulation setting are presented in Figure 2.8. Although the performance of our method was not as good as in the previous settings, our method was still much better than others in terms of both ROC curves and FDR control.

2.4 Real Data Analysis

We applied our method to two published datasets that study how the microbiome associated with plant root change across different conditions. The first dataset is from *Arabidopsis* (Lundberg et al., 2012) and the other one is from grass (Naylor et al., 2017). For the *Arabidopsis* study, we used a subset of samples for our analysis. Samples corresponding to the rhizosphere compartment and the soil type Mason Farm with soil and sand mixed in the proportion of 2:1 were considered in our current analysis. This gave us a total of 44 samples across 5 genotypes (treatment conditions). Then OTUs with low abundance (the proportion of non-zero counts samples was less than 10%) were excluded (Chen and Li, 2016). This resulted in 1853 remaining OTUs used in following differential abundance analysis. For the grass dataset, samples from endophytic compartment and soil type A were selected. The treatments considered in this study were the combination between two levels of watering regimes (drought and well-water control) and four cultivated rice varieties. Similar pre-processing procedure was conducted to filter lowly abundant OTUs. This leaded to 3447 OTUs



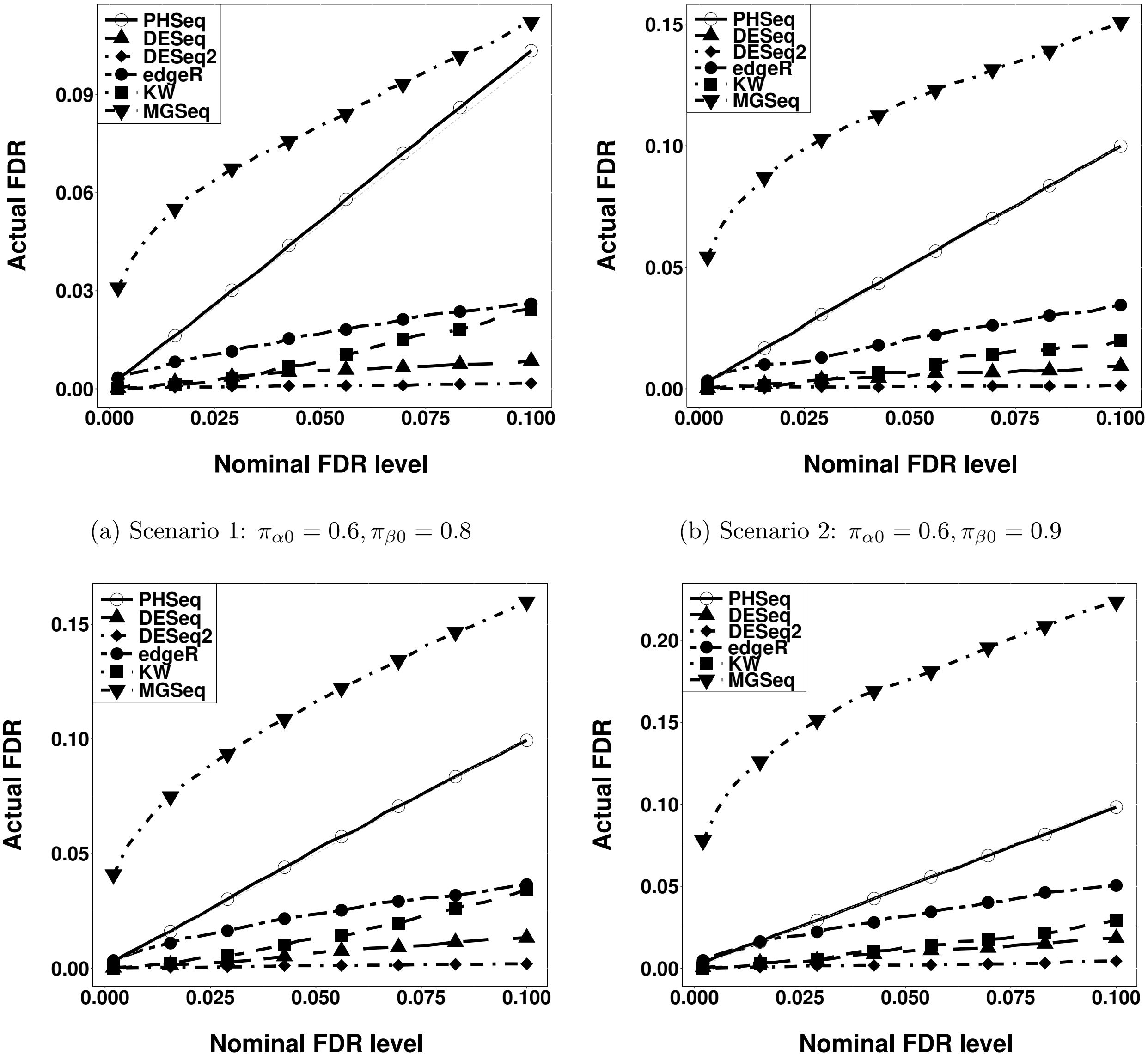
(a) Scenario 1: $\pi_{\alpha 0} = 0.6, \pi_{\beta 0} = 0.8$

(b) Scenario 2: $\pi_{\alpha 0} = 0.6, \pi_{\beta 0} = 0.9$

(c) Scenario 3: $\pi_{\alpha 0} = 0.8, \pi_{\beta 0} = 0.8$

(d) Scenario 4: $\pi_{\alpha 0} = 0.8, \pi_{\beta 0} = 0.9$

Figure 2.4: ROC curves for our proposed method and five other methods under comparison. Each ROC curve was an average over 50 simulated datasets. The AUC values were calculated as the percentage of the total area in the range of FPR < 0.1. The percentage in each parenthesis was the associated standard deviation among the 50 simulated datasets.



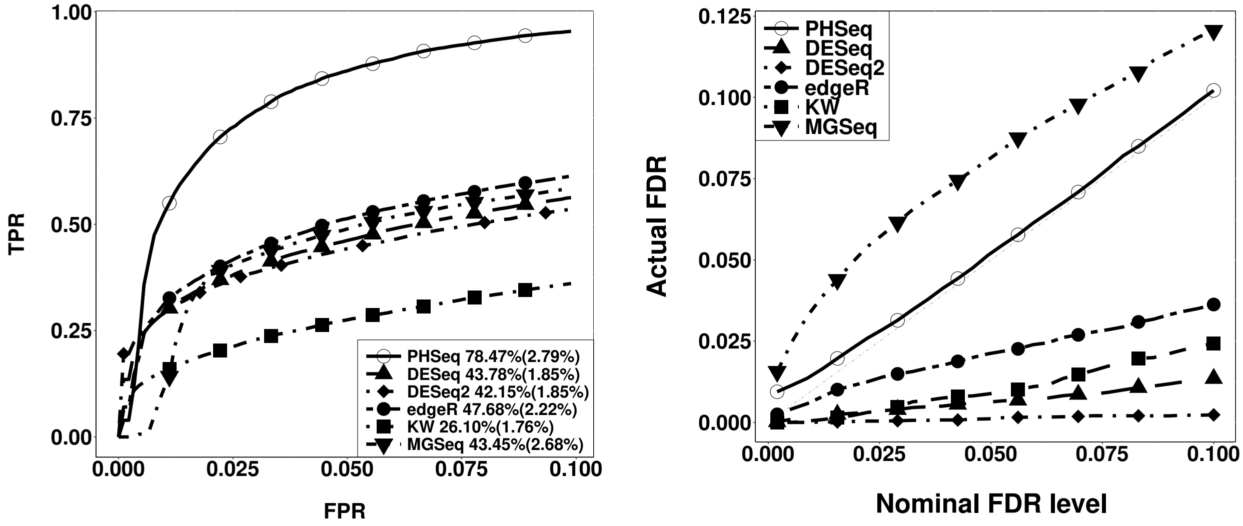
(a) Scenario 1: $\pi_{\alpha 0} = 0.6, \pi_{\beta 0} = 0.8$

(b) Scenario 2: $\pi_{\alpha 0} = 0.6, \pi_{\beta 0} = 0.9$

(c) Scenario 3: $\pi_{\alpha 0} = 0.8, \pi_{\beta 0} = 0.8$

(d) Scenario 4: $\pi_{\alpha 0} = 0.8, \pi_{\beta 0} = 0.9$

Figure 2.6: FDR plots. Each line corresponds to the averages of false discovery proportions across 50 datasets at each given level of FDR. The $Y = x$ line is shown as dashed grey lines in all four panels. Our proposed method (PHSeq) overlaps the $Y = x$ line in all four panels.



(a)

(b)

Figure 2.8: Robustness of the hierarchical Poisson Hurdle model under model mis-specification. (a) ROC curve were generated similarly as for Figure 2.4. Each ROC curve was an average over 50 simulated datasets. The AUC values were calculated as the percentage of the total area in the range of FPR < 0.1. The percentage in each parenthesis was the associated standard deviation among the 50 simulated datasets. (b) FDR plots. Each line corresponds to the average of false discovery proportions across 50 datasets at each given level of FDR. The $Y = x$ line is shown as dashed grey line.

from 32 samples remaining for the following analysis. The summary of those two datasets is listed in the Table 2.2.

Table 2.2: Number of samples, OTUs and treatments in each real dataset after filtering.

Dataset	Treatment/genotype	No. of OTUs after filtering	No. of samples
Arabidopsis	Five genotypes	1853	44
Grass	Eight treatments	3447	32

We performed our proposed approach and compared with Kruskal-Wallis, edgeR, DESeq, DESeq2 and metagenomeSeq methods. Table 2.3 exhibited the results of the numbers of detected OTUs as differentially abundant when FDR was controlled at 5%. As it shows, our proposed method found the largest number of differentially abundant OTUs. Figure 2.9 displays the Venn diagrams between Poisson Hurdle model, edgeR and metagenomeSeq methods applied on grass dataset.

Table 2.3: The numbers of OTUs for real data analysis detected as significant found by each method when control FDR at 5% level.

Dataset	Kruskal-Wallis	edgeR	DESeq	DESeq2	metagenomeSeq	PHSeq
Arabidopsis	0	2	0	0	3	187
Grass	0	83	9	0	1438	2733

2.5 Discussion

In this paper, we present a Bayesian approach based on hierarchical Poisson Hurdle model for differential abundance analysis of the microbiome count data. The mixture of a point mass at zero and a zero-truncated Poisson distribution for positive counts takes the natural characteristic of excess zeros in microbiome data into account. The Poisson Hurdle model also naturally handles data with zero-deflation and this provides more flexibility in modeling different OTUs than other methods. In addition, the hierarchical Bayesian framework allows information borrowing across OTUs.

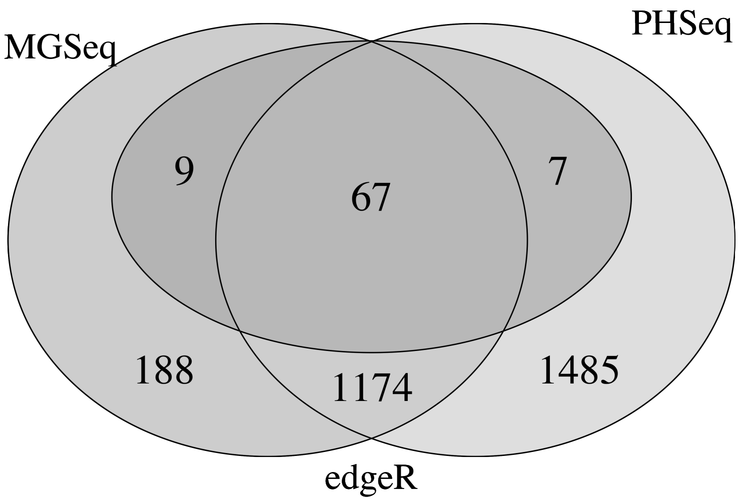


Figure 2.9: Number of OTUs detected as differentially abundant by Poisson Hurdle model (PHSeq), edgeR and metageomeSeq (MGSeq) methods when FDR controlled at 5% level in the analysis of the grass dataset.

Several simulation studies demonstrated that our approach had better performance compared to the conventional methods, which only analyze individual OTU separately, in terms of power and FDR control simultaneously.

A two-part method for modeling individual OTU at a time has been developed for OTU analysis in the literature (Wagner et al., 2011), where they only considered a two-group comparison problem. More recently, Chen and Li (2016) proposed the two-part mixed effect Beta regression model for microbiome. However, Chen and Li's (Chen and Li, 2016) method was designed for compositional data that is considered continuous rather than for the discrete count data. The compositional nature imposes the correlation among the OTUs (needs to satisfy the unit sum constraint) and also depends on the methods of normalization. This may affect the performance of the method especially the FDR control.

In our proposed method, we developed this two-part model based on count data without transformation. The OTUs were automatically clustered into two groups indicated by the latent variables in this hierarchical structure, which would be used to rank the importance of OTUs, providing some insights on the OTUs that are more likely to be associated with the treatment effects. Our extensive simulations show that our method is stable and powerful.

Our framework is very general and a few extensions are possible. In this manuscript, we focused on the test of treatment effects on both the proportion of zero counts and mean of the OTU abundance simultaneously. With the Bayesian inference, our proposed framework can be applied to test other biologically relevant null hypotheses by using the joint posterior samples. For example, we can test the treatment effects on the OTU abundance $H_0^g : \beta_{g1} = \cdots = \beta_{gI}$ and the treatment effects associated with proportion of zero count $H_0^g : \alpha_{g1} = \cdots = \alpha_{gI}$, respectively. In addition, our model is very flexible, which is able to include different covariates in two different components of the mixture model.

In conclusion, our proposed framework provides a powerful statistical tool for the differential abundance analysis of microbiome data. Although we focus on the count data from 16S rRNA sequencing, the proposed method is applicable to the shotgun metagenomic sequencing which is commonly used in microbiome studies.

CHAPTER 3. CAUSAL MEDIATION ANALYSIS OF HIGH-DIMENSIONAL MICROBIOME SEQUENCING DATA

The rapid development of high-throughput sequencing technologies has revolutionized the filed of metagenomics and provided researchers insights on studying the relationship between microbial communities, environment and associated biological outcome. The measured microbes are speculated to mediate the effects of environment on biological outcome. In agriculture, if the pathways between the microbes and the outcome can be elucidated, it might be possible to intervene upon the microbiome to maximize the plant performance. This motivates a causal mediation inference. However, existing causal mediation approaches cannot be directly applied to the setting with a large number of discrete mediators. In this project, we propose a testing procedure for the mediation effects of high-dimensional count mediators. To accommodate the setting with high-dimensional count mediators and a small sample size, we develop a novel screening procedure based on the ranking of indirect effect to reduce the dimension to a moderate size. Then the causal inference is conducted using a Bayesian variable selection framework that detects important mediators. By combining the ideas from independence screening process, Bayesian variable selection and causal inference, our proposed method sheds insights on disentangling indirect from direct effect pathway. Extensive simulation studies are carried out to assess the performance of our proposed approach. We also apply the method to a real dataset example and pinpoint a set of taxa significantly mediating the effect of treatment on the outcome.

3.1 Introduction

The recent rapid advancement of metagenomics, which is the study of genetic material of microbes from the environment without culturing (Handelsman et al., 1998; Thomas et al., 2012), provides a promising opportunity for researchers to investigate the roles of microbes in diverse fields.

With the advent of next generation sequencing technologies, millions of DNA reads are generated simultaneously. These sequence reads are further clustered based on their sequence similarity to form Operational Taxonomic Units (OTUs) (also known as taxa). Nowadays, researchers focus on not only demonstrating the components of microbial communities, but also how they can impact a outcome (Zhao, 2013; Fritz et al., 2013; Vorholt et al., 2017; Xia and Sun, 2017). From a statistical point of view, jointly modeling different types of genomic data and their relationships with the outcome and environmental conditions can be described using a causal diagram and be formalized using a causal mediation model (Lin et al., 2014b).

Our motivating example is a study for understanding of metabolic responses to different nitrogen levels in Sorghum. Our interest lies in the effect of environmental treatment conditions (measured in high vs low nitrogen levels) on biofuel feedstock yields (measured in metabolite). More specifically, we are interested in whether this effect might be mediated by the microbial communities at individual taxonomic levels (measured as the OTU counts). Previous studies have demonstrated that the production of plant metabolites might be impacted by several factors, including environmental conditions (Council et al., 2012; Gelli et al., 2014) and the complex microbial communities (Witowski and Baker, 2012; Chaparro et al., 2012; Kembel et al., 2014). Additionally, several studies have reported that different environmental conditions would assemble different types of microbes (Lundberg et al., 2012; Edwards et al., 2015). Therefore, identifying which taxa mediate the effect of environmental condition on the production of metabolite is an pivotal area of research as it can allow investigators to modify the microbiome abundance for the optimum plant growth or other favorable plant phenotypes in modern plant breeding programs (Tkacz and Poole, 2015).

To date, several statistical methods based on microbiome data have been proposed to study the relationships between microbial communities and their corresponding covariates, including environmental conditions and outcome of interest. In general, these approaches can be roughly categorized into two groups. The first group focuses on exploring the association between microbiome and outcome of interest. In this group, there are two common used approaches in the literature. One is comparing the microbial communities among different environmental conditions without adjusting for covariates, such as the permutation MANOVA (McArdle and Anderson, 2001; La Rosa et al., 2012). The other is a regression model approach where microbiome data are treated as either predictors or outcomes in the model (Lin et al., 2014a; Randolph et al., 2018; Shi et al., 2016; Tang et al., 2016; Xia et al., 2013; Chen and Li, 2013; Zhao et al., 2015). However, such association type of analysis suffers from one problem that it can only study the relationship between two variables at a time. In other words, it can only study the relationship between microbial communities and treatment conditions, between microbial communities and outcome, or between outcome and treatment conditions. These methods cannot be directly applied to a mediation analysis where three components (e.g., environmental conditions, microbial communities and outcome) need to be analyzed jointly.

The second group of approaches is the causal mediation type of analysis. Several methodologies have been developed recently for this purpose. Zhang et al. (2018) have proposed phylogeny-based distance metrics to test the overall mediation effect of the microbial communities instead of on individual taxonomic level. However, from a biological point of view, the ability to pinpoint specific individual taxon with mediation effect would help researchers to have a better understanding of the biological mechanisms and to target microbiome-based intervention in the future. A compositional mediation analysis has also been studied by Sohn and Li (2017). This approach analyzes mediation effect at taxonomic level but requires transforming the taxon count data into compositional data. Such transformation usually relies on a pre-decided reference baseline. Typically, the overall sum of sequence counts in each biological sample is the choice of reference. However, the total sum of reads is not robust to outliers. In addition, it is computationally expensive to make statistical inferences based on these two approaches as it is not possible to derive the asymptotic distribution for the test statistics. Hence, extra effort is needed, such as permutation used in Chen's paper (Zhang et al., 2018) and bootstrap used in Li's paper (Sohn and Li, 2017).

Therefore, there is an urgent need to have an appropriate mediation analysis method in place of testing the microbiome mediation effects. Nevertheless, despite these two recent studies which focused on microbiome mediation effects there are still several statistical challenges to overcome.

First of all, microbiome data are usually summarized as sequencing reads at taxonomic level, i.e. the nature of relative abundance of each taxon is an integer. Secondly, microbiome data are usually high-dimensional as millions of sequencing reads can be generated from a single sample. However, the sample size can be extremely small as it is still expensive to conduct biological experiments, i.e. it is a small n large p problem. For instance, in our motivating example, there are 34 biological samples and about 4000 OTUs. Taken together, it is crucial to take into account these challenges inherent in microbiome data analysis.

Additionally, most existing mediation analyses are concerned with single or a limited number of mediators (VanderWeele and Vansteelandt, 2014; Imai et al., 2010a; Robins and Greenland, 1992). However, there is very little information available on the topic about high-dimensional mediation effects. Although Zhang et al. (2016) and Huang and Pan (2016) study the causal mediation under high-dimensional setup, both studies assume that the mediating variables are continuous. Therefore, existing methods cannot be readily applicable to our OTU count data, as they do not account for sparsity (i.e. excessive zeros), high-dimensionality and the count nature of mediators.

