Summer 8-4-2021

Bayesian Statistical Modeling of Metagenomics Sequencing Data

Shuang Jiang
Southern Methodist University, shuangj@smu.edu

BAYESIAN STATISTICAL MODELING OF

METAGENOMICS SEQUENCING DATA

Approved by:

Dr. Xiaowei Zhan
Assistant Professor in Department of Population and Data Sciences, UTSW

Dr. Qiwei Li
Assistant Professor in Department of Mathematical Sciences, UTD

Dr. Xinlei Wang
Professor in Department of Statistical Science, SMU

Dr. Guanghua Xiao
Professor in Department of Population and Data Sciences, UTSW

Dr. Daniel F. Heitjan
Professor in Department of Statistical Science, SMU & Population and Data Sciences, UTSW

BAYESIAN STATISTICAL MODELING OF METAGENOMICS SEQUENCING DATA

A Dissertation Presented to the Graduate Faculty of the

Dedman College

Southern Methodist University

in

Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

with a

Major in Biostatistics

by

Shuang Jiang

B.S., Statistics, Shandong University
M.S., Statistics, Rice University

Auguest 4, 2021

ACKNOWLEDGMENTS

I would like to thank my advisors, Dr. Xiaowei Zhan and Dr. Qiwei Li, for their extraordinary guidance on my research in the past four years. Their knowledge, vision, motivation and patience have greatly inspired me to excel in my study and research. I have learned from them the essential skills to be a good statistician, from designing a comprehensive model to giving a clear presentation. I am certain that their influence will carry over into my future career.

This thesis would not have been possible without the assistance of my committee members, Dr. Xinlei Wang, Dr. Guanghua Xiao and Dr. Daniel F. Heitjan. I would like to acknowledge them for guiding me through all these years, They always provided insightful comments and encouragements on my research. I am also thankful for having the joyful company of my colleagues at the Quantitative Biomedical Research Center at UT Southwestern. They not only helped me sharpen my skills in developing statistical models, but also motivated me to dive into new research fields.

My deep and sincere gratitude to my family, my boyfriend, and friends, for their continuous and unparalleled love, help and support. Without their tremendous support and encouragement in the past few years, it would be impossible for me to complete my study.

Jiang, Shuang

B.S., Statistics, Shandong University
M.S., Statistics, Rice University

Bayesian Statistical Modeling of
Metagenomics Sequencing Data

Advisor: Assistant Professor Dr. Xiaowei Zhan

Co-advisor: Assistant Professor Dr. Qiwei Li

Doctor of Philosophy degree conferred Auguest 4, 2021

Dissertation completed May 12, 2021

Advances in next-generation sequencing technology have enabled the high-throughput profiling of metagenomes and accelerated the study of the microbiome. Microbiome count data are high-dimensional and usually suffer from uneven sampling depth, over-dispersion, and zero-inflation. In this thesis, we develop specialized analytical models for analyzing such count data. In Chapter 2, I develop a bi-level Bayesian hierarchical framework for microbiome differential abundance analysis. The bottom level is a multivariate count-generating process that links the observed counts in each sample to their latent normalized abundances. The top level is a mixture of Gaussian distributions with a feature selection scheme for differential abundance analysis. The model further employs Markov random field priors to incorporate taxonomic tree information to identify differentially abundant bacterial taxa at different taxonomic ranks. A simulation study on both simulated and synthetic data is conducted. A colorectal cancer case study demonstrates that a resulting diagnostic model trained by the selected microbial taxa can significantly improve the disease outcome prediction accuracy.

Along with identification of specific microbial taxa associated with diseases, recent scientific advancements provide mounting evidence that metabolism, genetics and environmental factors can all modulate microbial effects. In Chapter 3, I develop an integrative framework that can distinguish differentially abundant taxa across phenotypes while quantifying covariate-taxa effects. As an extension of the bi-level Bayesian hierarchical model

in Chapter 2, the new integrative model incorporates a regression framework to successfully integrate microbiome taxonomies and metabolomics in two real microbiome datasets to provide biologically interpretable findings.

Microorganisms form complex communities and collectively affect host health. In Chapter 4, I propose a general framework, HARMONIES, a Hybrid Approach foR MicrobiOme Network Inferences via Exploiting Sparsity, to infer a sparse microbiome network that describe the associations between microbial taxa. HARMONIES first utilizes a model-based approach to normalize the microbiome data. Then, it infers a sparse and stable network by imposing non-trivial regularizations based on the Gaussian graphical model. In comprehensive simulation studies, HARMONIES outperformed four other commonly used methods. When using published microbiome data from a colorectal cancer study, it discovered a novel community with disease-enriched bacteria.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

I dedicate this dissertation to my family, my friends and everyone I love.

CHAPTER 1

Introduction

1.1. Microbiome study

The human body hosts more than 100 trillion microorganisms. Collectively, the microorganism genomes contains at least 100 times as many genes as the human genome [7]. The microbes in a healthy body aid in digestion and metabolism, and prevent the colonization of pathogenic microorganisms [50]. Microbial communities have a profound impact on human health [121]. Recently, microbiome studies have identified disease-associated bacteria taxa in type 2 diabetes [54], liver cirrhosis [101], inflammatory bowel disease [47], and melanoma patients responsive to cancer immunotherapy [38]. An increasing number of research projects continue to systematically investigate the role of the microbiome in human diseases [51].

Advances in next-generation sequencing (NGS) technology, such as high-throughput 16S rRNA gene and metagenomic sequencing, have accelerated microbiome research by generating enormous amounts of low-cost sequencing data [91]. For instance, metagenomic shotgun sequencing (MSS) generates a massive amount of sequence reads that can provide species or isolate level taxonomic resolution [111]. The availability of massive data motivates the development of advanced analytical models to guide further scientific approaches in the microbiome field. For example, specialized models for microbiome differential abundance analysis ensure accurate identifications of microbiota-disease associations could facilitate the elucidation of disease etiology and lead to novel therapeutic approaches. Moreover, microbial taxa can be modulated by metabolites, antibiotics and host genetics. Ultimately, there may be a clinical need to quantify the associations be-

tween microbiome and clinical confounders [55, 82, 140]. Last but not least, understanding the structural organization of the human microbiome plays a vital role in revealing how the microbial taxa are collaborating or competing with each other under different physiologic conditions.

In sequencing-based microbial association studies, the enormous amount of NGS data can be summarized in a sample-by-taxon count table where each entry is a proxy to the underlying true abundance. However, there is no simple relationship between the true abundances and the observed counts. The microbiome count data are highly variable, both with respect to the number of total reads per sample and per taxonomic features. Hence, the distributions of observed counts are typically skewed and over-dispersed, since a large number of taxa are recorded at low frequencies whereas a few are recorded very frequently. Additionally, the sequencing count data are often summarized as counts of bacterial taxa at various taxonomic levels, since there is a natural hierarchy of biological organism classification, i.e., species, genus, family, order, class, etc. In analysis of microbiome data, it is also of biological interest to account for the dependent structure between bacterial taxa through the taxonomic tree. With these characteristics,the microbiome analysis requires specialized modeling frameworks [129]. A number of zero-inflated (ZI) models have been proposed to account for the inflated amount of zeros in microbiome data. [130] argued that these models have an advantage in controlling type 1 error in differential abundance analysis. On the other hand, statistical models based on either negative binomial or Dirichlet-multinomial distribution have been developed to directly model the overdispersed microbiome count data [21, 49, 97, 137].

There is a growing number of Bayesian methods proposed for analyzing microbiome data. For example, [122] developed a Bayesian framework for quantifying the association between microbiome taxa and KEGG orthology pathways. [57] proposed a Bayesian joint model for the identification of covariate associations in microbiome study and prediction of phenotypic outcomes. [136] constructed a Bayesian compositional regression model for

microbiome feature selection. These works have demonstrated that the Bayesian hierarchical model is an efficient technique to simultaneously handling the following aspects in microbiome study: 1) zero-inflated sequencing counts from thousands of bacterial taxa; 2) confounding effects from clinical covariates and other experimental factors; and 3) uncertainty in the parameter estimations. Besides, the Bayesian framework allows to flexibly incorporate biologically meaningful prior information in statistical inference. Therefore, we propose the following Bayesian hierarchical frameworks for three analyses in microbiome research.

1.2. Microbiome differential abundance analysis

Accurate identification of microbiota-disease associations could facilitate the elucidation of disease etiology and lead to novel therapeutic approaches. In microbiome differential abundance analysis, our goal is to identify disease-associated microbiota, for example, a set of taxa whose abundances significantly differ across clinical outcomes. To overcome the aforementioned limitations in microbiome data modeling: 1) zero-inflation, 2) uneven sampling depth, and 3) over-dispersion, we present a general Bayesian hierarchical framework to model microbiome count data for differential abundance analysis. It consists of two levels in order to allow flexibility. The bottom level is a multivariate count variable generating process, where a wide range of classic models can be plugged in, such as Dirichlet-multinomial (DM) model and zero-inflated negative binomial (ZINB) model. The mean parameters of the bottom-level model typically refer to the latent relative abundance. For the ZINB model, we further incorporate model-based normalization through Bayesian nonparametric prior distributions with stochastic constraints to infer the normalizing factors (i.e., sequencing depth). The top level is a mixture of Gaussian distributions to model the latent relative abundance with a feature selection scheme, which enables to identify a set of discriminatory taxa among different clinical groups. In addition, we introduce how to incorporate the taxonomic tree structure to jointly select biologically similar taxa.

1.3. Microbiome integrative analysis

We propose a Bayesian integrative model to analyze microbiome count data while adjusting for effects from confounding variables (e.g., metabolites, antibiotics, host genetics, etc.). Our model jointly identifies differentially abundant taxa among multiple groups and simultaneously quantifies the taxon-covariate associations. Our modeling construction includes several advantages. First, it characterizes the over-dispersion and zero-inflation frequently observed in microbiome count data by introducing an ZINB model. Second, it models the heterogeneity from different sequencing depths, covariate effects, and group effects via a log-linear regression framework on the ZINB mean components. Last, we propose two feature selection processes to simultaneously detect differentially abundant taxa and estimate the covariate-taxa associations using the *spike-and-slab* priors. We compute Bayesian posterior probabilities for these correlated features and provide the Bayesian false discovery rate (FDR).

1.4. Microbiome network analysis

Microbiota form complex community structures and collectively affect human health. Studying their relationship as a network can provide key insights into their biological mechanisms. While the number of discovered microbial taxa continues to increase, our knowledge of their interactive relationships is severely lacking. HARMONIES (a Hybrid Approach foR MicrobiOme Network Inferences via Exploiting Sparsity) is developed to infer the microbiome networks. It consists of two major steps: (1) normalization of the microbiome count data by fitting an ZINB model with the Dirichlet process prior (DPP), (2) application of Glasso to ensure sparsity and using a stability-based approach to select the tuning parameter in Glasso. The estimated network contains the information of both the degree and the direction of associations between taxa, which facilitates the biological interpretation.

1.5. Overview of projects

Chapters 2, 3, and 4 describe the modeling approaches for microbiome differential abundance analysis, integrative analysis, and network analysis, respectively.

In Chapter 2, Section 2.1 reviews the statistical methods developed for microbiome differential abundance analysis. Section 2.2 introduces the bi-level Bayesian modeling framework and discusses the prior formulations. Section 2.3 briefly describes the Markov chain Monte Carlo algorithm and the resulting posterior inference. In Section 2.4, we present comprehensive simulation studies using both simulated and synthetic data to illustrate the performance of the method. Section 2.5 consists of two case studies, using the proposed ZINB model. Section 2.6 concludes the chapter with remarks on future directions.

In Chapter 3, Section 3.1 briefly surveys the existing work in microbiome integrative analysis. Section 3.2 introduces the integrative hierarchical mixture model and the prior formulations. Section 3.3 supplies a brief discussion of the MCMC algorithm and the resulting posterior inference. In Section 3.4, we evaluate model performance on simulated data through a comparison study. Two real data analyses are shown in Section 3.5. Our conclusions are presented in Section 3.6.

In Chapter 4, Section 4.1 discusses the commonly used approaches in microbiome network analysis. Section 4.2 introduces the hybrid modeling framework of the network inference along with the simulation study setting. The results for the simulation study and the real microbiome data network analysis are presented in Section 4.3.6. Section 4.5 discusses the potential future directions.

Chapter 5 presents conclusions and outlines the future research directions.

CHAPTER 2

A Bayesian Model of Microbiome Differential Abundance Analysis

2.1. Literature review

Advances in NGS technology, such as high-throughput 16S rRNA gene and metagenomic profiling, have accelerated microbiome research by generating enormous amounts of low-cost sequencing data [91]. The availability of massive data motivates the development of specialized analytical models to identify disease-associated microbiota, for example, a set of taxa whose abundances significantly differ across clinical outcomes. Perhaps the simplest approach is to first convert the count data to its compositional version via dividing each read count by the total number of reads in each sample, and then apply the Wilcoxon rank-sum test (or its generalized version, the Kruskal–Wallis test) to each taxonomic feature individually [65]. Differential abundance analysis has also been extensively studied in other types of NGS data. Hence, several methods that were designed for RNA-Seq, including `edgeR` [105], `DESeq2` [81], and their modifications [83], have been used for analyzing microbiome count data. However, those methods result in strong biases since they neglect to account for the excess zeros observed in the microbiome data. The sparsity is usually due to both biological and technical phenomena: some microorganisms are found in only a small percentage of samples, whereas others are simply not detected owing to insufficient segueing depth [97].

A number of zero-inflated (ZI) models have been proposed to analyze zero-inflated microbiome count data for differential abundance analysis. For example, the ZI Gaussian model [97], ZI negative binomial model [138], and ZI beta regression model [99]. [130] argued that these models have an advantage in controlling type 1 error. However, all of

them require an *ad hoc* normalizing factor for each sample to reduce biases due to uneven sequencing depth. From a statistical perspective, the employment of pre-normalized quantities leads to non-optimal performance and limits the power of downstream analysis [88]. In addition, the microbiome count data are also highly variable, both with respect to the number of total reads per sample and per taxonomic features. Hence, the distributions of observed counts are typically skewed and over-dispersed, since a large number of taxa are recorded at low frequencies whereas a few are recorded very frequently. In order to take account of this characteristic, statistical models based on either negative binomial or Dirichlet-multinomial distribution have been developed [21, 49, 97, 137].

2.2. Model

In this section, we present a bi-level Bayesian framework for microbial differential abundance analysis. Section 2.2.1 introduces the bottom level, which estimates the normalized abundance of each taxon in each sample via a representative count generative model (i.e., ZINB model). Section 2.2.2 describes a Gaussian mixture model as the top level, which is used to select the differentially abundant taxa. Fig A.1 shows the graphical formulation of the proposed ZINB model.

Before introducing the main components, we summarize the observed data as follows. Let $\boldsymbol{Y}$ denote an n-by-p taxonomic abundance table of n subjects and p taxa, with $y_{ij} \in \mathbb{N}, i = 1, \ldots, n, j = 1, \ldots, p$ indicating the count of taxon j observed from subject i. Note that $\boldsymbol{Y}$ can be obtained from either 16S rRNA gene sequencing or MSS. For the sake of simplicity, we assume that the taxonomic features in $\boldsymbol{Y}$ are all at the lowest available taxonomic levels (i.e., OTU for 16S rRNA data, and species for MSS data). As the count matrix at a higher taxonomic level can be easily summed up from its lower level, we discuss how to integrate information from a taxonomic tree in Section 2.2.3. We use an n-dimensional vector $\boldsymbol{z} = (z_1, \ldots, z_n)^T$ to allocate the n subjects into K different groups (e.g., phenotypes), with $z_i = k, k = 1, \ldots, K$ indicating that subject i belongs to group k. In addition, we use the following notations throughout this chapter. For any n-by-p matrix

$\boldsymbol{X}$, we use $\boldsymbol{x}_{i\cdot} = (x_{i1}, \ldots, x_{ip})^T$ and $\boldsymbol{x}_{\cdot j} = (x_{1j}, \ldots, x_{nj})^T$ to denote the vector from i-th row and j-th column of $\boldsymbol{X}$, and use $X_{i\cdot} = \sum_{j=1}^{p} x_{ij}$ and $X_{\cdot j} = \sum_{i=1}^{n} x_{ij}$ to denote the sum of all counts in the i-th row and j-th column of $\boldsymbol{X}$.

2.2.1. Multivariate count variable generating processes

In the bottom level of the framework, we consider the multivariate counts in subject i, i.e., $\boldsymbol{y}_{i\cdot}$, as sampled from a probabilistic model $\mathcal{M}$. The model learns the normalized abundance of each taxon in each subject, and should characterize one or more attributes of microbiome count data. More importantly, it automatically accounts for measurement errors and uncertainties associated with the counts [68]. Without loss of generality, we write

$$\boldsymbol{y}_{i\cdot} \sim \mathcal{M}(\boldsymbol{\alpha}_{i\cdot}, \boldsymbol{\Theta}), \tag{2.1}$$

where the positive vector $\boldsymbol{\alpha}_{i\cdot} = (\alpha_{i1}, \ldots, \alpha_{ip})^T, \alpha_{ij} > 0$ denotes the normalized abundance for each taxon in subject i, and $\boldsymbol{\Theta}$ denotes all other model parameters. Table 2.1 provides a list of $\mathcal{M}$ and their features. In this chapter, we mainly focus on the choice of a ZINB model due to its flexibility. An alternative choice of Dirichlet-multinomial (DM) model and the related model fitting procedure are presented in Section A.1 and A.2.1.1 in the Appendix.

The excess zeros are often attributed to rare or low abundance microbial species that may be present in only a small percentage of samples, whereas others are not recorded owing to the limitations of the sampling effort. Thus, we consider modeling each taxonomic count using a ZINB model,

$$y_{ij} \sim \pi_i \mathsf{I}(y_{ij} = 0) + (1 - \pi_i)\mathsf{NB}(\lambda_{ij}, \phi_j), \tag{2.2}$$

where we constrain one of the two mixture kernels to be degenerate at zero, thereby allowing for zero-inflation. In model (2.2), $\pi_i \in (0, 1)$ can be viewed as the proportion

Table 2.1: List of multivariate count generating processes and their characterizations

	$\mathcal{M}(\boldsymbol{y}_{i\cdot}; \boldsymbol{\alpha}_{i\cdot}, \boldsymbol{\Theta})$	$\boldsymbol{\Theta}$	Uneven depth	Zero-inflation	Over-dispersion	Example
Multi	Multi$(\boldsymbol{y}_{i\cdot}; Y_{i\cdot}, \alpha_{i1}, \ldots, \alpha_{ip})$		•			
DM	DM$(\boldsymbol{y}_{i\cdot}; \alpha_{i1}, \ldots, \alpha_{ip})$		•		•	[64]
Poisson	$\prod_{j=1}^{p}$ Poi$(y_{ij}; s_i\alpha_{ij})$	$\{\boldsymbol{s}\}$	•			[12]
NB	$\prod_{j=1}^{p}$ NB$(y_{ij}; s_i\alpha_{ij}, \phi_j)$	$\{\boldsymbol{s}, \boldsymbol{\phi}\}$	•		•	[137]
ZIG	$\prod_{j=1}^{p} \pi_i(Y_{i\cdot})I(y_{ij}=0)+$ $(1-\pi_i(Y_{i\cdot}))$N$\left(\log(y_{ij}+1); \alpha_j, \sigma_j^2\right)$	$\{\boldsymbol{\sigma}, \boldsymbol{\pi}\}$	•	•	•	[97]
ZIP	$\prod_{j=1}^{p} \pi_i$I$(y_{ij}=0)+$ $(1-\pi_i)$Poi$(y_{ij}; s_i\alpha_{ij})$	$\{\boldsymbol{s}, \boldsymbol{\pi}\}$	•	•		[24]
ZINB	$\prod_{j=1}^{p} \pi_i$I$(y_{ij}=0)+$ $(1-\pi_i)$NB$(y_{ij}; s_i\alpha_{ij}, \phi_j)$	$\{\boldsymbol{s}, \boldsymbol{\phi}, \boldsymbol{\pi}\}$	•	•	•	[31]

Abbreviations: Multinomial (Multi); Dirichlet-multinomial (DM); Negative binomial (NB); Zero-inflated Gaussian (ZIG); Zero-inflated Poisson (ZIP); Zero-inflated negative binomial (ZINB).

of extra zero counts in sample i. Here we use NB$(\lambda, \phi), \lambda, \phi > 0$ to denote a negative binomial (NB) distribution, with expectation λ and dispersion $1/\phi$. With this parameterization of the NB model, the p.m.f. is written as $\frac{\Gamma(y+\phi)}{y!\Gamma(\phi)}\left(\frac{\phi}{\lambda+\phi}\right)^{\phi}\left(\frac{\lambda}{\lambda+\phi}\right)^{y}$, with the variance $\text{Var}(Y) = \lambda + \lambda^2/\phi$. Note that ϕ controls the degree of over-dispersion. A small value indicates a large variance to mean ratio, while a large value approaching infinity reduces the NB model to a Poisson model with the same mean and variance. Now we rewrite model (2.2) by introducing a latent indicator variable η_{ij}, which follows a Bernoulli distribution with parameter π_i, such that if $\eta_{ij} = 1$ then $y_{ij} = 0$, whereas if $\eta_{ij} = 0$ then $y_{ij} \sim$ NB(λ_{ij}, ϕ_j). The independent Bernoulli prior assumption can be further relaxed by formulating a Be(a_π, b_π) hyperprior on π_i, leading to a beta-Bernoulli prior of η_{ij} with expectation $a_\pi/(a_\pi + b_\pi)$. Setting $a_\pi = b_\pi = 1$ results in a noninformative prior on π_i. Lastly, we specify the same prior distribution for each dispersion parameter as $\phi_j \sim$ Ga(a_ϕ, b_ϕ). Small values, such as $a_\phi = b_\phi = 0.001$, result in a weakly informative gamma prior.

Multiplicative characterizations of the NB (or Poisson as a special case) mean are typical in both the frequentist [e.g., 15, 69, 127] and the Bayesian literature [e.g., 2, 9] to justify latent heterogeneity and over-dispersion in multivariate count data. Here, we parameterize the mean of the NB distribution as the multiplicative effect of two parameters, $\lambda_{ij} = s_i\alpha_{ij}$. We denote s_i as the size factor of sample i, reflecting the fact that samples are

sequenced in different depths. Once this global effect is accounted for, α_{ij} is interpreted as the normalized abundance for counts y_{ij}. Conditional on the parameters, the likelihood of observing the counts $\boldsymbol{y}_{i\cdot}$ can be written as

$$f_{\text{ZINB}}(\boldsymbol{y}_{i\cdot}|\boldsymbol{\alpha}_{i\cdot}, \boldsymbol{\eta}_{i\cdot}, \boldsymbol{\phi}, s_i) = \prod_{j=1}^{p} \mathsf{I}(y_{ij}=0)^{\eta_{ij}} \left(\frac{\Gamma(y_{ij}+\phi_j)}{y_{ij}!\,\Gamma(\phi_j)} \left(\frac{\phi_j}{s_i\alpha_{ij}+\phi_j} \right)^{\phi_j} \left(\frac{s_i\alpha_{ij}}{s_i\alpha_{ij}+\phi_j} \right)^{y_{ij}} \right)^{1-\eta_{ij}}. \tag{2.3}$$

To ensure identifiability between the normalized abundance α_{ij} and its relevant size factor s_i, one typical choice is to calculate $\boldsymbol{s} = (s_1, \ldots, s_n)$ based on the observed counts $\boldsymbol{Y}$, combined with some constraint such as $\sum_{i=1}^{n} s_i = 1$ or $\prod_{i=1}^{n} s_i = 1$ (i.e., $\sum_{i=1}^{n} \log s_i = 0$). Table A.1 in the Appendix summarizes the existing methods for estimating the size factors. The simplest approach is to set the size factor s_i to be proportional to the total sum of counts in the sample, i.e., $\hat{s}_i \propto Y_{i\cdot}$, although it does not account for heteroscedasticity and yields biased estimation on all other model parameters [27]. In practice, most methods were developed for mitigating the influence of extremely low and high counts in RNA-seq data, such as upper-quartiles (Q75) [14], relative log expression (RLE) [5], and weighted trimmed mean by M-values (TMM) [105]. However, these assumptions are likely not appropriate for highly diverse microbial environments [126]. [97] developed a so-called cumulative sum scaling (CSS) method. It is an adaptive extension of Q75, which is better suited for microbiome data. While convenient, the use of the plug-in estimates $\hat{s}_i$ has noticeable shortcomings. In a Bayesian framework, those plug-in estimates can be viewed as point mass priors. On one hand, the “double dipping" occurs as those informative priors are derived from the data before model fitting and thus the uncertainty quantification for estimation of s_i will not be reflected in the inference; on the other hand, a discontinuity on the point mass priors may bias the inference on other parameters.

To address the identifiability issue and allow flexibility in the estimation of the unknown normalizing factors s_i, [72] imposed a regularizing prior with a stochastic constraint on the

logarithmic scale of each size factor. They assumed that $\log s_i$ is drawn from a mixture of a two-component Gaussian mixture,

$$\log s_i \sim \sum_{m=1}^{M} \psi_m \left[t_m \, \mathsf{N}(\nu_m, \sigma_s^2) + (1 - t_m) \, \mathsf{N}\left(-\frac{t_m \nu_m}{1 - t_m}, \sigma_s^2 \right) \right], \tag{2.4}$$

with the weight of outer mixtures denoted by ψ_m ($0 < \psi_m < 1$, $\sum_{m=1}^{M} \psi_m = 1$), where M is an arbitrary large positive integer. The use of mixture distributions allows for flexible estimation of the posterior density of $\log s_i$. In order to satisfy the desired stochastic constraint (e.g., $\mathsf{E}[\log s_i] = 0$), each of M components is further modeled by a mixture of two Gaussian distributions with a constant mean of zero. The weight of each inner mixture is denoted by t_m ($0 < t_m < 1$). Note that if $M \to \infty$, model (2.4) can be interpreted as Bayesian nonparametric infinite mixtures. With the assumption that the weights ψ_m are defined by the stick-breaking construction, i.e., $\psi_1 = V_1, \psi_m = V_m \prod_{u=1}^{m-1}(1 - V_u)$, $m = 1, 2, \ldots$, it becomes a case of Dirichlet process mixture models, which have been extensively used in recent literature for flexible density estimation [see 63, 114, 118]. [67] have demonstrated the superiority of employing the Dirichlet process prior (DPP) in a Bayesian semiparametric regression model for joint analysis of ocean microbiome data. The authors claimed that DPP can accommodate various features in a distribution, such as skewness or multi-modality, while satisfying the mean constraint. We conclude the ZINB model by specifying the following hyper-prior distributions for DPP: $\nu_m \sim \mathsf{N}(0, \tau_\nu)$, $t_m \sim \mathsf{Be}(a_t, b_t)$, and $V_m \sim \mathsf{Be}(a_m, b_m)$. Note that ψ_m will be updated according to the stick-breaking construction. We assume that $\sigma_s^2 = 1$, which completes an automatic normalization of the size factors.