To address the aforementioned issues, in this paper, we develop a mediation analysis approach to identify taxa with mediation effect under high-dimensional setup using a fully Bayesian variable selection methodology. Our key ideas are to adopt a screening methodology to reduce the potential mediators to a moderate range, then to employ the Bayesian variable selection method. To test for mediation effects at individual mediator level, we introduce latent random variables indicating whether treatment conditions have an effect on the abundance of mediator and whether mediator is associated with outcome after adjusting for the treatment effects. Extensive simulation studies demonstrate that our proposed method has reasonable statistical power.

The rest of this paper is organized as follows. In section 2, we introduce the screening process and the high-dimensional mediation model and propose the testing and inference procedure. In section 3, we examine the performance of screening procedure and mediation model via extensive simulation studies. In section 4, we apply our proposed method to study the mediating effect of taxa on the causal effect of nitrogen levels on metabolites. Section 5 summaries the method and discusses some further research topics.

3.2 Methods

Let M_{gij} denote the observed read count for taxon g, treatment i, and biological replicate j, where $j = 1, 2, \cdots, n_i$, $i = 1, 2, \cdots, I$, $g = 1, 2, \cdots, G$, and n_i is the total number of replicates for treatment i. Y_{ij} is the continuous outcome of interest. T_{ij} denotes the treatment, which is randomly assigned to subjects. The treatment is assumed to have only two levels (i.e. $I = 2, T_{1j} = 0, T_{2j} = 1$) for simplicity. However, our proposed method can be easily generalized to situations where there are more than two treatment conditions. In our real data example, Y_{ij} denotes the concentration of metabolite on the logarithm scale. Treatment T_{ij} will be nitrogen level (full vs low nitrogen). Figure 3.1 depicts the high-dimensional multiple mediators in a conceptual form. Here we assume a parallel multiple mediators model (Hayes, 2017), meaning no mediator causally influences another even though they might be correlated. We assume that the effect of treatment T on outcome Y is mediated through the microbiome M, i.e. the path $T \rightarrow \boldsymbol{M} \rightarrow Y$ in Figure 3.1, where $\boldsymbol{M} = \{M_{1ij}, \cdots, M_{Gij}\}$. Thus the mediation effects can be assessed under the following general framework:

$$f(E(M_{gij})) = \beta_{gi} \tag{3.1}$$

$$Y_{ij}|\boldsymbol{M}, T_{ij} = \tau_0 + \tau_1 T_{ij} + \sum_{g=1}^{G} \gamma_g M_{gij} + \epsilon_{ij}, \quad i = 1, 2; g = 1, 2, \cdots, G, \tag{3.2}$$

where τ_1 represents the direct effect of the treatment T on the outcome Y; ϵ_{ij} are uncorrelated random errors with mean zero and variance one. $E(\cdot)$ denotes the expectation, and $f(\cdot)$ is a function that describes the relationship between the treatment T and the expectation of the abundance of microbiome data M.

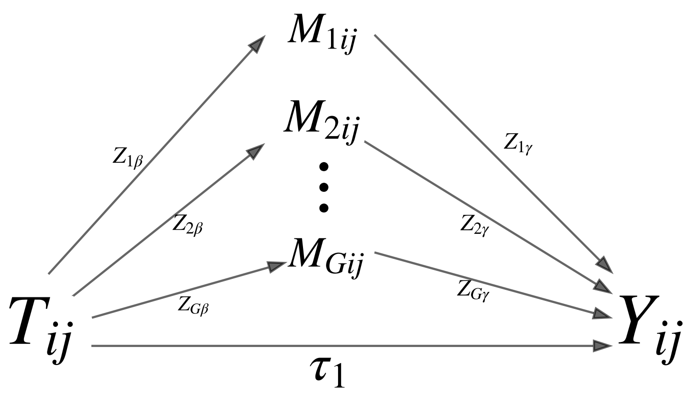


Figure 3.1: Illustration of the causal mediation model: $Z_{g\beta}$ is an indicator for whether treatment T_{ij} has an effect on the abundance of gth taxon, and and $Z_{g\gamma}$ is an indicator for whether gth taxon affects the outcome, $g = 1, 2, \cdots, G$. τ_1 is the direct effect of the treatment T_{ij} on the outcome Y_{ij}.

Our goal is to identify a set of taxa that would mediate the treatment effect on the outcome. To assess the mediation effects, we consider a two-stage statistical algorithm: a screening procedure is first conducted to reduce the dimensionality of taxa to a moderate size (but still larger than the sample size), then a Bayesian inference is performed on this reduced feature space to detect taxa with mediation effect.

3.2.1 First stage - screening methods

The purpose of the first stage is to screen out the non-informative taxa to reduce the dimensionality. As mentioned before, although the scale of the microbiome data is very large due to high-throughput characterizations of sequencing technology, a considerable amount of taxa have low abundance due to sequencing error or sampling process. In addition, much of the variation across samples may just be noise instead of biological signals. Therefore, dimension reduction is imperative, as a traditional regression approach cannot be applied to Equation (3.2) if the sample size is smaller than the number of taxa. To deal with the sparse feature of the microbiome data, we will consider two screening methods.

The first screening method (referred to as Screening method I in the rest of the paper) is to apply the sure independence screening (SIS) (Fan and Lv, 2008) by ranking the maximum marginal association of taxon with the outcome in Equation (3.2) to detect a set of potential taxa of interest. The details of this procedure are described as follows:

1. Identify a subset $\mathcal{I} = \{1 \leq k \leq G : M_{gij}$ is among top p taxon with largest effects on $Y\}$, where p is a pre-specified integer. More specifically, for each taxon g, conduct the following linear regression model:
$$y_{ij} = \alpha_0 + \alpha T_{ij} + \alpha_g M_{gij} + \epsilon_{ij}, \tag{3.3}$$
where and $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$. In practice, the taxa abundance M_{gij} are first standardized to have mean zero and unit variance across samples before fitting the regression models to ensure that the coefficients are in the same range.

2. Rank the absolute value of the ordinary least square estimation of $\hat{\alpha}_g$ decreasingly and then pick up the top p corresponding OTUs.

The SIS screening approach ranks OTU by marginal correlation with the outcome only, which does not consider the effect of treatment. Considering our goal for detecting taxa with mediation effect, i.e. the treatment has an effect on the abundance of microbiome then the microbiome impact the outcome, it is reasonable to incorporate the effect of treatment on OTUs (Equation (3.1)) into our screening stage. We thus propose another screening algorithm by ranking OTUs based on the magnitude of the indirect effect (referred to as Screening method II). The rationale is to rank the taxon based on the measured magnitude of the indirect effect, a function related to parameter β in Equation (3.1) and γ in (3.2).

1. For each taxon g, perform a Negative binomial regression model (considering the overdispersion feature in sequencing data) with treatment T as explanatory variable, i.e.
$$M_{gij} \sim \mathrm{NB}\left(\exp(\alpha_{g0} + \alpha_{g1} T_{ij}), \theta_g\right), \tag{3.4}$$

and let $\alpha_g^{(1)} = \exp(\alpha_{g0} + \alpha_{g1}) - \exp(\alpha_{g0})$, and the corresponding MLE (maximum likelihood estimates) is denoted as $\widehat{\alpha_g^{(1)}}$.

2. Then for each taxon g, conduct the following linear regression model:

$$y_{ij} = \alpha_0 + \alpha T_{ij} + \alpha_g^{(2)} M_{gij} + \epsilon_{ij}, \quad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2) \tag{3.5}$$

   The resulting OLS estimator for $\alpha_g^{(2)}$ is written as $\widehat{\alpha_g^{(2)}}$.

3. Let $\widehat{\alpha}_g = |\widehat{\alpha_g^{(1)}}\widehat{\alpha_g^{(2)}}|$, then rank the value of $\hat{\alpha}_g$ from largest to smallest and select the top p OTUs.

3.2.2 Second stage - mediation effect detection

The second stage of our approach is to detect the taxa that mediate the treatment effect on the outcome. To achieve this, we need to test whether there is a significant effect of the treatment T on mediators M, and whether these mediators are significantly associated with the outcome Y, controlling for the treatment T.

From examining real microbiome count data, we found that the variance of abundances is larger than corresponding expected mean even after the first stage screening (see Figure˜3.2). To account for the overdispersion issue, a generalized linear model with a Negative binomial distribution is used in Equation (3.1) to establish the relationship between microbial community and environmental conditions.

Assume p out of G taxa are picked up from the first stage. Under the assumption of the conditional independence among the taxonomic count, each read count M_{gij} can be modeled as:

$$M_{gij} \sim \mathrm{NB}\left(\mu_{gij}, \exp(\phi_g)\right), \tag{3.6}$$

where $\mathrm{NB}(\mu, \exp(\phi))$ indicates a Negative binomial distribution with mean μ and variance $\mu + \mu^2 \exp(\phi)$. As shown in Equation (3.6), it involves taxon-specific overdispersion parameters ϕ_g and mean parameters μ_{gij} that depend on the treatment i. These mean parameters are of our

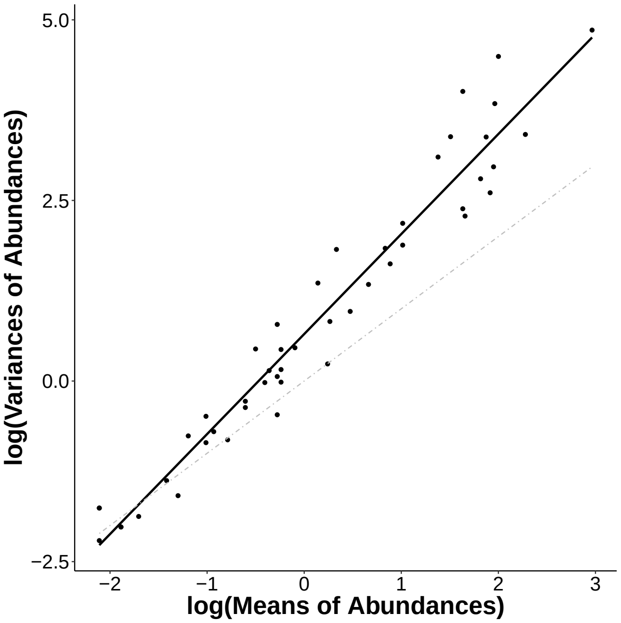


Figure 3.2: Scatter plot between means and variances of abundances from a real dataset after first screening stage. There are 33 samples, and $p = 50$ OTUs are selected in the first stage using screening method II. The gray dashed line represents the relationship where means equal to the variances of abundances. The fitted simple linear regression (solid line in the figure) is: $E\log(\sigma^2) = 0.6 + 1.37\log(\mu) + 0.03(\log(\mu))^2$ with $R^2 = 0.97$, where μ and σ^2 denote the mean and variance of abundances of these 50 OTUs, respectively. The p-value of the coefficient for the quadratic term $(\log(\mu))^2$ is 0.0003.

primary interest and can be further related to the treatment under the generalized linear regression framework. In other words, the treatment conditions could be incorporated into the model as:

$$\log\left(\mu_{gij}\right) = \beta_{gi} + \log(s_{ij}), \tag{3.7}$$

where s_{ij} represent the normalization factor for adjusting the bias owing to the variation in sequencing depth of replicate j in ith treatment. In our analysis, we used the geometric mean values (Anders and Huber, 2010), which is implemented in the DESeq Bioconductor package.

In order to share information across OTUs to improve estimation of taxonomic abundances, we propose the following hierarchical model:

$$\begin{aligned}
&Z_{g\beta} \mid \pi_{\beta 0} \overset{i.i.d}{\sim} \mathrm{Ber}(1-\pi_{\beta 0}),\\
&(\beta_{g1}, \cdots,\ \beta_{gI})^{'} = \begin{cases} \tilde{\beta}_{g0}(1,1,\cdots,1)^{'}, & \text{if } Z_{g\beta}=0,\\ \left(\tilde{\beta}_{g1},\cdots,\tilde{\beta}_{gI}\right)^{'}, & \text{if } Z_{g\beta}=1.\end{cases}\\
&\tilde{\beta}_{gi} \sim N(\theta_i, \sigma^2_{\beta i}),\\
&\phi_g \sim N(\eta_\phi, \sigma^2_\phi),
\end{aligned} \tag{3.8}$$

where ϕ_g and $\tilde{\beta}_{gi}$ are assumed to be independent with each other.

The association between mediators and outcome controlling for the treatment can be expressed as:

$$Y_{ij}|\mathbf{M_{ij}}, T_{ij} \sim N\left(\tau_0 + \tau_1 T_{ij} + \sum_{g=1}^{p} \gamma_g M_{gij}, \sigma^2_y\right), \tag{3.9}$$

where $\mathbf{M_{ij}} = (M_{1ij}, M_{2ij}, \cdots, M_{pij})^T$. We also assume σ^2_y, $\sigma^2_{\beta i}$ and σ^2_ϕ are mutually independent with each other.

In our proposed model, a latent random variable $Z_{g\beta}$ in Equation (3.8) is introduced for each OTU, indicating whether the treatment has an effect on the abundance of g-th OTU or not. Additionally, the hierarchical structure has the advantage of accommodating a general correlation between taxa. Moreover, τ_1 is the parameter relating treatment variable T_{ij} to outcome variable Y_{ij} through direct effect after adjusting for all potential mediators M_{ij} in Equation (3.9). Parameter

γ_g relates the gth mediator (taxon) to the outcome adjusting for the treatment effect T and the rest of the mediators. Thus the indirect effect, the effect of treatment on outcome through mediators, is denoted by the path $T \to M \to Y$ as in Figure 3.1.

In this paper, our main focus is on detecting whether a taxon has a mediation effect or not. A taxon is identified as a mediator if the following two conditions are met: (1) treatment T_{ij} is associated with M_{gij}; and (2) conditional on T_{ij} and on other taxon except M_{gij}, the outcome Y_{ij} is associated with M_{gij}. The overall null hypothesis that a taxon does not have a mediation effect can be further formalized in terms of two null hypotheses testing for each taxon, i.e.

$$H_{g0}^{(1)} : T_{ij} \perp M_{gij} \quad \text{or} \quad H_{g0}^{(2)} : Y_{ij} \perp M_{gij} | T_{ij}, \boldsymbol{M}_{(-g)ij}, \tag{3.10}$$

where $\boldsymbol{M}_{(-g)ij} = \big\{M_{1ij}, M_{1ij}, \cdots, M_{(g-1)ij}, M_{(g+1)ij}, \cdots, M_{pij}\big\}$.

It was suggested by Yuan and MacKinnon (2009) and Koopman et al. (2015) that the Bayesian approach would increase the statistical power in the mediation analysis, especially for cases with multiple mediators and small sample size. Considering the complexity of our data structure, we will perform a Bayesian approach for inference. Under the assumption that most of the taxa will not have mediation effect, we conduct a Bayesian variable selection using the shrinking and diffusing priors proposed by Narisetty et al. (2014). Under the Bayesian framework, the following independent priors are assigned for all parameters involved in Equations (3.8 - 3.9):

$$\begin{aligned}
&\pi_{\beta 0} \sim \text{Unif}(0,1),\\
&\theta_i \sim N(0,10^4), \quad \sigma^2_{\beta i} \sim \text{IG}(0.0001, 0.0001), i = 0,1,2,\\
&\eta_\phi \sim N(0,10^4), \quad \sigma^2_\eta \sim \text{IG}(0.0001, 0.0001),\\
&\tau_1 \sim N(0,10^4), \quad \sigma^2_y \sim \text{IG}(0.0001, 0.0001),\\
&\gamma_g \sim (1 - Z_{g\gamma})N(0, \sigma^2_y \sigma^2_{\gamma 0}) + Z_{g\gamma} N(0, \sigma^2_y \sigma^2_{\gamma 1}),\\
&Z_{g\gamma} \sim \text{Ber}(p_g), \quad p_g \sim \text{Beta}(0.2, 0.5),
\end{aligned} \tag{3.11}$$

where the default values in (Narisetty et al., 2014) are used for parameters $\sigma^2_{\gamma 0} = 1/n$, $\sigma^2_{\gamma 1} = \max\{100 \times \sigma^2_{\gamma 0}, \frac{0.1p}{(1-0.1)\rho}\}$, where $\rho = \phi(\sqrt{2.1 * \log(p+2)}; 0, 1)$ is the density function of the standard normal distribution evaluated at $\sqrt{2.1 * \log(p+2)}$, and p is the total number of taxa after screening.

Note that the hyperparameters p_g, $\sigma^2_{\gamma 0}$ and $\sigma^2_{\gamma 1}$ have important roles in the process of Bayesian variable selection. Parameter p_g reflects the sparsity of the model that relates OTUs to the outcome (i.e. Equation (3.9)), as the gth mediator would have impact on the outcome if $Z_{g\gamma}$ takes value one. The size of zero regression coefficients γ_g in Equation (3.9) is determined by the parameter $\sigma^2_{\gamma 0}$. Similarly, $\sigma^2_{\gamma 1}$ represents the size of nonzero regression coefficients γ_g. We place priors on p_g to allow the values to be informed by the data rather than setting prespecified values. The advantage of these data-driven hyperparameters will provide data-driven estimates of the posterior inclusion probabilities for the potential mediators. For the remaining hyperparameters, we assign conditional conjugate non-informative priors.

Consequently, under the proposed Bayesian framework, the null hypothesis for each taxon g described in Equation (3.10) can be rewritten in terms of latent variables $Z_{g\beta}$ and $Z_{g\gamma}$ together with the alternative hypothesis as:

$$H_{g0} : Z_{g\beta} Z_{g\gamma} = 0 \quad \text{vs} \quad H_{g1} : Z_{g\beta} Z_{g\gamma} \neq 0, \tag{3.12}$$

for $g = 1, 2, \cdots, p$.

3.2.3 Parameter estimation

Markov Chain Monte Carlo (MCMC) is the prominent approach used in Bayesian parameter estimation. In a nutshell, we use an overall Gibbs sampling framework and the univariate stepping-out slice sampler within Gibbs steps when a full conditional distribution is inaccessible (Neal, 2003; Cruz et al., 2014). The full conditional distributions are presented at the end of this chapter. We implemented the proposed algorithm using R software (R Core Team, 2017).

3.2.3.1 Markov chain monte carlo implementation

Based on the Equation (3.8), the full likelihood function for the differentially abundant related parameters can be written as:

$$
\begin{aligned}
&p\left(\boldsymbol{M}|\eta_\phi, \sigma_\phi, \theta_0, \sigma_{\beta0}, \theta_1, \sigma_{\beta1}, \theta_2, \sigma_{\beta2}, \pi_{\beta0}\right) \\
&\propto \prod_{g=1}^{p}\left\{\frac{1}{\sqrt{2\pi\sigma_\phi^2}}\exp\left(-\frac{(\phi_g-\eta_\phi)^2}{2\sigma_\phi^2}\right)\prod_{l=0}^{2}\frac{1}{\sqrt{2\pi\sigma_{\beta l}^2}}\exp\left(-\frac{(\tilde{\beta}_{gl}-\theta_l)^2}{2\sigma_{\beta l}^2}\right)(1-\pi_{\beta0})^{Z_{g\beta}}\pi_{\beta0}^{(1-Z_{g\beta})}\right\} \\
&\times \prod_{g=1}^{p}\prod_{i=1}^{2}\prod_{j=1}^{n_i}\left\{\left[f_{NB}\left(M_{gij};\exp(\tilde{\beta}_{g0}+\log(s_{ij})),\exp(\phi_g)\right)\right]^{1-Z_{g\beta}}\right. \\
&\qquad\qquad \left.\left[f_{NB}\left(M_{gij};\exp(\tilde{\beta}_{gi}+\log(s_{ij})),\exp(\phi_g)\right)\right]^{Z_{g\beta}}\right\},
\end{aligned}
\tag{3.13}
$$

where $f_{NB}(y;\mu,\theta)$ is the probability mass function of negative binomial distribution with mean μ and variance $\mu+\mu^2\theta$. Then the posterior distribution of the parameters can be updated using MCMC as follows:

1. Update η_ϕ. The full conditional distribution of η_ϕ is a normal distribution with mean $\frac{1}{1/10^4+p/\sigma_\eta^2}\frac{\sum_{g=1}^{p}\phi_g}{\sigma_\eta^2}$, and variance $\left(\frac{1}{10^4}+\frac{p}{\sigma_\eta^2}\right)^{-1}$.

2. Update σ_η^2. The full conditional distribution of σ_η^2 is an inverse gamma distribution with shape parameter $0.0001+0.5*p$ and rate parameter $0.0001+\frac{\sum_{g=1}^{p}(\phi_g-\eta_\phi)^2}{2}$.

3. Update θ_0. The full conditional distribution of θ_l is a normal distribution with mean
$$\frac{1}{\frac{1}{10^4}+\frac{\sum_{g=1}^{p}(1-Z_{g\beta})}{\sigma_{\beta0}^2}}\frac{\sum_{g=1}^{p}(1-Z_{g\beta})\tilde{\beta}_{g0}}{\sigma_{\beta0}^2}$$
and variance $\left(\frac{1}{10^4}+\frac{\sum_{g=1}^{p}(1-Z_{g\beta})}{\sigma_{\beta0}^2}\right)^{-1}$.

4. Update $\sigma_{\beta0}^2$. The full conditional distribution of $\sigma_{\beta0}^2$ is an inverse gamma distribution with shape parameter $0.0001+0.5*\sum_{g=1}^{p}(1-Z_{g\beta})$ and scale parameter $0.0001+\frac{\sum_{g=1}^{p}(1-Z_{g\beta})(\tilde{\beta}_{g0}-\theta_0)^2}{2}$.

5. Update θ_l, $l = 1, 2$. The full conditional distribution of θ_l is a normal distribution with mean $\frac{1}{\frac{1}{10^4}+\frac{\sum_{g=1}^{p} Z_{g\beta}}{\sigma^2_{\beta l}}} \frac{\sum_{g=1}^{p}(1-Z_{g\beta})\tilde{\beta}_{gl}}{\sigma^2_{\beta l}}$ and variance $\left(\frac{1}{10^4}+\frac{\sum_{g=1}^{p} Z_{g\beta}}{\sigma^2_{\beta l}}\right)^{-1}$.

6. Update $\sigma^2_{\beta l}$, $l = 1, 2$. The full conditional distribution of $\sigma^2_{\beta l}$ is an inverse gamma distribution with shape parameter $0.0001+0.5*\sum_{g=1}^{p} Z_{g\beta}$ and scale parameter $0.0001+\frac{\sum_{g=1}^{p} Z_{g\beta}(\tilde{\beta}_{gl}-\theta_l)^2}{2}$.

7. Independently update $\tilde{\beta}_{gl}, \phi_g$ using slice sampler, for $l = 0, 1, 2$ and $g = 1, 2, \cdots, p$.

8. Update $Z_{g\beta}$, $g = 1, 2, \cdots, p$. The full conditional distribution of $Z_{g\beta}$ is:

$$\begin{aligned} & P(Z_{g\beta} = 1|\pi_{\beta 0}) \\ & = \frac{(1-\pi_{\beta 0})\prod_{i=1}^{2}\prod_{j=1}^{n_i} f_{NB}\left(M_{gij};\mu_{gi},\exp(\phi_g)\right)}{(1-\pi_{\beta 0})\prod_{i=1}^{2}\prod_{j=1}^{n_i} f_{NB}\left(M_{gij};\mu_{gi},\exp(\phi_g)\right)+\pi_{\beta 0}\prod_{i=1}^{2}\prod_{j=1}^{n_i} f_{NB}\left(M_{gij};\mu_{g0},\exp(\phi_g)\right)}, \end{aligned} \tag{3.14}$$

where $\mu_{gi} = \exp(\tilde{\beta}_{gi} + \log(s_{ij}))$ and $\mu_{g0} = \exp(\tilde{\beta}_{g0} + \log(s_{ij}))$.

9. Update $\pi_{\beta 0}$. The full conditional distribution of $\pi_{\beta 0}$ is a beta distribution, i.e.

$$\pi_{\beta 0}|\cdot \sim \text{Beta}\left(1+\sum_{g=1}^{p}(1-Z_{g\beta}), 1+\sum_{g=1}^{p} Z_{g\beta}\right). \tag{3.15}$$

The implementation of the regression model was adapted from Narisetty et al. (2014). More specifically, the full conditional distribution of $\boldsymbol{\gamma}^\star = (\tau_0, \tau_1, \boldsymbol{\gamma}) = (\tau, \gamma_1, \cdots, \gamma_p)$ is expressed as:

$$\boldsymbol{\gamma}^\star|\cdot \sim N\left((M^{\star\prime}M^\star + D_k^{-1})^{-1}M^{\star\prime}Y, \sigma_y^2(M^{\star\prime}M^\star + D_k^{-1})^{-1}\right), \tag{3.16}$$

where $M^\star = [1, \boldsymbol{T}', M]$, D_k^{-1} is a diagonal matrix with diagonal $(10^4, 10^4, \sigma^2_{\gamma Z_{1\gamma}}, \cdots, \sigma^2_{\gamma Z_{p\gamma}})$. A block updating algorithm (Ishwaran et al., 2005) is utilized to speed up computation time.

The conditional distribution of $Z_{g\gamma}$ is given by

$$P(Z_{g\gamma} = 1|\gamma, \sigma_y^2, \sigma_\gamma^2) = \frac{p_g\phi(\gamma_g; 0, \sigma_y^2\sigma_{\gamma 1}^2)}{p_g\phi(\gamma_g; 0, \sigma_y^2\sigma_{\gamma 1}^2) + (1-p_g)\phi(\gamma_g; 0, \sigma_y^2\sigma_{\gamma 0}^2)}. \tag{3.17}$$

The full conditional distribution of σ_y^2 is the inverse gamma distribution with shape parameter $0.0001 + 0.5 * n + p * 0.5$ and the scale parameter $0.0001 + 0.5 * \gamma' D_k \gamma + (Y - \tau_0 - \tau_1 T - M\gamma)'(Y - \tau_0 - \tau_1 T - M\gamma) * 0.5$, where D_k is a diagonal matrix with diagonal $(\sigma^2_{\gamma Z_{1\gamma}}, \cdots, \sigma^2_{\gamma Z_{p\gamma}})$.

3.2.4 Controlling the false discovery rate

As multiple hypotheses are tested simultaneously, it is crucial to control the false discovery rate (FDR). In this paper, we adapt the approach proposed by Newton et al. (2004) to control the posterior expected false discovery rate. Briefly, let $d_g \in \{0, 1\}$ indicate the gth OTU having mediation effect. A natural decision rule would rely on the posterior probability of having mediation effect, i.e. $d_g = I(\hat{v}_g > \alpha)$, for the given threshold α, where $\hat{v}_g = P(Z_{g\beta} Z_{g\gamma} \neq 0|\text{data})$. Let $D = \sum_{g=1}^{p} d_g$ denote the number of OTUs with mediation effect. Then the posterior expected FDR, the fraction of false positives, can be expressed as $\frac{1}{D} \sum_{g=1}^{p} (1 - \hat{v}_g) d_g$. Newton et al. (2004) proposed to pick up the value α so that the posterior expected FDR could achieve a desired level, say 0.05, i.e. $\frac{1}{D} \sum_{g=1}^{p} (1 - \hat{v}_g) d_g \leq 0.05$.

3.3 Results

3.3.1 Simulation studies

In this section, we will conduct several simulation studies to examine the performance of the proposed method. The performance is assessed with the receiver operating characteristic (ROC) curve, and FDR control. ROC curve, a figure with true positive rate versus the false positive rate, is widely used to measure the ranking of a method for signal detection.

In the following simulation studies, the model parameters used to generate data are estimated based on our model Equation (3.8) - (3.9) from the real data. For simplicity, we fix the normalization factor s_{ij} to be constant value one for all samples. Let $\mathcal{A} \subset \{1, 2, \cdots, p\}$ denote the indices of the differentially abundant OTUs, $\mathcal{L} \subset \{1, 2, \cdots, p\}$ denote the indices of the OTUs affecting the outcome Y after adjusting the treatment T, and $\mathcal{M} = \mathcal{A} \cap \mathcal{L}$ denote the indices of the mediating OTUs. The size of each set is denoted as $| \cdot |$, e.g. the size of mediating OTUs is denoted as $|\mathcal{M}|$.

3.3.2 First stage simulation study

First, we conduct simulation studies to numerically compare the screening method I with the screening method II.

Consider $n_1 = n_2 = 25$, $G = 4000$, $\pi_{\beta 0} = 0.8$, i.e. there are 800 differentially abundant OTUs. Let the size of mediators ($|\mathcal{M}|$) vary from 100 to 300 and fix the number of OTUs associated with the outcome (i.e. $|\mathcal{L}|$) to be 1000. Without loss of generality, we assume that the first $|\mathcal{A}|$ OTUs are differentially abundant, i.e. $\mathcal{A} = \{1, \cdots, n_{|\mathcal{A}|}\}$, $\mathcal{M} = \{1, 2, \cdots, |\mathcal{M}|\}$, $\mathcal{L} = \{1, \cdots, |\mathcal{M}|, n_{|\mathcal{A}|} + 1, \cdots, n_{|\mathcal{A}|} + |\mathcal{L}| - |\mathcal{M}|\}$. All simulations are based on 100 independent simulated datasets. In each simulation, we use 5000 burn-in iterations for the Gibbs sampler followed by 5000 iterations for estimating the posterior probabilities.

The simulation procedure is described as follows:

1. Generate the OTU count matrix M. For each OTU $g \notin \mathcal{M}$:

$$M_{gij} \sim \begin{cases} NB\left(\exp(\beta_{g0}), \exp(\phi_g)\right), & \text{if } Z_g = 0 \\ NB\left(\exp(\beta_{gi}), \exp(\phi_g)\right), & \text{if } Z_g = 1, g \notin \mathcal{M} \end{cases}$$

   where two cases are considered here: case i)$\beta_{g0} \sim N(0.3, 1.4)$, $\beta_{g1} \sim N(0.2, 1.3)$, $\beta_{g2} \sim N(0.3, 1.4)$, $\phi_g \sim N(-2, 1.4)$, and case ii) $\beta_{g0} \sim N(0.3, 0.5)$, $\beta_{g1} \sim N(0.2, 0.5)$, $\beta_{g2} \sim N(0.3, 0.5)$, $\phi_g \sim N(-2, 1.4)$.

2. Generate the outcome Y.

$$Y_{ij} \sim N\left(\tau_1 T_{ij} + \sum_{g=1}^{p} \gamma_g M_{gij}, 1\right), \tag{3.18}$$

   where $\tau_1 = 1$ and

$$\gamma_g = \begin{cases} 0, & \text{if } g \notin \mathcal{L} \\ \sim N(0, 0.04), & \text{if } g \in \mathcal{L} \setminus \mathcal{M} \end{cases},$$

   where γ_g are the coefficients for the abundances of OTUs after standardizing (i.e. with mean zero and variance one).

3. For the rest $g \in \mathcal{M}$, generate $|\mathcal{M}|$ OTUs from $NB(\exp(\beta_{gi}^m), \exp(\phi_g^m))$ based on Table 3.1, where each set of parameters has $|\mathcal{M}|/10$ OTUs.

Table 3.1: Parameters values for OTUs with mediation effects. Those values were estimated from real data based on our proposed model.

β_{g0}^m	β_{g1}^m	γ_g
3.7	0.76	-0.12
3.6	0.66	0.12
1.45	0.07	-0.1
0.12	9.06	-0.13
1.01	0.003	-0.17
9.1	3.83	-0.16
1.13	0.005	0.2
1.39	0.049	0.12
10.39	1856	0.22
6.87	2.72	0.15

For each simulated dataset, we apply both screening method I and screening method II respectively and compute the number of mediators that are correctly identified by each method. Figure 3.3 depicts the simulation results. Overall, our proposed novel screening method II are able to select more mediators compared to screening method I, especially for the case ii) where the treatment effects on the mediators are larger compared to other differentially abundant taxa (see the second row in the Figure 3.3).

3.3.3 Second stage simulation study

In this section, we conduct several simulation studies to assess the performance of our proposed causal mediation model (Equation (3.8) - (3.9)). We assume there are p OTUs selected from the first stage. The same procedure as described before is utilized to generate the simulated dataset, except in the first step, the set of parameters are set to be: $\beta_{g0} \sim N(0.4, 3.0)$, $\beta_{g1} \sim N(0.2, 2.2)$, $\beta_{g2} \sim N(-0.7, 2.4)$. The reason that we change the parameters values is that these parameters were estimated based on our proposed model applied to the real data after first stage screening. The sample size is fixed to be $n = 50$, and three cases with different size of total number of OTUs (i.e.

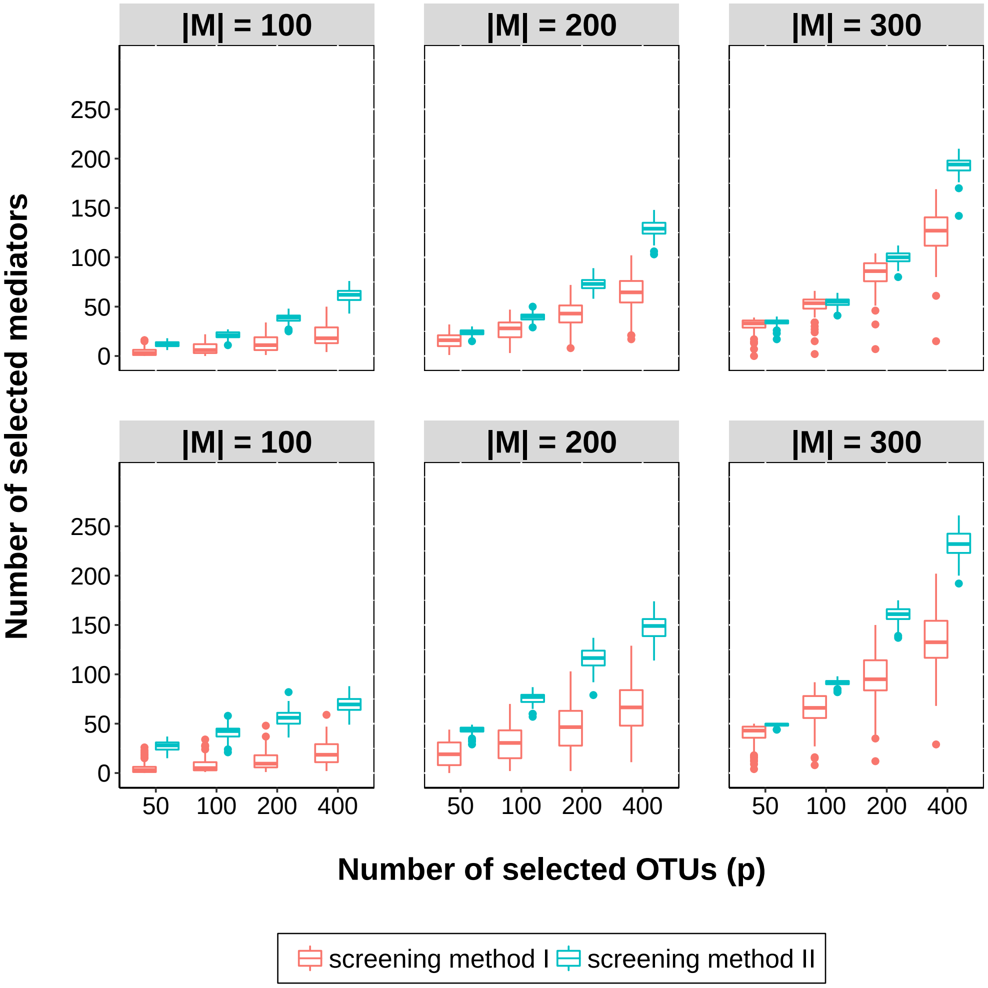


Figure 3.3: Compare the screening method I against screening method II varying size of mediators from 100 to 300. X-axis corresponds to the different size of p (i.e. the number of the OTUs prespecified to be selected) out of 4000 OTUs. Y-axis denotes the number of the mediating OTUs are selected. The first row shows the results for case i) and the second row presents the results for case ii)

$p = 50$, $p = 100$ and $p = 200$) are considered. For each case, we vary the size of mediators, number of differentially abundant OTUs, and the size of OTUs that are associated with the outcome. More specifically, we consider the following four scenarios: i) the treatment affects the relative abundance of some OTUs and all these differentially abundant OTUs impact the outcome, i.e. $|\mathcal{L} \setminus \mathcal{M}| = 0$ and $|\mathcal{A} \setminus \mathcal{M}| = 0$, ii) the treatment affects the relative abundance of some OTUs and there are additional five non-differentially abundant OTUs, except all differentially abundant OTUs, have an effect on the outcome, i.e. $|\mathcal{L} \setminus \mathcal{M}| = 5$ and $|\mathcal{A} \setminus \mathcal{M}| = 0$, iii) there are additional five differentially abundant OTUs that are not associated with the outcome, i.e. $|\mathcal{L} \setminus \mathcal{M}| = 0$ and $|\mathcal{A} \setminus \mathcal{M}| = 5$, and iv) there are additional five differentially abundant OTUs that are not associated with the outcome and additional five non-differentially abundant OTUs that have an effect on the outcome, i.e. $|\mathcal{L} \setminus \mathcal{M}| = 5$ and $|\mathcal{A} \setminus \mathcal{M}| = 5$. Hence, for each value p, there are total 12 simulation settings. For each setting, we conduct 100 simulations.

Figure 3.4 shows the ROC curves from our simulation studies. From the figure we can see that our proposed model works well in terms of ranking the taxa in all scenarios. As we increase the number of differentially abundant OTUs while keeping the size of mediators fixed, such as the third column and fourth column in Figure 3.4, the performance for $p = 50$ is slightly worse compared to the other cases $p = 100, 200$.

Figure 3.5 presents the summary of results from the FDR analysis. In all simulation settings, the false discovery proportions averaged over 100 simulated datasets are close or below the nominal FDR level, which indicates that FDR is well-controlled. As it shows that the performance of the proposed model with different size (i.e $p = 50, 100$ and $p = 200$) are similar in the case that all differentially abundant OTUs are mediators. However, as increasing the size of differentially abundant of OTUs, the FDR is not controlled but close to the desired level for $p = 200$ case, especially when the number of mediators increases to be 40.

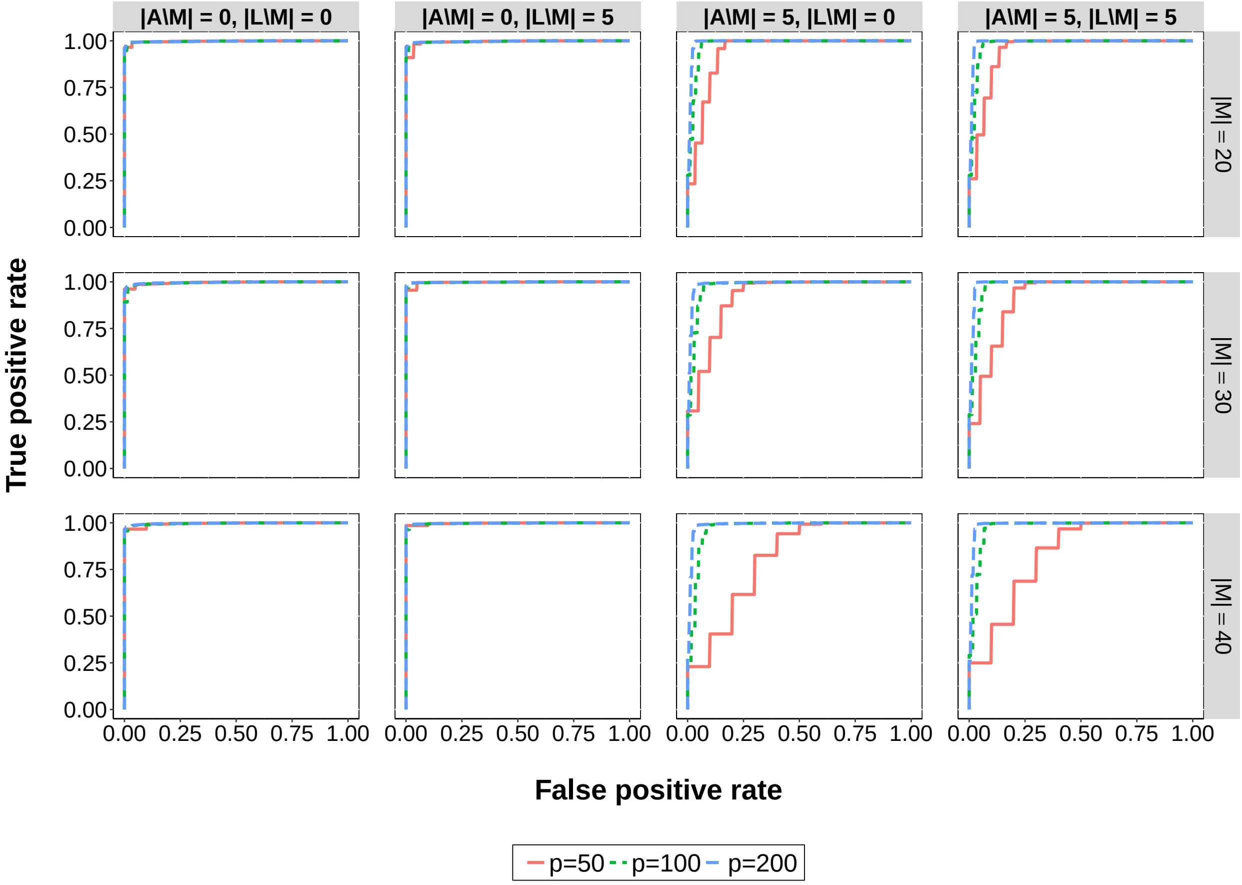


Figure 3.4: ROC curves averaged across 100 simulations for second stage under different scenarios. Each row corresponds to the size of mediators from 20 to 40. The first column denotes that $\mathcal{L} = \mathcal{A} = \mathcal{M}$, i.e. all differentially abundant OTUs have mediation effect and all OTUs that have an effect on outcome adjusting for treatment effect are mediators. The second column denotes that all differentially abundant OTUs have mediation effect and there are five additional non-differentially abundant OTUs that have an effect on outcome adjusting for treatment effect. The third column represents that there are five differentially abundant OTUs that are not associated with outcome adjusting for treatment effect. The forth column describes that there are five differentially abundant OTUs that are not associated with outcome adjusting for treatment effect and there are five non-differentially abundant OTUs that has effect on outcome adjusting for treatment effect.