2.2.2. Gaussian mixture models with feature selection

In the top level of our framework, we aim to identify a subset of taxa that are relevant to discriminating the n subjects into K distinct groups. We postulate the existence of a latent

binary vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^T$, with $\gamma_j = 1$ if taxon j is differentially abundant among the K groups, and $\gamma_j = 0$ otherwise. This assumption could be formulated as,

$$\log \alpha_{ij} | \gamma_j \sim \begin{cases} \mathsf{N}(\mu_{kj}, \sigma^2_{kj}) & \text{if } \gamma_j = 1 \text{ and } z_i = k \\ \mathsf{N}(\mu_{0j}, \sigma^2_{0j}) & \text{if } \gamma_j = 0 \end{cases}. \tag{2.5}$$

Note that the use of $\log$ transformation has two folds: 1) it ensures that the normalized abundance α_{ij}'s are not skewed; 2) it converts a positive value of α_{ij} to be either positive or negative, which is more appropriate for Gaussian fittings. A common choice for the prior of the binary latent vector $\boldsymbol{\gamma}$ is independent Bernoulli distributions on each individual component with a common hyperparameter ω, i.e., $\gamma_j \sim \text{Bernoulli}(\omega)$. It is equivalent to a binomial prior on the number of discriminatory taxa, i.e., $p_\gamma = \sum_{j=1}^{p} \gamma_j \sim \mathsf{Bin}(p, \omega)$. The hyperparameter ω can be elicited as the proportion of taxa expected *a priori* to be differentially abundant among the K groups. This prior assumption can be further relaxed by formulating a $\mathsf{Be}(a_\omega, b_\omega)$ hyperprior on ω, which leads to a beta-binomial prior on p_γ with expectation $pa_\omega/(a_\omega + b_\omega)$. [115] suggest a vague prior of ω, by imposing the constraint $a_\omega + b_\omega = 2$.

Taking a conjugate Bayesian approach, we impose a normal prior on μ_0 and each μ_k, and an inverse-gamma (IG) prior on σ^2_{0j} and each σ^2_{kj}; that is, $\mu_{0j} \sim \mathsf{N}(0, h_0\sigma^2_{0j})$, $\mu_{kj} \sim \mathsf{N}(0, h_k\sigma^2_{kj})$, $\sigma^2_{0j} \sim \mathsf{IG}(a_0, b_0)$, and $\sigma^2_{kj} \sim \mathsf{IG}(a_k, b_k)$. This parameterization setting is standard in most Bayesian normal models. It allows for creating a computationally efficient feature selection algorithm by integrating out means (i.e., μ_{0j} and μ_{kj}) and variances (i.e., σ^2_{0j} and σ^2_{kj}). The integration leads to marginal non-standardized Student's t-distributions on $\log \alpha_{ij}$. Consequently, we can write the likelihood of observing the normalized abun-

dances of taxon j as,

$$p(\boldsymbol{\alpha}_{\cdot j}|\gamma_j) = (2\pi)^{-\frac{n}{2}} \times$$

$$\begin{cases} \prod_{k=1}^{K}(n_k h_k + 1)^{-\frac{1}{2}} \frac{\Gamma\left(a_k + \frac{n_k}{2}\right)}{\Gamma(a_k)} \frac{b_k^{a_k}}{\left\{b_k + \frac{1}{2}\left[\sum_{\{i:z_i=k\}} \log \alpha_{ij}^2 - \frac{\left(\sum_{\{i:z_i=k\}} \log \alpha_{ij}\right)^2}{n_k + \frac{1}{h_k}}\right]\right\}^{a_k + \frac{n_k}{2}}} \\ \qquad \text{if } \gamma_j = 1 \\ (nh_0 + 1)^{-\frac{1}{2}} \frac{\Gamma\left(a_0 + \frac{n}{2}\right)}{\Gamma(a_0)} \frac{b_0^{a_0}}{\left\{b_0 + \frac{1}{2}\left[\sum_{i=1}^{n} \log \alpha_{ij}^2 - \frac{\left(\sum_{i=1}^{n} \log \alpha_{ij}\right)^2}{n + \frac{1}{h_0}}\right]\right\}^{a_0 + \frac{n}{2}}} \\ \qquad \text{if } \gamma_j = 0 \end{cases}, \tag{2.6}$$

where n_k is the number of subjects belonging to group k. To specify the IG hyperparameters of σ_{0j}^2 and σ_{kj}^2, we recommend a weakly informative choice by setting the shape parameters a_0 and a_k's to 2, and the scale parameters b_0 and b_k to 1, following [71]. To specify the hyperprior on h_0, we suggest setting it to a large value so as to obtain a fairly flat distribution over the region where the data are defined. According to [113], a large value of h_k allows for mixtures with widely different component means and typically encourages the selection of relatively large effects (e.g., those taxa of large effect size among groups), whereas a small value encourages the selection of small effects. We carried out a sensitivity analysis in Section A.3 and found that the ZINB model performed reasonably well if the value of h_k ranged from 10 to 100.

2.2.3. Markov random field prior model to incorporate taxonomic tree

One feature of microbiome data is that the count matrix can be summarized at different taxonomic levels, since there is a natural hierarchy of biological organism classification, i.e., species, genus, family, order, class, etc. Given a count table $\boldsymbol{Y}$ at the bottom-most level, we can aggregate the counts into any upper level based on the taxonomic tree. A tree is an undirected graph where any two vertices are connected by exactly one path.

Thus, we describe the taxonomic tree by using the adjacent matrix in graph theory. Suppose the relationship between taxa in different levels are represented by a $p' \times p'$ symmetric matrix $\boldsymbol{G}$, with $g_{jj'} = 1$ if taxon j and j' have a direct link in the tree. Let $l, 1 \leq l \leq L$ index the taxonomic level in the order of $\{\text{species}, \text{genus}, \text{family}, \text{order}, \text{class}, \text{phylum}, \text{kingdom}\}$. Given the count matrix at a lower level, each element of the count matrix at the upper level can be calculated by $y_{ij}^{(l)} = \sum_{\{j':g_{jj'}=1\}} y_{ij'}^{(l-1)}$.

We assume that the size factor estimation should be irrelevant to the choices of microbiome count data at different taxonomic levels. Therefore, we consider the following scheme for a joint inference: 1) fit the bottom level model to the microbiome count matrix $\boldsymbol{Y}^{(1)}$ and infer the corresponding normalized abundance matrix $\boldsymbol{A}^{(1)}$, as well as the sample-specific size factor $\boldsymbol{s}$; 2) fit the bottom level model to the microbiome count matrices at each upper level with fixed $\boldsymbol{s}$, and obtain the corresponding normalized abundance matrices $\boldsymbol{A}^{(2)}, \ldots, \boldsymbol{A}^{(L)}$; and 3) fit the top level model to all the abundance matrices from level 1 to L, independently.

The last implementation is to individually fit the top level model to the abundance matrix at each taxonomic level, although some efforts could be made to sharpen the inference. One proposal is to replace the independent Bernoulli prior with a Markov random field (MRF) prior, which incorporates information from the taxonomic classification system, on the selection of discriminatory microbial features. This could encourage two connected taxa in the taxonomic tree to be both selected. In particular, we consider the MRF prior on each γ_j at level l as

$$p(\gamma_j^{(l)} | \boldsymbol{\gamma}^{(l-1)}, \boldsymbol{\gamma}^{(l+1)}) = \frac{\exp\left(\gamma_j^{(l)}\left(d + f\sum_{l' \in \{l-1, l+1\}} \sum_{j':g_{jj'}=1} \gamma_{j'}^{(l')}\right)\right)}{1 + \exp\left(d + f\sum_{l' \in \{l-1, l+1\}} \sum_{j':g_{jj'}=1} \gamma_{j'}^{(l')}\right)}, \tag{2.7}$$

with hyperparameters d and f to be chosen. According to (2.7) those taxa that have a direct evolutionary relationship are more likely to be jointly selected. The hyperparameter d controls the sparsity of the prior model, while f affects the probability of selection of a

feature according to the status of its connected taxa.

2.3. Model fitting and posterior inference

In this section, we briefly describe the Markov chain Monte Carlo (MCMC) algorithm for posterior inference. Our inferential strategy allows us to simultaneously infer the latent relative abundance of each taxon j (at different taxonomic levels indexed by l) in each subject i, while identifying the discriminating taxa through $\boldsymbol{\gamma}^{(l)}$'s.

Our primary interest lies in the identification of discriminating taxa via the selection vectors $\boldsymbol{\gamma}^{(l)}$'s, or $\boldsymbol{\gamma}$ if no taxonomic tree available. To serve this purpose, a MCMC algorithm is designed based on Metropolis search variable selection algorithms [13, 44]. As discussed in Section 2.2.2, we have integrated out the mean and variance components in Equation (2.5). This step helps us speed up the MCMC convergence and improve the estimation of $\boldsymbol{\gamma}^{(l)}$'s. The detailed MCMC implementation is available in Section A.2 in the Appendix.

An efficient summarization of $\boldsymbol{\gamma}^{(l)}$ is to select the taxa based on their marginal distributions. In particular, we estimate marginal posterior probabilities of inclusion (PPI) of a single taxon by $\text{PPI}_j^{(l)} = \sum_{b=1}^{B} \left(\gamma_j^{(l)} \text{ at iteration } b\right)/B$, where B is the total number of iterations after burn-in. The marginal PPI represents the proportion of MCMC samples in which a taxon is selected to be discriminatory. A set of differentially abundant taxa can be picked based on their PPIs. For example, the selection can be done by including those taxa with marginal PPIs greater than a pre-specified value such as 0.5. Alternatively, we can choose the threshold that controls for multiplicity [94], which guarantees the expected Bayesian false discovery rate (FDR) to be smaller than a number. The Bayesian FDR is calculated as follows,

$$\text{FDR}(c_\gamma) = \frac{\sum_{l=1}^{L}\sum_{j=1}^{p}(1-\text{PPI}_j^{(l)})\text{I}(1-\text{PPI}_j^{(l)} < c_\gamma)}{\sum_{l=1}^{L}\sum_{j=1}^{p}\text{I}(1-\text{PPI}_j^{(l)} < c_\gamma)}. \tag{2.8}$$

Here c_γ is the desired significance level, with $c_\gamma = 0.05$ being generally used in other parametric/nonparametric test settings for microbiome studies.

2.4. Simulation

We use both simulated and synthetic data to assess the performance of the Bayesian framework embedded with the bottom level model of DM and ZINB. We demonstrate the advantage of our models against alternative approaches. We also investigate how the prior choices affect the posterior inference.

Let $\boldsymbol{Y}_{n\times p}$ denote the simulated count table, where the number of features $p = 1,000$, and the sample size $n = 24$ or 108. We do not consider the taxonomic structure among the p features in all simulation settings. We set the number of truly discriminatory taxonomic features $p_\gamma = 50$ among $K = 2$ or 3 groups, helping us test the ability of our method to discover relevant features in the presence of a good amount of noise.

2.4.1. Generative model for simulated data

We generated simulated datasets that favor the proposed bi-level frameworks. For the normalized abundance α_{ij} of a discriminating feature ($\gamma_j = 1$), we drew its logarithmic value from a two-component Gaussian mixture distribution,

if $K = 3$. Each permutation of $\{d_{1j}, \ldots, d_{Kj}\}$ followed an arithmetic progression with unit mean and difference σ, i.e., $\{1 - \sigma/2, 1 + \sigma/2\}$ if $K = 2$, and $\{1 - \sigma, 1, 1 + \sigma\}$ if $K = 3$. For the scenario of $K = 2$, σ can be interpreted as the between-group standard deviation or the effect size in the logarithmic scale. We considered two scenarios of $\sigma = 1$ or 2, and set the within-group standard deviation $\sigma_{\text{within}} = \sigma/10$. For a non-discriminating feature ($\gamma_j = 0$), we generated its logarithmic value from a normal distribution with zero mean and variance 4, i.e., $\log \alpha_{ij} | \gamma_j = 0 \sim \mathsf{N}(0, 4)$. For the bottom level of the DM model, we first sampled the underlying fractional abundances for sample i from a Dirichlet distribution with parameters $\boldsymbol{\alpha}_{i\cdot}$, i.e., $\boldsymbol{\psi}_{i\cdot} \sim \mathsf{Dir}(\boldsymbol{\alpha}_{i\cdot})$. Then, their corresponding observed counts $\boldsymbol{y}_{i\cdot}$

were drawn from a multinomial distribution, i.e., Multi$(N_i, \boldsymbol{\psi}_{i\cdot})$, where the total counts N_i was randomly selected from a discrete uniform distribution U$(50,000, 100,000)$. As for the ZINB model, we sampled the size factors s_i from a uniform distribution U$(0.5, 4)$, and the dispersion parameters ϕ_j from an exponential distribution with mean 10, i.e., Exp$(1/10)$. Next, each observed count y_{ij} was generated from NB$(s_i\alpha_{ij}, \phi_j)$. Lastly, we randomly selected half of the counts and forced their values to zero in order to mimic the excess zeros seen in the real data. Combined with the two bottom level kernels ($\{$DM, ZINB$\}$), the two choices of the sample size ($n \in \{24, 108\}$), the number of groups ($K \in \{2, 3\}$) and the $\log$ effect size ($\sigma \in \{1, 2\}$), there were $2^4 = 16$ scenarios in total. For each of the scenarios, we independently repeated the above steps to generate 50 datasets.

2.4.2. Generative model for synthetic data

To evaluate the performance of the proposed methods on the count data that are different from the model assumptions, we also generated synthetic datasets based on multinomial models that characterize a real taxa abundance distribution. A brief description of the data-generating scheme is given below, while detailed information can be found in the supplement of [126]. Let $\boldsymbol{O} = (O_1, \ldots, O_{p_\gamma/2}, O_{p_\gamma/2+1}, \ldots, O_{p_\gamma}, O_{p_\gamma+1} \ldots, O_p)^T$ be a count vector, where $(O_1, \ldots, O_{p_\gamma/2}) = (O_{p_\gamma/2+1}, \ldots, O_{p_\gamma})$, and each $O_j, p_\gamma/2 < j \leq p$ was the sum of OTU counts for one randomly selected taxon (without replacement) from all the skin or feces samples in a real microbiome study [16]. We defined two p-by-1 vectors, $\boldsymbol{P}$ and $\boldsymbol{Q}$, as

$$P_j = \begin{cases} \exp(\sigma)O_j & \text{for } 1 \leq j \leq p_\gamma/2 \\ O_j & \text{otherwise} \end{cases}, \text{ and } Q_j = \begin{cases} \exp(\sigma)O_j & \text{for } p_\gamma/2 < j \leq p_\gamma \\ O_j & \text{otherwise} \end{cases},$$

where σ represented the $\log$ effect size. Note that $\sum_{j=1}^p P_j = \sum_{j=1}^p Q_j$. We further drew the observed counts $\boldsymbol{y}_{i\cdot}$ from a multinomial model Multi$(N_i, \boldsymbol{\psi}_{i\cdot})$, where $N_i = 10,000$ and $\boldsymbol{\psi}_{i\cdot} = \mathsf{I}\left(1 \leq i \leq \frac{n}{2}\right) \frac{\boldsymbol{P}}{\sum_{j=1}^p P_j} + \mathsf{I}\left(\frac{n}{2} < i \leq n\right) \frac{\boldsymbol{Q}}{\sum_{j=1}^p Q_j}$. This would yield the first p_γ taxa to

be truly discriminating between the two equally sized groups. Finally, we permuted the columns of the data matrix, $\boldsymbol{Y}$, to disperse the taxa. Combined with the two types of samples ($\{\text{Skin}, \text{Feces}\}$), the two choices of the sample size ($n \in \{24, 108\}$), and the log effect size ($\sigma \in \{1, 2\}$), there were $2^3 = 8$ scenarios in total. For each of the scenarios, we repeated the steps above to generate 50 independent datasets.

2.4.3. Prior and algorithm settings

For prior specification in the top level of the proposed Bayesian framework, we used the following default settings. We set the hyperparameters that control the selection of discriminatory features, $\omega \sim \text{Be}(a_\omega = 0.2, b_\omega = 1.8)$, resulting in the proportion of taxa expected *a priori* to discriminate among the K groups to be $a_\omega/(a_\omega + b_\omega) = 10\%$. As for the inverse-gamma priors on the variance components σ_{0j}^2 and σ_{kj}^2, we set the shape parameters $a_0 = a_1 = \ldots = a_k = 2$ and the scale parameters $b_0 = b_1 = \ldots = b_k = 1$ to achieve a fairly flat distribution with an infinite variance. We further set the default values of h_0 and h_k to 100, as our sensitivity analysis in Appendix A.3 showed the posterior inference on $\boldsymbol{\gamma}$ remained almost the same when those values were in the range of 10 to 100. As indicated by [113], larger values of these hyperparameters would encourage the selection of only very large effects, whereas smaller values would encourage the selection of smaller effects. For the bottom level of the ZINB model, we used the following weakly informative settings. The hyperparameters that controlled the percentage of extra zeros *a priori* were set to $\pi \sim \text{Be}(a_\pi = 1, b_\pi = 1)$. As for the gamma prior on the dispersion parameters, i.e., $\phi_j \sim \text{Ga}(a_\phi, b_\phi)$, we set both a_ϕ and b_ϕ to small values such as 0.001, which led to a vague prior with expectation and variance equal to 1 and $1,000$. For the Dirichlet priors on the size factors s_i, we followed [72] by specifying $M = n/2,\ \sigma_s = 1,\ \tau_\eta = 1,\ a_t = b_t = 1$, and $a_m = b_m = 1$. For each dataset, we ran a MCMC chain with $10,000$ iterations (first half as burn-in). The chain was initialed from a model with 5% randomly chosen γ_j set to 1. Note that the DM model does not have any parameters needing to be

specified in the bottom level.

2.4.4. Evaluation metrics

To quantify the accuracy of identifying discriminatory features via the binary vector γ, we consider two widely used measures of the quality of binary classifiers: 1) area under the curve (AUC) of the receiver operating characteristic (ROC); and 2) Matthews correlation coefficient (MCC) [86]. The former considers both true positive (TP) and false positive (FP) rates across various threshold settings, while the latter balances TP, FP, true negative (TN), and false negative (FN) counts even if the true zeros and ones in γ are of very different sizes. MCC is defined as

$$\frac{(\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}.$$

In differential analysis settings, the number of truly discriminatory features are usually assumed to be a small fraction of the total. Therefore, MCC is more appropriate to handle such an imbalanced scenario. Note that the AUC yields a value between 0 to 1 that is averaged by all possible thresholds used to select discriminatory features based on PPI, and the MCC value ranges from -1 to 1 to pinpoint a specified threshold. The larger the index, the more accurate the inference.

2.4.5. Alternative methods

To demonstrate the superiority of the proposed Bayesian models, particularly the ZINB-DPP model, we compare ours with other general approaches for microbial differential abundance analysis, all of which can be implemented in `R`. They are: 1) Analysis of variance (ANOVA); 2) Kruskal-Wallis test; 3) `WaVE-edgeR` [105]; 4) `WaVE-DESeq2` [81]; and 5) `metagenomeSeq` [97]. The first two are parametric/nonparametric methods for testing whether samples originate from the same distribution, after converting each $\boldsymbol{y}_{i\cdot}$ into a

compositional vector of proportions by dividing each count by the total number of reads $Y_{i\cdot}$. Note that the aim is to determine whether there is a significant difference among the abundance means/medians of multiple groups for each individual taxonomic feature. The third and fourth are the modified versions of `edgeR` and `DESeq2`, using a ZINB-based wanted variation extraction (WaVE) strategy to downweight the inflated amount of zeros in microbiome data. `edgeR` implements an exact binomial test generalized for over-dispersed counts, while `DESeq2` employs a Wald test by adopting a generalized linear model based on an NB kernel. The last one, `metagenomeSeq`, assumes a zero-inflated Gaussian model on the log-transformed counts, and performs a multiple groups test on moderated F-statistics. All these competitors produce p-values. In order to control for the FDR, i.e., the rate of type-I errors in these null hypothesis testings, we further adjusted their p-values using the Benjamini-Hochberg (BH) method [10]. We independently generated 50 replicates for each of the 16 simulated data scenarios, and each of the eight synthetic data scenarios. For each dataset, we ran the DM and ZINB-DPP models, and the five competitors, and computed their individual AUC and MCC.

2.4.6. Results

We first describe posterior inference on the parameters of interest, the latent binary vector $\boldsymbol{\gamma}$ and the size factor $\boldsymbol{s}$, on a single simulated dataset (bottom level kernel=ZINB, $n = 24$, $K = 2$, $\sigma = 2$). The results are obtained by fitting the ZINB-DPP model. As for the feature selection, Fig 2.1(a) shows the marginal PPI of each feature, $p(\gamma_j|\cdot)$. The red dots indicate the truly discriminatory features and the horizontal dashed line corresponds to a threshold that ensures an expected Bayesian FDR of 5%. This threshold results in a model that includes 55 features, 45 of which are in the set of truly discriminatory features. As for the size factors $\boldsymbol{s}$, Fig 2.1(b) shows the true values against the estimated ones by different normalization techniques. One advantage of the use of DPP is that it can output the uncertainty in estimating the size factor $\boldsymbol{s}$. It clearly shows that all of the true

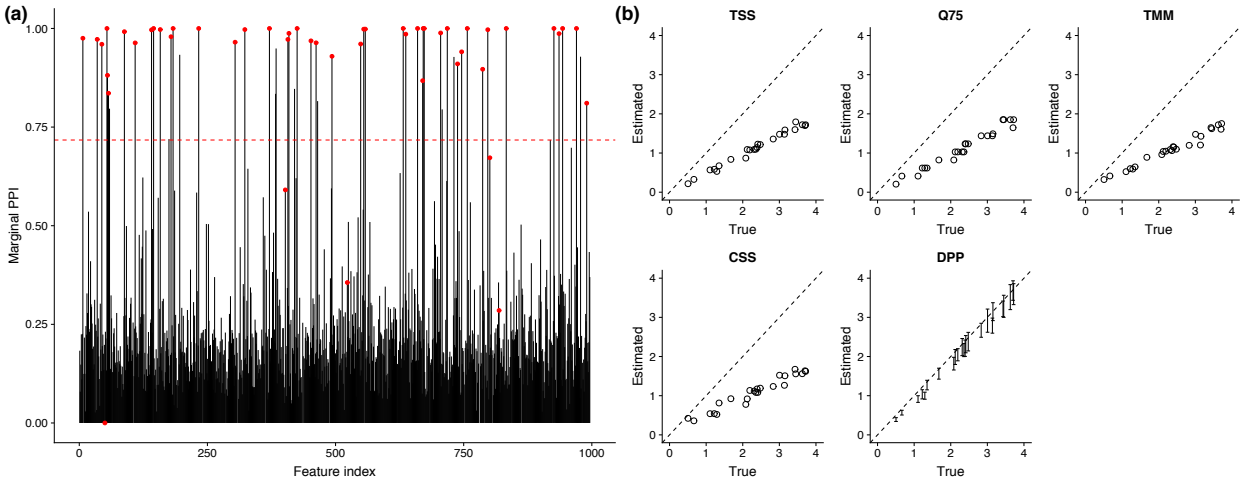


Figure 2.1: Simulated data: (a) Marginal posterior probabilities of inclusion (PPI), $p(\gamma_j|\cdot)$, with the red dots indicating truly discriminatory features and the horizontal red dashed line indicating a threshold for a 5% Bayesian FDR; (b) The scatter plots of the true and estimated size factors (s_i's) obtained by different normalization methods summarized in Table A.1. Note that RLE is not shown here because a large number of zeros in the data made the geometric means (the key component to calculating the size factors by RLE) of a few features inadmissible.

values are within the 95% credible intervals derived by our method. Note that the true size factors are generated from $\text{U}(0.5, 4)$ instead of the mixture model that DPP assumes. In comparison, the alternative normalization techniques with constraint $\prod_{i=1}^{n} s_i = 1$ yielded biased estimations.

Fig 2.2(a) and 2.2(c) display the average AUCs by different methods over 50 simulated datasets under the same group number ($K = 3$) and different sample sizes and effect sizes, (n, σ). It shows that all methods perform reasonably well for the data generated by the DM model when either the sample size or the effect size was fairly large. However, for a small sample size ($n = 24$) and a small effect size ($\sigma = 1$), the performance of `WaVE-edgeR`, `WaVE-DESeq2`, and `metagenomeSeq` significantly drop. For the data generated by the ZINB model, the results show that the ZINB-DPP model always achieves the highest AUC values. Decreasing either the sample size or the effect size would lead to greater disparity between the ZINB-DPP and the others. Fig 2.2(b) and 2.2(d) show

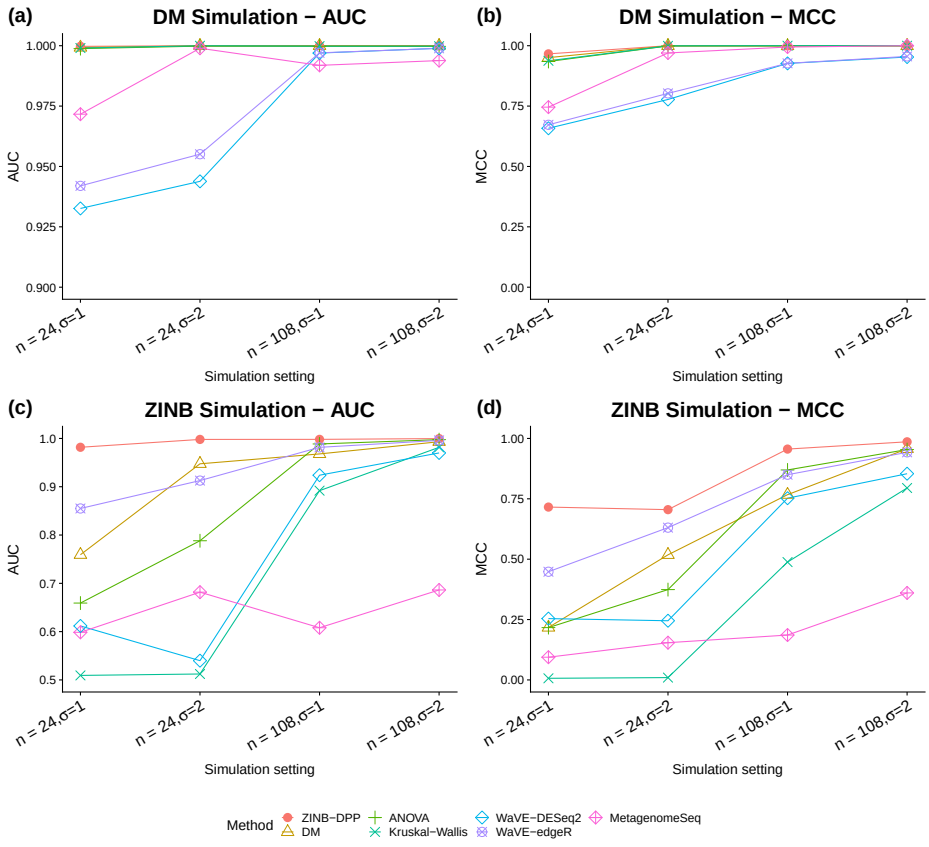


Figure 2.2: Simulation study: Averaged AUCs and MCCs achieved by the proposed framework with DM and ZINB models, and the five competitors: ANOVA, Kruskal-Wallis, *WaVE-edgeR*, *WaVE-DESeq2*, and `metagenomeSeq`. A and B are plotted from the simulated data generated by the DM model. C and D are plotted from the simulated data generated by the ZINB model.

the comparison in terms of MCC. To make a fair comparison between the methods that output p-values and those that output probability measures such as PPI, we picked only the top 50 significant features from each method on each dataset, and computed their individual MCC in the corresponding scenarios. These two plots confirm the overall best performance of the ZINB-DPP model. Notice that we also evaluated the ZINB model with different size factor estimation methods mentioned in Table A.1. The full numerical results are summarized in Tables A.2 and A.3, which show the model performance with respect to AUC and MCC on the simulated data generated from the DM and ZINB models. Here, the ZINB-RLE fail to produce any results on data generated by the ZINB model. This is be-

cause a large number of zeros is likely to make the geometric means (the key component to calculating the size factors by RLE) of a few features inadmissible.