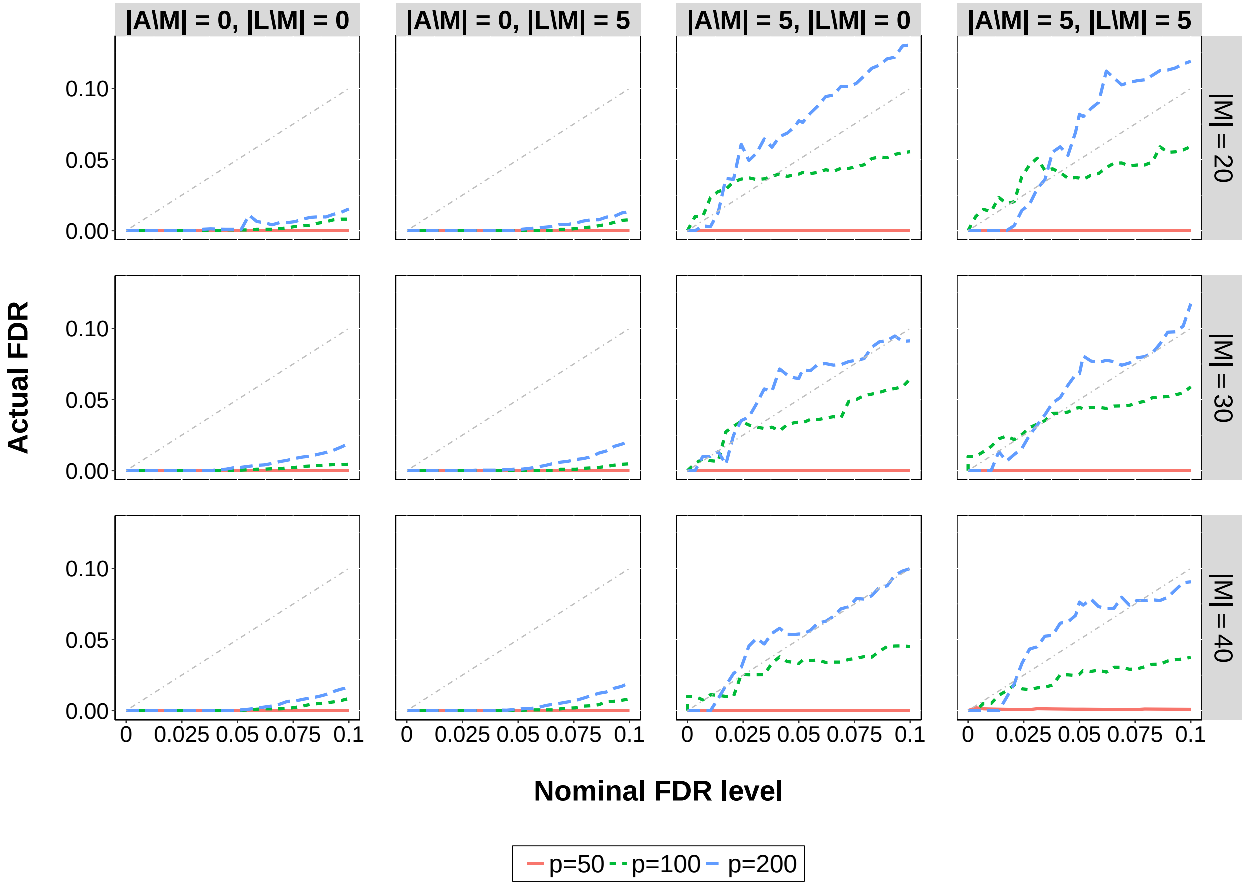


Figure 3.5: Examination of FDR control. Each row correspond to a different size of mediators (20, 30, and 40). Each line corresponds to the plot of average false discovery proportions across 100 simulated datasets versus the nominal level of FDR. The first column denotes that $\mathcal{L} = \mathcal{A} = \mathcal{M}$, i.e. all differentially abundant OTUs have meditation effect and all OTUs that have an effect on outcome adjusting for treatment effect are mediators. The second column denotes that all differentially abundant OTUs have mediation effect and there are five non-differentially abundant OTUs that have an effect on outcome adjusting for treatment effect. The third column represents that there are five differentially abundant OTUs that are not associated with outcome adjusting for treatment effect. The forth column describes that there are five differentially abundant OTUs that are not associated with outcome adjusting for treatment effect and there are five non-differentially abundant OTUs that have an effect on outcome adjusting for treatment effect.

Table 3.2: Number of potential mediating taxa detected using screening method I and screening method II in the real dataset when FDR was controlled at 5% level.

		Screening method I			Screening method II		
Metabolite	Collection date	$p = 50$	$p = 100$	$p = 200$	$p = 50$	$p = 100$	$p = 200$
Abscisic acid	July	0	0	0	17	5	0
Abscisic acid	September	3	0	0	13	80	0
Indole 3 acetic acid	July	0	0	0	19	0	0
Indole 3 acetic acid	September	0	4	0	0	0	0
Indole 3 carboxylic July	July	0	0	0	0	7	0
Jasmonic acid	July	1	0	0	0	25	0
Jasmonic acid	September	0	0	0	0	14	11
Phaseic acid	July	1	0	0	0	14	39
Salicylic acid	July	0	0	0	7	6	0
trans zeatin riboside	July	1	0	0	17	0	0
X12 oxo phytodienoic acid	September	0	0	0	0	20	0

3.4 Real Data Application

We applied our proposed approach to a real data example. In this study, Sorghum were grown under low and full nitrogen conditions and roots were sampled at two time points. Eight different metabolites were examined. In this project, to reduce the variability in the time and metabolites, we conducted separate mediation analysis for each metabolite at each time point. First, we filtered out the low abundance OTUs with zero counts across all samples at each time point. Then we applied our proposed procedure: selected the top p ($p = 50, 100, 200$) OTUs using screening method I and screening method II separately. Then applied causal mediation model to the selected OTUs with mediation effect on the the metabolite. Table 3.2 displayed the number of potential mediating OTUs that were detected using screening method I and screening II respectively when FDR was controlled at 5% level. Table 3.2 did not include that cases that no OTUs were detected with mediation effect. The results in Table 3.2 indicated that screening method II was able to select more number of potential mediators compared to screening method I.

3.5 Discussion

In this project, we developed a novel two-stage statistical approach to jointly test for the mediation effects. We proposed an original screening procedure followed with Bayesian variable selection to account for the high-dimensional overdispersed count nature of microbiome data. We evaluated the proposed method using simulation studies and applied to a real data example. Applying mediation analysis in microbiome data analysis can provide an important tool for helping researchers to identify novel taxa associated with observed outcomes. Other investigators also have assessed approaches to study the causality for the microbiome data; however we are unaware of others that have tested the mediation effects on individual taxa count level. Therefore, it is difficult to compare our method to others.

Although the proposed method focuses on testing the mediation effects, the estimation of indirect effect for individual OTU can also be obtained from the posterior distributions. Our method can be easily generalized to situations where there are more than two treatment levels, or where there are more than one treatment factors. In addition, typically, making causal inference relies on a couple of assumptions, such as no unmeasured confounders. It is usually difficult to assess these assumptions for complicated experiment design. Therefore, sensitivity analysis can be conducted to partially addresses such concern (Imai et al., 2010b).

CHAPTER 4. A HIERARCHICAL BAYESIAN LATENT CLASS MIXTURE MODEL WITH CENSORING FOR DETECTION OF TEMPORAL CHANGES IN ANTIBIOTIC RESISTANCE

The control of antimicrobial resistance (AMR) is a high priority for researchers and public health officials. One critical component of this control effort is surveillance for emerging or increasing resistance, as evidenced by the growth in the number and scale of surveillance programs around the world. Traditional detection of temporal changes in antibiotic resistance relies only on the analyses of proportion of resistance based on dichotomized Minimum Inhibitory Concentration (MIC) values, which ignores changes in the mean MIC below or above the MIC cutoff. We develop and validate a hierarchical Bayesian latent class mixture model approach which is able to detect temporal changes in the mean $\log_2$(MIC) as well as proportion of resistance together. Our model appropriately addresses challenges in analyses of AMR MIC values, including the left-, right- or interval-censoring and the latent class mixture distribution nature of the observed MIC values. We show that our method has less bias in mean estimation and more power in detection of changes compared to naive method ignoring censorship. We demonstrate our method with application to analyses of *Salmonella* enterica I,4,[5],12:i:- and *Salmonella* serotype Typhimurium with the antibiotic chloramphenical in CDC NARMS human dataset and show that evidence of temporal changes in mean $\log_2$(MIC) exist in spite of no changes or changes of adverse direction in the proportion of resistance.

4.1 Introduction

The control of antimicrobial resistance (AMR) is a high priority for researchers and public health officials. One critical component of this control effort is surveillance for emerging or increasing resistance, as evidenced by the number and scale of surveillance programs around the world

(Deckert et al., 2010; Gagliotti et al., 2011). In the United States (US), the National Antimicrobial Monitoring System (NARMS) has been in place since 1996 and includes data collected by the US Centers for Disease Control and Prevention (CDC), the US Food and Drug Administration (FDA) and the US Department of Agriculture (USDA) (US Centers for Disease Control , US Food and Drug Administration, US Department of Agriculture, 2016b). NARMS collects isolates of *Salmonella* spp., *Escherichia coli* and *Campylobacter* spp. . The antimicrobial resistance data for the CDC surveillance program are obtained from bacteria isolated from patients who attend public health departments or hospitals that are part of the CDC NARMS surveillance network. These bacterial isolates are tested for the ability to grow in the presence of antibiotics. Generally, the approach to AMR determination can be described as follows: each antibiotic is serially diluted and incubated with the bacterium, and the lowest dilution that inhibits bacterial growth is called the Minimum Inhibitory Concentration (MIC) (US Centers for Disease Control , US Food and Drug Administration, US Department of Agriculture, 2016b); then after determination of the MIC value, each isolate is categorized as susceptible or resistant to the tested antibiotic based on a clinical break-point value. Some antibiotics also have an intermediate category. In the US, breakpoints for NARMS data are provided by the Clinical and Laboratory Standards Institute. When breakpoints do not exist, the categorizations may be developed by other methods (Clinical and Laboratory Standards Institute, 2015; US Centers for Disease Control , US Food and Drug Administration, US Department of Agriculture, 2016b).

It is critically important for public health to monitor trends in antimicrobial resistance using AMR data generated by surveillance programs. One aim of surveillance programs is to enable detection of emerging resistance in a timely manner and to enable antimicrobial stewardship programs to be implemented properly and accurately (Atlanta and Human Service, 2016). To date, the predominate approach to assessing changes in AMR has focused on assessing changes in the proportion of resistant isolates for a particular antibiotic over time. Several statistical methods have been employed and all use testing approaches such as the Cochran-Armitage trend test, logistic regression model with time as a co-variate (Aerts et al., 2011; Cummings et al., 2016; Hanon et al., 2015), or the Mann-Kendall non-parametric method to test monotonic trend over time (US Centers for Disease Control , US Food and Drug Administration, US Department of Agriculture, 2016a). These statistical methods are based on dichotomized MIC data: resistant vs. non-resistant (susceptible and, when applicable, intermediate combined). However as Mazloom et al. (2017) pointed out, methods based on categorizations cause information loss. Further, the focus on proportion changes, means that the change in mean MIC of isolates occurring above or below the resistant breakpoints (MIC creep/decline) are not part of current surveillance monitoring (Ruiz et al., 2016). Similarly, reliance on dichotomized MIC data, means that correlations in mean MIC can not be readily monitored, despite the fact that such information would aid in the identification of emerging joint resistance patterns. Currently, these aspects of AMR surveillance data are not routinely monitored because accurate and robust approaches to estimation of the mean MIC are not available.

Appropriate mean MIC estimation must address the natural characteristic of MIC values, which are obtained from serial dilution experiments (Figure 4.1). Observed MIC values are often interval censored, for example, an observed MIC of 8 mg/ml for the organism A in Figure 4.1 actually implies the true MIC is > 4 mg/ml and ≤ 8 mg/ml. Observed MIC values might also be left or right censored at the starting or ending dilutions (Figure 4.2). This means the exact MIC values are unknown (Hamilton and Rinaldi, 1988). Estimation of the MIC mean without adjusting for censoring is biased and likely overestimates bacterial resistance to an antibiotic (Annis and Craig, 2005). An additional issue is the modeling of the underlying distribution of true MIC values. With respect to MIC data, bacteria samples typically consist of a mixture of two components, which weakly correspond to resistant and non-resistant populations. The true MIC value is believed to follow a log normal distribution (Mouton, 2002) for each component.

Others have previously reported approaches to estimation of the mean MIC. Van de Kassteele et al. (2012) proposed a model for estimation of mean $\log_2$(MIC) that incorporated the censored nature of the data and adjusted for such bias using the interval censored normal distribution as the underlying distribution. However, they did not consider the mixture of resistant and non-resistant

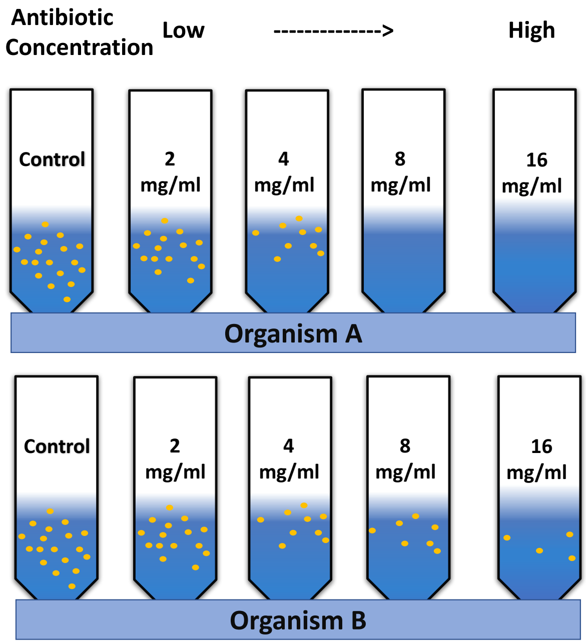


Figure 4.1: Schematic of Minimum Inhibitory Concentration determination. Yellow dots are bacteria growing in the blue broth with increasing concentrations at antibiotic(mg/ml). The growth of organism A is stopped by the concentration of 8 mg/ml i.e., the observed MIC is 8. The growth of organism B is not inhibited by even the highest dilution (dots in the 16 tube), so the observed MIC is "> 16".

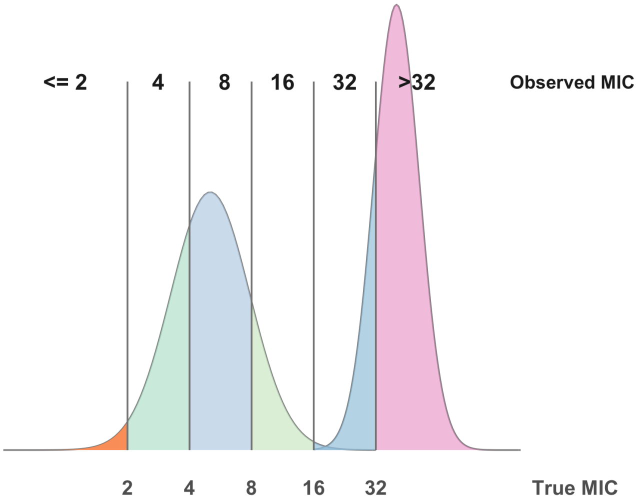


Figure 4.2: Illustration of censored data. Under the assumption that the underlying distribution of MIC value follows a bi-modal mixture normal distributions, the observed MIC value "<= 2" indicates the observed MIC value is left censored. In this example "> 32" indicates that the MIC value belongs to the right censored category. If the observed MIC value is reported as exactly equal to a concentration value, such as 2, 3, 8, 16, it is then considered as interval censored.

populations in observed data. Craig (2000) suggested that the underlying distribution of $\log_2$(MIC) can be modeled by a mixture of normal distributions, which weakly represents resistant and non-resistant populations. Jaspers et al. (2014b,a, 2016) described a similar approach to estimate the distribution of wild type and non-wild type bacteria populations determined by epidemiological cut-off rather than clinical breakpoints. These previously published approaches suggested that $\log_2$(MIC) mean could be estimated, however, none of the above approaches evaluated the temporal trend in mean $\log_2$(MIC) which is clearly a critical need for surveillance programs.

Therefore, building upon this prior work, the objective of this report is to describe an approach to detect temporal change in mean $\log_2$(MIC) using a Bayesian hierarchical model with a mixture of normal distributions for mean $\log_2$(MIC) from censored MIC data. The proposed model enables testing of temporal trends in mean $\log_2$(MIC) in resistant and non-resistant populations while retaining the ability to assess changes in the proportion of resistant bacteria over time. We illustrate our approach to assessing temporal changes in the mean $\log_2$(MIC) and compare the results to an approach ignoring censorship which we refer to as a "naive" approach. We also illustrate that the model can be used to assess changes in the proportion of resistant data. The examples use data obtained from *Salmonella* I,4,[5],12:i:- and *Salmonella* Typhimurium collected from 1996 to 2014 by the CDC NARMS surveillance programs. Inclusion of such analyses into current surveillance programs would add additional dimensions to monitoring AMR and increase the value of information extracted from surveillance systems.

4.2 Methods

4.2.1 Bayesian hierarchical model

To account for censoring statistically, each observed MIC value is assumed to represent an interval of true MIC values rather than a single discrete point value (see Figure 4.2). Here we suppose the observed MIC values can be transformed into a $\log_2$ scale.