The ZINB-DPP model also shows very competitive performance on the synthetic data. The results of the average AUCs and MCCs are presented in Fig 2.3(a) and 2.3(b). Note that we generate the synthetic datasets from the multinomial model whose parameters are estimated by using the skin/feces samples collected by [16]. Therefore, they ought to favor our DM model. However, the ZINB-DPP model, again, maintains the highest MCCs across all scenarios, and the DM model perform the second-best in general. Additionally, all methods show great improvement when either the sample size or the effect size increases, which is expected. The full numerical results are summarized in Table A.4, which compares the AUCs and MCCs for all methods implemented on the synthetic data.

2.4.7. Impact of the MRF prior

To evaluate how different tree information via the MRF prior model affects the discriminatory taxa identifications, we used the real taxonomic tree structure of the CRC dataset analyzed in Section 5 in the main text to generate new simulated data. We set the number of species $p^{(1)} = 276$ (according to the real taxonomic tree) and the number of samples $n = 24$ (equally splitting into $K = 2$ groups). We further selected $p_\gamma = 20$ differentially abundance species, with half of them enriched in one group and the remaining enriched in the other group. We considered three scenarios of true $\boldsymbol{\gamma}$, as shown in Fig 2.4, to comprehensively examine the proposed MRF prior model. In the strong scenario, all the 20 discriminatory species were from a single genus branch, i.e. *Bacteroides*. Note that it is the only genus in the tree with more than 20 species within. In the mild scenario, only half of them were from the *Bacteroides* branch, while the rest were randomly selected from all other branches. In the weak scenario, all discriminatory species were randomly selected. Following Section 2.4.1, we drew $\log \alpha_{ij}$'s of a discriminating feature $\gamma_j = 1$ from a two-component Gaussian mixture distribution, where the logarithmic effect size $\sigma \sim \mathrm{U}(1, 2)$.

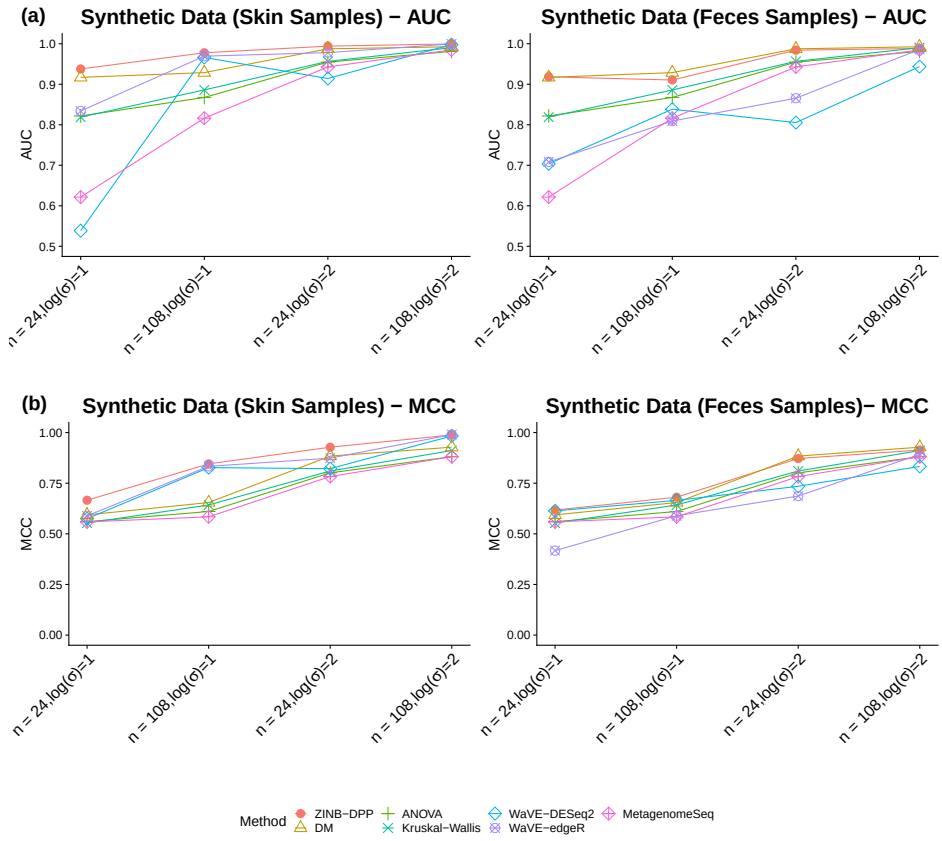


Figure 2.3: Simulation study: The average AUCs (a) and MCCs (b) achieved by the proposed framework with DM model and ZINB model, and the five competitors: ANOVA, Kruskal-Wallis, *WaVE-edgeR*, *WaVE-DESeq2*, and `metagenomeSeq`. Results are plotted from the synthetic data generated by the multinational model of skin/feces samples.

For each of the scenarios, we repeated the steps above to generate 100 independent datasets.

We set the hyperparameters that control the MRF prior model to $d = -2.2$, which means that if neither the upper nor lower-level neighbor of a taxa is discriminating, then its prior probability of being a discriminatory taxon is $\exp(-2.2)/(1 + \exp(-2.2)) \approx 0.1$. For the choice of f, we consider three settings $f \in \{0, 0.5, 1, 2\}$. Note that larger values of f would induce stronger joint selection effects. For all the other prior and algorithm specification, we use the default setting as described in Section 2.4.3.

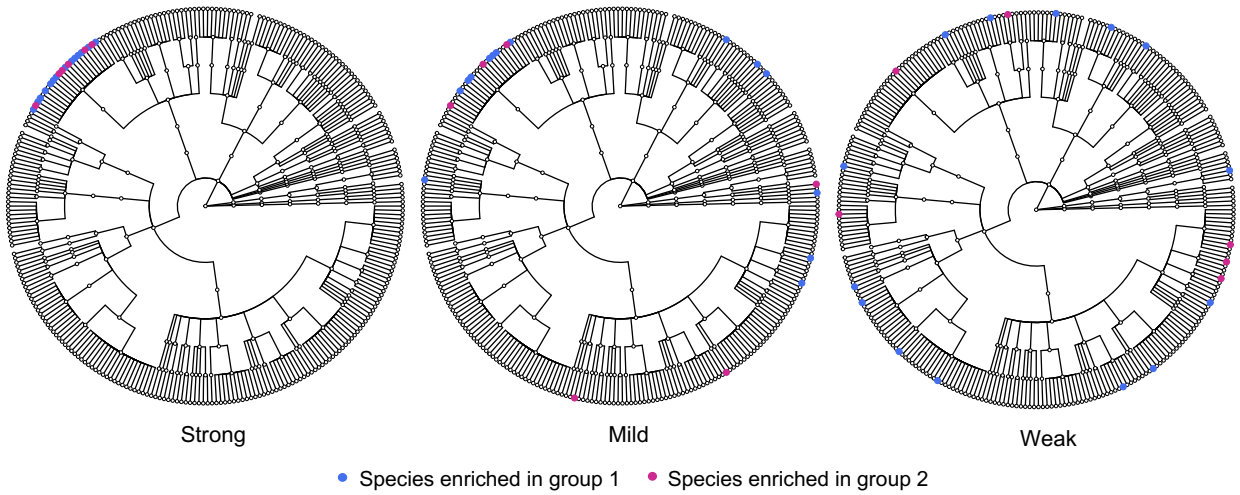
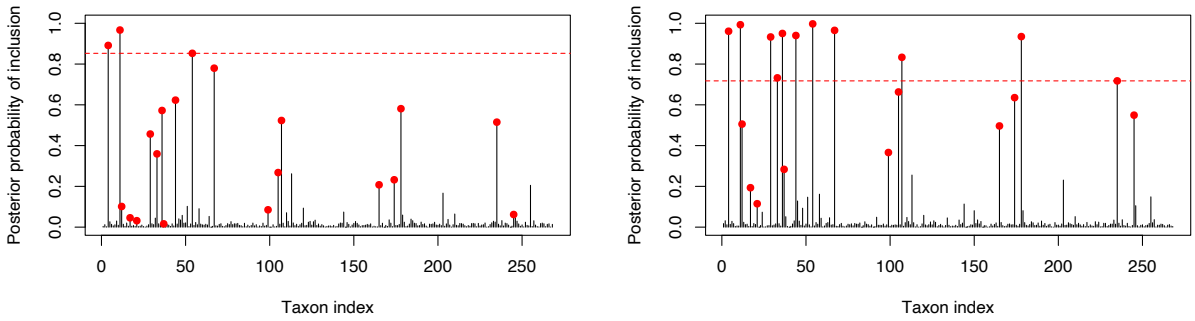


Figure 2.4: Simulation study: The three scenarios of true discriminatory species γ used to generate the simulated data with taxonomic tree information. Blue and red dots indicate the differential abundant species enriched in group 1 and 2, respectively.

We first examined the accuracy of identifying discriminatory species with and without using the MRF prior to a single simulated dataset randomly selected from the strong scenario. Fig 2.5 shows the marginal posterior probabilities of inclusion (PPI) of $p^{(1)} = 276$ species when using the independent Bernoulli prior, which is equivalent to MRF$(d = -2.2, f = 0)$, and the MRF$(d = -2.2, f = 2)$ prior. As we can see from this comparison, a PPI threshold that controls the same FDR of 5% resulted in three and 11 identified discriminatory species, all of which were truly discriminating between the two groups. We also found that the PPIs of all 20 true discriminatory species obtained with the MRF prior tended to be higher than those obtained with the independent Bernoulli prior. Consequently, the MRF prior model boosted both AUC (from 0.970 to 0.998) and MCC (from 0.375 and 0.693). The benefit of using the MRF prior was also noticed on almost all other simulated datasets.

Then, we conducted an overall comparison. Fig 2.6 shows the AUCs and MCCs from 100 replicated datasets generated under the strong, mild, and weak scenarios, respectively. To demonstrate that the improvement was indeed significant, we further applied the paired t-test to compare the results between the independent Bernoulli prior (when



(a) Without using the MRF prior

(b) With the MRF$(d = -2.2, f = 2)$ prior

Figure 2.5: Simulation study: The marginal posterior probabilities of inclusion (PPI) of all species, $p\left(\gamma_j^{(1)} = 1|\cdot\right)$, with the red dots indicating those true discriminatory species and the horizontal dashed lines indicating the threshold controlling a Bayesian FDR of 5%, using (a) the independent Bernoulli prior with $\omega = \frac{\exp(-2.2)}{1+\exp(-2.2)}$ and (b) the Markov random field (MRF) prior with $d = -2.2$ and $f = 2$.

$f = 0$) and each of the MRF prior settings. In terms of both AUC and MCC, the model with the MRF prior performed no worse than the model with the Bernoulli prior. It is worthwhile to mention that the MRF prior model would not be advantageous in the weak scenario because the joint selection could hardly be achieved under this setting.

2.5. Colorectal cancer case study

Colorectal Cancer (CRC) is the third most common cancer diagnosed in both men and women in the United States [6]. It is among the most studied diseases implicated with the gut microbiota [33, 110]. We applied our model to a CRC microbiome dataset released by [135][1]. The cohort consisted of 199 individuals from France and Germany. The disease status were confirmed by intestinal biopsy. We used curatedMetagenomicData [96] to obtain the taxonomic abundance table of all subjects with 3940 detected taxa. After the quality control procedure (detailed in Appendix A.4.1), a total of $n = 182$ subjects (104 non-CRC controls and 78 CRC patients), with abundance measurements over $p = 492$

[1]The original metagenomic sequence data from the fecal samples are available in the European Nucleotide Archive Database (accession number ERP005534).

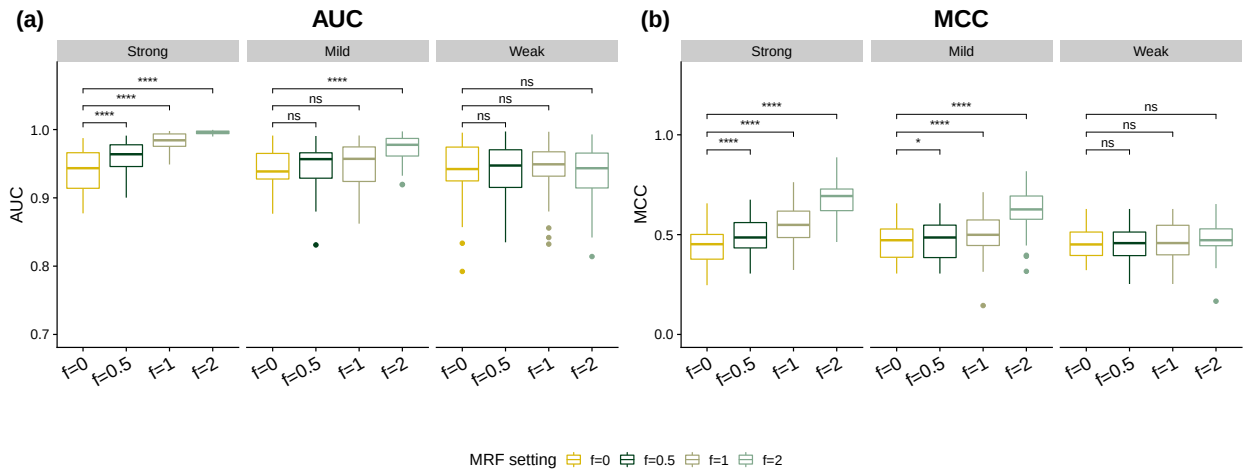


Figure 2.6: Simulation study: The box plots of (a) AUCs and (b) MCCs achieved by ZINB-DPP under the three scenarios of true discriminatory species as shown in Fig 2.4. Note that $f = 0$ corresponds to the independent Bernoulli prior, while the other three settings corresponds to the Markov random field (MRF) prior with different choices of f. The paired t-test was performed to compare each pair of settings, with ****, *, and ns indicating a p-value ≤ 0.0001, ≤ 0.05, and > 0.05, respectively.

taxa (276 species, 113 genus, 51 family, 22 order, 16 class, 10 phylum, and 4 kingdom levels), constituted the analytic dataset.

We applied the ZINB-DPP model to identify the differentially abundant taxa between the non-CRC group ($k = 1$) and CRC group ($k = 2$). We set the hyperparameters that control the selection of discriminatory features, $\omega \sim \text{Be}(a_\omega = 0.2, b_\omega = 1.8)$, resulting in the proportion of taxa expected *a priori* to discriminate between the two groups to be $a_\omega/(a_\omega + b_\omega) = 0.1$. We set the shape parameters $a_0 = a_1 = a_2 = 2$ and the scale parameters $b_0 = b_1 = b_2 = 1$ for those variance components $\sigma^2_{0j}, \sigma^2_{1j}, \sigma^2_{2j}$. Next, we set $h_0 = h_1 = h_2 = 50$. Our sensitivity analysis in Appendix A.3 shows that the posterior inference on $\boldsymbol{\gamma}$ remained almost the same when the values of h_k were in the range of 10 to 100. Note that larger values of $h_1, \ldots, h_K$ would encourage the selection of only very large effects whereas smaller values would encourage the selection of smaller effects [113]. We set $d = -2.2$ and $f = 0.5$ as the default choice of the MRF prior, indicating that if a taxon does not have any neighbor as a discriminatory taxon in the taxonomic tree,

its prior probability of being discriminating is equal to $\exp(-2.2)/(1+\exp(-2.2)) \approx 0.1$. Finally, we specified $M = n/2,\ \sigma_s = 1,\ \tau_\eta = 1,\ a_t = b_t = 1,\ a_m = b_m = 1,\ a_\phi = b_\phi = 0.001$ and $a_\pi = b_\pi = 1$ as the default DPP setting. Our inference used four independent MCMC chains with $20,000$ iterations each (first $10,000$ as burn-in). We calculated the PPIs for all chains and found their pairwise Pearson correlation coefficients ranged from 0.954 to 0.964, which suggested good MCMC convergence. We then averaged the outputs of all chains as final results and selected the discriminating taxa by controlling a Bayesian FDR of 1%.

The marginal posterior probabilities of inclusion (PPIs) of γ is presented in Fig 2.7(a), where all taxa were arranged in descending order of taxonomic levels. Based on those probabilities, a 1% Bayesian FDR threshold corresponded to a cut-off probability of 0.881 and selected 33 differentially abundant taxa, one third of which were at the species level. We also labeled those 33 taxa in a cladogram, as shown in Fig 2.7(b), and reported their estimated marginal posterior logarithmic effect sizes in Fig 2.7(c). Table 2.2 lists the 11 species with their significance and supporting evidence. Among them, *Fusobacterium nucleatum* (*Fn*) had the largest PPI value as well as the largest effect size. *Fn* is a well-known taxon associated with CRC reported by a series of studies. [19] observed that the over-abundance of *Fn* was associated with CRC tumor specimens, and they suggested that *Fn* can invade colonic mucosa and thus induce local inflammations. Later, [58] and [107] confirmed the causative role of *Fn* in CRC, and they experimentally showed that *Fn* invasion would replenish tumor-infiltrating immune cells and generate a tumorigenic microenvironment to promote colorectal neoplasia. Recently, [74] employed both targeted qPCR and metagenomic approaches to assess the diagnostic ability of five candidate taxa to predict CRC and found that *Fn* performed very well as a predictive biomarker. Interestingly, when they performed a linear combination of *Fn* and several other candidate taxa, including *Clostridium hathewayi*, which was also identified in our analysis, the diagnostic performance at predicting CRC was enhanced. [128] reported a novel mechanism of epigenetic regulation of tumor suppressor genes in host cells that is regulated by *Fn* and

other taxa through activation of DNA methyltransferases. Our model also chose *Gemella morbillorum*, which is a gram-positive species. Several reports have found that it and other species belonging to the genus, *Gemella*, are associated with rare forms of infective endocarditis. Meanwhile, it has been also observed in the stool and tissues of CRC patients at high abundance [23, 123, 133]. Another two species, *Porphyromonas asaccharolytica* and *Peptostreptococcus stomatis*, have been recently reported to be associated with CRC [36, 100]. Interestingly, all three CRC-enriched species were suggested to be the most important CRC predictors by [135]. Besides, our model reported a few species enriched in the non-CRC group, such as *Eubacteriaceae* and *Pseudoflavonifractor*. The former was found to be underrepresented in CRC tissue [84], while the latter was found to be over-represented in healthy control samples in a CRC study [124]. To summarize, six of those 11 differentially abundant species were previously reported as potentially important microbial features in CRC pathophysiology.

As a result of utilizing the MRF prior on γ, those selected taxa exhibited clustering in the taxonomic tree, as shown in Fig 2.7(b). For instance, our model chose all taxa in the branch from the phylum *Fusobacteria* to the species *Fn*, suggesting those biologically similar sequences under *Fusobacteria* were all positively associated with CRC. Our model also detected a branch of gram-negative bacteria from the class *Epsilonproteobacteria* to the genus *Campylobacter*. Interestingly, [124] reported significant co-occurrence of a number of species under both the genus *Fusobacterium* and *Campylobacter* in individual CRC tumors.

2.5.1. Comparative analysis

We compared the above ZINB-DPP result with the one from each alternative approach evaluated in the simulation study. Since the KW test is the most common differential analysis tool for metagenomics data, we focused on the comparison between ZINB-DPP and KW here, while the comparisons with all other methods, including ANOVA, `WaVE-DESeq2`,

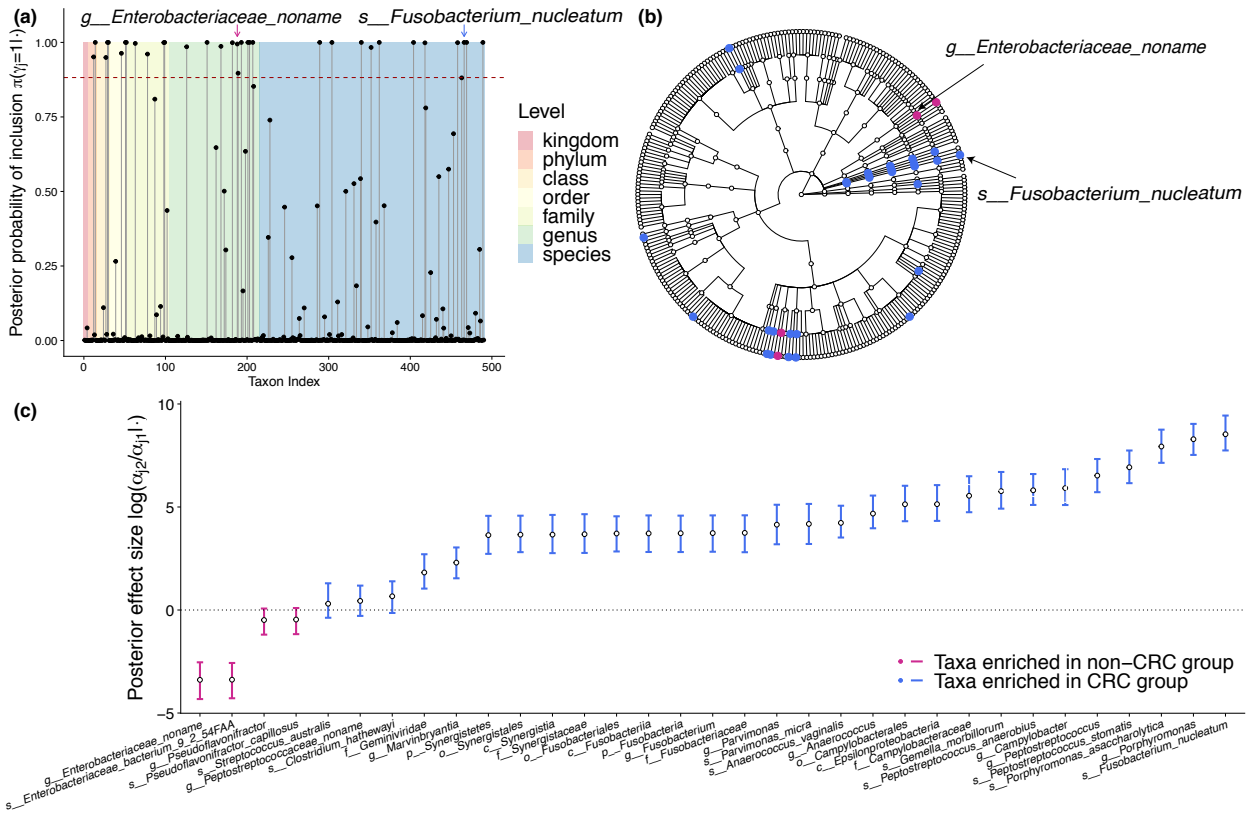


Figure 2.7: CRC study: (a) Marginal posterior probabilities of inclusion (PPI) of all taxa, $p(\gamma_j = 1|\cdot)$, with the horizontal dashed line indicating the threshold controlling a Bayesian FDR of 1%; (b) Cladogram of all taxa at different taxonomic levels, with marked dots indicating the 33 discriminating taxa identified by our ZINB-DPP; (c) 95% credible intervals of marginal posterior logarithmic effect sizes $\log(\alpha_{j2}/\alpha_{j1}|\cdot)$ of the 33 discriminating taxa identified by our ZINB-DPP.

`WaVE-edgeR`, and `MetagenomeSeq` are presented in Fig A.3 in the Appendix. The KW test reported 30 taxa below the 1% significance level threshold on the Benjamini-Hochberg (BH)-adjusted p-values, 19 of which were claimed by the ZINB-DPP model controlled by the same FDR. There were 14 taxa identified by ZINB-DPP but not KW. We chose four of them and plotted their relative abundance distributions and logarithmic normalized abundance (i.e., $\log \alpha_{ij}$) distributions, respectively, for both non-CRC and CRC groups. As we can see in Fig 2.8, the median differences of logarithmic normalized abundances between these two groups are notably visible. This is because our ZINB-DPP model could properly adjust for sample heterogeneity and zero-inflation based on information shared

Table 2.2: CRC study: List of 11 discriminating species identified by our ZINB-DPP

Species name	PPI	Distribution plot	Supporting biological evidence
s_Fusobacterium_nucleatum	1.000		[19, 58, 107]
s_Clostridium_hathewayi	1.000		
s_Gemella_morbillorum	1.000		[62]
s_Peptostreptococcus_stomatis	1.000		[29, 100]
s_Peptostreptococcus_anaerobius	1.000	Fig 2.8(b)	[119]
s_Porphyromonas_asaccharolytica	1.000		[36]
s_Streptococcus_australis	1.000		
s_Anaerococcus_vaginalis	0.999	Fig 2.8(d)	
s_Enterobacteriaceae_bacterium_9_2_54FAA	0.999		
s_Pseudoflavonifractor_capillosus	0.999		
s_Parvimonas_micra	0.983		[29, 100]

by all samples and taxa, while KW only examines the relative abundance of a taxon at a time. It is worth noting that *Synergistaceae* and *Peptostreptococcus anaerobius* (shown in Fig 2.8(a) and (b)) have been recently validated to be associated with CRC [25, 119], while *Enterobacteriaceae* and *Anaerococcus vaginalis* (shown in Fig 2.8(c) and (d)) are novel findings with evidence exhibited at their higher taxonomic levels [4, 35]. Among the 11 taxa that were identified by KW but not ours, we chose four of them and plotted their distributions for both non-CRC and CRC groups in Fig 2.9. Take *Clostridium symbiosum* (shown in Fig 2.9(a)) for instance: the resulting BH-adjusted p-value by KW is 2×10^{-4}, indicating its relative abundance was significantly different between the two groups. We found that such a small p-value was driven by a large proportion of zeros in the non-CRC group, which was 65%. However, many of them might be false zeros owing to the limitations of the sampling effort. Our flexible modeling framework could minimize the false zero impact on the identification of discriminating taxa. For the remaining three species (shown in Fig 2.9(b)–(d)), the violin plots of either relative abundances or logarithmic normalized abundances fail to show noticeable median differences between the two groups.

2.5.2. Principal components analysis and unsupervised clustering analysis

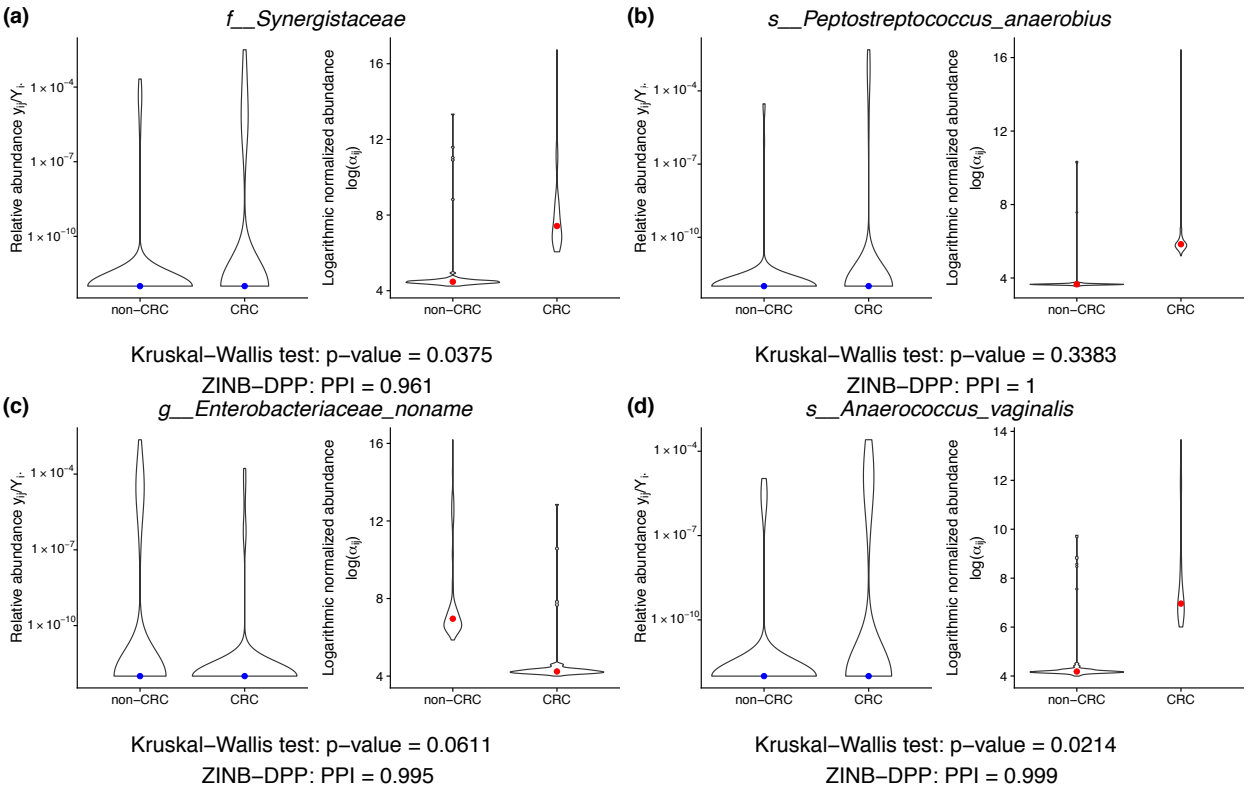


Figure 2.8: CRC study downstream analysis I: Violin plots of relative abundance and logarithmic normalized abundance, $\log \alpha_{ij}$, under different groups for the four taxa identified by our ZINB-DPP but not the Kruskal-Wallis (KW) test: (a) *Synergistaceae*; (b) *Peptostreptococcus anaerobius*; (c) *Enterobacteriaceae*; (d) *Anaerococcus vaginalis*, with the colored dots indicating the group medians.

The proposed ZINB-DPP model identified a subgroup of taxa that were differentially abundant between the two groups, in which 11 were at the species level (listed in Table 2.2). We first conducted a principle component analysis (PCA) to demonstrate that using only the 11 species can result in a better separation of CRC patients from controls than using all 276 available species that passed the quality control procedure. Specifically, we first normalized the species abundances in each sample (i.e., each row of $\boldsymbol{Y}^{(1)}$) into a compositional vector. Next, we applied the centered log-ratio (CLR) transformation [3] to each sample. Then, we rescaled each feature (i.e., each column of $\boldsymbol{Y}^{(1)}$) to ensure it had zero-mean and unit-variance. The PCA projections, with the 95% confidence ellipse of each group, are shown in Fig 2.10(a) and (b). When we performed the PCA based on all

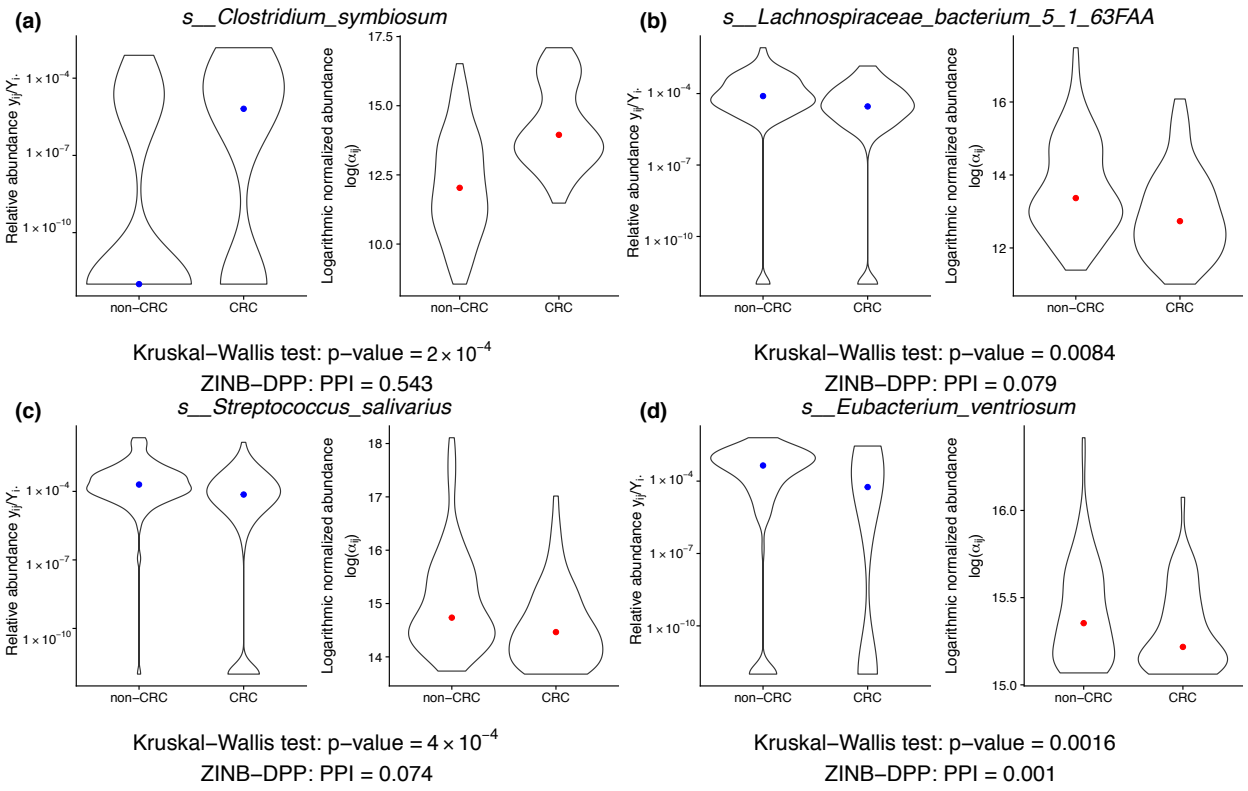


Figure 2.9: CRC study downstream analysis I: Violin plots of relative abundance and logarithmic normalized abundance, $\log \alpha_{ij}$, under different groups for the four taxa identified by the Kruskal-Wallis (KW) test but not our ZINB-DPP: (a) *Clostridium symbiosum*; (b) *Lachnospiraceae bacterium*; (c) *Streptococcus salivarius*; (d) *Eubacterium ventriosum*, with the colored dots indicating the group medians.

276 species, the first two principal components (PCs) accounted for 5.5% and 4.3% of total variance, respectively, while only the first one was associated with CRC status (Wilcoxon rank-sum test, PC1: p-value $= 0.011$; PC2: p-value $= 0.070$). When we performed the PCA based on the 11 selected species by ZINB-DPP, the first two PCs accounted for 30.2% and 14.2% of total variance, respectively, while both of them were associated with CRC status (Wilcoxon rank-sum test, PC1: p-value $= 4 \times 10^{-15}$; PC2: p-value $= 1 \times 10^{-4}$). The comparison between the two sub-plots clearly reveals a more significant separation between non-CRC and CRC subjects along with the first two PCs if using only the 11 ZINB-DPP-identified species only. In addition, as shown in Fig 2.10(b), the CRC group seemed to be less concentrated, which may be due to 1) CRC patients in this study were

from different cancer stages (1 stage 0, 24 stage I, 13 stage II, 13 stage III, and 27 stage IV patients); 2) CRC is a heterogeneous disease that can be stratified into different subtypes based on microbial community profiling.

Furthermore, we performed a model-based clustering analysis on the PC1 constructed by the 11 ZINB-DPP-identified species. Specifically, we implemented a Gaussian mixture model [37] to cluster subjects using PC1 as displayed in Fig 2.10(b). To estimate the number of clusters that best represents the data, we plotted the Bayesian information criterion (BIC) values against the number of clusters from 1 to 9, as shown in Fig 2.10(c). It shows that clustering all subjects into two groups achieves the best fit of the data measured by BIC, where the group sample sizes were 53 and 129, respectively. Next, we used a contingency table to visualize the clustering result, following the BIC plot. The χ^2 test of independence reported a p-value of < 0.001 and the Rand index, a measure of the similarity between two data clusters (a value between 0 and 1, with the former indicating that the two data clusters do not agree at all and the latter indicating that the data clusters are exactly the same), reported a value of 0.649, both demonstrating a strong association between the clustering result and the true labels. In conclusion, these two clustering analyses confirmed that the discriminating species identified by the proposed ZINB-DPP model can be used as potential biomarkers for CRC detection.

2.5.3. Predictive performance in an independent cohort

Utilizing the microbial features that were extracted by different approaches, we developed multiple diagnostic models to predict CRC status and compared their predictive performances. All models were independently validated in another CRC dataset provided by [33][2]. This cohort consisted of 154 individuals from Austria (108 non-CRC controls and 46 CRC patients). First, we applied the same data preprocessing as described in [135] on

[2]The original metagenomic shotgun sequencing data from the fecal samples are available in the European Bioinformatics Institute Database (accession number ERP008729).

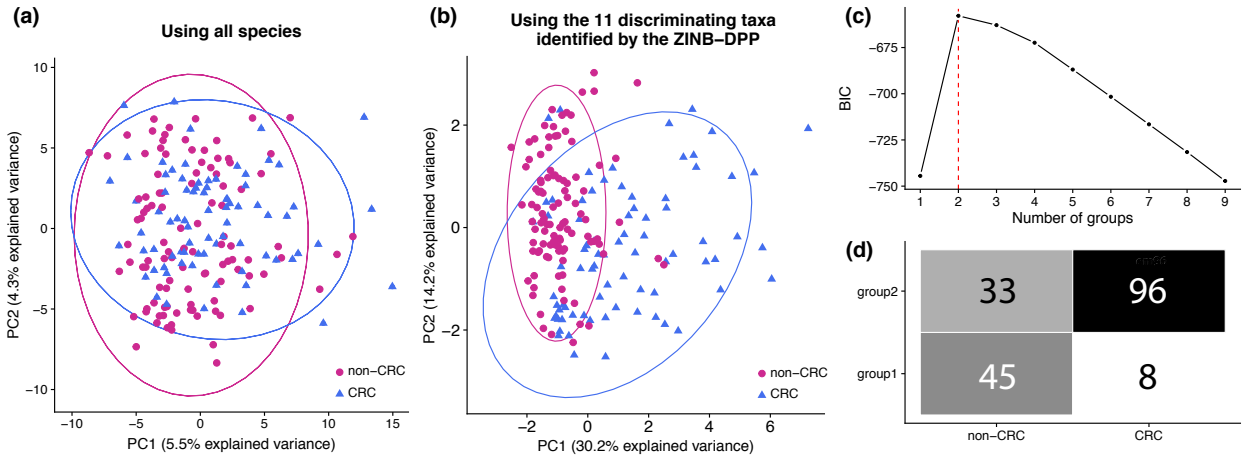


Figure 2.10: CRC study downstream analysis II: The scatter plots of the second against first principal component (PC2 vs. PC1) for the 104 non-CRC (red circles) and 78 CRC (blue triangles) samples using (a) all 276 available species that passed the quality control procedure, and (b) 11 ZINB-DPP-identified species; (c) The BIC plot of the model-based clustering on PC1 for all samples using the 11 ZINB-DPP-identified species and the contingency table of the model-based clustering results against the truth.

both training and test datasets. It included the following: 1) removing subjects with large adenoma; 2) taking logarithms of relative abundances at the species level for each sample; 3) scaling each feature (i.e., species) to ensure zero-mean and unit-variance. Next, we trained seven classifiers using an L_2-penalized logistic regression model in the CRC dataset provided by [135]. Each regression model had a unique set of species as predictors. They are as follows: 1) the 22 species suggest by [135] as potential CRC microbial signatures; 2) and 3) the 11 and 10 differentially abundant species identified (whose PPIs were above a threshold controlling a Bayesian FDR of 1%) by our ZINB-DPP and DM, respectively; 4) – 7) the 12, 145, 34, and 8 differentially abundant species (whose BH-adjusted p-values were below 0.01) identified by the KW test, `WaVE-DESeq2`, `WaVE-edgeR`, and `metagenomeSeq`, respectively. Then, we determined the penalty coefficient through 10-fold cross-validation in the training dataset, and repeated the process 10 times (implemented in `R` with `caret` package). Last, we predicted CRC status (i.e., non-CRC vs. CRC) for each subject in the test dataset using each of the seven classifiers. Their performances were compared in terms of area under the receiver operating characteristic (ROC) curve

(AUC) in that this metric considers both true and false positive rates at various threshold settings. Fig 2.11(a) and (b) shows the performance of all seven classifiers. It clearly shows that the diagnostic model based on the 11 ZINB-DPP-identified species achieved the highest AUCs, where the median is 0.853, while the model based on the 22 species originally reported by [135] performed the worst, with a AUC median of 0.781. We also implemented the logistic regression models and L_1-penalized logistic regression models. ZINB-DPP still outperformed all others under these two scenarios. Fig 2.11(c) shows the variable importance values of the diagnostic model based on the 11 biomarkers identified by our ZINB-DPP. The top three have been experimentally verified to be associated with CRC (see Table 2.2). In conclusion, this downstream analysis demonstrated the generalizability of the diagnostic model based on ZINB-DPP-identified taxonomic features to other CRC cohorts.

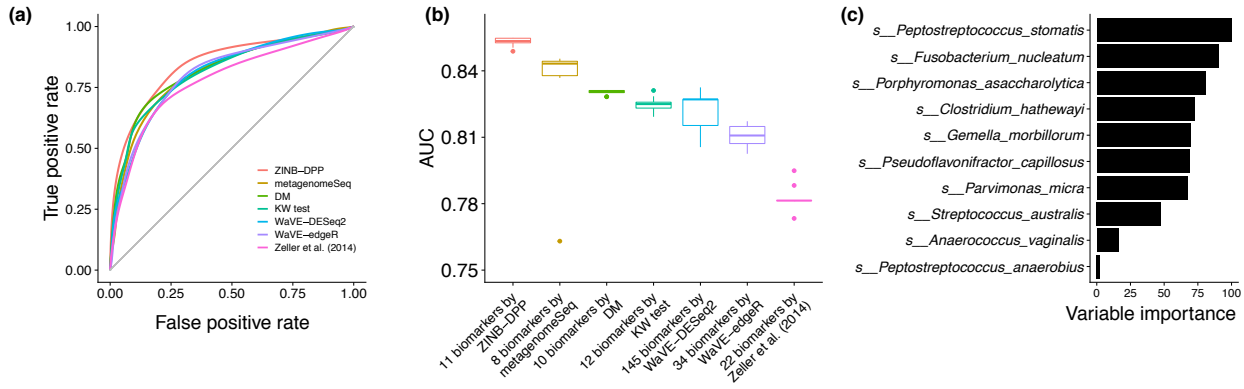


Figure 2.11: CRC study downstream analysis III: (a) The ROC curves and (b) The AUC box plots achieved by L_2-penalized logistic regression models built on different sets of species in an independent dataset; (c) The bar plot of variable importance values by the diagnostic model built on the 11 ZINB-DPP-identified species.

2.6. Discussion

In this chapter, we have proposed a Bayesian hierarchical framework for microbiome differential abundance analysis. Our bi-level framework offers flexibility to choose different normalization models and differential abundance analysis models, in distinct levels. Under this framework, we showed that our Bayesian nonparametric prior with stochastic constraints can reduce estimation bias and improve the posterior inferences of the other parameters of interests. Notably, our application of the Dirichlet process prior is not restricted to microbiome data analysis, and it is generally applicable to other types of heterogeneous sequence data [72]. Moreover, our model can jointly analyze multiple microbes at different taxonomic levels while offering well-controlled Bayesian false discovery rates.

As a summary of model performance, the ZINB-DPP model consistently outperforms commonly used methods in model-based simulations, synthetic data simulations and two real data analyses. The advantages become more obvious as either the sample size or the effect size decreases. In two case studies, our results were consistent with the current biological literature. We noticed that the sparsity observed in microbiome data could impair the statistical power of ANOVA. Meanwhile, `WaVE-edgeR` and `WaVE-DESeq2` tend to have higher false positive rates, whereas `metagenomeSeq` produces relatively conservative results compared to our model. These findings are consistent with [126] and are helpful to future microbiome data analysis.

CHAPTER 3

A Bayesian Model for the Microbiome Integrative Analysis

3.1. Literature review

Microbial abundance can be affected by covariates, such as metabolites, antibiotics and host genetics. These confounding variables need to be adjusted for more accurate differential abundance analysis. Ultimately, there may be a clinical need to quantify the associations between microbiome and clinical confounders [55, 82, 140]. One common approach is to calculate pairwise correlations between all taxa and covariates [70], but this method may be significantly underpowered. Other model-based methods [21, 122] have been proposed to detect covariate-taxa associations, but the taxon-outcome associations have been ignored. Recently, [73] developed a multivariate zero-inflated logistic-normal model to quantify the associations between microbiome abundances and multiple factors (e.g. disease risk factors or health outcomes) based on microbiome compositional data instead of the count data.

3.2. Model

Our model starts with a high-dimensional count matrix where each entry represents the count of sequence reads belonging to a taxonomy such as bacterial species. Specifically, we denote $\boldsymbol{Y}_{n\times p}$ (usually $n \ll p$) be a microbial abundance matrix, with $y_{ij} \in \mathbb{N}, i = 1, \ldots, n, j = 1, \ldots, p$ representing the observed count of the i-th sample and j-th taxon out of the total n samples and p taxa (features). Note that the proposed model can also be applied to an operational taxonomic unit (OTU) count table obtained via 16S metagenomic approaches. For an OTU table, each feature would be a taxonomic unit of a bacteria species or genus depending on the sequence similarity threshold (e.g. 97%).We also denote a covariate matrix $\boldsymbol{X}_{n\times R}$ where each entry x_{ir} represents the measurement of the

r-th covariate on the i-th sample. The graphical formulation of the proposed model is summarized in Figure B.3 and B.4.

3.2.1. Count generating process

In practice, the microbial abundance matrix $\boldsymbol{Y}$ is characterized by an inflated amount of zeros, resulting from insufficient sampling depth. Meanwhile, the abundance matrix usually consists of extremely large counts. Based on these two facts, we assume that each count is sampled from a zero-inflated negative binomial (ZINB) distribution so as to simultaneously account for both zero-inflation and over-dispersion presented in $\boldsymbol{Y}$:

$$y_{ij}|\pi, \lambda_{ij}, \phi_j \quad \sim \quad \pi \mathsf{I}(y_{ij} = 0) + (1 - \pi)\mathsf{NB}(y_{ij}; \lambda_{ij}, \phi_j), \tag{3.1}$$

where $\pi \in [0, 1]$ represents the weight of generating extra zeros, $\mathsf{I}(\cdot)$ is an indicator function, and $\mathsf{NB}(y; \lambda, \phi)$ denotes a negative binomial distribution for random variable y with the expectation λ and dispersion $1/\phi$. Under this parameterization, the variance of y is $\lambda + \lambda^2/\phi$. A small value of ϕ allows modeling of extra variation. Note that increasing ϕ towards infinity yields a Poisson distribution with both expectation and variance equal to λ. We assume a Gamma prior $\mathsf{Ga}(a_\phi, b_\phi)$ for the dispersion parameter ϕ.

An equivalent way to model this count generating process is to introduce a latent binary variable r_{ij}, such that

$$y_{ij}|r_{ij}, \lambda_{ij}, \phi_j \begin{cases} \sim & \mathsf{NB}(\lambda_{ij}, \phi_j) \quad \text{if } r_{ij} = 0 \\ = & 0 \qquad\qquad\quad \text{if } r_{ij} = 1 \end{cases}, \tag{3.2}$$

where r_{ij} is from a Bernoulli distribution with parameter π, i.e. $\sim \mathsf{Bernoulli}(\pi)$. We further impose $\pi \sim \mathsf{Beta}(a_\pi, b_\pi)$, which leads to a Beta-Bernoulli prior for r_{ij} with expectation $a_\pi/(a_\pi + b_\pi)$.

3.2.2. Integrative modeling with feature selection

Microbiome count data is characterized by high variability in the number of reads among samples from different groups (due to distinct biological conditions), or even the same group (due to uneven sequencing depths). To accommodate this setting, we parameterize the mean parameter λ_{ij} of the negative binomial distribution as the multiplicative effects of two positive random effects: 1) the size factor s_i reflects how the sequencing depth affects counts across all taxa observed in the i-th sample; 2) the normalized abundance α_{ij} for the j-th taxon in the i-th sample once the sample-specific variability has been accounted for. Our goal is to find a subset of p taxa that enables us to discriminate the n samples from K distinct groups. We introduce a binary latent vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$, with $\gamma_j = 1$ indicating that the j-th taxon has significantly differential abundances among the K groups, and $\gamma_j = 0$ otherwise. Therefore, conditional on $r_{ij} = 0$, we reparameterize the negative binomial kernel of Equation (3.2) as follows:

$$y_{ij} \mid r_{ij} = 0, \gamma_j, s_i, \alpha_{ijk}, \alpha_{ij0}, z_i \quad \sim \quad \begin{cases} \text{NB}(y_{ij}; s_i\alpha_{ij0}, \phi_j) & \text{if } \gamma_j = 0 \\ \text{NB}(y_{ij}; s_i\alpha_{ijk}, \phi_j) & \text{if } \gamma_j = 1 \text{ and } z_i = k \end{cases} . \quad (3.3)$$

Here, z_i is the sample allocation indicator. Collectively, $\boldsymbol{z}_{n\times 1} = (z_1, z_2, \ldots, z_n)^T$ indicates the membership for each sample, where $z_i = k, k \in \{1, \ldots, K\}$ reveals that the i-th sample belongs to the k-th group. s_i is the size factor of the i-th sample, which can be estimated from the data (see Section 3.2.3). We assume an independent Bernoulli prior $\gamma_j \sim \text{Bernoulli}(\omega)$ for each taxon j, and further impose a beta hyperprior on ω to formulate a Beta-Bernoulli prior, i.e. $\omega \sim \text{Beta}(a_\omega, b_\omega)$. The choice of a_ω and b_ω incorporates the prior belief that a certain percentage of taxa are discriminatory.

We further specify a log-link function to integrate the covariates into the modeling construction for each normalized abundance:

$$\begin{cases} \log \alpha_{ij0} &= \mu_{0j} + \boldsymbol{x}_i \boldsymbol{\beta}_j^T & \text{if } \gamma_j = 0 \\ \log \alpha_{ijk} &= \mu_{0j} + \mu_{kj} + \boldsymbol{x}_i \boldsymbol{\beta}_j^T & \text{if } \gamma_j = 1 \text{ and } z_i = k \end{cases}, \tag{3.4}$$

where μ_{0j} is a feature-specific baseline parameter for taxon j. Note that $\exp(\mu_{0j})$'s can be considered as scaling factors adjusting for feature-specific levels across all samples. The group-specific parameter μ_{kj} captures the baseline shift between the k-th group and the reference group. We set $\mu_{kj} = 0$ if the k-th group is the reference group to avoid identifiability problems arising from the sum of the components. $\boldsymbol{x}_i$, the i-th row of covariate matrix $\boldsymbol{X}$, contains all the covariate measurements for sample i. Here, $\boldsymbol{\beta}_j$ is a 1-by-R vector, with each element β_{rj} modeling the global effect of the r-th covariate on the observed counts for the j-th taxon. In practice, not all of the covariates are related to the abundance of a taxon. Therefore, we allow different sets of covariates to affect different taxa by specifying a *spike-and-slab* prior [13, 53] as $\beta_{rj} \sim (1 - \delta_{rj}) I(\beta_{rj} = 0) + \delta_{rj} N(0, \sigma^2_{\beta j})$, where $\delta_{rj} = 1$ indicates the r-th covariate is associated with the normalized abundance for the j-th feature, and $\delta_{rj} = 0$ otherwise. This modeling approach allows us to identify significant covariate-taxa associations, via the selection of the nonzero β_{rj} coefficients, for all discriminatory and non-discriminatory taxonomic features. We complete the model by setting $\mu_{0j} \sim N(0, \sigma^2_{0j})$, $\mu_{kj} \sim N(0, \sigma^2_{\mu j})$, and $\delta_{rj} \sim$ Beta-Bernoulli(a_p, b_p). Letting $\sigma^2_{0j} = 10^2$ for all j yields a vague prior for the feature-specific baseline parameter. An inverse-gamma (IG) hyperprior IG(a, b) is shared by $\sigma^2_{\mu j}$ and $\sigma^2_{\beta j}$.

3.2.3. Size factor estimation

The parameterization of the negative binomial mean, as shown in Equation (3.3), is a product of the size factor and the normalized abundance. It is typical to normalize the size factor first to ensure model identifiability. Hence, the plug-in estimator (equivalent to

a point-mass prior) of s_i is adopted to facilitate the inference based on the normalized abundance α_{ij} as shown in Equation (3.4). The plug-in estimators can be calculated from the observed count matrix $\boldsymbol{Y}$. There have been a number of proposals to estimate the size factors in the context of RNA-seq data analyses. Both [127] and [72] conducted a comprehensive literature review. However, the assumptions of many existing methods for RNA-seq are likely not appropriate for highly diverse microbial environments [126]. A so-called cumulative sum scaling (CSS) method has been developed by [97] as $\hat{s}_i^{\text{CSS}} \propto \sum_{j=1}^{p} y_{ij} I(y_{ij} \leq q_i^{l_{\text{CSS}}})$, where the default value of l_{CSS} is 50. CSS can be viewed as an adaptive extension of [14], and it is better suited for microbiome data. Moreover, a new normalization method named geometric mean of pairwise ratios (GMPR) has been proposed by [22], aiming to handle the zero-inflated sequencing data. GMPR calculates the size factor s_i based on the median count ratio of nonzero counts between the i-th sample and the remaining samples. It has been shown to be robust to differential and outlier OTUs. Combining this with some constraints such as $\sum_{i=1}^{n} \log \hat{s}_i = 0$ (i.e. $\prod_{i=1}^{n} \hat{s}_i = 1$), we are able to obtain a set of identifiable values. In this chapter, both CSS and GMPR are considered.

3.3. Model fitting and posterior inference

Our model space consists of $(\boldsymbol{R}, \boldsymbol{\phi}, \boldsymbol{\mu}_0, \boldsymbol{M}, \boldsymbol{B}, \boldsymbol{\gamma}, \boldsymbol{\Delta}, \omega, \pi)$ with the extra zero indicators $\boldsymbol{R} = (r_{ij}, i = 1, \ldots, n, j = 1, \ldots, p)$, the dispersion parameters $\boldsymbol{\phi} = (\phi_j, j = 1, \ldots, p)$, the feature-specific baselines $\boldsymbol{\mu}_0 = (\mu_{0j}, j = 1, \ldots, p)$, the group-specific baselines $\boldsymbol{M} = (\mu_{kj}, k = 1, \ldots, K, j = 1, \ldots, p)$, the covariate effects $\boldsymbol{B} = (\beta_{rj}, r = 1, \ldots, R, j = 1, \ldots, p)$, the discriminatory taxa indicators $\boldsymbol{\gamma} = (\gamma_j, j = 1, \ldots, p)$, and the association indicators $\boldsymbol{\Delta} = (\delta_{rj}, r = 1, \ldots, R, j = 1, \ldots, p)$. We explore the posterior distribution via a Markov chain Monte Carlo (MCMC) algorithm based on stochastic search variable selection with within-model updates [108]. Full details can be found in Appendix B.1.