Let y^*_{ij} denote the observed $\log_2$(MIC) value of the jth isolate at the ith year. Table 4.1 presents the conversion of observed MIC values to a continuous scale interval (l_{ij}, u_{ij}) for each isolate and each antibiotic (on log scale) with censorship. The corresponding unobserved true MIC value in the interval (l_{ij}, u_{ij}) is denoted as y_{ij}, where $i = 1, 2, \cdots, I$ and $j = 1, 2, \cdots, n_i$. Here n_i is the total number of isolates tested at the ith year, and I is the total number of years of interest (in CDC NARMS dataset $I = 19$). The distribution of y_{ij} can be considered as a mixture of the two bacterial populations (resistant and non-resistant). Let c_{ij} denote the unobserved random variable representing the bacterial population from which the MIC value y_{ij} is draw. A hierarchical model is proposed to fit the data:

$$\begin{aligned} c_{ij}\big|p_i &\overset{ind}{\sim} \text{Ber}(p_i), \\ y_{ij}\big|c_{ij}, \beta_{1i}, \beta_{0i}, \sigma_1^2, \sigma_0^2 &\overset{ind}{\sim} \begin{cases} N(\beta_{1i}, \sigma_1^2), & \text{if } c_{ij} = 1, \\ N(\beta_{0i}, \sigma_0^2), & \text{if } c_{ij} = 0, \end{cases} \end{aligned} \tag{4.1}$$

where $i = 1, 2, \cdots, I, j = 1, 2, \cdots, n_i$. Ber$(p)$ denotes a Bernoulli distribution with probability p, and $N(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2. $c_{ij} = 1$, with probability p_i, if y_{ij} comes from the resistant population, and $c_{ij} = 0$, with probability $1-p_i$, if y_{ij} comes from the non-resistant population. The parameters β_{1i} represent the mean $\log_2$(MIC) for the resistant component in ith year, and the parameters β_{0i} denotes the mean $\log_2$(MIC) for the non-resistant in ith year. Considering the heterogeneity of bacteria isolates in the CDC NARMS dataset, which might be caused due to different sampling collection methods from year to year or different labs were used to test isolates, a hierarchical modeling methodology is adopted to borrow information about $\log_2$(MIC) mean values across years and to integrate uncertainty gained from each individual year. In this way, it would adjust for multiple comparison testing as well based on Gelman's methodology (Gelman et al., 2012).

We model these populations averages with two independent normal distributions, respectively:

$$\begin{aligned} \beta_{1i}\big|\mu_1, \tau_1^2 &\overset{i.i.d}{\sim} N(\mu_1, \tau_1^2), \\ \beta_{0i}\big|\mu_0, \tau_0^2 &\overset{i.i.d}{\sim} N(\mu_0, \tau_0^2), \end{aligned} \tag{4.2}$$

where $i = 1, 2, \cdots, I$. The proportion of resistant population at ith year is modeled through a logit link function:

$$\begin{aligned} \text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \alpha_i, \quad i = 1, 2, \cdots, I, \\ \alpha_i \overset{iid}{\sim} N(\theta, \nu^2). \end{aligned} \tag{4.3}$$

Table 4.1: Examples of converting the Minimum inhibitory Concentration (MIC) reported in the CDC NARMS human data to the corresponding censorship in our model

MIC values(mg/ml)	y_{ij}^*	Censor Type	l_{ij}	u_{ij}
≤ 2	≤ 1	left censored	$-\infty$	1
$= 4$	$= 2$	interval censored	1	2
> 16	> 4	right censored	4	∞

4.2.1.1 Prior distributions for the Bayesian hierarchical model parameters

The full Bayesian analysis requires the joint prior distribution of all unknown parameters in the hierarchical model expression (4.1), (4.2) and (4.3). The vector of parameters of interest is $(\mu_1, \mu_0, \tau_1^2, \tau_0^2, \theta, \nu^2, \sigma_0^2, \sigma_1^2)^T$. Inverse gamma distribution is assigned to the variance of hierarchical part τ_1^2, τ_0^2, σ_1^2, σ_0^2 and ν^2 with both shape and scale parameters to be 0.0001. We assume an independent non-informative prior for the mean parameter for hierarchical part $\mu_l \sim N(0, 10^4)$, $l = 0, 1$, and $\theta \sim N(0, 10^4)$.

Bayesian hierarchical model can be fitted using two distinct data approaches. An all-data approach where data from all years are used to estimate the posterior distribution of the parameters, or a precedent-data approach where only data precedent to the years of interest is used. In the most recent year of any surveillance program these methods are equivalent, however, retrospective analysis of yearly changes will be different for the methods.

4.2.1.2 Implementation of Bayesian hierarchical model

Bayesian hierarchical model was implemented in a full Bayesian framework using Markov Chain Monte Carlo (MCMC) Gibbs sampling methodology (see Full Conditional Distribution Derivation). Gibbs sampling algorithm was adapted for censorship in finite mixture model (Komárek, 2009; Dey et al., 2000). The statistical inference for the linear model with time trend was based on joint posterior distributions. All computation was implemented using R software version 3.3.2 (R Core Team, 2016).

The initial values were estimated from AMR data collected over 19 years (1996 - 2014) for the NARMS program (`https://wwwn.cdc.gov/narmsnow/`), which is administered by the US Centers for Disease Control (CDC) NARMS collects data from numerous bacteria: *Salmonella* spp., *E.Coli*, *Campylobacter* spp., and *Shigella*. However, we limited our analysis to data related to *Salmonella* spp. . The isolate average of $\log_2$(MIC) for non-resistant population ignoring censorship in each tested year was calculated as the initial values for β_{0i}, and the isolate average of $\log_2$(MIC) values for resistant population in each year was evaluated as initial values for β_{1i} $(i = 1, 2, \cdots, I)$. The sample means and sample standard deviations of β_{1i} $(i = 1, 2, \cdots, I)$ were calculated as the initial values for μ_1 and τ_1, respectively. Similarly, the initial values for μ_0 and τ_0 were obtained from β_{0i} $(i = 1, 2, \cdots, I)$. The initial values for the proportion of resistant population p_i were calculated by the number of resistant isolates divided by the total number of isolates in each year. The initial values for α_i $(i = 1, 2, \cdots, I)$ were taken by performing an logit inverse transformation on those values of proportion of resistant population p_i.

10000 MCMC iterations were ran and after burn-in 3000 MCMC iterations were collected to make inference. The parameters in the model were estimated by the mean of posterior distribution. The 2.5th and 97.5th percentiles of those 3000 samples of posterior estimates were used for determination of the 95% credible interval (CI).

4.2.1.3 Full Conditional Distribution Derivation

We outlined the Gibbs sampling procedure as follows. Without specification, use notation $|\cdot$ to indicate full conditional distribution, i.e. conditional on all other parameters and the data:

1. Obtain draws of latent unobserved continuous observation y_{ij} by inverse cumulative distribution function sampling method, i.e. sampling from the full conditional normal distribution and constrained by the limits of interval l_{ij} and u_{ij}, $i = 1, 2, \cdots, I, j = 1, 2, \cdots, n_i$. More specifically,

   - if the observation is interval censored, y_{ij} is updated via

     $$y_{ij} = \Phi^{-1}\{\Phi(l_{ij}) + U[\Phi(u_{ij}) - \Phi(l_{ij})]\} \tag{4.4}$$

     where $U \sim \text{Unif}(0, 1)$, $\Phi(\cdot)$ is the Cumulative distribution function (cdf) function for standard normal distribution, and $\Phi^{-1}(\cdot)$ is the inverse of cdf function.

   - if the observation is right censored, y_{ij} is updated via

     $$y_{ij} = \Phi^{-1}\{\Phi(l_{ij}) + U[1 - \Phi(l_{ij})]\} \tag{4.5}$$

   - if the observation is left censored, y_{ij} is updated via

     $$y_{ij} = \Phi^{-1}\{\Phi(l_{ij}) + U\Phi(l_{ij})\} \tag{4.6}$$

2. Draw samples of c_{ij} from their full conditional distribution

   $$c_{ij}|\cdot \overset{ind}{\sim} \text{Ber}(h_{ij}), \tag{4.7}$$

where $h_{ij} = \frac{p_i\phi_1(y_{ij}|\beta_{1i},\sigma_1^2)}{p_i\phi_1(y_{ij}|\beta_{1i},\sigma_1^2)+(1-p_i)\phi_0(y_{ij}|\beta_{0i},\sigma_0^2)}$, $j = 1, 2, \cdots, n_i, i = 1, 2, \cdots, I$. $\phi(x|m, v)$ denotes the density function of normal distribution with mean m and variance v.

3. Sample the hierarchical MIC value μ_l from full conditional distribution

$$\mu_l|\cdot \overset{ind}{\sim} N(m, v), \tag{4.8}$$

where $v = \left(\frac{I}{\tau_l^2} + \frac{1}{c'}\right)^{-1}$, and $m = v\left(\frac{1}{\tau_l^2}\sum_{i=1}^{I}\beta_{li} + \frac{\mu'}{c'}\right)$, $l = 1, 0$, $\mu' = 0, c' = 10^4$.

4. Sample true MIC population mean in each year β_{li} from full conditional distribution

$$\beta_{li}|\cdot \sim N(M, V), \tag{4.9}$$

$$M = V\left(\frac{\mu_l}{\tau_l^2} + \frac{\sum_{j=1}^{n_i} I(c_{ij} = l)y_{ij}}{\sigma_l^2}\right), V = \left[\frac{\sum_{j=1}^{n_i} I(c_{ij} = l)}{\sigma_l^2} + \frac{1}{\tau_l^2}\right]^{-1}, \tag{4.10}$$

where $l = 1, 0$.

5. Sample variance for continuous variable y_{ij} σ_l^2 from full conditional distribution

$$\sigma_l^2|\cdot \sim \text{ IG}\left(\frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{n_i} I(c_{ij} = l), \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{n_i} I(c_{ij} = l)(y_{ij} - \beta_{li})^2\right), \tag{4.11}$$

where $l = 1, 0$.

6. Sample variance for hierarchical part τ_l^2 from full conditional distribution

$$\tau_l^2|\cdot \sim \text{ IG}\left(a + \frac{I}{2}, b + \frac{1}{2}\sum_{i=1}^{I}(\beta_{li} - \mu_l)^2\right), \tag{4.12}$$

where $l = 1, 0$, $a = 0.0001, b = 0.0001$.

7. Sample parameter θ from full conditional distribution

$$\theta|\cdot \sim N(M, V), \quad V = \left(\frac{I}{\nu^2} + \frac{1}{10^4}\right)^{-1}, M = V\left(\frac{\sum_{i=1}^{I}\alpha_i}{\nu^2}\right). \tag{4.13}$$

8. Sample parameter ν^2 from full conditional distribution

$$\nu^2|\cdot \sim \text{ IG}\left(\frac{I}{2}, \frac{1}{2}\sum_{i=1}^{I}(\alpha_i - \theta)^2\right). \tag{4.14}$$

9. Sample parameter for log-odds for proportion α_i for $i = 1, 2, \cdots, I$ using random walk

$$\pi(\alpha_i) \propto \sum_{j=1}^{n_i}\left(\frac{\exp(\alpha_i I(c_{ij} = 1))}{1 + \exp(\alpha_i)}\right)\exp\left(-\frac{(\alpha_i - \theta)^2}{2 * \nu^2}\right) \tag{4.15}$$

4.2.2 A Bayesian hierarchical model with linear time trend assumption

In this section, we are going to consider a simplifier model for the temporal trend. More specifically, a linear relationship with respect to time is assumed for the $\log_2$(MIC) value in the non-resistant and resistant population, and proportion of bacteria in the resistant population. The linear time trend model, using the same notation as the Bayesian hierarchical model (expression (4.1) and expression (4.3)) at the sample unit level, however, the mean $\log_2$(MIC) values are modeled as:

$$\begin{aligned}\beta_{1i} &\overset{ind}{\sim} N(\mu_{10} + \mu_{11} t_i, \tau_1^2) \\ \beta_{0i} &\overset{ind}{\sim} N(\mu_{00} + \mu_{01} t_i, \tau_0^2) \\ \alpha_i &= \gamma_0 + \gamma_1 t_i\end{aligned} \tag{4.16}$$

where $i = 1, 2, \cdots, I$. The parameter γ_1 in expression (4.16) can be interpreted as the odds ratio (OR) of proportion of bacteria in resistant population on logarithmic scale. In other words, $\exp(\gamma_1) > 1$ indicates the odds of being in the resistant population are increasing with time while $\exp(\gamma_1) < 1$ indicates the odds of isolates being in the resistant population are decreasing with time. The parameter μ_{11} represents the slope for the mean of $\log_2$(MIC) for the resistant population, while μ_{01} representes the slope on time for the mean $\log_2$(MIC) for non-resistant population. For either parameter, a positive value indicates the corresponding mean $\log_2$(MIC) is increasing with time.

Similarly, we conducted the Bayesian inference using this linear time trend model expression (4.16).

4.2.3 Model validation using simulation

4.2.3.1 Bias in mean $\log_2$(MIC) estimation

Simulation studies were conducted to document the extent of bias in mean $\log_2$(MIC) estimation using expression (4.1) compared to an approach to mean $\log_2$(MIC) estimation that ignores censorship. The exercise simulated a population consisting of observations either left-censored or interval censored i.e., a non-resistant population. More specifically, considering a hypothetical dataset normally distributed with known mean and variance, each realization/observed value is censored based on a serial experiment as would occur for AMR data. If a realization was less than the starting dilution, this realization was considered to be left censored, otherwise, it is interval censored type. For these data, we applied expression (4.1) to estimate the mean $\log_2$(MIC) and a naive method of mean $\log_2$(MIC) calculation that ignores censoring for each simulated data set. This procedure was repeated for 5000 independent datasets. Two metrics were examined, average bias and root mean square error. Average bias was the difference between estimated mean value and the true normal mean averaging across 5000 datasets. This metric measures the average repeated estimates deviates from the true value. Root mean square error (RMSE) was the square root of the averaging the squared difference between the estimated mean value and true normal mean. This metric indicates how close, on average, the estimate is to the true value. For both metrics, smaller value indicates less bias.

4.2.3.2 Detecting mean $\log_2$(MIC) differences

Here we assessed the power for detecting mean $\log_2$(MIC) difference using proposed Bayesian method and naive t-test at 5% significance level. Specifically, the 95% quantile credible interval from the posterior distribution using the proposed Beyesian method were used to make inference about changes in the mean $\log_2$(MIC). These results were compared to that obtained from a two-sample t-test where the mean $\log_2$(MIC) and standard error where obtained by naive calculation of the MIC using the censored data. 5000 simulation datasets were generated, with each containing five years of data, with equal sample size per year. Four cases of underlying normal distributions were considered: case 1: normal distributions with mean from 1.7 to 2.1 increased by 0.1 with standard deviance 1; case 2: normal distributions with mean from 1.7 to 2.7 increased by 0.25 with standard deviance 0.5; case 3: normal distributions with mean from 1.0 to 3.0 increased by 0.5 with standard deviance 1; case 4: normal distributions with mean from 1.0 to 3.0 increased by 0.5 with standard deviance 0.5. In all cases the cutoff value were set to be fixed at 2. Sample size varied from 5 to 30 in our simulation study. To summarize the comparison for each sample size we sum the number of t-tests with p value less than 0.05, and the number of credible intervals for the mean difference that did not included zero over 5000 simulated datasets, and reported the results as a proportion out of 5000 simulations.

4.2.4 Testing for temporal changes in mean $\log_2$(MIC) using NARMS data

A key promise of the approach proposed is the ability to monitor changes in mean $\log_2$(MIC) and test hypotheses about changes in the mean $\log_2$(MIC) through evaluating contrasts of mean $\log_2$(MIC) values with associated 95% credible intervals between consecutive years. Here we describe a variety of approaches to assessing temporal changes in $\log_2$(MIC) mean using the proposed models.

4.2.4.1 Comparison of consecutive year mean difference estimates

The mean $\log_2$(MIC) in the resistant and non-resistant populations respectively can be compared between consecutive years using the posterior sample distribution and 95% credible interval. If this interval contains zero, this indicates insufficient evidence to conclude a meaningful difference of the means between those two years. We illustrate this approach using all available data and then with only precedent data.

To evaluate changes in the mean $\log_2$(MIC) over years, we calculate the posterior mean and 95% CI for the mean difference between consecutive years, $\beta_{0i} - \beta_{0i-1}$, $i = 2, \cdots, I$. If this 95% CI contains zero for a particular consecutive year pair, implies that there is insufficient evidence to reject the null hypothesis of no difference in mean $\log_2$(MIC) of non-resistant population between that particular consecutive years. If all values in this 95% CI are greater than zero for a particular year pair, indicates that the mean $\log_2$(MIC) in non-resistant population is increasing, and vice versa. Concerns about multiple comparisons are already addressed with due to the incorporation of the Gelman methodology. The mean of the posterior estimates of the proportion of bacteria in the resistant population between consecutive years, $p_i - p_{i-1}$, for $i = 2, \cdots, I$ and 95% CIs, can be calculated to test the null hypothesis that no changes in proportion of resistant population between each consecutive year pairs, $H_{0i} : p_i - p_{i-1} = 0$, for $i = 2, \cdots, I$.

We compare the mean differences obtained from above approaches, to the results of a two-sample t-test calculated based on the naive method which calculates the mean of the observed $\log_2$(MIC) value and its standard error. Specifically, in order to compare the MIC value between two consecutive years, the Bayesian approach used the associated posterior sample distribution and 95% quantile credible interval to make inference based on inclusion of zero in the interval, while the naive t-test would just estimate the mean and standard error only using the MIC values from those two years under consideration.

4.3 Results

4.3.1 Estimating mean $\log_2$(MIC) using the Bayesian hierarchical model

Here we illustrate the results of the mean estimation using Bayesian hierarchical approach with all available data for two organisms from the CDC NARMS human data. The first organism is *Salmonella* enterica I,4,[5],12:i:- with the antibiotic chloramphenicol. Figure 4.3 displays the observed naive mean $\log_2$(MIC) for the 1029 *Salmonella* enterica I,4,[5],12:i:- isolates collected from 1996 to 2014. In Figure 4.3, the naive mean $\log_2$(MIC) was calculated by averaging the observed $\log_2$(MIC) values of non-resistant population based on breakpoints in each year. The observed MIC value for each isolate was approximated by ignoring censorship. For example, if the observed MIC value was ≤ 2, then value 2 was treated as the corresponding MIC value. As shown by the simulation exercises (more detail in the following section), these estimates of the mean $\log_2$(MIC) created without adjustment for censoring are biased and overestimates bacterial resistance to an antibiotic (Annis and Craig, 2005). More valid estimates of mean $\log_2$(MIC) based on the Bayesian model are presented in Table 4.2. Consistent with the observed data, we see that the proportion of bacteria classified as resistant does not show a consistent pattern. However, unique to our approach also provides estimates of the mean $\log_2$(MIC) in the non-resistant and resistant populations. Based on visual inspection, the mean $\log_2$(MIC) appears to be increasing. As these estimates are not biased by ignoring censoring, unlike those in Figure 4.3, these means estimates can be used for testing hypotheses about trend.

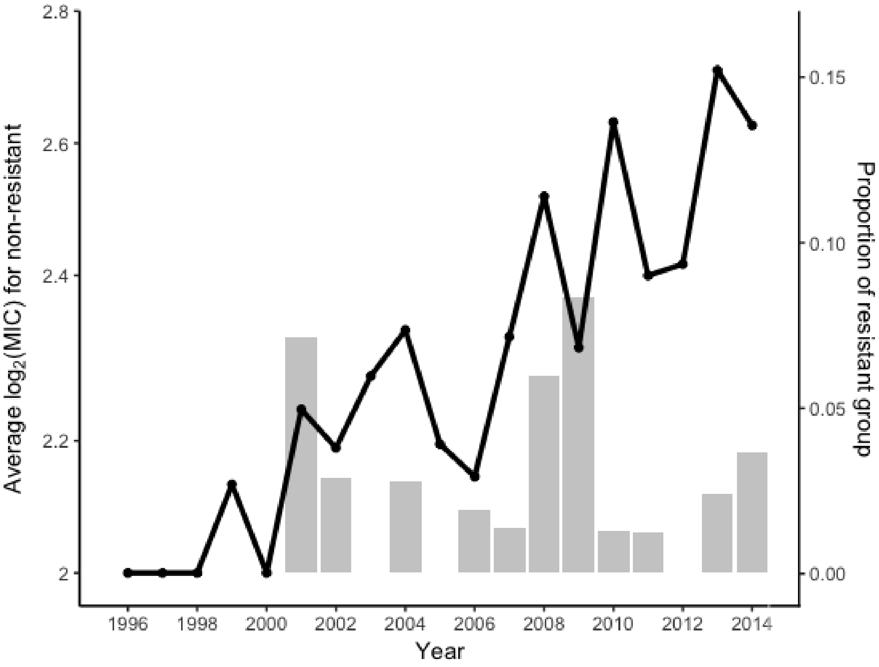


Figure 4.3: The changes in naive mean $\log_2$(MIC) (left y -axis) and proportion of bacteria in resistant group (right y-axis) for *Salmonella* enterica I,4,[5],12:i:- and antibiotic chloramphenicol from 1996 - 2014 (x-axis). The grey bars represent the percentage of resistant group in each year. Each dot is the naive mean value of $\log_2$(MIC) in the non-resistant group in each year ignoring censoring, i.e. if the observed MIC value was ≤ 2, value 2 was treated as the MIC value (Data from CDC NARMS).