We are interested in distinguishing taxa that are differentially abundant among different groups, via $\boldsymbol{\gamma}$, as well as their associations with covariates, via $\boldsymbol{\Delta}$. One way to summa-

rize the posterior distributions of these binary parameters is via the marginal posterior probability of inclusion (PPI). Suppose $t = 1, \ldots, T$ index the MCMC iterations after burn-in. Then PPI of each γ_j and δ_{rj} can be written as $\text{PPI}(\gamma_j) = \frac{1}{T}\sum_{t=1}^{T}\gamma_j^{(t)}$ and $\text{PPI}(\delta_{rj}) = \frac{1}{T}\sum_{t=1}^{T}\delta_{rj}^{(t)}$, respectively. Subsequently, important features and covariates can be selected based on a given PPI threshold. Following [94], we choose a threshold that controls the Bayesian FDR. Specifically, we solve the following equations to determine the thresholds: $\text{FDR}_{\boldsymbol{\gamma}}(c_\gamma) = \frac{\sum_{j=1}^{p}(1-\text{PPI}_{(\gamma_j)})\mathsf{I}_{(1-\text{PPI}_{(\gamma_j)}<c_\gamma)}}{\sum_{j=1}^{p}\mathsf{I}_{(1-\text{PPI}_{(\gamma_j)}<c_\gamma)}}$, $\text{FDR}_{\boldsymbol{\Delta}}(c_\delta) = \frac{\sum_{r=1}^{R}\sum_{j=1}^{p}(1-\text{PPI}_{(\delta_{rj})})\mathsf{I}_{(1-\text{PPI}_{(\delta_{rj})}<c_\delta)}}{\sum_{r=1}^{R}\sum_{j=1}^{p}\mathsf{I}_{(1-\text{PPI}_{(\delta_{rj})}<c_\delta)}}$, where $\mathsf{I}(\cdot)$ is an indicator function. A well-accepted setting is to set both $\text{FDR}_{\boldsymbol{\gamma}}$ and $\text{FDR}_{\boldsymbol{\Delta}}$ equal to 0.05, which corresponds to an expected FDR of 5%.

3.4. Simulation

3.4.1. Generative model for simulated data

In this section, we evaluated the proposed model using simulated data. In particular, we considered two methods (CSS and GMPR) introduced in Section 3.2.3 for estimating the size factor s_i's. We also compared our model with other existing methods described in the prior microbiome studies. In order to mimic metagenome sequencing data from real data applications (Section 3.5), we chose the parameters as follows: we set n samples for $K = 2$ groups with balanced group size $n_1 = n_2 = n/2$. We chose a large number of candidate features by setting the number of taxa $p = 300$, and randomly selected 20 true discriminant features to evaluate our model performance. Each row of $\boldsymbol{Y}$, denoted as $\boldsymbol{y}_i$, was generated from a Dirichlet-Multinomial distribution as described in [122]. For $i = 1, \ldots, n$, we let $\boldsymbol{y}_i \sim \text{Multinomial}(N_i,\ \boldsymbol{\pi}_i)$ with the row sum $N_i \sim \text{Discrete Uniform}(2 \times 10^7, 6 \times 10^7)$ and $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{ip}) \sim \text{Dirichlet}(\boldsymbol{a}_i)$. We further incorporated the feature and covariate effects through $\boldsymbol{a}_i = (a_{i1}, \ldots, a_{ip})$ by setting $a_{ij} = \exp(a_{ij}^*)$ with $a_{ij}^* \sim \text{Normal}(\mu_{0j} + \mu_{kj} + \boldsymbol{x}_i\boldsymbol{\beta}_j^T,\ \sigma_e^2)$. Here, a larger value of σ_e^2 corresponded to a higher noise level. Compared with Equation (3.3), this data generating pro-

cess is different from the assumption of the proposed model. We set $\mu_{0j} \sim$ Uniform$(8, 10)$, $\mu_{1j} = 0$ for all j and $\mu_{2j} = \pm 2$ for all selected discriminating features and 0 otherwise. Then for the covariate effects, we first obtained the covariate matrix $\boldsymbol{X}_{n\times R}$ by sampling each row $\boldsymbol{x}_i$ from the covariate matrix of the liver cirrhosis study in Section 3.5.1 (with $n = 237$ and $R = 7$). In particular, we sampled $n/2$ covariate records from healthy and disease groups respectively. For each taxon j, we then randomly selected $m \in \{0, 2, 4, 6\}$ out of R covariates and let the corresponding $\beta_{rj} \sim \pm$Uniform$(0.5, 1)$ while setting the rest $\beta_{rj} = 0$. Lastly, we randomly set $\pi_0 np$ counts in $\boldsymbol{Y}$ to be zeros to mimic the zero-inflation in the real data. To summarize, we varied the following settings in order to comprehensively examine the model performance: 1) sample size per group $n/2 = 10$ or 30; 2) noise level $\sigma_e = 0.5$, 1.0, or 1.5; 3) zero proportion $\pi_0 = 30\%$, 40%, or 70%. In the main text, we present the results obtained from the simulated datasets that $n/2 = 30, \sigma_e^2 = 1$, and $\pi_0 = 40\%$, and the remaining results can be found in Appendix B.2.

3.4.2. Prior and algorithm settings

The hyperparameters were specified using the following default settings. For the binary variables with Beta-Bernoulli priors $\gamma_j \sim$ Beta-Bernoulli(a_ω, b_ω), $\delta_{rj} \sim$ Beta-Bernoulli(a_p, b_p) and $r_{ij} \sim$ Beta-Bernoulli(a_π, b_π), we set $a_\omega = 0.2,\ b_\omega = 1.8,\ a_p = 0.4,$ and $b_p = 0.6$, which means that 10% of the taxa are expected to be discriminant features, and 20% of the covariate coefficients to be nonzero. We chose $a_\pi = b_\pi = 1$ assuming that about half of the zeros are truly missing. For the dispersion parameter with Ga(a_ϕ, b_ϕ) prior, we set $a_\phi = 1,\ b_\phi = 0.01$ to obtain a vague gamma prior with mean of 100 and variance of 10,000. Next, we specified a flat prior IG$(a = 2,\ b = 10)$ for the variance term $\sigma_{\mu j}^2$ and $\sigma_{\beta j}^2$. The sensitivity analysis reported in Appendix B.3 contains more details on the choice of a and b. When implementing our model on a dataset, we ran four independent chains with different starting points where each feature or covariate was randomly initialized to have $\gamma_j = 1$ or $0, \delta_{rj} = 1$ or 0. We set $20,000$ iterations as the default and discarded the first half

as burn-in. To assess the concordance between four chains, we looked at all the pairwise correlation coefficients between the marginal PPI of $\boldsymbol{\gamma}$ and $\boldsymbol{\Delta}$. As mentioned by [113], high values of correlation suggest that MCMC chains are run for a satisfactory number of iterations. After ensuring convergence, we assessed our model performance based on the averaged result over four chains.

3.4.3. Alternative methods

Our goal was to identify the discriminating features (e.g. taxa) and the significant feature-covariate associations (i.e. all nonzero γ_j and δ_{rj} in our model). We thus obtained the PPI for all γ_j and δ_{rj}, and visualized the accuracy in feature selection using the receiving operating characteristic (ROC) curve. We also computed the false positive rate when all feature-covariate associations were zero. We further considered two types of competitors for model comparison. The first type, similar to the proposed model, can simultaneously identify discriminating features and detect the feature-covariate associations. Here, we compared with the multivariate zero-inflated logistic-normal (MZILN) regression model proposed by [73]. The MZILN model treats the sample allocation vector as an observed covariate for each sample. Therefore we combined the group label with other observed covariates to create a new covariate matrix, and the MZILN model gave a regularized estimation of the regression coefficient between each feature and covariate. The selected discriminating features and feature-covariate associations corresponded to the nonzero coefficient estimations. The second type of method achieves the same goal in two separate stages. The first stage consists of four methods to select discriminating features based on p-values, including the Wilcoxon rank-sum test (Wilcoxon test) and three differential expression analysis methods implemented by the `R` packages `metagenomeSeq` [97], `edgeR` [105] and `limma` [104]. Specifically, `metagenomeSeq` assumes a zero-inflated Gaussian model, `edgeR` models count data using a negative binomial distribution, and `limma` adopts a linear model for the log-transformed count data. Then, the discriminating

features were selected to be those with BH [10] adjusted p-values smaller than 0.05. To make a head to head comparison in the first stage, we also included a simplified version of the ZINB model by excluding the covariate term $\boldsymbol{x}_i\boldsymbol{\beta}_j^T$ in Equation (3.4). In the second stage, we considered the following feature selection strategies for each p-value based method. They are: 1) correlation test, 2) lasso regression, 3) random forest, and 4) multivariate linear regression. We centered the selected discriminating features by group, and the rest across all samples. For the correlation test, the Pearson correlation coefficients were calculated between the log scaled compositional data and the covariate measurements for each outcome group. Next, a Fisher z-transformation [34] was applied to obtain the p-values for testing the significance of correlation. For lasso regression, we calculated the true positive rates and the false positive rates with respect to a range of lasso penalty. For the last two, we fitted a random forest model or a multivariate linear regression model between each feature and the covariate matrix X, which yielded variable importance measures or p-values. In all, we have four choices in the first stage {Wilcoxon test, `metagenomeSeq`, `edgeR`, `limma`} and four choices in the second stage {correlation test, lasso regression, random forest, multivariate linear regression}, with $4 \times 4 = 16$ choices in total. For clear visualization of the result, we excluded `limma` in the second stage due to its relatively inferior performance in the first stage. We also dropped random forest and linear regression since they showed similar performance as the lasso regression. Besides, all the p-values generated using different methods were adjusted using the BH method to control the FDR.

We also demonstrate that our model can estimate the association between a taxonomic feature and a covariate by adjusting for the remaining confounders. As a comparison, current approaches rely on correlation analysis between the pairwise microbiome and covariates. Specifically, those analyses converted each observed taxonomic count to a fraction (or termed percentages, intensities) by sample. Next the Pearson correlation coefficients were calculated between the log scaled fractions and the covariate measurements for each outcome group. Lastly, a Fisher z-transformation [34] was applied to obtain

the p-values for testing the significance of correlation.

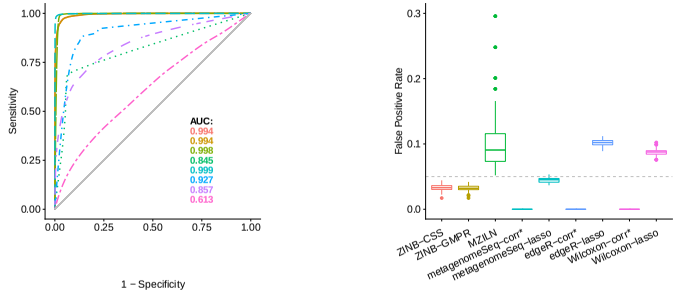
3.4.4. Evaluation metrics

We quantify the accuracy of identifying discriminatory features via the binary vector γ by calculating the area under the curve (AUC) of the receiver operating characteristic (ROC). AUC considers both true positive (TP) and false positive (FP) rates across various threshold settings.

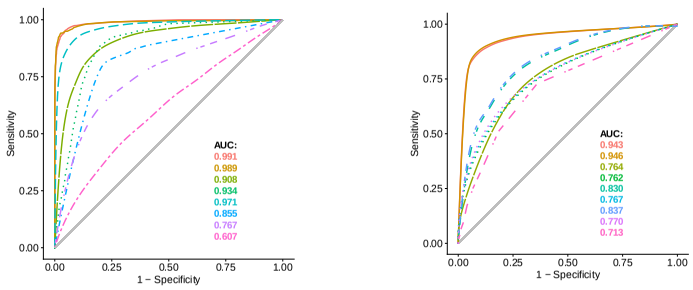
As for detecting feature-covariate associations, our model constructs a regression framework to quantify the relationship between the normalized abundance α_{ijk} and covariates through the Equation (3.4). Based on Equation (3.4), given a feature j and a covariate r of interest, we first normalized the observed abundance using CSS and performed logarithmic transformation. Next, to calculate $x_{ir}\hat{\beta}_{rj}$ and the group shift $\hat{\mu}_{kj}$, we subtracted the estimated feature-specific influence $\hat{\mu}_{0j}$ and other covariates' impact $\sum_{r' \neq r} x_{ir'}\hat{\beta}_{r'j}$ from the transformed abundance. Lastly, we could evaluate whether our model provided a reasonable estimation ($\hat{\beta}_{rj}$) of the feature-covariate association between covariate r and the normalized and adjusted observations of feature j.

3.4.5. Results

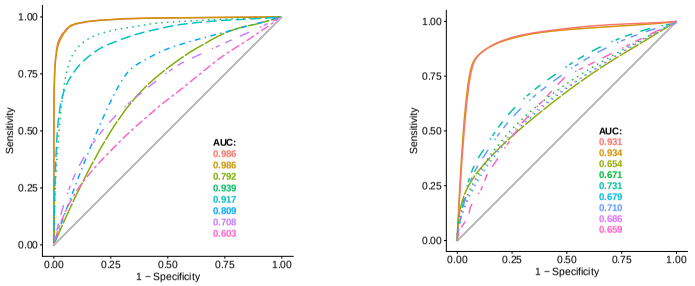
For each of the four scenarios, Figure 3.1 compares the model performance through the averaged ROC curve over 100 simulated datasets. For detecting discriminating features, the proposed method consistently shows high AUC (> 0.98) across all scenarios, and similar results for capturing the feature-covariate associations (AUC > 0.90). Moreover, the proposed method maintains a low FDR even when all β_{rj} are 0. The correlation-based method shows low false positive rates in the case where the true number of contributing covariate is 0, but has low power when $\{2, 4, 6\}$ out of 7 covariates have nonzero contribution. In addition, the proposed model achieves the highest true



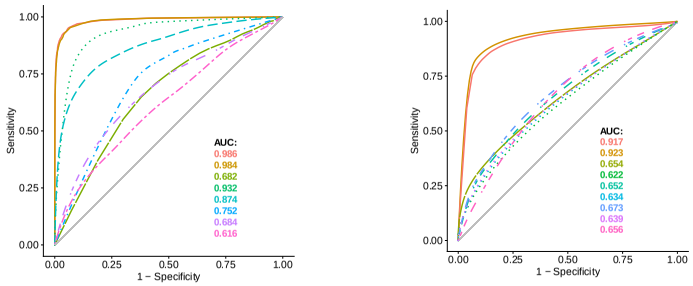
(a) ROC curves for γ (left) and false positive rates for Δ (right) with all covariate coefficients β_{rj} as 0



(b) ROC curves for γ (left) and Δ (right) with a number of nonzero covariate coefficients β_{rj} of 2



(c) ROC curves for γ (left) and Δ (right) with a number of nonzero covariate coefficients β_{rj} of 4



(d) ROC curves for γ (left) and Δ (right) with a number of nonzero covariate coefficients β_{rj} of 6

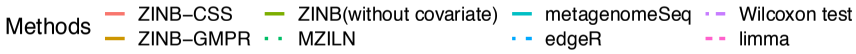
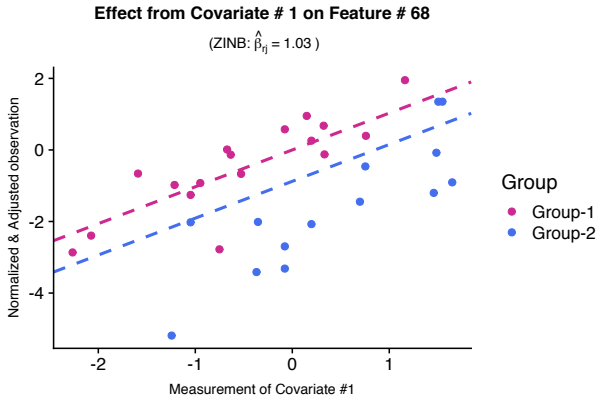


Figure 3.1: Averaged ROC curves for the discriminating feature indicator γ (left) and the feature-covariate association indicator Δ (right) with respect to different numbers of nonzero covariate coefficients, i.e. (a) 0, (b) 2, (c) 4, and (d) 6 out of 7, over 100 replicates in each scenario.

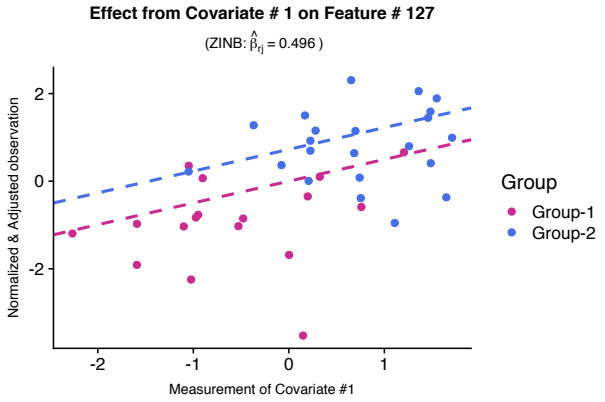
*The correlation-based methods showed low false positive rates in the case where there is no truly nonzero covariate coefficients.

positive fraction under a fixed small value of FDR in all scenarios, with the MZILN model and `metagenomeSeq` performing the second and third best in estimating the discriminating feature indicator γ (shown in the left column of Figure 3.1, Figures B.3-B.5). We also noticed that the MZILN model could not outperform the two-stage methods in estimating the feature-covariate association indicator $\mathbf{\Delta}$. The above conclusions hold for using either CSS or GMPR to estimate the plug-in size factors. To test if our model is robust to the choice of size factor estimation methods, we further conducted a sensitivity analysis in Appendix B.3. The result, as shown in Figure B.6, suggests that our model is considerably robust to the choice of different normalization methods, while CSS and GMPR have a marginal performance improvement. Furthermore, we reach the same conclusion with varying group sizes, log-scale noise levels, and zero proportions. In particular, the proposed ZINB model is robust to a larger amount of extra zeros. Either decreasing the group size or increasing the noise level hampers the performance of all the methods. Nevertheless, the ZINB model still consistently outperforms the alternative approaches in estimating γ and $\mathbf{\Delta}$. Results are summarized in Figures SB.3-B.5.

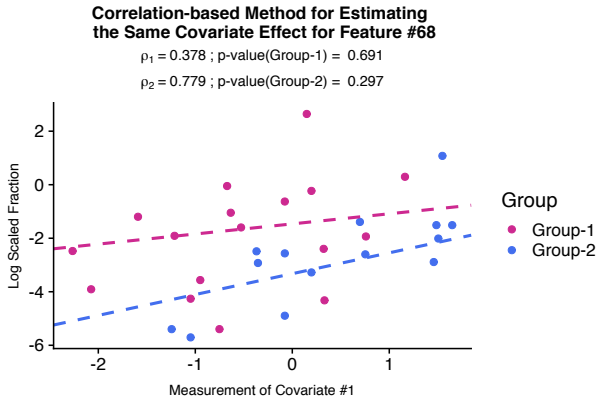
Next, we demonstrated the advantages of the proposed model in estimating the feature-covariate association over the correlation-based method through simulation. For each feature, we randomly selected four out of seven covariates to have nonzero linear effects on the latent abundances, and generated a simulated dataset following the description in Section 3.4. We kept the same prior and algorithm settings to obtain the estimations for all parameters of interest. We chose a 5% Bayesian FDR for estimating $\mathbf{\Delta}$. Among all feature-covariate combinations, the proposed model achieved sensitivity and specificity rates of 82.9% and 86.7% respectively. We randomly chose several pairs of feature and covariate and compared the proposed method and the correlation-based method. Figure 3.2 displays the results of 2 examples, where the true values of δ_{rj} were 1. The two dashed lines in Figure 3.2a or 3.2b have the same slope of $\hat{\beta}_{rj}$ as our estimated covariate effect. Both plots suggest that the proposed model is able to capture the feature-covariate relationship. Notice that we did not adjust for the group-specific effect. Hence the differ-



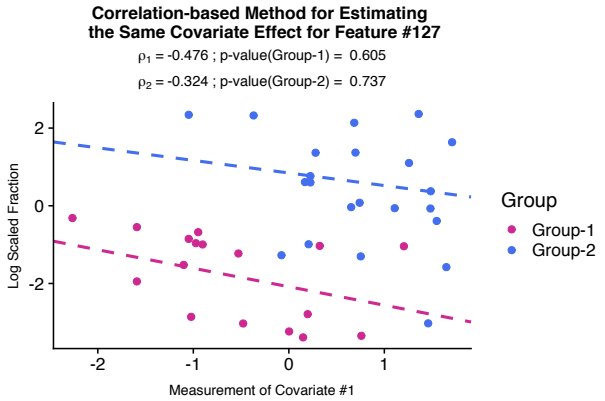
(a) Estimation of $\hat{\beta}_{1,68}$ from the ZINB model



(b) Estimation of $\hat{\beta}_{1,127}$ from the ZINB model



(c) Pearson correlation coefficient between feature 68 and covariate 1(with corresponding p-value)



(d) Pearson correlation coefficient between feature 127 and covariate 1(with corresponding p-value)

Figure 3.2: Feature-Covariate Association Analysis: comparison of the results given by the proposed method ((a) and (b)) and correlation-based method((c) and (d)) from the simulated dataset, where the two features shown (randomly selected for illustration) were truly discriminating with the covariate effect $\beta_{1,68} > 0$ and $\beta_{1,127} < 0$ by simulation. The proposed method provided a reasonable estimation ($\hat{\beta}_{rj}$) of the feature-covariate association.

ences between two dashed lines represents the group-specific parameter $\hat{\mu}_{kj}$. These results illustrated the advantages in simultaneously detecting the discriminating features and quantifying the feature-covariate associations. Furthermore, we also validated that the proposed model had correctly captured the direction of covariate effects in both cases. Figure 3.2b and 3.2d show the results from the correlation-based model. The slope of each dashed line represents the Pearson correlation coefficient. However, there was no significant result as all p-values were greater than 0.05, suggesting that the correlation test can be underpowered. The correlation-based method failed to isolate the covariate of interest from the confounders, and it might suggest a wrong direction of covariate effect, as shown in Figure 3.2d.

3.5. Real data analysis

We applied the proposed model on two real data sets: one with hundreds of samples and the other with only 24 samples. Compared with the analysis methods used in the original publications, our model demonstrates better performance in detecting differentially abundant bacteria. In addition, our model supports adjusting for biologically meaningful covariates. When adjusting for the metabolic pathway quantities (or metabolites through metabolomics technology) as covariates, our model estimates the association between taxa and metabolism-related functions (or metabolites).

3.5.1. Liver cirrhosis case study

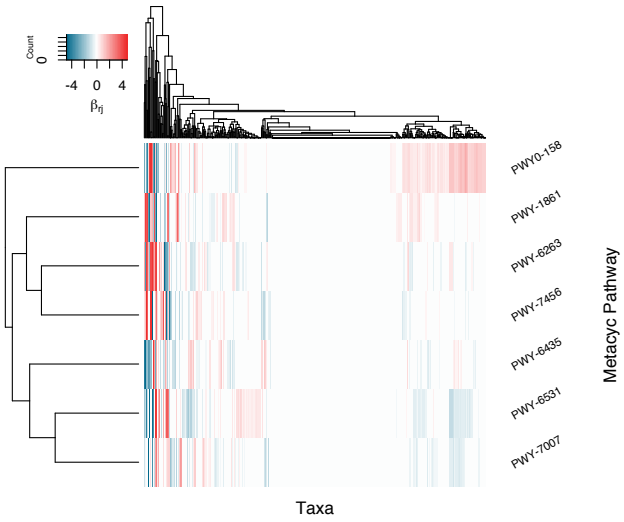
Cirrhosis is a late-stage condition of scarring or fibrosis of the liver caused by liver disease such as hepatitis B, hepatitis C, and non-alcoholic fatty liver disease [1]. The liver is connected to the gastrointestinal tract via the hepatic portal and bile secretion systems. Interestingly, distinct gut microbiota signatures have been associated with both early-stage liver diseases and end-stage liver cirrhosis [11, 41, 131]. We applied our model on a gut microbiome dataset from a liver cirrhosis study carried out by [101]. All metagenome sequenced samples were available from the NCBI Short Read Archive and

the curated microbial abundance matrix was accessible from ExperimentHub [96].

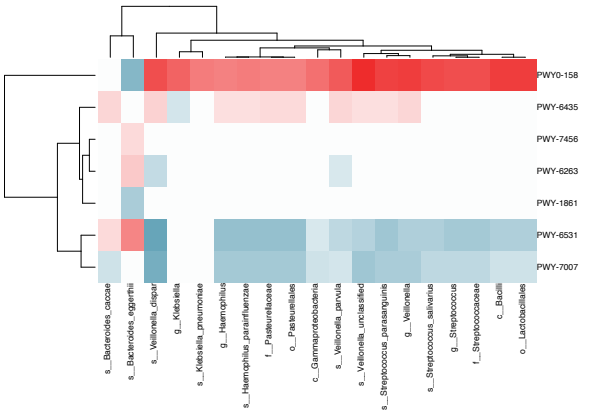
The full dataset includes 237 samples with their observed microbial abundance matrix $\boldsymbol{Y}$ profiled from the gut microbiome at the species taxonomic level. The study has two patient groups, including 114 healthy controls and 123 liver cirrhosis patients. We filtered out the taxa with extremely low abundance before the analysis as suggested in [122]. We obtained 528 taxa that had at least 2 observed counts in both groups for further analysis. As for the covariate information, we used MetaCyc, a collection of microbial pathways and enzymes involved in metabolism for an extensive amount of organisms [18]. We incorporated the 529 MetaCyc pathway measurements for 237 individuals in the study, and reduced the high correlation among the pathways by average linkage clustering on their correlation matrix [122]. Specifically, we kept the pathway with the largest fold-change between two groups in each cluster, and decided the number of clusters such that the correlations between the resulting pathways were less than 0.5. Logarithmic transformation and normalization (zero mean and unit variance) were applied to the selected covariates to ensure the zero mean and unit standard deviation. After the pre-processing step, we had seven covariates representing metabolic functions.

In [101], differential analyses were based on the Wilcoxon test and the p-values were corrected by the BH method. Although a stringent threshold of significance level (0.0001) was used, the authors discovered 79 differentially abundant species and had to restrictively report the 30 top candidates in each group. Figure 3.5a is the cladogram of the discriminating taxa selected by different methods, with blue dots representing the results by [101] and red dots reported by the ZINB model. As suggested in our simulation study, these results may reflect a high FDR as covariate effects were not factored in the analysis. In addition, the Wilcoxon test cannot account for the pathway effects and thus the associations between bacteria and metabolic pathways were not identified.