Table 4.2: Posterior means and 95% credible intervals for the mean $\log_2$(MIC) (β_0 and β_1) and proportion (p) from Bayesian hierarchical model using *Salmonella* serotype I 4,[5],12:i:- and the antibiotic chloramphenicol. p_i represents the proportion of resistant component at ith year, β_{0i} represents the MIC mean on $\log_2$ scale for non-resistant population at ith year, and β_{1i} represents the MIC mean on $\log_2$ scale for resistant population at ith year

Year	$\widehat{\beta_{0i}}$	**95% CIs of** $\widehat{\beta_{0i}}$	$\widehat{\beta_{1i}}$	**95% CIs of** $\widehat{\beta_{1i}}$	$\widehat{p_i}$	**95% CIs of** $\widehat{p_i}$
1996	1.265	(0.240, 1.971)	6.257	(4.455, 8.264)	0.027	(0.004, 0.071)
1997	1.433	(0.745, 1.996)	6.243	(4.455, 8.188)	0.026	(0.005, 0.065)
1998	1.436	(0.328, 2.308)	6.267	(4.436, 8.386)	0.025	(0.003, 0.060)
1999	1.546	(1.100, 1.940)	6.235	(4.389, 8.263)	0.024	(0.003, 0.059)
2000	1.040	(0.069, 1.618)	6.236	(4.427, 8.208)	0.024	(0.004, 0.054)
2001	1.710	(1.414, 1.981)	6.670	(6.130, 7.891)	0.031	(0.008, 0.086)
2002	1.699	(1.533, 1.862)	5.651	(5.128, 7.874)	0.026	(0.006, 0.060)
2003	1.799	(1.647, 1.955)	6.206	(4.262, 8.185)	0.021	(0.003, 0.045)
2004	1.847	(1.691, 2.003)	7.100	(6.078, 8.188)	0.026	(0.007, 0.058)
2005	1.706	(1.536, 1.873)	6.224	(4.379, 8.214)	0.021	(0.003, 0.042)
2006	1.659	(1.567, 1.752)	6.391	(6.071, 8.017)	0.023	(0.007, 0.042)
2007	1.849	(1.741, 1.960)	7.119	(6.014, 7.813)	0.022	(0.005, 0.044)
2008	2.025	(1.921, 2.129)	6.168	(5.442, 8.474)	0.037	(0.016, 0.078)
2009	1.835	(1.725, 1.950)	5.809	(5.394, 7.914)	0.046	(0.019, 0.108)
2010	2.121	(2.016, 2.227)	5.672	(5.164, 7.759)	0.022	(0.005, 0.044)
2011	1.928	(1.822, 2.031)	6.483	(5.992, 8.064)	0.021	(0.005, 0.041)
2012	1.943	(1.861, 2.027)	6.171	(4.062, 8.086)	0.017	(0.002, 0.033)
2013	2.192	(2.109, 2.275)	5.792	(5.466, 8.363)	0.025	(0.009, 0.043)
2014	2.127	(2.040, 2.214)	5.895	(5.547, 7.513)	0.029	(0.011, 0.056)

We conducted the same analysis on *Salmonella* serotype Typhimurium with the antibiotic chloramphenicol. Figure 4.4 displays the observed estimates of the naive mean $\log_2$(MIC) for the 6773 isolates collected between 1996 and 2014 from the CDC NARMS human dataset. Table 4.3 presents the estimated mean $\log_2$(MIC) values and 95% CIs for the non-resistant population and the resistant population, as well as the proportion estimates of resistant population in each year. For example, in Table 4.3, for 1996 the naive estimate of mean $\log_2$(MIC) 2.215, while using the Bayesian model the estimated mean $\log_2$(MIC) for the non-resistant population was 1.662.

4.3.2 Model validation using simulation

4.3.2.1 Bias in mean $\log_2$(MIC) estimation

Here we compare our proposed method with the naive method of estimating the mean $\log_2$(MIC) ignoring censorship in terms of bias in estimation. Representative results of the first simulation

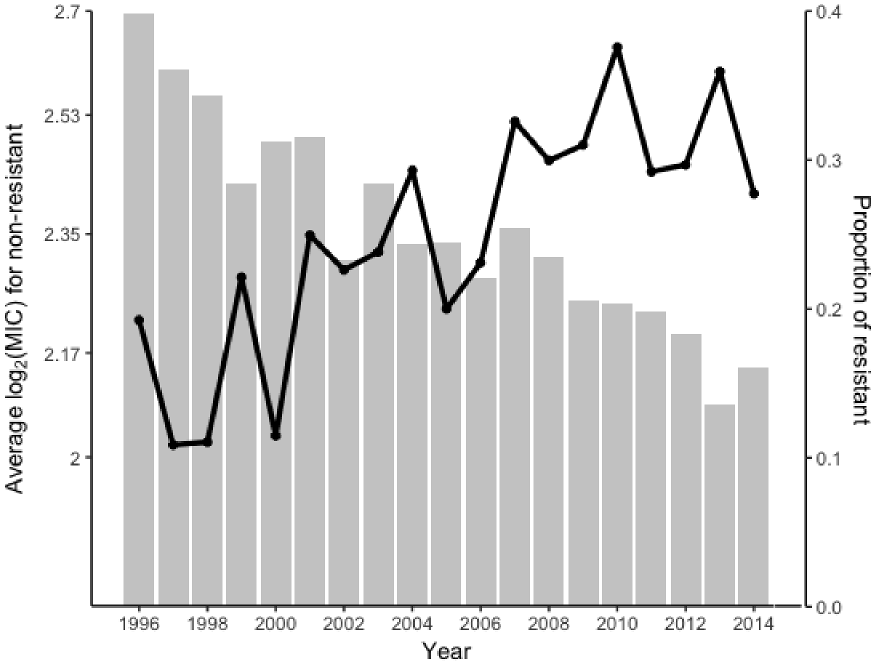


Figure 4.4: The changes in naive mean $\log_2$(MIC) (left y -axis) and proportion of bacteria in resistance group (right y-axis) for *Salmonella* serotype Typhimurium and the antibiotic chloramphenicol from 1996 - 2014 (x-axis). The grey bars represent the percentage of resistant group calculated using a naive method that ignores censoring in each year. Each dot is the naive mean $\log_2$(MIC) value in the non-resistant group in each year (Data from CDC NARMS).

Table 4.3: Posterior means and 95% credible intervals for the mean $\log_2$(MIC) (β_0 and β_1) and proportion (p) from Bayesian hierarchical model using *Salmonella* serotype Typhimurium and the antibiotic chloramphenicol. p_i represents the proportion of resistant component at ith year, β_{0i} represents the MIC mean on $\log_2$ scale for non-resistant population at ith year, and β_{1i} represents the MIC mean on $\log_2$ scale for resistant population at ith year

Year	$\widehat{\beta_{0i}}$	**95% CIs of** $\widehat{\beta_{0i}}$	$\widehat{\beta_{1i}}$	**95% CIs of** $\widehat{\beta_{1i}}$	$\widehat{p_i}$	**95% CIs of** $\widehat{p_i}$
1996	1.662	(1.564, 1.758)	5.301	(5.294, 5.310)	0.374	(0.327, 0.428)
1997	1.464	(1.357, 1.570)	5.006	(5.001, 5.012)	0.328	(0.279, 0.376)
1998	1.042	(0.864, 1.189)	5.145	(5.132, 5.178)	0.335	(0.290, 0.376)
1999	1.710	(1.630, 1.788)	5.219	(5.209, 5.242)	0.279	(0.239, 0.331)
2000	1.332	(1.209, 1.449)	5.013	(5.007, 5.028)	0.293	(0.247, 0.344)
2001	1.797	(1.715, 1.880)	5.088	(5.077, 5.099)	0.307	(0.262, 0.360)
2002	1.799	(1.734, 1.863)	5.301	(5.288, 5.309)	0.234	(0.198, 0.274)
2003	1.837	(1.773, 1.903)	5.075	(5.065, 5.085)	0.277	(0.236, 0.317)
2004	1.948	(1.883, 2.015)	5.208	(5.201, 5.215)	0.244	(0.202, 0.287)
2005	1.759	(1.697, 1.824)	5.028	(5.021, 5.035)	0.242	(0.202, 0.278)
2006	1.810	(1.748, 1.871)	5.069	(5.046, 5.083)	0.224	(0.184, 0.268)
2007	2.043	(1.980, 2.105)	5.118	(5.111, 5.130)	0.249	(0.212, 0.290)
2008	1.964	(1.899, 2.028)	5.076	(5.066, 5.084)	0.235	(0.192, 0.274)
2009	1.991	(1.926, 2.056)	5.053	(5.037, 5.064)	0.210	(0.174, 0.250)
2010	2.137	(2.072, 2.205)	5.143	(5.130, 5.167)	0.209	(0.169, 0.249)
2011	1.947	(1.879, 2.016)	5.159	(5.143, 5.198)	0.205	(0.162, 0.250)
2012	1.972	(1.897, 2.045)	5.124	(5.106, 5.139)	0.190	(0.148, 0.235)
2013	2.107	(2.040, 2.174)	5.097	(5.084, 5.133)	0.149	(0.116, 0.188)
2014	1.929	(1.855, 2.003)	5.012	(5.003, 5.032)	0.172	(0.129, 0.216)

exercise are presented in Table 4.4. As expected the results document that use of naive method to calculate the mean $\log_2$(MIC) is associated with more biased estimation of the mean with the extent of bias increasing as the proportion of censored observations increases. In this simulation, one sample of $\log_2$(MIC) values was generated from a normal distribution with mean increasing from 1.5 to 1.9, and left censored at value 1. As the sample size increased from 10 to 200, the average bias of the estimate from our Bayesian model decreases. In contrast, the average bias from the naive method was larger (between 0.52 and 0.59) than the Bayesian method (between -0.17 and 0.00), and did not decrease with sample size. The bias in the naive estimate is also always positive i.e. overestimation, compared the the Bayesian methods which tends to slightly underestimate the mean (negative estimates). As the mean $\log_2$(MIC) value increased from 1.5 to 1.9, which is equivalent to increasing the proportion of interval censored observations, the extent of average bias did not change meaningfully for either the Bayesian and naive method. Table 4.5 reports the performance of the model using root mean square error (RMSE), which measures the variation in bias. In all simulations, the RMSE for the Bayesian method becomes substantially smaller as the sample size increases, but this was not true for naive method.

Table 4.4: Assessment of estimation average bias for naive method ignoring censoring and the proposed Bayesian method. Average bias values based on 5000 simulated datasets. The variance of the normal distribution and the cutoff value for censoring are both set to be 1. The mean of normal distribution is increased from 1.5 to 1.9 by 0.1

	Mean 1.5		Mean 1.6		Mean 1.7		Mean 1.8		Mean 1.9	
Sample Size	naive method	Bayesian	naive method	Bayesian	naive method	Bayesian	naive method	Bayesian	naive method	Bayesian
10	0.573	-0.176	0.560	-0.143	0.557	-0.096	0.540	-0.084	0.534	-0.062
15	0.573	-0.078	0.557	-0.070	0.555	-0.046	0.537	-0.048	0.529	-0.036
20	0.569	-0.060	0.560	-0.043	0.547	-0.037	0.540	-0.028	0.533	-0.022
25	0.576	-0.035	0.561	-0.033	0.548	-0.026	0.537	-0.024	0.531	-0.018
30	0.572	-0.033	0.561	-0.025	0.548	-0.021	0.543	-0.013	0.532	-0.013
50	0.574	-0.015	0.560	-0.014	0.549	-0.011	0.538	-0.012	0.535	-0.003
100	0.573	-0.009	0.562	-0.004	0.550	-0.004	0.540	-0.003	0.529	-0.006
150	0.573	-0.006	0.559	-0.005	0.548	-0.004	0.536	-0.006	0.533	0.000
200	0.574	-0.003	0.559	-0.004	0.548	-0.002	0.540	-0.001	0.529	-0.003

Table 4.5: Assessment of root mean square error (RMSE) for naive method ignoring censoring and the proposed Bayesian method. RMSE value based on 5000 simulated datasets. The variance of the normal distribution and the cutoff value for censoring are both set to be 1. The mean of normal distribution is increased from 1.5 to 1.9 by 0.1

	Mean 1.5		Mean 1.6		Mean 1.7		Mean 1.8		Mean 1.9	
Sample Size	naive method	Bayesian	naive method	Bayesian	naive method	Bayesian	naive method	Bayesian	naive method	Bayesian
10	0.642	0.716	0.635	0.691	0.634	0.588	0.619	0.463	0.616	0.411
15	0.619	0.355	0.609	0.360	0.607	0.319	0.594	0.320	0.585	0.296
20	0.604	0.290	0.598	0.272	0.587	0.263	0.583	0.259	0.577	0.254
25	0.605	0.245	0.592	0.244	0.580	0.227	0.571	0.226	0.566	0.224
30	0.595	0.215	0.586	0.212	0.575	0.207	0.571	0.203	0.561	0.200
50	0.589	0.162	0.575	0.159	0.566	0.157	0.555	0.155	0.553	0.154
100	0.580	0.112	0.570	0.110	0.558	0.108	0.549	0.106	0.538	0.107
150	0.578	0.091	0.564	0.088	0.553	0.088	0.542	0.090	0.539	0.087
200	0.578	0.080	0.563	0.078	0.552	0.076	0.544	0.076	0.534	0.075

4.3.2.2 Detecting mean MIC differences

The results of a second simulation exercise, which aimed to evaluate if the Bayesian model was associated with greater power to detect changes in mean $\log_2$(MIC), are presented in Table 4.6. The proposed Bayesian model has greater power to detect true difference between consecutive mean $\log_2$(MIC) when compared to naive t-test ignoring censorship. This is evidenced by the higher proportion of the 5000 simulated datasets where the true differences in mean $\log_2$(MIC) were detected for the Baysian model compared to the naive t-test. For example, when the sample size was 10, the difference in means was detected in 6.1% of simulations using Bayesian method compared to only 3.7% of simulations when means calculated using the naive method and tested using a t-test. This advantage in power of difference detection of the Bayesian model is due to the fact that the model considered all data together to provide a more precise estimation of the means and difference between means, whereas the naive method only uses the two years being compared. This is another advantage of the proposed method compared to naive analyses ignoring censorship, in addition to having smaller bias in estimation.

Table 4.6: Power for detecting mean MIC difference using proposed Bayesian method and naive t-test at 5% significance level. 5000 simulation datasets were generated, with each containing five years of data, with equal sample size per year. Four cases of underlying normal distributions were considered: case 1: normal distributions with mean from 1.7 to 2.1 increased by 0.1 with standard deviance 1; case 2: normal distributions with mean from 1.7 to 2.7 increased by 0.25 with standard deviance 0.5; case 3: normal distributions with mean from 1.0 to 3.0 increased by 0.5 with standard deviance 1; case 4: normal distributions with mean from 1.0 to 3.0 increased by 0.5 with standard deviance 0.5. In all cases the cutoff value were set to be fixed at 2. Reported power in table is the proportion of cases that has been detected as significant between Year 2 and Year 1 by each method among 5000 simulated datasets

	Sample size 10		Sample size 15		Sample size 20		Sample size 25		Sample size 30	
	tTest	Bayes	tTest	Bayes	tTest	Bayes	tTest	Bayes	tTest	Bayes
Case 1	0.037	0.061	0.055	0.062	0.061	0.065	0.102	0.110	0.121	0.124
Case 2	0.110	0.178	0.180	0.205	0.234	0.245	0.294	0.302	0.332	0.348
Case 3	0.066	0.144	0.143	0.224	0.194	0.244	0.242	0.280	0.279	0.317
Case 4	0.076	0.209	0.166	0.224	0.195	0.233	0.247	0.275	0.300	0.332

4.3.3 Estimation of temporal changes in mean $\log_2$(MIC) using NARMS data

4.3.3.1 Estimation of mean difference between consecutive years

In this section we illustrate the utility of our proposed method of mean $\log_2$(MIC) estimation for surveillance programs by assessing changes in the mean $\log_2$(MIC) in non-resistant bacteria populations. The estimation and testing results for the changes in mean $\log_2$(MIC) value and the proportion of resistant sub population between consecutive years for *Salmonella* enterica I,4,[5],12:i:- with the antibiotic chloramphenical are presented in Table 4.7. For changes in mean $\log_2$(MIC) value of the non-resistant population, we present a comparison results from three testing methods: the naive two sample t-test, the proposed Bayesian model with all data up to date, and the proposed Bayesian model with only data precedent to the years of interest. The purpose of the analyses with precedent data is to provide an alternative comparison to naive t-test for timely identification of the yearly trend. These results are presented in the 2^{nd}, 3^{rd} and 4^{th} columns of Table 4.7. If the associated 95% CI of a difference in mean $\log_2$(MIC) did not contain the null value zero, this would be interpreted to imply evidence that the mean $\log_2$(MIC) differs between the consecutive years. It can be observed that although the inference based on Bayesian models would be similar to the two sample t-test for most of the consecutive year comparisons, the Bayesian model did identify a positive change as early as between 2000 and 2001, whereas the t-test only detected the trend from year 2006 - 2007. It is also critical to recall that the two sample t-test approach is based upon a biased estimate of the mean and variance of the mean calculated using the naive method. Although we present this comparison, it is important to acknowledge that the means used as the basis for this t-test results are biased and should not be employed for testing at all. The 5^{th} to 8^{th} columns presents additional results from the Bayesian model with all-data regarding the resistant population. We can see that there was insufficient evidence to reject the null hypothesis of no change in the proportion of resistant population between consecutive years in *Salmonella* I 4,[5],12:i:- and chloramphenicol as all of those intervals containing zero. These results agree with the patterns observed in Figure 4.3 that the proportion of resistance did not show any obvious trend while the mean $\log_2$(MIC) in non-resistant population had an obviously increasing trend over years. The Bayesian model was able to identify this increasing trend in mean MIC creep as early as in year 2001.

We conducted the same analyses on *Salmonella* serotype Typhimurium with the antibiotic chloramphenicol and results are presented in Table 4.8. For all but one pair, there was no evidence of meaningful changes between consecutive years detected for the proportion of bacteria in the resistant population since the associated 95% credible intervals contains zero. The proportion of bacteria in the resistant population was detected as decreasing from 2001 to 2002 as the associated

95% CI lies on the left side of zero. The mean $\log_2$(MIC) value for non-resistant population increased by 0.13 from 2012 to 2013 and decreased again in 2014 based on the Model 1 estimation. The results in Table 4.8 also implied that the mean $\log_2$(MIC) for non-resistant population was increasing between 2001 - 2002, 2006 - 2007 and 2009 - 2010 since all 95% CIs were greater than 0. The same increasing or decreasing trends within those time periods were also observed in Figure 4.4.

4.3.3.2 Estimation of linear time trend for mean $\log_2$(MIC) and comparison of linear time trend assessment for proportion resistant

Table 4.9 presents the results using the linear time trend model for I 4,[5],12:i:-. The estimated slope for the co-variate variable time for mean $\log_2$(MIC) value in non-resistant population was $0.058(0.029, 0.096)$, indicating that mean $\log_2$(MIC) for non-resistant population was increasing from 1996 to 2014, while there was no statistically significant evidence that the proportion of bacteria in the resistant population was changing over the years since the 95% credible interval for the coefficient for time for proportion of resistant population including zero $(0.006(-0.101, 0.109))$. With respect to the proportion analysis, the results of the Model 2 were consistent with the results from logistic regression model and in the Mann-Kendall test in terms of proportion of resistance. The two-sided p value provided from Mann-Kendall test was 0.09, implying that there was no enough evidence to reject the null hypothesis that there is no monotonic trend of the proportion of resistant sub population over the years at 5% significant level.

For Typhimurium with the antibiotic chloramphenicol, the results for linear time trend model are presented on Table 4.10. From Table 4.10, it can be seen that the results from our linear time trend, logistic regression, Mann-Kendall test, provide the same conclusion that the proportion of bacteria in the resistant population was decreasing from 1996 - 2014. using our linear time trend model, was also able to detect that the mean $\log_2$(MIC) value for non-resistant population was increasing over the years ($\hat{\mu}_{01} = 0.037(0.021, 0.054)$), such information could not be obtained from current methods which do not estimate mean MIC.