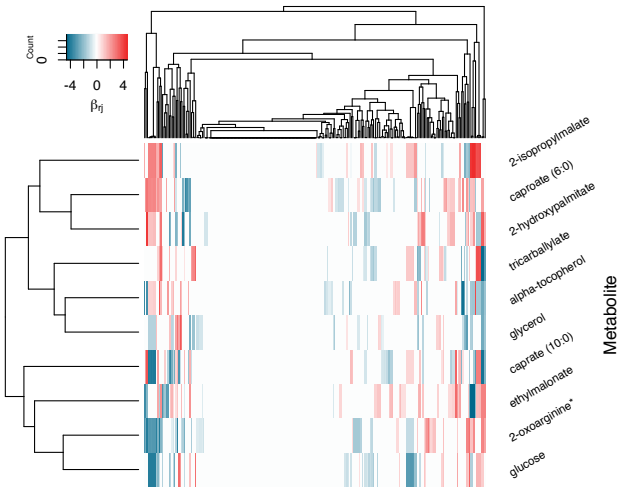
We applied the proposed Bayesian ZINB model to simultaneously analyze the microbial abundance matrix of bacteria and their metabolic pathway abundance. We set a



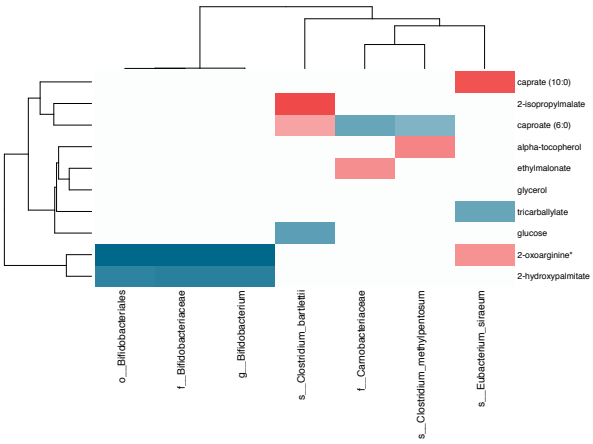
(a) Liver cirrhosis study: heatmap of covariate effect for all taxa



(b) Liver cirrhosis study: heatmap of covariate effect for selected discriminating taxa



(c) Metastatic melanoma study: heatmap of covariate effect for all taxa

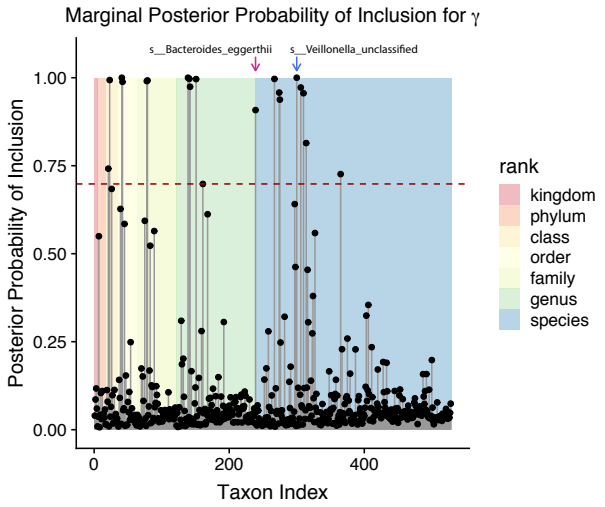


(d) Metastatic melanoma study: heatmap of covariate effect for selected discriminating taxa

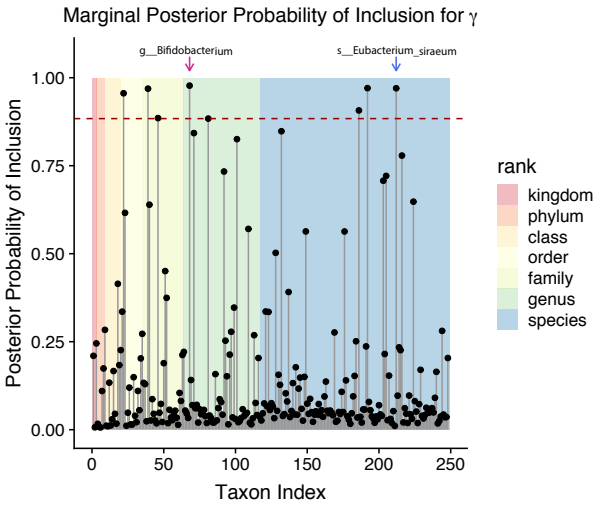
Figure 3.3: Real Data Analysis: Heatmap showing the effect from covariates, the MetaCyc pathway abundances, in two studies. (a)(b), we use a liver cirrhosis dataset and show the effect between covariate effects and all microbiome, or differential abundant microbiome, respectively; (c)(d), we use metastatic melanoma dataset and show the effect between covariate effects and all microbiome, or differential abundant microbiome, respectively;)

similar hyperparameter setting as discussed in Section 3.4 by first specifying $a_\mu = a_\beta = 2$ and $b_\mu = b_\beta = 10$. Next, we set $a_\omega,\ b_\omega,\ a_p,\ b_p,\ a_\phi,\ b_\phi$ to be the same as their default values discussed in Section 3.4. We ran four independent Markov chains with different starting points. Each chain had 40,000 iterations with the first half discarded as burn-in. We checked the convergence visually and calculated the pairwise Pearson correlation for PPIs, which ranged from 0.988 to 0.994 for γ's and from 0.982 to 0.989 for δ's. These concluded highly consistent results. Figure 3.4a shows the PPIs for all 528 taxa, where the dashed line represents the threshold corresponding to an expected FDR of 0.05. We identified 19 differentially expressed taxa, the majority of which are more abundant in the liver cirrhosis group.

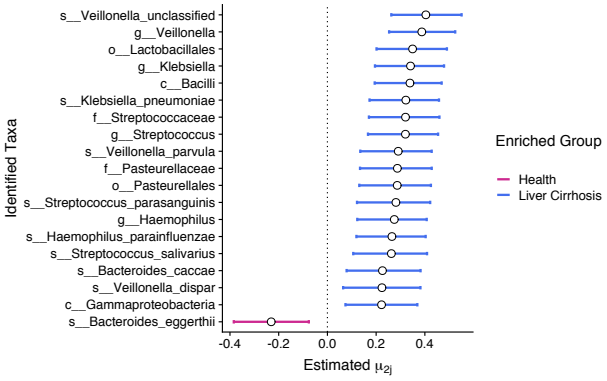
Figure 3.4c shows the posterior mean of μ_{2j} for all identified discriminating taxa, and Table B.1 contains all the detailed parameter estimations for those taxa. Interestingly, two clear taxonomic branches are distinguished by our model (as indicated by red dots, Figure 3.5a): the genera *Veillonella* and *Streptococcus*, both of which can originate from the oral cavity. Of note, oral commensal bacteria are able to colonize the distal intestinal tract in liver cirrhosis patients [101], probably due to bile acid changes. *Veillonella* spp. and *Streptococcus* spp. have been identified as more abundant in patients with primary biliary cholangitis [116], another hepatic disorder which shares pathophysiologic features with liver cirrhosis [103]. Figure 3.3a and 3.3b show the identified associations between microbiota and metabolic pathways. For example, L-alanine biosynthesis (PWY0-1061) is positively correlated with *Veillonella*. Alanine is a gluconeogenesis precursors in liver metabolism, and increased alanine is thought to induce pyruvate kinase in *Veillonella*. Thus, this connection between alanine synthesis and *Veillonella* is intriguing and potentially novel, and biologic validation experiments might offer further clarification.



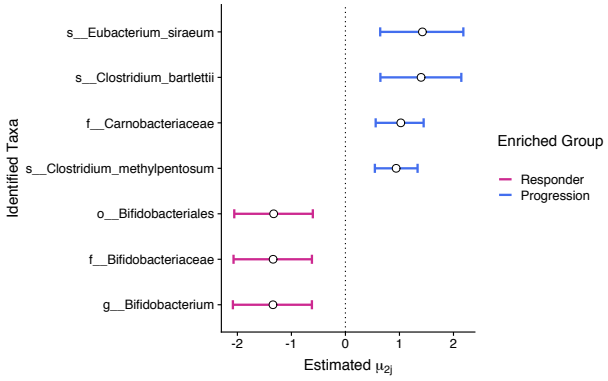
(a) Liver cirrhosis study: plot of γ PPI



(b) Metastatic melanoma study: plot of γ PPI



(c) Liver cirrhosis study: credible interval for μ_{2j}



(d) Metastatic melanoma study: credible interval for μ_{2j}

Figure 3.4: Real Data Analysis: Plots for γ PPI and credible interval. The horizontal dashed line in the PPI plot represents the threshold controlling the Bayesian false discovery rate < 0.05. All taxa whose PPI pass the threshold are included in (c) and (d), where each horizontal bar is the 95% credible interval for μ_{2j} (group-specific parameter) with posterior mean shown in circle. Each arrow in (a), (b) points out the taxon with largest absolute value of μ_{2j} in one patient group as shown in Figure 3.4c and 3.4d.

3.5.2. Metastatic melanoma case study

The proposed Bayesian ZINB model can perform integrative analysis of microbiome taxonomic data and other omics datasets. In this section, we applied this model to simultaneously analyze microbiome and metabolomics data from a study of advanced stage melanoma patients receiving immune checkpoint inhibitor therapy (ICT) [38]. The data were collected using MSS and unbiased shotgun metabolomics. Here, we aim at identifying unique microbiome taxonomic and metabolomic signatures in those patients who responded favorably to ICT.

A subset of patients in this study ($n = 24$) were treated with ipilimumab and nivolumab(IN), a combination therapy that has been shown to be more efficacious than therapy with anti-PD1 or anti-CTLA4 therapy alone. 16 patients responded to treatment and 8 patients had progression. We performed quality control steps on MSS reads and profiled them using MetaPhlAn [111] as described in [38]. We filtered out taxa with at most one observation in either patient group, which left $p = 248$ taxa from species to kingdom level. For the same fecal samples, we performed metabolomics profiling and quantified 1,901 patients' metabolite compounds as the covariate matrix $\boldsymbol{X}$. We are interested in statistically assessing how the biochemical volumes between patient groups are associated with bacteria burden or quantities. We adopted the same strategy mentioned in section 3.5.1 to reduce the correlation between covariates, which resulted in a 24×9 matrix as the covariate matrix of $\boldsymbol{X}$.

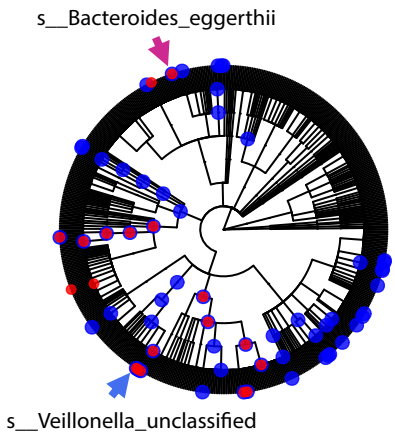
In the model fitting stage, for prior specification, we used $a_p = 0.2,\ b_p = 1.8$ to obtain a sparser covariate effect due to the small sample size, and it suggested about 10% of taxa-covariate associations were significant. We kept the same default setting for the rest of the hyperparameters. Next, we ran four independent chains with different starting points, and discarded the first half of 40,000 iterations for each chain. Although the small sample size ($n = 24$) posed challenges for parameter estimation, the results showed high pairwise

Pearson correlations of PPIs for γ (ranging from 0.989 to 0.992) and δ (ranging from 0.927 to 0.953). Figure 3.4b shows the PPIs for all taxa, and Figure 3.4b(d) illustrates the posterior means of the selected taxa. Table B.2 includes detailed parameter estimations of the taxa in Figure 3.4b(d).

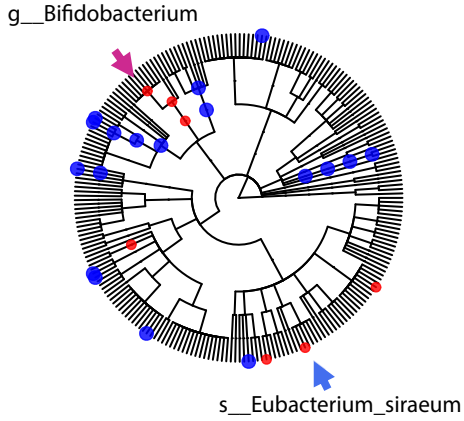
Our model jointly identified differentially abundant taxa and revealed the microbiome-metabolite associations. First, among all seven taxa identified, it is of specific interest to investigate the responder-enriched taxon *Bifidobacterium* (genus level), *Bifidobacteriaceae* (family level). *Bifidobacterium*, nesting within *Bifidobacteriaceae*, is a genus of gram-positive, nonmotile, often branched anaerobic bacteria [109]. *Bifidobacteria* are one of the major genera of bacteria that make up the gastrointestinal tract microbiota in mammals. This result about *Bifidobacterium* is supported by recent melanoma studies. [112] compared melanoma growth in mice harboring specific microbiota, and used sequencing of the 16S ribosomal RNA to identify *Bifidobacterium* as associated with the antitumor effects. They also found that oral administration of *Bifidobacterium* augmented ICT efficacy. Moreover, [85] detected significant association between several species from *Bifidobacterium* with patients' outcomes in an immunotherapy treatment study for metastatic melanoma. Both studies showed consistent direction of effect, as did our model. The responder-enriched taxon *Bifidobacterium* were estimated to negatively correlate with 2-oxoarginine and 2-hydroxypalmitate in Figure 3.4(d). The suppression of these fatty-acid metabolites may induce better cancer treatment as they were shown to have the oncogenic signaling role in cancer cells [79].

3.6. Discussion

In this chapter, we presented a Bayesian ZINB model for the integrative analysis of high-throughput sequencing microbiome data. Our method is novel in simultaneously incorporating the effect from measurable genetic covariates and identifying differentially abundant taxa for multiple patient groups in one statistical framework. This allows for integrative analysis of microbiome data and other omics data. Our method is flexible, as



(a) Liver cirrhosis study: Cladogram for discriminating taxa



(b) Metastatic melanoma study: Cladogram for discriminating taxa

Figure 3.5: Real Data Analysis: Cladograms of the identified discriminating taxa (shown in dots). Red dots: taxa found by the proposed model; Blue dots: taxa found by methods reported in the original studies. Each arrow in (a), (b) points out the taxon with the largest absolute value of μ_{2j} (group-specific parameter) in one patient group, as shown in Figure 3.4c and 3.4d.

it allows for identification and estimation of the association between covariates and each taxon's abundance. These results could potentially guide clinical decisions for precision shaping of the microbiome, although results would need to be validated in preclinical models first. In addition, our method is computationally efficient in posterior inferences. We implemented the MCMC algorithm to analyze the data from two MSS studies with results readily available in minutes.

In real data analysis, the identified differentially abundant taxa by our model are often cluttered in the same phylogenetic branch. These results are achieved without imposing the phylogenetic structures in the model. This highlights that the results from our model are biologically interpretable and thus capable of guiding further biological mechanism studies. Our results on the metastatic melanoma study uncover novel relationships between taxa and metabolites which merit further experimental investigation.

CHAPTER 4

A Hybrid Model for Microbiome Networks Analysis

4.1. Literature review

There are two major categories of statistical methods that are often used to infer microbial abundance networks. The first type is based on a taxa abundance covariance structure. For example, [32] and [125] used pairwise Pearson correlations to represent edge weights. This simple inference could be problematic since two variables (i.e., taxa) may be connected in the network due to their confounding variables [45]. The other type aims to estimate taxa abundance partial correlations, removing confounding effects. [61] proposed a statistical model for inferring microbial ecological network, which is based on estimating the precision matrix (via exploiting sparsity) of a Gaussian multivariate model and relies on graphical lasso (Glasso) [39]. However, their data normalization step needs to be improved to account for unique characteristics observed in microbiome count data.

Microbiome sequencing data usually have an inflated amount of zeros, uneven sequencing depths across samples, and over-dispersion. As initial attempts of constructing microbial association networks with this type of data, [8, 77] first transformed the microbiome sequencing counts into their compositional formula. Specifically, a count was normalized to its proportion in the respective sample. Then, each sample were transformed by a choice of log-ratio transformations to remove the unit-sum constraint of the compositional data. While this type of normalization is simple to implement and preserves the original ordering of the counts in a sample, it fails to capture the sample to sample variation and it overlooks the excess zeros in the microbiome data. Note that these zeros can be attributed to biological or technical reasons: either certain taxa are not present

among samples, or they are not sequenced due to insufficient sequencing depths. As the existing logarithmic transformation neglects the difference between these two types of zeros, it can lead to a biased estimation of the network structure.

4.2. Model

4.2.1. Microbiome count data normalization

Let $\boldsymbol{Y}$ denote the n-by-p taxonomic count matrix obtained from either the 16S rRNA or the metagenomic shotgun sequencing (MSS) technology. Each entry $y_{ij}, i = 1, \ldots, n, j = 1, \ldots, p$ is a non-negative integer, indicating the total reads related to taxon j observed in sample i. It is recommended that all chosen taxa should be at the same taxonomic level (e.g., OTU for 16S rRNA or species for MSS) in that mixing different taxonomic levels in the proposed model could lead to improper biological interpretation . As the real microbiome data are characterized by zero-inflation and over-dispersion, we model y_{ij} through a zero-inflated negative binomial (ZINB) model as

$$y_{ij} \sim \pi_i \mathsf{I}(y_{ij} = 0) + (1 - \pi_i)\mathsf{NB}(\lambda_{ij}, \phi_j). \tag{4.1}$$

The first component in the Equation (4.1) models whether zeros come from a degenerate distribution with a point mass at zero. It can be interpreted as the “extra" zeros due to insufficient sequencing effort. We can assume there exists a true underlying abundance for the taxon in its sample, but we fail to observe it with the mixture probability π_i representing the proportion of “extra" zeros in sample i. The second component, NB(λ_{ij}, ϕ_j), models the “true" zeros and all the nonzero observed counts. i.e., counts generated from a negative binomial (NB) distribution with the expectation of λ_{ij} and dispersion $1/\phi_j$. Here, “true" zero refers to a taxon that is truly absent in the corresponding sample. The variance of the random variable from NB distribution, under the current parameterization equals to

$\lambda_{ij} + \lambda_{ij}^2/\phi_j$. Smaller values of ϕ_j can lead to over-dispersion.

To avoid explicitly fixing the value of π_i's and ϕ_j's, we use a Bayesian hierarchical model for parameter inference. First, we rewrite the model (2.2) by introducing a binary indicator variable $\eta_{ij} \sim$ Bernoulli(π_i), such that $y_{ij} = 0$ if $\eta_{ij} = 1$, and $y_{ij} \sim$ NB(λ_{ij}, ϕ_j) if $\eta_{ij} = 0$. Then, we formulate a beta-Bernoulli prior of η_{ij} by assuming $\pi_i \sim$ Beta(a_π, b_π), and we let $a_\pi = b_\pi = 1$ to obtain a non-informative prior on η_{ij}. We specify independent Gamma prior Ga(a_ϕ, b_ϕ) for each dispersion parameter ϕ_j. Letting $a_\phi = b_\phi = 0.001$ results in a weakly informative gamma prior.

The mean parameter of the NB distribution, λ_{ij}, contains the key information of the true underlying abundance of the corresponding count. As λ_{ij} is affected by the varying sequencing effort across samples, we use a multiplicative characterization of the NB mean to justify the latent heterogeneity in microbiome sequencing data. Specifically, we assume $\lambda_{ij} = s_i \alpha_{ij}$. Here, s_i is the sample-specific size factor that captures the variation in sequencing depth across samples, and α_{ij} is the normalized abundance of taxon j in sample i.

In parameter estimation, one need to ensure identifiability between s_i and α_{ij}. For example, s_i can be the reciprocal of the total number of reads in sample i. The resulted α_{ij} is often called relative abundance, which represents the proportion of taxon j in sample i. In this setting, the relative abundances of all the taxa in one sample always sum up to 1. Similarly, other methods have been proposed with different constraints for normalizing the sequencing data [5, 14, 97, 106]. Some normalization methods can perform better than the others in the downstream analysis (e.g.„ the differential abundance analysis) under certain settings. From a Bayesian perspective, fixing the values of s_i's imposes a strongly informative prior in model inference. Hence, all these methods could bias the estimations of other model parameters and degrade the performance of downstream analyses. We thus propose a regularizing prior with a stochastic constraint for estimating s_i's. Our method can simultaneously infer the size factor and other model parameters. In

particular, we adopt the following mixture model for s_i,

$$\log s_i \sim \sum_{m=1}^{M} \psi_m \left[t_m \mathsf{N}(\nu_m, \sigma_s^2) + (1 - t_m) \mathsf{N}\left(-\frac{t_m \nu_m}{1 - t_m}, \sigma_s^2 \right) \right], \tag{4.2}$$

where ψ_m is the weight for outer mixtures of the mth component. The inner mixture of the mth component consists of two Gaussian distributions with t_m and $1 - t_m$ as weights, respectively. It is straightforward to see that the inner mixture has a mean of zero and thus ensuring the stochastic constraint of $\mathsf{E}(\log s_i) = 0$. For the outer mixtures, M is an arbitrary large positive integer. Letting $M \to \infty$ and defining the weight ψ_m by the stick-breaking procedure (i.e., $\psi_1 = V_1, \psi_m = V_m \prod_{u=1}^{m-1}(1 - V_u), m = 1, 2, \ldots$) makes model (4.2) a special case of Dirichlet process mixture models. This class of Bayesian nonparametric infinite mixtures is widely used in quantifying the model uncertainty and allowing for flexibility in parameter estimation [63, 114]. In particular, this Dirichlet process prior (DPP) has been used to account for sample heterogeneity since it is able to capture multi-modality and skewness in a distribution [67, 72]. In practice, we set M to be a large positive integer, and adopt the following hyper-prior distributions for the parameters in (4.2) such that $\nu_m \sim \mathsf{N}(0, \tau_\nu), t_m \sim \mathsf{Beta}(a_t, b_t)$, and $V_m \sim \mathsf{Beta}(a_m, b_m)$. We further set $\sigma_s^2 = 1$ to complete the parameter specification in the DPP prior.

In our model, the normalized abundance matrix $\boldsymbol{A} = \{\alpha_{ij}\}$ represents the true underlying abundance of the original count matrix. We further assume $\log \alpha_{ij} \sim \mathsf{N}(\mu_j, \sigma_j^2)$. This variance-stabilizing transformation on each α_{ij} not only reduces the skewness of the normalized abundance, but converts the nonnegative α_{ij} to a real number. We apply the following conjugate setting to specify the priors for μ_j and $\sigma_j^2, j = 1, \ldots, p$. We let $\mu_j \sim \mathsf{N}(0, h_0 \sigma_0^2)$ and $\sigma_j^2 \sim$ inverse-gamma(a_0, b_0). After integrating out μ_j and σ_j^2, the prior of the normalized abundances of taxon j follows a non-standardized Student's

t-distribution, i.e.,

$$p(\boldsymbol{\alpha}_{\cdot j}) = (nh_0+1)^{-\frac{1}{2}} \frac{\Gamma\left(a_0+\frac{n}{2}\right)}{\Gamma(a_0)} \frac{b_0^{a_0}}{\left\{b_0+\frac{1}{2}\left[\sum_{i=1}^n \log \alpha_{ij}^2 - \frac{\left(\sum_{i=1}^n \log \alpha_{ij}\right)^2}{n+\frac{1}{h_0}}\right]\right\}^{a_0+\frac{n}{2}}}. \tag{4.3}$$

The logarithmic scale of $\boldsymbol{A}$, denoted as $\boldsymbol{Z} = \log \boldsymbol{A}$, represents the normalized microbiome abundances on the log scale. We use Markov chain Monte Carlo (MCMC) algorithm for model parameter estimation (see details in the Appendix C.1), and calculate the posterior mean of $\boldsymbol{Z}$ to fit the Gaussian graphical model in the next step. Since the observed zero counts may not always represent the absence of taxa in the samples, we treat these zeros differently in the matrix $\boldsymbol{Z}$. We categorize the two types of zeros ("extra" and "true" zeros) based on the estimated η_{ij} for each observed $y_{ij} = 0$ in the data. In particular, suppose that we observe L zeros in total. We calculate the marginal posterior probability of being 1 for each $\eta_l, l = 1, \ldots, L$ as $p_l = \sum_{b=1}^B \mathsf{I}\left(\eta_l = 1\right)/B$, where $\mathsf{I}(\cdot)$ is the indicator function, and B is the number of MCMC iteration after burn-in. This marginal posterior probability p_l represents the proportion of MCMC iterations in which the lth 0 is essentially a missing value rather than the lowest count in the corresponding sample. Then, the observed zeros can be dichotomized by thresholding the L probabilities. The zeros with p_l greater than the threshold are considered as "true" zeros in the data, whereas the rest are imputed by the corresponding posterior mean of $\log \boldsymbol{\alpha_{\cdot j}}$. We used the method proposed by [94] to determine the threshold that controls the Bayesian false discovery rate (FDR) to be smaller than c_η. Specifically, we first specify a small number c_η, which is analogue to the significance level in the frequentist setting. Then we compute the threshold following Equation (4.4), which guarantees the imputed zeros have a Bayesian FDR to be smaller than c_η,

$$\text{Bayesian FDR} = \frac{\sum_{l=1}^L (1-p_l)\,\mathsf{I}\left(1-p_l < c_\eta\right)}{\sum_{l=1}^L \left(1-p_l < c_\eta\right)}. \tag{4.4}$$

In practice, a choice of $c_\eta = 0.01$ guarantees that the Bayesian FDR to be at most 0.01. We set $c_\eta = 0.05$ for the simulation study and $c_\eta = 0.01$ for the real data analysis.

4.2.2. Graphical model for inferring taxa-taxa association

Based on the normalized microbial abundances, we estimate their partial correlation matrix in order to construct the microbiome network under the Gaussian graphical model (GGM) framework. An undirected graph $G = (V, E)$ is used to illustrate the associations among vertices $V = \{1, \ldots, p\}$, representing the p microbial taxa. $E = \{e_{mk}\}$ is the collection of (undirected) edges, which is equivalently represented via a p-by-p adjacency matrix with $e_{mk} = 1$ or 0 according to whether vertices m and k are directly connected in G or not. GGM assumes that the joint distribution of p vertices is multivariate Gaussian N$(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, yielding the following relationship between the dependency structure and the network: a zero entry in the precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ indicates the corresponding vertices are conditional independent, and there is no edge between them in graph G. Hence, a GGM can be defined in terms of the pairwise conditional independence. If $\boldsymbol{X} \sim$ N$(\boldsymbol{\mu}, \boldsymbol{\Omega})$, then

$$\omega_{mk} = 0 \Leftrightarrow X_m \perp X_k | X_{V \setminus \{m,k\}} \Leftrightarrow \rho_{mk} = 0,$$

where $\rho_{mk} = -\omega_{mk}/\sqrt{\omega_{mm}\omega_{kk}}$ is the partial correlation between vertices m and k, representing the degree and direction of association between two vertices, conditional on the rest variables. Consequently, learning the network is equivalent to estimating the precision matrix $\boldsymbol{\Omega}$. For real microbiome data, we set the taxa (on the same taxonomic level) as vertices. Hence, a zero partial correlation in the precision matrix can be interpreted as no association between the corresponding pair of taxa, while a nonzero partial correlation can be interpreted as cooperative or competing associations between that taxa pair.

In biological applications, we often require a sparse and stable estimation of the precision matrix Ω. The sparsity can be achieved by imposing l_1-penalized log-likelihood,

$$\hat{\Omega} = \underset{\Omega \succ 0}{\text{argmin}} \log \det \Omega - \text{trace}(S\Omega) - \lambda \|\Omega\|_1 , \tag{4.5}$$

where S is the sample covariance matrix. The coordinate descent algorithm can iteratively solve p. The estimated precision matrix is sparsistent (i.e., all the parameters that are zeros would be estimated as zero with probability one) [66], as Glasso theoretically guarantees a consistent recovery of the sparse graph for the p vertices. When $p >> n$, the computational efficiency is often satisfactory, and thus Glasso is widely used in studying large-scale biological networks [90, 95, 139]. We employ a stability-based approach to select the tuning parameter in the Glasso, which is named Stability Approach to Regularization Selection (StARS) [75]. This method is an improved algorithm for estimating the tuning parameter λ in (4.5). The StARS selects the optimal sparsity parameter according to the graph reproduciblity under the subsampling of the original data. In general, for each λ along the sparsity parameter path, we first obtain random subsamples from the original data. Then we estimate the graph for each subsample using the Glasso. Next, for each sparsity parameter, we calculate the overall edge selection instability from all the graphs constructed by the subsamples. Finally, the optimal sparsity parameter λ^* is chosen such that it corresponds to the smallest amount of regularization and still results in a graph instability to be lower than the pre-specified tolerance level. [75] showed that StARS could provide the “sparsistent" network estimation that includes all the true associations with probability one. Further, the StARS has been widely used in biological network studies [61, 117, 139]. Due to its excellent performance, here we adopt the StARS to select the tuning parameter for Glasso. In summary, we use the normalized abundances (on the log scale) as inputs, calculate the sparse estimation of the precision matrix using the Glasso, and use the StARS method to select λ in problem (4.5) to obtain the estimated graph that represents the microbiome network.