Table 4.7: Results for consecutive years comparison using data of *Salmonella* enterica I,4,[5],12:i:- with the antibiotic chloramphenical. Results for consecutive years comparison of mean $\log_2$(MIC) in non-resistant population ($\beta_{0i} - \beta_{0,i-1}$) were reported using the proposal Bayesian hierarchical model with all data, the proposed model with precedent data only (data up to the year of comparison) and naive t-test ignoring censorship. Other results regarding resistant population ($\beta_{1i} - \beta_{1,i-1}$) and proportion of resistant population ($p_i - p_{i-1}$) were based on all data. The Fisher exact test was used to test the null hypothesis that there is no difference in the proportion of resistant population between the ith year and the $i-1$th year using the breakpoints in CDC NARMS human data. Results significant at 5% level were made bold in the table

Year	$\widehat{\beta_{0i} - \beta_{0,i-1}}$ **all data**	$\widehat{\beta_{0i} - \beta_{0i-1}}$ **precedent-data**	**T-test of** $\beta_{0i} - \beta_{0i-1}$	$\widehat{\beta_{1i} - \beta_{1,i-1}}$	$\widehat{p_i - p_{i-1}}$	**Fisher exact test p value**	**Fisher odds ratio (95% confidence interval)**
1997-1996	0.168(-0.808,1.299)	Not estimable	Not estimable	-0.014(-2.329,2.358)	-0.001(-0.050,0.046)	1.000	Not estimable
1998-1997	0.003(-1.190,1.107)	Not estimable	Not estimable	0.024(-2.328,2.300)	-0.001(-0.048,0.040)	1.000	Not estimable
1999-1998	0.110(-0.867,1.274)	Not estimable	Not estimable	-0.033(-2.451,2.305)	-0.001(-0.039,0.037)	1.000	Not estimable
2000-1999	-0.506(-1.519,0.241)	Not estimable	Not estimable	0.001(-2.352,2.371)	-0.001(-0.039,0.033)	1.000	Not estimable
2001-2000	**0.670(0.018, 1.629)**	**1.204(0.007, 3.575)**	0.08	0.434(-1.450,2.190)	0.008(-0.029,0.067)	1.000	0.000(0.000,48.714)
2002-2001	-0.010(-0.324,0.320)	-0.029(-0.269,0.232)	0.70	-1.018(-1.418,1.341)	-0.005(-0.060,0.034)	0.506	2.450(0.030,201.388)
2003-2002	0.100(-0.126,0.323)	0.081(-0.095,0.268)	0.32	0.555(-1.944,2.425)	-0.005(-0.046,0.024)	1.000	Not estimable
2004-2003	0.047(-0.173,0.259)	0.08(-0.087,0.262)	0.58	0.894(-1.114,2.879)	0.005(-0.023,0.045)	1.000	0.000(0.000,40.084)
2005-2004	-0.140(-0.371,0.093)	-0.153(-0.336,0.015)	0.16	-0.876(-2.780,1.178)	-0.005(-0.044,0.020)	1.000	Not estimable
2006-2005	-0.047(-0.236,0.147)	-0.043(-0.0186,0.103)	0.55	0.167(-1.567,2.044)	0.002(-0.022,0.027)	1.000	0.000(0.000,17.395)
2007-2006	**0.190(0.047, 0.329)**	**0.162(0.057,0.277)**	**0.00**	0.728(-1.235,1.192)	-0.001(-0.027,0.023)	1.000	1.388(0.071,83.103)
2008-2007	**0.175(0.023, 0.329)**	**0.147(0.032,0.272)**	**0.01**	-0.952(-1.762,1.236)	0.015(-0.012,0.063)	0.222	0.232(0.005,2.140)
2009-2008	**-0.189(-0.337, -0.039)**	**-0.159(-0.281,-0.043)**	**0.01**	-0.358(-0.722,1.634)	0.008(-0.031,0.067)	0.757	0.716(0.165,2.946)
2010-2009	**0.286(0.128, 0.442)**	**0.253(0.125,0.386)**	**0.00**	-0.138(-1.528,0.878)	-0.024(-0.092,0.008)	0.063	6.436(0.753,302.291)
2011-2010	**-0.193(-0.339, -0.045)**	**-0.179(-0.31,-0.053)**	**0.01**	0.811(-0.991,1.573)	-0.001(-0.028,0.024)	1.000	1.051(0.013,83.463)
2012-2011	0.015(-0.119,0.150)	0.012(-0.108,0.135)	0.84	-0.312(-2.491,1.489)	-0.004(-0.028,0.013)	0.415	Not estimable
2013-2012	**0.249(0.130, 0.368)**	**0.243(0.13,0.361)**	**0.00**	-0.379(-2.155,1.738)	0.008(-0.011,0.034)	0.249	0.000(0.000,2.678)
2014-2013	-0.065(-0.193,0.054)	-0.065(-0.193,0.054)	0.29	0.103(-1.202,0.904)	0.004(-0.020,0.036)	0.708	0.651(0.093,3.937)

Table 4.8: Results for consecutive years comparison using data of *Salmonella* serotype Typhimurium with the antibiotic chloramphenical. Results for consecutive years comparison of mean $\log_2$(MIC) in non-resistant population ($\beta_{0i} - \beta_{0i-1}$) were reported using the naive t-test, the proposal Bayesian hierarchical model with all data, and the proposed model with precedent data only (data up to the year of comparison). Other results regarding resistant population ($\beta_{1i} - \beta_{1,i-1}$) and proportion of resistant population ($p_i - p_{i-1}$) were based on all data. The Fisher exact test was used to test the null hypothesis that there is no difference in the proportion of resistant population between the ith year and the $i-1$th year using the breakpoints in CDC NARMS human data. Results significant at 5% level were made bold in the table

Year	$\widehat{\beta_{0i} - \beta_{0,i-1}}$ **all data**	$\widehat{\beta_{0i} - \beta_{0i-1}}$ **precedent-data**	**T-test of** $\beta_{0i} - \beta_{0i-1}$	$\widehat{\beta_{1i} - \beta_{1,i-1}}$	$\widehat{p_i - p_{i-1}}$	**Fisher exact test p value**	**Fisher odds ratio (95% confidence interval)**
1997-1996	**-0.199(-0.343, -0.053)**	**-0.42(-0.649,-0.179)**	**0.00**	**-0.296(-0.303, -0.286)**	-0.047(-0.116,0.026)	0.545	1.103(0.811,1.501)
1998-1997	**-0.422(-0.619, -0.239)**	0.016(-0.29,0.258)	0.76	**0.140(0.124, 0.169)**	0.007(-0.057,0.072)	0.769	1.050(0.778,1.415)
1999-1998	**0.668(0.499, 0.857)**	**0.382(0.189,0.648)**	**0.00**	**0.073(0.053, 0.106)**	-0.056(-0.117,0.006)	0.229	1.211(0.892,1.648)
2000-1999	**-0.378(-0.526, -0.238)**	**-0.3(-0.577,-0.119)**	**0.00**	**-0.205(-0.221, -0.197)**	0.013(-0.053,0.080)	0.569	0.908(0.651,1.266)
2001-2000	**0.465(0.319, 0.617)**	**0.499(0.21,0.888)**	**0.00**	**0.075(0.056, 0.088)**	0.014(-0.053,0.081)	1.000	0.990(0.709,1.381)
2002-2001	0.002(-0.105,0.106)	-0.034(-0.125,0.058)	0.22	**0.213(0.202, 0.227)**	**-0.073(-0.136, -0.009)**	0.063	1.355(0.974,1.888)
2003-2002	0.038(-0.050,0.129)	0.027(-0.06,0.113)	0.54	**-0.226(-0.243, -0.205)**	0.043(-0.012,0.098)	0.213	0.819(0.595,1.126)
2004-2003	**0.111(0.015, 0.203)**	**0.127(0.04,0.212)**	**0.01**	**0.133(0.121, 0.146)**	-0.033(-0.090,0.027)	0.350	1.168(0.850,1.607)
2005-2004	**-0.189(-0.282, -0.097)**	**-0.206(-0.287,-0.127)**	**0.00**	**-0.180(-0.188, -0.172)**	-0.002(-0.059,0.050)	1.000	0.997(0.722,1.375)
2006-2005	0.050(-0.039,0.139)	0.068(-0.016,0.15)	0.07	**0.040(0.020, 0.055)**	-0.018(-0.076,0.034)	0.528	1.107(0.802,1.531)
2007-2006	**0.233(0.144, 0.323)**	**0.213(0.13,0.295)**	**0.00**	**0.049(0.038, 0.083)**	0.025(-0.032,0.084)	0.380	0.867(0.625,1.202)
2008-2007	-0.079(-0.170,0.011)	-0.06(-0.141,0.022)	0.16	**-0.042(-0.059, -0.033)**	-0.014(-0.069,0.042)	0.633	1.083(0.783,1.500)
2009-2008	0.027(-0.064,0.119)	0.029(-0.053,0.111)	0.58	**-0.023(-0.045, -0.007)**	-0.025(-0.082,0.032)	0.445	1.143(0.807,1.622)
2010-2009	**0.147(0.055, 0.237)**	**0.143(0.061,0.227)**	**0.00**	**0.090(0.071, 0.120)**	-0.001(-0.060,0.055)	1.000	1.010(0.699,1.461)
2011-2010	**-0.191(-0.286, -0.095)**	**-0.185(-0.271,-0.099)**	**0.00**	0.016(-0.011,0.036)	-0.004(-0.066,0.053)	0.925	1.026(0.699,1.510)
2012-2011	0.025(-0.074,0.123)	0.012(-0.076,0.098)	0.82	**-0.035(-0.076, -0.018)**	-0.015(-0.076,0.043)	0.689	1.086(0.718,1.647)
2013-2012	**0.135(0.038, 0.236)**	**0.132(0.044,0.22)**	**0.00**	-0.027(-0.040,0.010)	-0.041(-0.099,0.012)	0.192	1.347(0.859,2.121)
2014-2013	**-0.178(-0.276, -0.078)**	**-0.178(-0.276, -0.078)**	**0.00**	**-0.086(-0.115, -0.059)**	0.022(-0.035,0.076)	0.488	0.845(0.523,1.365)

Table 4.9: Linear model for time trend analysis for *Salmonella* serotype I, 4,[5],12:i:- and the antibiotic chloramphenicol from 1996 to 2014. The first column is posterior mean with corresponding 95% credible intervals for the parameters in the linear model to capture time trend. The second column is the generalized linear model with logit link when we treat the isolates as dichotomous based on breakpoints. The third column is the p value using Mann-Kendall trend test on the proportion of resistant componenet based on breakpoints. The 95% credible interval for the slope for susceptible β_{01} was greater than zero indicating the MIC mean values susceptible component were increasing over years, while 0 lied in the 95% credible interval for the slope for resistant component implies that there was no statistically changes in resistant component over years. Both our linear time trend model and the logistic regression model based on breakpoints didn't provide statistical significant evidence on the proportion of resistant component changes over years since the credible interval or confidence interval for slope for proportion of resistant component contained 0

	Posterior Mean (95%CIs)	**Ests(95% confidence interval)**	**Mann-Kendall test p value**
Intercept for susceptible	1.1267(0.6297, 1.4915)	NA	NA
Slope for susceptible	0.0576(0.0292, 0.0961)	NA	NA
Intercept for resistant	5.4721(4.8059, 7.4893)	NA	NA
Slope for resistant	0.0006(-0.0679, 0.1149)	NA	NA
Intercept for proportion for resistant	-3.7933(-5.1993,-2.3848)	-3.774(-5.273, -2.275)	NA
Slope for proportion for resistant	0.0073(-0.0878, 0.1110)	0.008(-0.093, 0.111)	0.0934

Table 4.10: Linear model for time trend analysis for *Salmonella* serotype Typhimurium and the antibiotic chloramphenicol from 1996 to 2014. The first column is posterior mean with corresponding 95% credible intervals for the parameters in the linear model to capture time trend. The second column is the generalized linear model with logit link when we treat the isolates as dichotomous based on breakpoints. The third column is the p value for Mann-Kendall trend test on the proportion of resistant componenet based on breakpoints in CDC NARMS human data. Again the 95% credible interval for the slope for susceptible component didn't contain zero, actually was greater than 0, indicated that the mean MIC values were increasing over the years. The 95% credible interval for slope for proportion for resistant component was lesser than zero, indicated that the proportions of resistant component were decreasing over years, which were consistent with logistic regression and Mann-kendall test

	Posterior Mean (95%CIs)	**Ests(95% confidence interval)**	**Mann-Kendall test p value**
Intercept for susceptible	1.4368(1.2367, 1.6222)	NA	NA
Slope for susceptible	0.0367(0.0207, 0.0543)	NA	NA
Intercept for resistant	5.1695(5.0209, 5.3258)	NA	NA
Slope for resistant	0.0114(-0.0020, 0.0247)	NA	NA
Intercept for proportion for resistant	-0.5202(-0.6259,-0.4242)	-0.48(-0.604, -0.376)	NA
Slope for proportion for resistant	-0.0608(-0.0705,-0.0509)	-0.06(-0.073, -0.051)	9.6827e-07

4.4 Discussion

Our goal with this project is to address a deficiency in the available approaches to analysis of AMR data using MIC values. Antibiotic resistance is a worldwide serious issue, and enormous resources are being devoted to monitoring the changes in MIC occurring. We propose here a Bayesian hierarchical model, and document that it enables additional information about changes in mean MIC to be monitored by surveillance while still allowing monitoring of the proportion of resistant bacteria. Therefore, the approach enables increased information to be gathered by surveillance programs. The proposed approach is founded on the concept of finding a valid and robust estimate of the mean MIC that addresses the censored nature of MIC data. We have documented that ignoring MIC creates a systematically biased estimate of the mean, i.e., the bias is not reduced by increasing sample size. In this paper, we proposed an approach that is not only able to detect the prevalence of resistant bacteria changes over time, but also to monitor the mean MIC value of the non-resistant and resistant populations over time. In this way, the proposed methods provide public health officials with an analysis approach to detect MIC creep or decline in the subpopuations of resistant or non-resistant populations. Considering the variation of the MIC values shown in the dataset, which might result from testing methods being different across states or labs, we allowed the mean $\log_2$(MIC) to vary across different years and applied the Bayesian hierarchical model to yield a more robust estimation via shrinking the estimates toward a common mean value in our Bayesian hierarchical model. This would lead to a more conservative conclusion when we are trying to make inference.

Under our framework, there are actually two levels of latent parameters. One is due to the censorship, the true underlying continuous values for each censored observed MIC value are unknown; the other is sub population of bacteria does the isolate arise from - resistant or non-resistant. By incorporating the censorship and heterogeneity of the data, the Bayesian hierarchical model proposed here would result in a more accurate estimation. This is a fundamental step to further statistical analysis and inference making involved.

However, there are also some assumptions in our analysis. We assumed normal distribution for non-resistant and resistant populations, although this assumption is supported by data observed and prior data (Craig, 2000). This assumption might be violated in some situations, and in those situations a more flexible semi-parametric or non-parametric methods, such as spline fitting, might be used to replace normal assumptions. Another assumption we made is the independence in proportion of resistant population and the mean of MIC value in each sub-population between years. If this assumption is not appropriate, then it would be possible to include a correlation structure between those variables.

In conclusion, we proposed a framework of analyzing such longitudinal $\log_2$(MIC) value data using Bayesian hierarchical approach, and not only estimated the mean of $\log_2$(MIC) values properly and accurately, but also conducted the hypothesis testing of $\log_2$(MIC) changes over years for given bacteria and antibiotics. Actually, our proposed framework can be easily extended to other interesting topics, such as studying the correlation between multiple antimicrobial (multi-drug resistance), and antibiotics resistance patterns among *Salmonella* isolated from different species, like swine, chicken and beef, and so on. In those cases, rather than a univariate mixture model, a multivariate mixed normal model would be further developed.

CHAPTER 5. GENERAL CONCLUSION

5.1 Summary

In this dissertation, we develop several statistical methods in metagenomics and antimicrobial resistance analysis. In chapter 2 and chapter 3, we present two statistical methods proposed for 16S rRNA microbiome data, aiming to answer the two central themes in metagenomics studies. One is to detect the potential microbiome whose relative abundance has been affected by the surrounding environmental conditions and the other is to establish the relationship between microbes and biological phenotype. While we study the application of the models to the specific microbiome data, the proposed frameworks will be applicable to fields requiring analysis of sparse high-dimensional count data. In addition, the mixture latent model developed to monitor the antimicrobial resistance that changes over time is also well suited to other areas involving censorship. This is discussed in chapter 4.

In chapter 2, we utilize a Hurdle model to address the excessive zeros issue in the microbiome data. To deal with the overdispersion issue, we employ the Poisson log normal hierarchical model to borrow information across taxa. Such hierarchical structure also accounts for the inherent variation and correlation between taxa which boosts the statistical power. Two independent latent indicators are introduced in the model to adjust for the multiple testing problems. Through comprehensive simulation studies, our proposed method outperforms the existing methods in terms of statistical power and false discovery rate control.

In chapter 3, we explore the relationship between microbial community, biological outcome, and environmental condition through a causal mediation approach. To accommodate the setting with a large number of count mediators and a small sample size, we develop a novel sure independence screening procedure and then adopt a Bayesian variable selection strategy to select key differentially abundant taxa that are associated with the outcome after adjusting the treatment effect. Simulation studies illustrate the performance of the proposed method.

In chapter 4, we propose a hierarchical Bayesian latent class mixture model to monitor the temporal trends of the prevalence of both resistant and non-resistant bacteria using minimum inhibitory concentration values. By taking the censorship into account, simulation studies demonstrate that our method has less bias in mean estimation and more power in detection of changes compared to a naive method ignoring censorship.

5.2 Future Work

Although we have demonstrated our proposed statistical methods using extensive simulation studies, there are still enumerous open problems that need further investigation.

In chapter 2, in order to classify OTUs into two groups (differentially abundant vs non-differentially abundant), two independent latent indicators are generated for each OTU. One is to indicate whether the occurrence of OTU is impacted by the treatment conditions, while the other represents whether the treatment conditions have an effect on the OTU abundance. This states that, for an OTU of interest, if two independent Bernoulli trials with two different prior success probabilities both take value 1, then this OTU is differentially abundant regardless of the results for any other OTUs. However, from a biological perspective, taxa interact with each other in the community to perform biological functions. That implies encoding such interaction into the hypothesis testing is feasible. A potential solution would be use the idea of the Markov random-field model proposed by Wei and Pan (2010). In their paper, they used the Gaussian-Markov random field model to incorporate the gene network interaction information in the hypothesis testing. It might be useful to adapt a this Markov network idea here because the microbiome network is similar to gene network (Layeghifard et al., 2017). The second extension of this project is from a computation point of view. Although Bayesian methodology is attractive in the analysis of metagenomics, a computationally efficient algorithm to speed up the Bayesian analysis is necessary. Therefore, parallelized computation algorithm is of importance in practice.

As for chapter 3, there are two extensions which might be further developed that are motivated by the real data example. The first one is to build a multivariate causal mediation model. In our data example, eight metabolites were measured from the same root. Therefore it is desirable to develop an appropriate multivariate causal mediation framework while accounting for the high-dimensional overdispersed count nature of microbiome data. The second extension is to dynamically test the mediation effect. As the microbiome composition varies over time, as observed in the real data, it is necessary to develop new methods to capture the time effect for each taxon. In addition, from a theoretical perspective, there are some possible directions for future research. For instance, no unmeasured confounding assumption is crucial for identifiability in the causal inference framework. Due to the biological complexity and the inherent hierarchical phylogenetic structure among taxa, the correlation between taxa sets might violate the identifiability assumptions. To address the potential confounding issue, further statistical methodology is required to account for the high-dimensional count mediators as the traditional theory proposed by Imai et al. (2010b) is limited to a single mediator case.

The Bayesian model proposed in chapter 4 can be also extended to study the correlation between multiple antimicrobial (i.e. multi-drug resistance problem), and antibiotic resistance patterns among *Salmonella* isolated from different species, like pork, chicken, beef, and so on.

BIBLIOGRAPHY

Aerts, M., Faes, C., and Nysen, R. (2011). Development of statistical methods for the evaluation of data on antimicrobial resistance in bacterial isolates from animals and food. *EFSA Supporting Publication*, 8(12):EN–186, 77 pp.

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

Annis, D. H. and Craig, B. A. (2005). Statistical properties and inference of the antimicrobial MIC test. *Stat. Med.*, 24(23):3631–3644.

Atlanta, G. U. D. o. H. and Human Service, C. (2016). National Antimicrobial Resistance Monitoring System for Enteric Bacteria (NARMS): human isolates surveillance final report for 2014.

Baron, R. M. and Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51(6):1173.

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.

Berg, G., Rybakova, D., Grube, M., and Köberl, M. (2015). The plant microbiome explored: implications for experimental botany. *Journal of experimental botany*, 67(4):995–1002.

Berk, R., Brown, L., Buja, A., Zhang, K., Zhao, L., et al. (2013). Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837.

Caprioli, A., Busani, L., Martel, J. L., and Helmuth, R. (2000). Monitoring of antibiotic resistance in bacteria of animal origin: epidemiological and microbiological methodologies. *International Journal of Antimicrobial Agents*, 14(4):295–301.

Chaparro, J. M., Sheflin, A. M., Manter, D. K., and Vivanco, J. M. (2012). Manipulating the soil microbiome to increase soil health and plant fertility. *Biology and Fertility of Soils*, 48(5):489–499.

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, J., King, E., Deek, R., Wei, Z., Yu, Y., Grill1, D., and Ballman, K. (2017). An omnibus test for differential distribution analysis of microbiome sequencing data. *Bioinformatics*, page btx650.

Chen, J. and Li, H. (2013). Kernel methods for regression analysis of microbiome compositional data. In *Topics in Applied Statistics*, pages 191–201. Springer.

Clemente, J. C., Jansson, J., and Valiente, G. (2011). Flexible taxonomic assignment of ambiguous sequencing reads. *BMC bioinformatics*, 12(1):8.

Clinical and Laboratory Standards Institute (2015). Performance standards for antimicrobial susceptibility testing; twenty-fifth informational supplement (M100-S25). http://www.facm.ucl.ac.be/intranet/CLSI/CLSI-2015-M100-S25-original.pdf.

Council, N. R. et al. (2007). *The new science of metagenomics: revealing the secrets of our microbial planet*. National Academies Press.

Council, N. R. et al. (2012). *Renewable fuel standard: potential economic and environmental effects of US biofuel policy*. National Academies Press.

Craig, B. A. (2000). Modeling approach to diameter breakpoint determination. *Diagn. Microbiol. Infect. Dis.*, 36(3):193–202.

Cruz, M. G., Peters, G. W., and Shevchenko, P. V. (2014). *Fundamental aspects of operational risk and insurance analytics: A handbook of operational risk*. John Wiley & Sons.

Cummings, K. J., Perkins, G. A., Khatibzadeh, S. M., Warnick, L. D., Aprea, V. A., and Altier, C. (2016). Antimicrobial resistance trends among Salmonella isolates obtained from horses in the northeastern United States (2001–2013). *Am. J. Vet. Res.*, 77(5):505–513.

Deckert, A., Gow, S., Rosengren, L., Leger, D., Avery, B., Daignault, D., Dutil, L., Reid-Smith, R., and Irwin, R. (2010). Canadian integrated program for antimicrobial resistance surveillance (CIPARS) farm program: results from finisher pig surveillance. *Zoonoses Public Health.*, 57(s1):71–84.

Dey, D. K., Ghosh, S. K., and Mallick, B. K. (2000). *Generalized linear models: A Bayesian perspective*. CRC Press.

Dini-Andreote, F. and Raaijmakers, J. M. (2018). Embracing community ecology in plant microbiome research. *Trends in plant science*.

Do, K.-A., Müller, P., and Tang, F. (2005). A bayesian mixture model for differential gene expression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):627–644.

Edwards, J., Johnson, C., Santos-Medellín, C., Lurie, E., Podishetty, N. K., Bhatnagar, S., Eisen, J. A., and Sundaresan, V. (2015). Structure, variation, and assembly of the root-associated microbiomes of rice. *Proceedings of the National Academy of Sciences*, 112(8):E911–E920.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*, volume 38. Siam.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Fierer, N. (2017). Embracing the unknown: disentangling the complexities of the soil microbiome. *Nature Reviews Microbiology*, 15(10):579.

Fritz, J. V., Desai, M. S., Shah, P., Schneider, J. G., and Wilmes, P. (2013). From meta-omics to causality: experimental models for human microbiome research. *Microbiome*, 1(1):14.

Gagliotti, C., Balode, A., Baquero, F., Degener, J., Grundmann, H., Gür, D., Jarlier, V., Kahlmeter, G., Monen, J., Monnet, D., et al. (2011). Escherichia coli and Staphylococcus aureus: bad news and good news from the European Antimicrobial Resistance Surveillance Network (EARS-Net, formerly EARSS), 2002 to 2009. *Euro. Surveill.*, 16(17):20–24.

Gelli, M., Duo, Y., Konda, A. R., Zhang, C., Holding, D., and Dweikat, I. (2014). Identification of differentially expressed genes between sorghum genotypes with contrasting nitrogen stress tolerance by genome-wide transcriptional profiling. *BMC genomics*, 15(1):179.

Gelman, A., Hill, J., and Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *J. Res. Educ. Eff.*, 5(2):189–211.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472.

Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):499–517.

Gilbert, R. O. (1987). *Statistical methods for environmental pollution monitoring.* John Wiley & Sons.

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

Hamilton, M. A. and Rinaldi, M. G. (1988). Descriptive statistical analyses of serial dilution data. *Stat. Med.*, 7(4):535–544.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., and Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & biology*, 5(10):R245–R249.

Hanon, J. B., Jaspers, S., Butaye, P., Wattiau, P., Méroc, E., Aerts, M., Imberechts, H., Vermeersch, K., and Van der Stede, Y. (2015). A trend analysis of antimicrobial resistance in commensal Escherichia coli from several livestock species in Belgium (2011–2014). *Prev. Vet. Med.*, 122(4):443–452.

Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach.* Guilford Publications.

Heilbron, D. (1989). Generalized linear models for altered zero probabilities and overdispersion in count data. *Unpublished Technical report, University of California, San Francisco, Department of Epidemiology and Biostatistics.*

Hilbe, J. M. (2011). *Negative binomial regression.* Cambridge University Press.

Huang, Y.-T. and Pan, W.-C. (2016). Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. *Biometrics*, 72(2):402–413.

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., Creasy, H. H., Earl, A. M., FitzGerald, M. G., Fulton, R. S., et al. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207.

Imai, K., Keele, L., and Tingley, D. (2010a). A general approach to causal mediation analysis. *Psychological methods*, 15(4):309.

Imai, K., Keele, L., and Yamamoto, T. (2010b). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical science*, pages 51–71.

Ishwaran, H., Rao, J. S., et al. (2005). Spike and slab variable selection: frequentist and bayesian strategies. *The Annals of Statistics*, 33(2):730–773.

Jaspers, S., Aerts, M., Verbeke, G., and Beloeil, P. A. (2014a). Estimation of the wild-type minimum inhibitory concentration value distribution. *Stat. Med.*, 33(2):289–303.

Jaspers, S., Aerts, M., Verbeke, G., and Beloeil, P. A. (2014b). A new semi-parametric mixture model for interval censored data, with applications in the field of antimicrobial resistance. *Comput. Stat. Data Anal.*, 71:30–42.

Jaspers, S., Lambert, P., and Aerts, M. (2016). A Bayesian approach to the semiparametric estimation of a minimum inhibitory concentration distribution. *Ann. Appl. Stat.*, 10(2):906–924.

Ji, H. and Liu, X. S. (2010). Analyzing omics data using hierarchical models. *Nature biotechnology*, 28(4):337–340.

Jiang, H., An, L., Lin, S. M., Feng, G., and Qiu, Y. (2012). A statistical framework for accurate taxonomic assignment of metagenomic sequencing reads. *PLoS One*, 7(10):e46450.

Judd, C. M. and Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation review*, 5(5):602–619.

Kembel, S. W., OConnor, T. K., Arnold, H. K., Hubbell, S. P., Wright, S. J., and Green, J. L. (2014). Relationships between phyllosphere bacterial communities and plant functional traits in a neotropical forest. *Proceedings of the National Academy of Sciences*, 111(38):13715–13720.

Köberl, M., Müller, H., Ramadan, E. M., and Berg, G. (2011). Desert farming benefits from microbial potential in arid soils and promotes diversity and plant health. *PLoS One*, 6(9):e24452.

Komárek, A. (2009). A new R package for Bayesian estimation of multivariate normal mixtures allowing for selection of the number of components and interval-censored data. *Comput. Stat. Data Anal.*, 53(12):3932–3947.

Koopman, J., Howe, M., Hollenbeck, J. R., and Sin, H.-P. (2015). Small sample mediation testing: Misplaced confidence in bootstrapped confidence intervals. *Journal of Applied Psychology*, 100(1):194.

Kristiansson, E., Hugenholtz, P., and Dalevi, D. (2009). Shotgunfunctionalizer: an r-package for functional comparison of metagenomes. *Bioinformatics*, 25(20):2737–2738.

Krull, J. L. and MacKinnon, D. P. (1999). Multilevel mediation modeling in group-based intervention studies. *Evaluation review*, 23(4):418–444.

La Rosa, P. S., Brooks, J. P., Deych, E., Boone, E. L., Edwards, D. J., Wang, Q., Sodergren, E., Weinstock, G., and Shannon, W. D. (2012). Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PloS one*, 7(12):e52078.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

Layeghifard, M., Hwang, D. M., and Guttman, D. S. (2017). Disentangling interactions in the microbiome: a network perspective. *Trends in microbiology*, 25(3):217–228.

Lempers, F. B. (1971). Posterior probabilities of alternative linear models.

Li, H. (2015). Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application*, 2:73–94.

Lin, W., Shi, P., Feng, R., and Li, H. (2014a). Variable selection in regression with compositional covariates. *Biometrika*, 101(4):785–797.

Lin, X., Genest, C., Banks, D. L., Molenberghs, G., Scott, D. W., and Wang, J.-L. (2014b). *Past, Present, and Future of Statistical Science*. CRC Press.

Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014). A significance test for the lasso. *Annals of statistics*, 42(2):413.

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.

Lundberg, D. S., Lebeis, S. L., Paredes, S. H., Yourstone, S., Gehring, J., Malfatti, S., Tremblay, J., Engelbrektson, A., Kunin, V., Del Rio, T. G., et al. (2012). Defining the core arabidopsis thaliana root microbiome. *Nature*, 488(7409):86.

MacKinnon, D. P., Fairchild, A. J., and Fritz, M. S. (2007a). Mediation analysis. *Annu. Rev. Psychol.*, 58:593–614.

MacKinnon, D. P., Lockwood, C. M., Brown, C. H., Wang, W., and Hoffman, J. M. (2007b). The intermediate endpoint effect in logistic and probit regression. *Clinical Trials*, 4(5):499–513.

Marston, H. D., Dixon, D. M., Knisely, J. M., Palmore, T. N., and Fauci, A. S. (2016). Antimicrobial resistance. *Jama*, 316(11):1193–1204.

Mazloom, R., Jaberi-Douraki, M., Comer, J. R., and Volkova, V. (2017). Potential information loss due to categorization of minimum inhibitory concentration frequency distributions. *Foodborne Pathog. Dis.*, 15(1):44–54.

McArdle, B. H. and Anderson, M. J. (2001). Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology*, 82(1):290–297.

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

Medalla, F., Gu, W., Mahon, B. E., Judd, M., Folster, J., Griffin, P. M., and Hoekstra, R. M. (2017). Estimated incidence of antimicrobial drug–resistant nontyphoidal salmonella infections, united states, 2004–2012. *Emerging infectious diseases*, 23(1):29.

Mendes, R., Garbeva, P., and Raaijmakers, J. M. (2013). The rhizosphere microbiome: significance of plant beneficial, plant pathogenic, and human pathogenic microorganisms. *FEMS microbiology reviews*, 37(5):634–663.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

Moe, L. A. (2013). Amino acids in the rhizosphere: from plants to microbes. *American journal of botany*, 100(9):1692–1705.

Morgan, X. C. and Huttenhower, C. (2012). Human microbiome analysis. *PLoS computational biology*, 8(12):e1002808.

Mouton, J. W. (2002). Breakpoints: current practice and future perspectives. *Int. J. Antimicrob. Agents.*, 19(4):323–331.

Narisetty, N. N., He, X., et al. (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, 42(2):789–817.

Naylor, D., DeGraaf, S., Purdom, E., and Coleman-Derr, D. (2017). Drought and host selection influence bacterial community dynamics in the grass root microbiome. *The ISME Journal*.

Neal, R. M. (2003). Slice sampling. *Annals of statistics*, pages 705–741.

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5(2):155–176.

Parks, D. H. and Beiko, R. G. (2010). Identifying biologically relevant differences between metagenomic communities. *Bioinformatics*, 26(6):715–721.

Parks, D. H., Tyson, G. W., Hugenholtz, P., and Beiko, R. G. (2014). Stamp: statistical analysis of taxonomic and functional profiles. *Bioinformatics*, 30(21):3123–3124.

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.

Plummer, M. et al. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, page 125. Vienna, Austria.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Randolph, T. W., Zhao, S., Copeland, W., Hullar, M., Shojaie, A., et al. (2018). Kernel-penalized regression for analysis of microbiome data. *The Annals of Applied Statistics*, 12(1):540–566.

Ridout, M., Demétrio, C. G., and Hinde, J. (1998). Models for count data with many zeros. In *Proceedings of the XIXth international biometric conference*, volume 19, pages 179–192.

Riesenfeld, C. S., Schloss, P. D., and Handelsman, J. (2004). Metagenomics: genomic analysis of microbial communities. *Annu. Rev. Genet.*, 38:525–552.

Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, pages 143–155.

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.

Rodriguez-Brito, B., Rohwer, F., and Edwards, R. A. (2006). An application of statistics to comparative metagenomics. *BMC bioinformatics*, 7(1):162.

Rolli, E., Marasco, R., Vigani, G., Ettoumi, B., Mapelli, F., Deangelis, M. L., Gandolfi, C., Casati, E., Previtali, F., Gerbino, R., et al. (2015). Improved plant resistance to drought is promoted by the root-associated microbiome as a water stress-dependent trait. *Environmental microbiology*, 17(2):316–331.

Ruiz, J., Villarreal, E., Gordon, M., Frasquet, J., Castellanos, A., and Ramirez, P. (2016). From MIC creep to MIC decline: Staphylococcus aureus antibiotic susceptibility evolution over the last 4 years. *Clin. Microbiol. Infect.*, 22(8):741–742.

Scallan, E., Hoekstra, R. M., Angulo, F. J., Tauxe, R. V., Widdowson, M.-A., Roy, S. L., Jones, J. L., and Griffin, P. M. (2011). Foodborne illness acquired in the united statesmajor pathogens. *Emerging infectious diseases*, 17(1):7.

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., and Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. *Genome biology*, 12(6):R60.

Shi, P., Zhang, A., Li, H., et al. (2016). Regression analysis for microbiome compositional data. *The Annals of Applied Statistics*, 10(2):1019–1040.

Sohn, M. B., Du, R., and An, L. (2015). A robust approach for identifying differentially abundant features in metagenomic samples. *Bioinformatics*, 31(14):2269–2275.

Sohn, M. B. and Li, H. (2017). Compositional mediation analysis for microbiome studies. *bioRxiv*, page 149419.

Spellberg, B., Guidos, R., Gilbert, D., Bradley, J., Boucher, H. W., Scheld, W. M., Bartlett, J. G., Edwards Jr, J., and of America, I. D. S. (2008). The epidemic of antibiotic-resistant infections: a call to action for the medical community from the infectious diseases society of america. *Clinical Infectious Diseases*, 46(2):155–164.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498.

Sun, J. and Dudeja, P. K. (2018). *Mechanisms Underlying Host-microbiome Interactions in Pathophysiology of Human Diseases*. Springer.

Tang, Z.-Z., Chen, G., Alekseyenko, A. V., and Li, H. (2016). A general framework for association analysis of microbial communities on a taxonomic tree. *Bioinformatics*, 33(9):1278–1285.

Thomas, T., Gilbert, J., and Meyer, F. (2012). Metagenomics-a guide from sampling to data analysis. *Microbial informatics and experimentation*, 2(1):3.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Tkacz, A. and Poole, P. (2015). Role of root microbiota in plant productivity. *Journal of experimental botany*, 66(8):2167–2175.

Tringe, S. G. and Rubin, E. M. (2005). Metagenomics: Dna sequencing of environmental samples. *Nature reviews genetics*, 6(11):805.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The human microbiome project. *Nature*, 449(7164):804.

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S., and Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37.

US Centers for Disease Control , US Food and Drug Administration, US Department of Agriculture (2016a). *NARMS Integrated Report*.

US Centers for Disease Control , US Food and Drug Administration, US Department of Agriculture (2016b). The National Antimicrobial Resistance Monitoring System: Manual of Laboratory Methods.

Van de Kassteele, J., van Santen-Verheuvel, M. G., Koedijk, F. D., van Dam, A. P., van der Sande, M. A., and de Neeling, A. J. (2012). New statistical technique for analyzing MIC-based susceptibility data. *Antimicrob. Chemother.*, 56(3):1557–1563.

VanderWeele, T. and Vansteelandt, S. (2014). Mediation analysis with multiple mediators. *Epidemiologic methods*, 2(1):95–115.

Vorholt, J. A., Vogel, C., Carlström, C. I., and Müller, D. B. (2017). Establishing causality: opportunities of synthetic communities for plant microbiome research. *Cell host & microbe*, 22(2):142–155.

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

Waldron, L. (2018). Data and statistical methods to analyze the human microbiome. *MSystems*, 3(2):e00194–17.

Wei, P. and Pan, W. (2010). Network-based genomic discovery: application and comparison of markov random-field models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(1):105–125.

Weston, L. A. and Mathesius, U. (2013). Flavonoids: their structure, biosynthesis and role in the rhizosphere, including allelopathy. *Journal of chemical ecology*, 39(2):283–297.

White, J. R., Nagarajan, N., and Pop, M. (2009). Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS computational biology*, 5(4):e1000352.

Witowski, C. and Baker, B. (2012). Enhanced secondary metabolite production by microbial cocultures. *Planta Medica*, 78(11):PI95.

Wooley, J. C. and Ye, Y. (2010). Metagenomics: facts and artifacts, and computational challenges. *Journal of computer science and technology*, 25(1):71–81.

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.

Xia, Y. and Sun, J. (2017). Hypothesis testing and statistical analysis of microbiome. *Genes & Diseases*, 4(3):138–148.

Xu, L., Paterson, A. D., Turpin, W., and Xu, W. (2015). Assessment and selection of competing models for zero-inflated microbiome data. *PLoS One*, 10(7):e0129606.

Yang, J., Kloepper, J. W., and Ryu, C.-M. (2009). Rhizosphere bacteria help plants tolerate abiotic stress. *Trends in plant science*, 14(1):1–4.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Yuan, Y. and MacKinnon, D. P. (2009). Bayesian mediation analysis. *Psychological methods*, 14(4):301.

Zhang, H., Zheng, Y., Zhang, Z., Gao, T., Joyce, B., Yoon, G., Zhang, W., Schwartz, J., Just, A., Colicino, E., et al. (2016). Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics*, 32(20):3150–3154.

Zhang, J., Wei, Z., and Chen, J. (2018). A distance-based approach for testing the mediation effect of the human microbiome. *Bioinformatics*.

Zhang, X., Mallick, H., Tang, Z., Zhang, L., Cui, X., Benson, A. K., and Yi, N. (2017). Negative binomial mixed models for analyzing microbiome count data. *BMC bioinformatics*, 18(1):4.

Zhao, L. (2013). The gut microbiota and obesity: from correlation to causality. *Nature Reviews Microbiology*, 11(9):639.

Zhao, N., Chen, J., Carroll, I. M., Ringel-Kulka, T., Epstein, M. P., Zhou, H., Zhou, J. J., Ringel, Y., Li, H., and Wu, M. C. (2015). Testing in microbiome-profiling studies with mirkat, the microbiome regression-based kernel association test. *The American Journal of Human Genetics*, 96(5):797–807.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.