4.3. Simulation

We compare the performance of the HARMONIES and several widely used methods for inferring microbiome networks. These methods include SPIEC-EASI [61], CClasso [30] and correlation-based network estimation used in [32, 125] . While the proposed model and SPIEC-EASI infer the network structure from sparse precision matrices, CClasso, and the correlation-based method utilize sparse correlation matrices to represent the network. We generated both simulated and synthetic datasets that mimic the real microbiome sequencing count data. We use $\boldsymbol{Y}_{n\times p}$ to denote the generated count matrix. For a comprehensive comparison, we varied the sample size and the number of taxa as $n \in \{60, 100, 200, 500\}$, and the number of taxa $p \in \{40, 60\}$.

4.3.1. Generative model for simulated data

We generated the simulated datasets from a Dirichlet-multinomial (DM) model using the following steps: (1) to generate the binary adjacency matrix; (2) to simulate the precision matrix and the corresponding covariance matrix; (3) to generate n multivariate Gaussian variables based on the covariance matrix to represent the true $n \times p$ underlying taxonomic abundances, denoted as $\boldsymbol{D}$; (4) to simulate the count table $\boldsymbol{Y}_{n\times p}$ from a DM model, with its parameters being $\exp(\boldsymbol{D})$; (5) to mimic the zero-inflation in real microbiome data by randomly setting part of entries in the count table to zeros. Note that the data generative scheme is different from the model assumption, which is given in Equation (2.2). The detailed generative models are described below.

We began with simulating a p-by-p adjacency matrix for the p taxa in the network. Here, the adjacency matrix was generated according to an Erdős–Rényi (ER) model. An ER model ER(p, ρ) generates each edge in a graph G with probability ρ independently from every other edge. Therefore, all graphs with p nodes and M edges have equal probability of $\rho^M(1-\rho)^{\binom{p}{2}-M}$. All the edges in graph G correspond to the 1's in the resulted binary adjacency matrix. Next, we simulated the precision matrix $\boldsymbol{\Omega}$ following [98]. We started by

setting all the diagonal elements of $\mathbf{\Omega}$ to be 1. Then, for the rest elements that correspond to the 1s in the adjacency matrix, we sampled their values independently from a uniform distribution Unif$([-0.1, 0] \cup [0, 0.1])$. To ensure positive definiteness of the precision matrix, we followed [98] by dividing each off-diagonal element by 1.5 times the sum of the absolute value of all the elements in its row. Finally, we averaged the rescaled precision matrix with its transpose and set the diagonal elements to 1. This process ensured the preceding matrix was positive definite and symmetric. The corresponding covariance matrix was set as $\mathbf{\Sigma} = \mathbf{\Omega}^{-1}$.

Next, we simulated n multivariate Gaussian variables from MN($\boldsymbol{\mu}, \mathbf{\Sigma}$) to represent the true underlying abundances $\boldsymbol{D}$. To obtain a count matrix that fully mimics the microbiome sequencing data, we generated counts from a DM model with parameter $\exp(\boldsymbol{D})$. Specifically, we first sampled the underlying fractional abundances for the ith sample from a Dirichlet distribution. The ith underlying fractional abundances was then denoted as $\boldsymbol{\psi}_i \sim$ Dirichlet$(\exp(\boldsymbol{D}_{i\cdot}))$. Next, the counts in the ith sample were generated from Multinomial$(N_i, \boldsymbol{\psi}_i)$. Finally, we randomly selected π_0% out of $n \times p$ counts and set them to zeros to mimic the zero-inflation observed in the real microbiome data. In general, the generative process had different assumptions from the proposed method. Under the appropriate choice of parameters, the simulated count data was zero-inflated, overdispersed, and the total reads varied largely between samples. In practice, we let $\rho = 0.1$ in the ER model. The mean parameter $\boldsymbol{\mu}$ of the underlying multivariate Gaussian variable was randomly sampled from a uniform distribution Unif$[0, 10]$. The number of total counts across samples $N_i,\ i = 1, \ldots, n$ was sampled from a discrete uniform distribution with range $[50,000, 100,000]$. Under each combination of n, p, and π_0, we generated 50 replicated datasets by repeating the process above.

4.3.2. Generative model for synthetic data

We generated synthetic data following the Normal-to-Anything (NorTA) approach proposed in [61]. NorTA was designed to generate multivariate random variables with an arbitrary marginal distribution from a pre-specified correlation structure [17]. Given the observations of p taxa from a real microbiome dataset, the NorTA generates the synthetic data with n samples as follows: (1) to calculate the p-by-p covariance matrix $\mathbf{\Sigma}_0$ from the input real dataset; (2) to generate an n-by-p matrix, denoted by $\boldsymbol{Z}_0$, from a multivariate Gaussian distribution with mean of $\mathbf{0}_{1\times p}$ and the covariance matrix of $\mathbf{\Sigma}_0$; (3) to use standard normal cumulative distribution function to scale values in each column of $\boldsymbol{Z}_0$ within $[0, 1]$; (4) to apply the quantile function of a ZINB distribution to generate count data from those scaled values in each column of $\boldsymbol{Z}_0$. In practice, we used `R` package `SPIEC-EASI` to implement the above data generative scheme, where the real data were from those healthy control subjects in our case study presented in Section 4.4. Under each combination of n and p, we generated 50 replicated datasets.

4.3.3. Prior and algorithm settings

The hyperparameters were specified using the following default settings. As for the fixed parameters a_0, b_0, h_0 and σ_0^2, we follow [71] and set $a_0 = 2, b_0 = 1$ to obtain a weakly informative prior for σ_j^2. We fix $\sigma_0^2 = 1$ and let $h_0 = 10$ such that the normal prior on μ_j is fairly flat. We adopt the following prior specification for the rest model parameters. First, we assume an noninformative prior for each π_i by letting $a_\pi = b_\pi = 1$. Next, we specify $a_\phi = b_\phi = 0.001$ in the Gamma prior distribution for all ϕ_j's. Then, we apply the following prior setting for the DPP: $M = n/2,\ \sigma_s = 1,\ \tau_\nu = 1,\ a_t = b_t = 1$, and $a_m = b_m = 1$. We set $20,000$ iterations as the default and discarded the first half as burn-in.

4.3.4. Alternative methods

We considered the four commonly used network learning methods. The first two methods, SPIEC-EASI-Glasso and SPIEC-EASI-mb, use the transformed microbiome abundances which are different from the normalized abundances estimated by HARMONIES. Both infer the microbial network by estimating a sparse precision matrix. The former method (SPIEC-EASI-Glasso) measures the dependency among microbiota by their partial correlation coefficients, and the latter method (SPIEC-EASI-mb) uses the “neighborhood selection" introduced by [89] to construct the network. The third method, denoted as Pearson-corr, calculates Pearson’s correlation coefficients between all pairs of taxa. In its estimated network, the edges correspond to large correlation coefficients. To avoid arbitrarily thresholding the correlation coefficients, the fourth method, CClasso [30], directly infers a sparse correlation matrix with l_1 regularization. However, as discussed in Section 4.1, representing the dependency structure by the correlation matrix may lead to the detection of spurious associations.

4.3.5. Evaluation metrics

We quantified the model performances on the simulated data by computing their receiver operating characteristic (ROC) curves and area under the ROC curve (AUC). For the HARMONIES or SPIEC-EASI, the network inference was based on the precision matrix. Hence, under each tuning parameter of Glasso, we calculated the number of edges being true positive (TP) by directly comparing the estimated precision matrix against the true one. More specifically, we considered an edge between taxon m and taxon k to be true positive if $\omega_{mk} \neq 0$, $\hat{\omega}_{mk} \neq 0$, and $\hat{\omega}_{mk}$ shared the same sign with ω_{mk}. We calculated the number of true negative (TN), false positive (FP), and false negative (FN) in a similar manner. Therefore, each tuning parameter defined a point on an ROC curve. As for the correlation-based methods, we started with ranking the absolute values in the estimated

correlation matrices, denoted as $\hat{C}$. Next, we used each value as a threshold and set all the entries in $\hat{C}$ having their absolute values smaller than the current threshold to be zeros. Then, the number of TP, TN, FP, or FN was obtained by comparing the sparse $\hat{C}$ against the true partial correlation matrix. Therefore, each unique absolute value in the original estimated correlation matrix defined a point on the ROC curve.

We further used the Matthew's correlation coefficient (MCC) to evaluate results from the simulated data. The MCC is defined as

$$\frac{(\text{TP} \times \text{TN} - \text{FP} \times \text{FN})}{\sqrt{(\text{TP+FP})(\text{TP+FN})(\text{TN+FP})(\text{TN+FN})}}.$$

Here, the MCC was particularly suitable for evaluating network models. As the number of conditionally independent taxa pairs was assumed to be much greater than the number of dependent pairs in a sparse network, MCC was preferable to quantify the performances under such an imbalanced situation. Note that MCC ranges from $[-1, 1]$, with a value close to 1 suggesting a better performance. Since each value of MCC was calculated using a given set of TP, TN, FP, and FN, we adopted the optimal choice of tuning parameter for the HARMONIES or SPIEC-EASI (with either Glasso or MB for network inference), given by StARS. As for the correlation-based methods, CClasso outputted a sparse correlation matrix. We used the result to calculate TP, TN, FP, and FN directly. For Pearson-corr, we set the threshold such that the resulted number of nonzero entries in the sparse correlation matrix was the same as the number of nonzero entries in the true sparse partial correlation matrix. In fact, this choice could favor the performance of Pearson-corr for larger sample size, as shown in Section 4.3.6.

To assess model performances on the synthetic datasets, we followed [61] to use a metric called area under the precision-recall curves (AUPR), in addition to AUC. Briefly speaking, the AUPR and AUC were calculated as follows: (1) to rank all possible edges according to their confidence values; (2) to generate the precision-recall curve and the

ROC curve by comparing edge inclusions against the true sparse precision matrix; (3) to calculate the area under the precision-recall curve or the ROC curve. Note that the confidence values were chosen as the edge stabilities under the optimal choice of the tuning parameter selected by StARS for HARMONIES, SPIEC-EASI-Glasso, and SPIEC-EASI-mb, while for CClasso and Pearson-corr, p-values were used.

4.3.6. Results

Figure 4.1 and 4.2 compare the AUCs and MCCs on the simulated data under various scenarios, including varying sample sizes ($n = 60, 100, 200$, or 500), total numbers of taxa ($p = 40$ or 60), extra percentages of zeros added ($\pi_0 = 10\%$, or 20%). In each subfigure, the HARMONIES outperformed the alternative methods in terms of both AUC and MCC, and it maintained this advantage even with the number of sample size greatly increases. Further, a smaller sample size, a larger proportion of extra zeros added ($\pi_0 = 20\%$), as well as a larger number of taxa in the network ($p = 60$), would hamper the performance of all the methods, as we expected. Two modes of SPIEC-EASI, SPIEC-EASI-Glasso, and SPIEC-EASI-mb, showed very similar performances under all the scenarios, with SPIEC-EASI-Glasso having only a marginal advantage over the other. Further, we observed that the Pearson-corr method yielded higher AUCs even than the precision matrix based methods, especially when there was a lager proportion of extra zeros or larger number of taxa in the network. This result suggested that the Pearson-corr could capture the overall rank of the signal strength in the actual network. However, under a fixed cut-off value that gave a sparse correlation network, the MCCs from the Pearson-corr were always smaller than the precision matrix based methods. Note that the cut-off value we specified for Pearson's correlation method indeed favored its performance. In general, the alternative methods considered here were able to reflect the overall rank of the signal strength by showing reasonable AUCs. However, they failed to give an accurate estimation of the network under a fixed cut-off value.

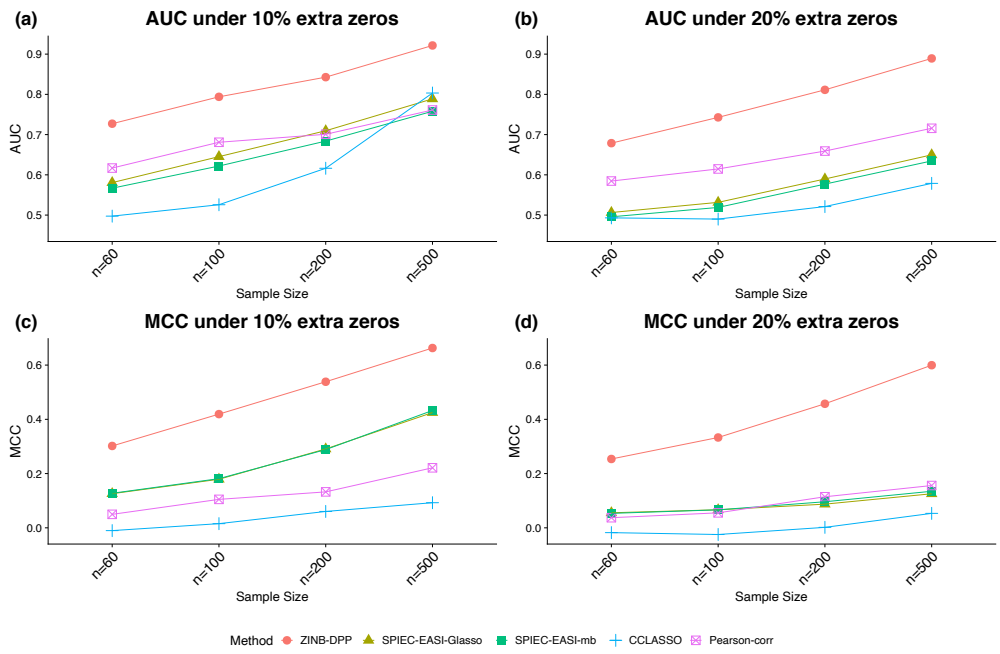


Figure 4.1: Simulated data: (a) and (b) area under the ROC curves (AUCs) and (c) and (d) area under the precision-recall curves (AUPRs) achieved by different methods under the number of taxa $p = 40$ and different sample sizes and zero proportions, averaged over 50 replicates.

Figure 4.3 demonstrates that our model outperformed all others on the synthetic datasets. The performances in terms of AUC under different scenarios are summarized in Figure 4.3(a) and (b), while those in terms of AUPR are displayed in (c) and (d). As we can see, either increasing the sample size n or decreasing the number of features p would improve the performance of all methods and lead to greater disparity between partial and pairwise correlation-based methods. In general, our HARMONIES maintained the best in all simulation and evaluation settings except for one case, where the SPIEC-EASI-mb only showed a marginal advantage (see $n = 60$ in Figure 4.3(c)). Interestingly, our observation confirmed a finding mentioned by [61], that is, the SPIEC-EASI-mb was slightly better than SPIEC-EASI-Glasso in terms of AUPR under the optimal choice of the tuning parameter. As for the two correlation-based methods, we found that Pearson-corr outperformed CClasso in most of the scenarios.

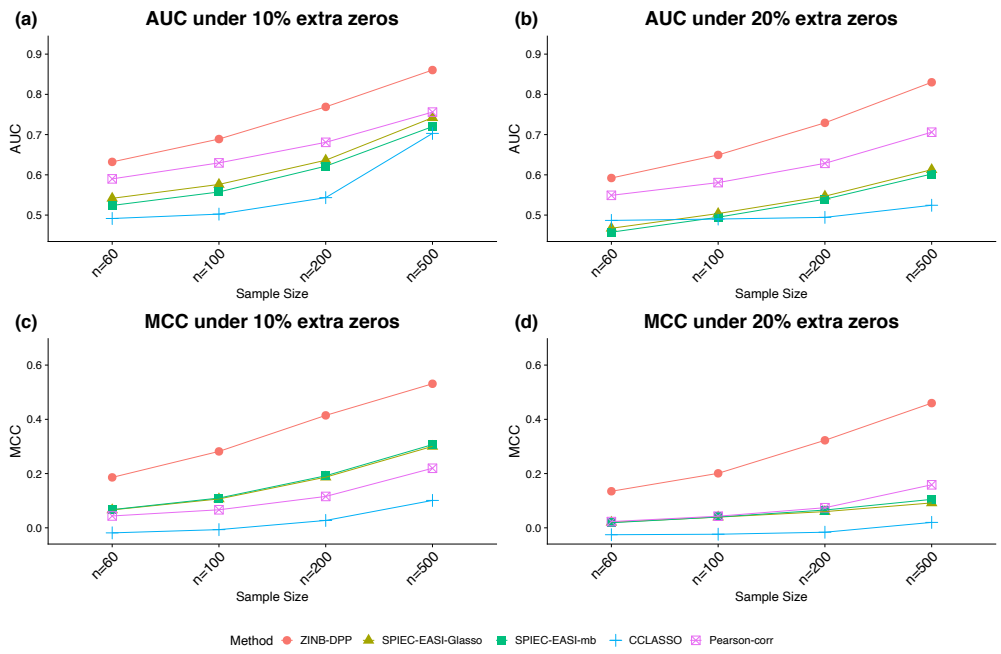


Figure 4.2: Simulated data: (a) and (b) area under the ROC curves (AUCs) and (c) and (d) area under the precision-recall curves (AUPRs) achieved by different methods under the number of taxa $p = 60$ and different sample sizes and zero proportions, averaged over 50 replicates.

4.4. Colorectal cancer case study

Colorectal cancer (CRC) is the third most common cancer diagnosed in both men and women in the United States [6]. Increasing evidence from the recent studies highlights a vital role for the intestinal microbiota in malignant gastrointestinal diseases including CRC [28, 80, 110]. In particular, studies have reported that dysbiosis of specific microbiota is directly associated with CRC [36, 58, 84]. The current microbiome research interests have gone beyond the discovery of disease-related microbiota, with a growing number of studies investigating the interactive associations among the microbial taxa. Using the proposed model, we interrogated the microbiome profiling data of a CRC study to determine the microbiome network structures.

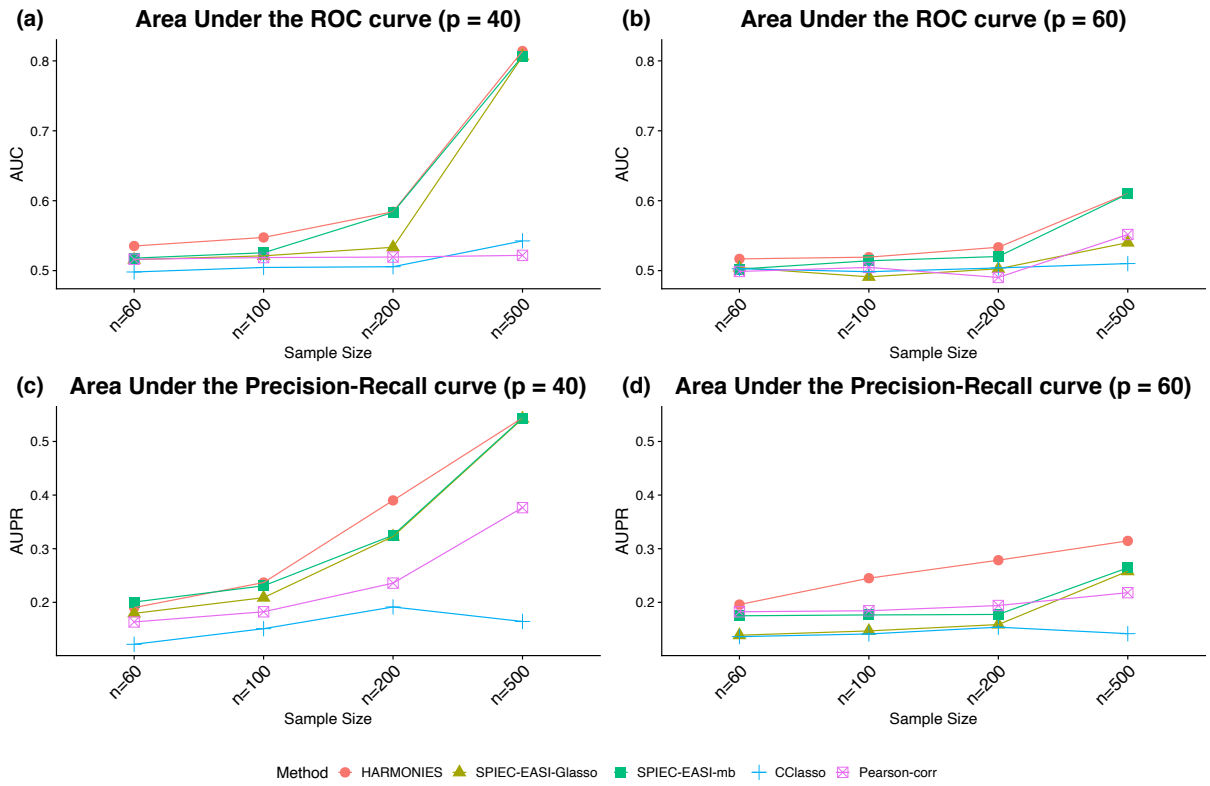


Figure 4.3: Synthetic data: (a) and (b) area under the ROC curves (AUCs) and (c) and (d) area under the precision-recall curves (AUPRs) achieved by different methods under different sample sizes and taxa numbers, averaged over 50 replicates.

We analyzed the gut microbiome dataset of a CRC study published by [33]. We extracted from the original cohort[1] the 43 CRC patients and the 58 healthy controls. The original sequencing data at the genus level were quantified using curatedMetagenomicData [96]. We had $p = 187$ genera for both the 43 CRC patients and the 58 healthy controls. We implemented the HARMONIES as follows. For the CRC group, we first applied the ZINB model to obtain the normalized abundance matrix $\boldsymbol{A}$, utilizing the specifications detailed in Section 4.2.1. We then took the logarithmic transformation of the normalized abundance and imputed the missing values. Before implementing the proposed method, we filtered out the low abundant genera with zeros occurring more than half samples. Removing low abundant taxa is a common step in microbiome research

[1]The original metagenomic shotgun sequencing data from the fecal samples are available in the European Bioinformatics Institute Database (accession number ERP008729)

[see e.g., 59, 61, 101, 122, 132, 135]. The rationale being that these “zero-abundant" taxa may be less important in a network, which was also confirmed by our simulation study. This filtering process left 51 and 36 genera in the CRC and control group, respectively.

Figure 4.4 (a) and (b) display the estimated networks for the CRC and the control group, respectively. Each node, corresponding to a genus, was named after its phylum level. All the genera shown in Figure 4.4 belong to six phyla in total. By using their phylum name to further categorize these distinct genera, we aimed at exploring interesting patterns among them at a higher taxonomic level. Figure C.1 displays the same network using the actual genus name on each node . The node sizes are proportional to its normalized abundances in the logarithmic scale. The green or red edge indicates a positive or a negative partial correlation, respectively. And the width of an edge is proportional to the absolute value of the partial correlation coefficient. To make a clear comparison, we intentionally kept the nodes and their positions to be consistent between the two subfigures. In either of the two groups, we included a node in the current plot if there exists an edge between it with any nodes in at least one group. In general, the two groups share several edges with the same direction of partial correlations, but the majority of edges are unique within each group.

Network estimation of the CRC group demonstrated several microbial communities. For example, three genera: *Fusobacterium*, *Peptostreptococcus*, and *Parvimonas* consisted of a unique subnetwork as highlighted in Figure 4.4 (a). These three genera were isolated in the control group’s network, as shown in Figure 4.4 (b). Interestingly, specific species under these three genera have been reported as enriched taxa in CRC and related to worse clinical outcome [78, 92, 134]. A previous CRC study by [58] supported the causal role of species *Fusobacterium nucleatum* (*F. nucleatum*) by showing that *F. nucleatum* promotes tumor progression by increasing both tumor multiplicity and tumor-infiltrating myeloid cells in a preclinical CRC model. Further, a recent study [78] demon-

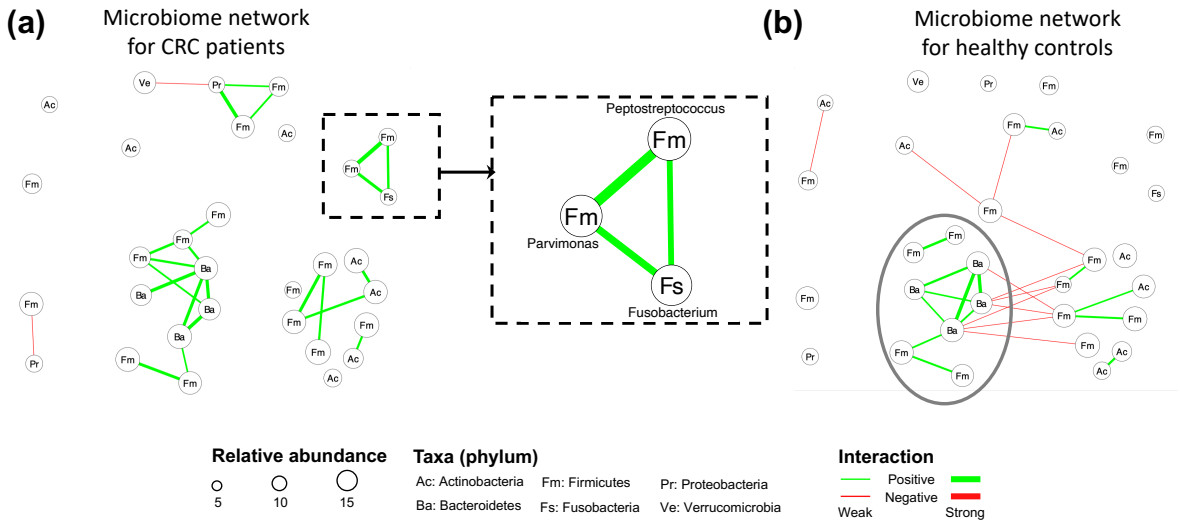


Figure 4.4: CRC case study: The estimated networks by HARMONIES for (a) CRC patients and (b) healthy controls. Increased abundances of species under the three genera (*Fusobacterium, Peptostreptococcus, Parvimonas*) in the dashed rectangular box in (a) were reported to be associated with the disease. CRC patients and healthy controls shared a similar subnetwork (composed of eight genera) circled in (b). Each node here represents a genus labeled by its phylum name. The version with distinct genus names is available in Figure C.1 in the Appendix.

strated that *Peptostreptococcus anaerobius* (*P. anaerobius*) accelerated colorectal tumorigenesis in a murine CRC model. This study suggested that *P. anaerobius* directly interacted with colonic epithelial cells and also promoted CRC by modifying the tumor immune microenvironment. While the causal role of the species *Parvimonas micra* (*P. micra*) has not been biologically validated, multiple clinical studies reported an elevated level of *P. micra* in CRC patients [26, 100, 134]. Of interest, *Parvimonas* were closely associated with animal-based diets, which have previously been shown to be significantly associated with increased risk for CRC [20]. The previous studies only investigated those CRC-related taxa individually, whereas a novel finding by HARMONIES analysis suggested that all the three genera were co-aggregating in CRC patients as their pairwise associations are all positive. Interestingly, in a prior study direct positive associations between *Fusobacterium* and *Peptostreptococcus*, as well as *Peptostreptococcus* and *Parvimonas*, were identified [48]. However, there was no direct association between *Fusobacterium* and *Parvimonas*.

Similarly, another study [29] found a direct co-occurrence pattern between two species: *F. nucleatum* and *P. micra*. Using HARMONIES, we could jointly identify the relationship among each pair of the three genera, conditional on all other genera. This novel subcommunity of three CRC-enriched genera formulated a recurring module, and may function as a cooperative group in CRC patients. A closer investigation of their co-occurrence pattern could potentially elucidate both their contributions to CRC and the basic biology under their relationships. Two additional novel taxa interactions were identified by HARMONIES analysis: *Streptococcus* and *Veillonella*, and *Streptococcus* and *Haemophilus*. In fact, previous CRC studies showed enrichment of these three genera or their species in CRC patients [see e.g., 43, 56, 60, 120], but had not detected these novel interactions. In conclusion, HARMONIES may reveal how multiple CRC-related taxa could potentially promote disease progression together.

Having shared edges between the two networks suggests that the HARMONIES is robust to the edge selection. We observed that the shared edges tended to appear for those more abundant genera. For example, we circled eight genera in Figure 4.4 (b), and the HARMONIES suggested multiple positive partial correlations among them. For these eight genera, we observed six shared edges between the CRC and healthy control networks. Notice that all the shared edges were consistent in the association directions, and they also corresponded to the relatively stronger association in both networks (wider in the edge width). We found these shared edges tend to connect those more abundant genera (node with larger size). Indeed, the eight genera considered here belong to phyla *Bacteroidetes* and *Firmicutes*, both were in the top three most abundant phyla for CRC patients and healthy controls reported by [40, 93]. Therefore, it was more likely that the highly abundant genera shared similar association patterns between the two groups, and the HARMONIES demonstrated its robustness by preserving these relatively stronger partial correlations among these genera. On the other hand, the network of the control group contained more negative partial correlations as shown in Figure 4.4 (b). Furthermore, the two edges linked to *Streptococcus* were different from the CRC group. Here, *Strep-*

tococcus had a negative association with *Subdoligranulum* and a positive association with *Rothia*. There has been no evidence suggesting these two genera are CRC-related. Hence a further investigation is merited. Additionally, the CRC group has another distinct small subnetwork formed by the four genera, two from *Firmicutes*, one from *Proteobacteria*, and one from *Verrucomicrobia*. These group-specific associations were never reported. Lastly, we observed several interesting patterns between the two groups when summarizing the genera to their phylum levels. Genera in Firmicutes (labeled as “Fm" in Figure 4.4) showed more positive associations in the case group than in the control group, whereas negative associations between Firmicutes and Bacteroidetes (labeled as “Ba" in Figure 4.4) were more common in the control group. Again, these novel patterns still need further biological validations to elucidate their functions.

4.5. Discussion

With the advent of next-generation sequencing technology, microbiome research now has the opportunity to explore microbial community structure and to characterize the microbial ecological association for different populations or physiology conditions [61]. In this chapter, we introduce HARMONIES as a statistical framework to infer sparse networks using microbiome sequencing data. It models the original count data by a zero-inflated negative binomial distribution to capture the large amount for zeros and over-dispersion, and it further implements Dirichlet process priors to account for sample heterogeneity. In contrast, current methods for microbiome network analyses rely on the compositional data, which could cause information loss due to ignoring the unique characteristics of the microbiome sequencing count data. Following the data normalization step, the HARMONIES explores the direct connections in the network by estimating the partial correlations. The results from the simulation study have demonstrated the advantage of the HARMONIES over alternative approaches under various conditions. When applied to an actual microbiome dataset, the HARMONIES suggests all the nodes be taxa at the same taxonomic level, such as species, genus, and family. This ensures proper biologically interpretations of those detected associations. When applied to a real CRC study, the HARMONIES

revealed an intriguing community among three CRC-enriched genera. Further, shared patterns between the CRC and the control networks suggest a common community pattern of disease neutral genera. Additional studies validating the biological relevance of these microbial associations, however, will need to be conducted.

Both the simulated and synthetic data showed that a larger sample size improved the performance of all the network learning methods. In practice, many disease-related microbiome studies, especially those studying rare diseases, always have small sample sizes. This limitation directly affects the estimation of the normalized matrix $\boldsymbol{A}$ from the ZINB model. Notice that for a taxon j, a small sample size could result in a large variance in the posterior distribution of $\log \boldsymbol{\alpha}_{\cdot \boldsymbol{j}}$. However, many disease studies include reference groups where the measurements on the same taxonomic features are available. The additional information from the subjects in the reference group can potentially help improve the posterior inference of the normalized abundances. We generalized the proposed ZINB model to handle two groups, with the goal of borrowing information between groups in estimating the normalized abundances. These detailed model formula and implementation were included in the Appendix (see Appendix C.2: Infer the normalized abundances for multiple groups). Our current method can infer the normalized abundances for two groups, and we provided the details steps in the Appendix C.2. However, an integrated differential network can be expected to better study the differential microbial community structure and link the communities to human health status.

CHAPTER 5

Conclusions and Future Directions

This thesis presents three statistical frameworks developed for analyzing metagenomics sequencing data in the context of human microbiome research. These models have taken into account the important characteristics of the microbiome sequencing count data, including the high-dimensionality, zero-inflation, and overdispersion of the taxa counts.

In the microbiome differential abundance analysis, our proposed method has shown great promise in detecting the differentially abundant signatures across phenotype groups, as illustrated in both simulation and synthetic data analyses. The consistency between taxa identified by the ZINB-DPP model and the current biological literature have further supported the reliability of our model. When it comes to predicting the subject's phenotype using the differentiating taxa detected by different methods, the model built by the ZINB-DPP detected taxa shows significantly lower prediction error compared with others. Our model framework can be naturally extended to other analysis scenarios. For example, the inferred latent abundance can be treated within a sample normalized distribution. It is thus applicable to longitudinal analysis, which can capture the dynamic structure in microbiome studies; or to differential network analysis, which can investigate the complex interactions among microbial taxa. In all, the proposed Bayesian framework provides more powerful microbiome differential abundance analyses and is suitable for multiple types of microbiome data analysis.

The microbiome integrative analysis associates the microbiome abundances with genetic covariates. Our method is novel in simultaneously identifying differentially abundant taxa for multiple patient groups and incorporating the effects from measurable covariates in one statistical framework. Our model allows for several extensions. For example, the

current method supports two phenotype groups. If there are multiple groups (e.g., the intermediate phenotypes), the current model can incorporate group-specific parameters while holding the other parameters unchanged (e.g., the normalized microbiome abundance can be inferred in the same way). Then the same posterior inferences can be applied. The proposed model based on a regression framework considers the microbiome normalized abundance as the response and integrates the omics data, e.g., metabolite compounds, as predictors. Similarly, [76] used the microbial abundance as the response and a type of genomics data (i.e., single nucleotide polymorphism, SNP) as predictors to identify several inflammatory bowel disease (IBD)-associated host-microbial interactions. Both methods focus on the omics effect on microbial abundance. However, the interaction between the microbiome and the host is bidirectional. Therefore, it is worthwhile to consider using the microbial features as predictors to investigate their modulations on any biological process with quantitative omics measurements. For instance, [102] explored how microbial abundances induced changes in chromatin accessibility and transcription factor binding of host genetics. Another interesting extension would be to analyze correlated covariates such as longitudinal clinical measurements [137].

In the microbiome network analysis, we propose HARMONIES, a hybrid approach that explores the direct connections between taxa in the network by estimating the partial correlations. Our hybrid approach for microbiome network inference can be extended. One future direction is to incorporate the differential network analysis into the existing framework. It jointly considers the association strengths between each pair of taxa from different groups, and it compares the estimated individual networks to capture the significantly different connectivities.

APPENDIX A

APPENDIX of CHAPTER 2

A.1. Dirichlet-multinomial model

One commonly used candidate of the multivariate count variable generating process $\mathcal{M}$ is the Dirichlet-multinomial (DM) model [see e.g. 21, 49, 64, 122]. To illustrate the model, we start by modeling the counts observed in subject i with a multinomial distribution $\boldsymbol{y}_{i\cdot}|\boldsymbol{\psi}_{i\cdot} \sim \text{Multi}(Y_{i\cdot}, \boldsymbol{\psi}_{i\cdot})$. The p-dimensional vector $\boldsymbol{\psi}_{i\cdot} = (\psi_{i1}, \ldots, \psi_{ip})^T$ is defined on a p-dimensional simplex (i.e. $\psi_{ij} > 0, \forall j$ and $\sum_{j=1}^{p} \psi_{ij} = 1$), and represents the underlying taxonomic abundances. The p.m.f. is $Y_{i\cdot} \prod_{j=1}^{p} \psi_{ij}^{y_{ij}} / y_{ij}!$, with the mean and variance of each component, $\text{E}(Y_{ij}) = \psi_{ij} Y_{i\cdot}$ and $\text{Var}(Y_{ij}) = \psi_{ij}(1 - \psi_{ij}) Y_{i\cdot}$, respectively.

We further impose a Dirichlet prior on the multinomial parameter vector to allow for over-dispersed distributions, $\boldsymbol{\psi}_{i\cdot}|\boldsymbol{\alpha}_{i\cdot} \sim \text{Dir}(\boldsymbol{\alpha}_{i\cdot})$, where each element of the p-dimensional vector $\boldsymbol{\alpha}_{i\cdot} = (\alpha_{i1}, \ldots, \alpha_{ip})^T$ is strictly positive. Due to the conjugacy between the Dirichlet distribution and the multinomial distribution, we can integrate $\boldsymbol{\psi}_{i\cdot}$ out, $p(\boldsymbol{y}_{i\cdot}|\boldsymbol{\alpha}_{i\cdot}) = \int p(\boldsymbol{y}_{i\cdot}|\boldsymbol{\psi}_{i\cdot}) p(\boldsymbol{\psi}_{i\cdot}|\boldsymbol{\alpha}_{i\cdot}) d\boldsymbol{\psi}_{i\cdot}$. The resulting DM model: $\boldsymbol{y}_{i\cdot}|\boldsymbol{\alpha}_{i\cdot} \sim \text{DM}(\boldsymbol{\alpha}_{i\cdot})$, has the following p.m.f.

$$f_{\text{DM}}(\boldsymbol{y}_{i\cdot}|\boldsymbol{\alpha}_{i\cdot}) = \frac{\Gamma(Y_{i\cdot} + 1)\Gamma(A_{i\cdot})}{\Gamma(Y_{i\cdot} + A_{i\cdot})} \prod_{j=1}^{p} \frac{\Gamma(y_{ij} + \alpha_{ij})}{\Gamma(y_{ij} + 1)\Gamma(\alpha_{ij})},$$

where $Y_{i\cdot} = \sum_{j=1}^{p} y_{ij}$ and $A_{i\cdot} = \sum_{j=1}^{p} a_{ij}$. The variance of each count variable is $\text{Var}(Y_{ij}) = (Y_{i\cdot} + A_{i\cdot})/(1 + A_{i\cdot})\text{E}(\psi_{ij})(1 - \text{E}(\psi_{ij}))Y_{i\cdot}$. Comparing this with the multinomial model, we see that the variance of the DM is inflated by a factor of $(Y_{i\cdot} + A_{i\cdot})/(1 + A_{i\cdot})$, Thus, the DM distribution can explicitly model extra variation. Note that $A_{i\cdot} = \sum_{j=1}^{p} \alpha_{ij}$ controls the

degree of over-dispersion. A small value of $A_{i\cdot}$ results in large over-dispersion, while a large value approaching infinity reduces the DM model to a multinomial model. Although the DM model offers more flexibility than the multinomial model in terms of modeling over-dispersion, neither models accounts for zero-inflation.

A.2. Details of the MCMC algorithms

We show the details of the MCMC algorithms of the proposed Bayesian framework, where taxonomic structure is taken into account. For a simple cases where the data are only available at genus or OTU level for 16S rRNA sequencing data, or at species level for metagenomic shotgun sequencing data, please ignore the superscript $(2), \ldots, (l)$.

A.2.1. Bottom level

A.2.1.1. Dirichlet-multinomial (DM) model

We start by writing the likelihood for each sample $i, i = 1, \ldots, n$, where the microbiome abundance is summarized at the bottom-most taxonomic levels, i.e. $l = 1$,

$$f_{\mathsf{DM}}(\boldsymbol{y}_{i\cdot}^{(1)}|\boldsymbol{\alpha}_{i\cdot}^{(1)}) = \frac{\Gamma(Y_{i\cdot}+1)\Gamma(A_{i\cdot})}{\Gamma(Y_{i\cdot}+A_{i\cdot})}\prod_{j=1}^{p^{(1)}}\frac{\Gamma(y_{ij}^{(1)}+\alpha_{ij}^{(1)})}{\Gamma(y_{ij}^{(1)}+1)\Gamma(\alpha_{ij}^{(1)})}.$$

Note that $Y_{i\cdot} = \sum_{j=1}^{p^{(1)}} y_{ij}^{(1)} \cdots = \cdots \sum_{j=1}^{p^{(L)}} y_{ij}^{(L)}$ and $A_{i\cdot} = \sum_{j=1}^{p^{(1)}} a_{ij}^{(1)} \cdots = \cdots \sum_{j=1}^{p^{(L)}} a_{ij}^{(L)}$; that is, the total read counts and the total normalized abundance should be unchanged, regardless of the choice of taxonomic levels.

A.2.1.2. Zero-inflated negative binomial (ZINB) model

We start by writing the likelihood for each sample $i, i = 1, \ldots, n$, where the microbiome abundance is summarized at level l,

$$f_{\text{ZINB}}(\boldsymbol{y}_{i\cdot}^{(l)}|\boldsymbol{\alpha}_{i\cdot}^{(l)}, \boldsymbol{\eta}_{i\cdot}, \boldsymbol{\phi}^{(l)}, s_i) = \prod_{j=1}^{p^{(l)}} f_{\text{ZINB}}(y_{ij}^{(l)}|\alpha_{ij}^{(l)}, \eta_{ij}, \phi_j^{(l)}, s_i),$$

where

$$\begin{aligned} & f_{\text{ZINB}}(y_{ij}^{(l)}|\alpha_{ij}^{(l)}, \eta_{ij}, \phi_j^{(l)}, s_i) \\ =& \text{I}(y_{ij}^{(1)} = 0)^{\eta_{ij}} \left(\frac{\Gamma(y_{ij}^{(l)} + \phi_j^{(l)})}{y_{ij}^{(l)}! \, \Gamma(\phi_j^{(l)})} \left(\frac{\phi_j^{(l)}}{s_i \alpha_{ij}^{(1)} + \phi_j^{(l)}} \right)^{\phi_j^{(l)}} \left(\frac{s_i \alpha_{ij}^{(l)}}{s_i \alpha_{ij}^{(l)} + \phi_j^{(l)}} \right)^{y_{ij}^{(l)}} \right)^{1-\eta_{ij}}. \end{aligned}$$

Update of zero-inflation indicator η_{ij}: We update each $\eta_{ij}, i = 1, \ldots, n, j = 1, \ldots, p^{(1)}$ that corresponds to $y_{ij}^{(1)} = 0$ by sampling from the normalized version of the following conditional:

$$p(\eta_{ij}|\cdot) \propto f_{\text{ZINB}}(y_{ij}^{(1)}|\alpha_{ij}^{(1)}, \eta_{ij}, \phi_j, s_i) \cdot \text{Bern}(\eta_{ij}; \pi_i).$$

After the Metropolis-Hasting steps for all η_{ij}, we use a Gibbs sampler to update each $\pi_i, i = 1, \ldots, n$:

$$\pi_i|\cdot \sim \text{Be}(a_\pi + \sum_{j=1}^{p^{(1)}} \eta_{ij}, b_\pi + p^{(1)} - \sum_{j=1}^{p^{(1)}} \eta_{ij}).$$

Update of dispersion parameter $\phi_j^{(l)}$: We update each $\phi_j^{(l)}, j = 1, \ldots, p^{(l)}, l = 1, \ldots, L$ by using a random walk Metropolis-Hastings algorithm. We first propose a new $\phi_j^{(l)^*}$ from $\text{Ga}(\phi_j^{(l)^2}/\tau_\phi, \phi_j^{(l)}/\tau_\phi)$ and then accept the proposed value $\phi_j^{(l)^*}$ with probability $\min(1, m_{\text{MH}})$,

where

$$m_{\mathsf{MH}} = \frac{\prod_{i=1}^{n} f_{\mathsf{ZINB}}(y_{ij}^{(l)}|\alpha_{ij}^{(l)}, \eta_{ij}, \phi_j^{(l)}, s_i)}{\prod_{i=1}^{n} f_{\mathsf{ZINB}}(y_{ij}^{(l)}|\alpha_{ij}^{(l)}, \eta_{ij}, \phi_j^{(l)}, s_i)} \frac{\mathsf{Ga}(\phi_j^{(l)^*}; a_\phi, b_\phi)}{\mathsf{Ga}(\phi_j^{(l)}; a_\phi, b_\phi)} \frac{J(\phi_j^{(l)}; \phi_j^{(l)^*})}{J(\phi_j^{(l)^*}; \phi_j^{(l)})}.$$

Here we use $J(\cdot|\cdot)$ to denote the proposal probability distribution for the selected move. Note that the last term, which is the proposal density ratio, can be canceled out for this random walk Metropolis update.

Update of size factor s_i: We can rewrite Equation (4) in the main text, i.e.

$$\log s_i \sim \sum_{m=1}^{M} \psi_m \left[t_m \,\mathsf{N}(\nu_m, \sigma_s^2) + (1 - t_m)\,\mathsf{N}\left(-\frac{t_m \nu_m}{1 - t_m}, \sigma_s^2 \right) \right]$$

by introducing latent auxiliary variables to specify how each sample (in terms of $\log s_i$) is assigned to any of the inner and outer mixture components. More specifically, we can introduce an $n \times 1$ vector of assignment indicators $\boldsymbol{g}$, with $g_i = m$ indicating that $\log s_i$ is a sample from the m-th component of the outer mixture. The weight ψ_m determines the probability of each value $g_i = m$, with $m = 1, \ldots, M$. Similarly, we can consider an $n \times 1$ vector $\boldsymbol{\epsilon}$ of binary elements ϵ_i, where $\epsilon_i = 1$ indicates that, given $g_i = m$, $\log s_i$ is drawn from the first component of the inner mixture, i.e. $\mathsf{N}(\nu_m, \sigma_s^2)$ with probability t_m, and $\epsilon_i = 0$ indicates that $\log s_i$ is drawn from the second component of the inner mixture, i.e. $\mathsf{N}\left(-\frac{t_m \nu_m}{1 - t_m}, \sigma_s^2 \right)$, with probability $1 - t_m$. Thus, the Dirichlet process prior (DPP) model can be rewritten as

$$\log s_i | g_i, \epsilon_i, \boldsymbol{t}, \boldsymbol{\nu} \sim \mathsf{N}\left(\epsilon_i \nu_{g_i} + (1 - \epsilon_i) \frac{-t_{g_i} \nu_{g_i}}{1 - t_{g_i}}, \sigma_s^2 \right),$$

where $\boldsymbol{t}$ and $\boldsymbol{\nu}$ denote the collections of t_m and ν_m, respectively. Therefore, the update of the size factor $s_i, i = 1, \ldots, n$ can proceed by using a random walk Metropolis-Hastings algorithm. We propose a new $\log s_i^*$ from $\mathsf{N}(\log s_i, \tau_s^2)$ and accept it with proba-

bility $\min(1, \mathsf{m}_{\mathsf{MH}})$, where

$$\mathsf{m}_{\mathsf{MH}} = \frac{\prod_{j=1}^{p^{(1)}} f_{\mathsf{ZINB}}(y_{ij}^{(1)}|\alpha_{ij}^{(1)}, \eta_{ij}, \phi_j^{(1)}, s_i^*)}{\prod_{j=1}^{p^{(1)}} f_{\mathsf{ZINB}}(y_{ij}^{(1)}|\alpha_{ij}^{(1)}, \eta_{ij}, \phi_j^{(1)}, s_i)} \frac{\mathsf{N}(\log s_i^*; \epsilon_i \nu_{g_i} + (1-\epsilon_i)\frac{-t_{g_i}\nu_{g_i}}{1-t_{g_i}}, \sigma_s^2)}{\mathsf{N}(\log s_i; \epsilon_i \nu_{g_i} + (1-\epsilon_i)\frac{-t_{g_i}\nu_{g_i}}{1-t_{g_i}}, \sigma_s^2)} \times \frac{J(\log s_i; \log s_i^*)}{J(\log s_i^*; \log s_i)}.$$

Note that the last term, which is the proposal density ratio, equals 1 for this random walk Metropolis update. Since $\boldsymbol{g}$, $\boldsymbol{\epsilon}$, $\boldsymbol{t}$, and $\boldsymbol{\nu}$ have conjugate full conditionals, we use Gibbs samplers to update them one after another:

- Gibbs sampler for updating $g_i, i = 1, \ldots, n$, by sampling from the normalized version of the following conditional:

$$p(g_i = m|\cdot) \propto \psi_m \mathsf{N}\left(\log s_i; \epsilon_i \nu_m + (1-\epsilon_i)\frac{-t_m \nu_m}{1-t_m}, \sigma_s^2\right).$$

- Gibbs sampler for updating $\epsilon_i, i = 1, \ldots, n$, by sampling from the normalized version of the following conditional:

$$p(\epsilon_i|\cdot) \propto \begin{cases} (1-t_m)\mathsf{N}\left(\log s_i; -\frac{t_m\nu_m}{1-t_m}, \sigma_s^2\right) & \textit{if } \epsilon_i = 0 \\ t_m \mathsf{N}\left(\log s_i; \nu_m, \sigma_s^2\right) & \textit{if } \epsilon_i = 1 \end{cases}.$$

- Gibbs sampler for updating $t_m, m = 1, \ldots, M$:

$$t_m|\cdot \sim \mathsf{Be}(a_t + \sum_{i=1}^n \mathsf{I}(g_i = m)\mathsf{I}(\epsilon_i = 1), b_t + \sum_{i=1}^n \mathsf{I}(g_i = m)\mathsf{I}(\epsilon_i = 0)).$$

- Gibbs sampler for updating $\nu_m, m = 1, \ldots, M$:

$$\nu_m|\cdot \sim \mathsf{N}\left(\frac{c_m/\sigma_s^2}{e_m/\sigma_s^2 + 1/\tau_\nu^2}, \frac{1}{e_m/\sigma_s^2 + 1/\tau_\nu^2}\right),$$

where $c_m = \sum_{\{i:g_i=m, \epsilon_i=1\}} \log s_i - \frac{t_m}{1-t_m}\sum_{\{i:g_i=m,\epsilon_i=0\}} \log s_i$ and $e_m = \sum_{i=1}^n \mathsf{I}(g_i =$

$m)\mathsf{I}(\epsilon_i = 1) + \sum_{\{i:g_i=m,\epsilon_i=0\}} \left(\frac{t_m}{1-t_m}\right)^2$.

- Gibbs sampler for updating $\psi_m, m = 1, \ldots, M$ by stick-breaking process [52]:

$$\begin{aligned}
\psi_1 &= v_1, \\
\psi_2 &= (1 - v_1)v_2, \\
&\vdots \\
\psi_M &= (1 - v_1) \cdots (1 - v_{M-1})v_M,
\end{aligned}$$

where $v_m|\boldsymbol{\nu} \sim \mathsf{Be}\left(a_m + \sum_{i=1}^n \mathsf{I}(g_i = m), b_m + \sum_{i=1}^n \mathsf{I}(g_i > m)\right)$.

A.2.2. Top level

Both of the DM model and the ZINB model share the same process to update the normalized abundance matrix at the bottom-most taxonomic level, i.e. $\boldsymbol{A}^{(1)}$, and to select the discriminatory taxa at different levels, i.e. $\boldsymbol{\gamma}^{(1)}, \ldots, \boldsymbol{\gamma}^{(L)}$. For the sake of convenience, we have copied Equation (2.6) in the main text here,

$$p(\boldsymbol{\alpha}_{\cdot j}^{(l)}|\gamma_j^{(l)}) = (2\pi)^{-\frac{n}{2}} \times$$
$$\begin{cases}
\prod_{k=1}^K (n_k h_k + 1)^{-\frac{1}{2}} \frac{\Gamma\left(a_k + \frac{n_k}{2}\right)}{\Gamma(a_k)} \frac{b_k^{a_k}}{\left\{b_k + \frac{1}{2}\left[\sum_{\{i:z_i=k\}} \log \alpha_{ij}^{(l)^2} - \frac{\left(\sum_{\{i:z_i=k\}} \log \alpha_{ij}^{(l)}\right)^2}{n_k + \frac{1}{h_k}}\right]\right\}^{a_k + \frac{n_k}{2}}} \\
\qquad\qquad \text{if } \gamma_j^{(l)} = 1 \\
(nh_0 + 1)^{-\frac{1}{2}} \frac{\Gamma\left(a_0 + \frac{n}{2}\right)}{\Gamma(a_0)} \frac{b_0^{a_0}}{\left\{b_0 + \frac{1}{2}\left[\sum_{i=1}^n \log \alpha_{ij}^{(l)^2} - \frac{\left(\sum_{i=1}^n \log \alpha_{ij}^{(l)}\right)^2}{n + \frac{1}{h_0}}\right]\right\}^{a_0 + \frac{n}{2}}} \\
\qquad\qquad \text{if } \gamma_j^{(l)} = 0
\end{cases}.$$

Update of normalized abundance at the bottom-most level $a_{ij}^{(1)}$: We update each $\alpha_{ij}^{(1)}, i = 1, \ldots, n, j = 1, \ldots, p^{(1)}$ by using a Metropolis-Hastings random walk algorithm.

We first propose a new $\alpha_{ij}^{(1)^*}$ from $\mathsf{N}(\alpha_{ij}^{(1)}, \tau_\alpha^2)$, and then accept the proposed value with probability $\min(1, \mathsf{m_{MH}})$, where

$$\mathsf{m_{MH}} = \frac{f_{\mathcal{M}}(\boldsymbol{y}_{i\cdot}^{(1)}|\boldsymbol{\alpha}_{i\cdot}^{(1)^*}, \cdot)}{f_{\mathcal{M}}(\boldsymbol{y}_{i\cdot}^{(1)}|\boldsymbol{\alpha}_{i\cdot}^{(1)}, \cdot)} \frac{p\left(\boldsymbol{\alpha}_{\cdot j}^{(1)^*}|\gamma_j^{(1)}\right)}{p\left(\boldsymbol{\alpha}_{\cdot j}^{(1)}|\gamma_j^{(1)}\right)} \frac{J\left(\alpha_{ij}^{(1)}; \alpha_{ij}^{(1)^*}\right)}{J\left(\alpha_{ij}^{(1)^*}; \alpha_{ij}^{(1)}\right)}.$$

Here we use $\mathcal{M}$ to denote the bottom level model, which should be chosen from $\{\mathsf{DM}, \mathsf{ZINB}\}$. Note that the last term, which is the proposal density ratio, equals 1 for this random walk Metropolis update.

Update of differentially abundant taxon indicator $\gamma_j^{(l)}$: We update each $\gamma_j^{(l)}, j = 1, \ldots, p^{(l)}, l = 1, \ldots, L$ via an *add-delete* algorithm. In this approach, a new candidate vector, say $\boldsymbol{\gamma}^{(l)^*}$, is generated by randomly choosing an element within $\boldsymbol{\gamma}^{(l)}$, say j, and changing its value to $1 - \gamma_j^{(l)}$. Then, this proposed move is accepted with probability $\min(1, \mathsf{m_{MH}})$, where the Hastings ratio is

$$\mathsf{m_{MH}} = \frac{p\left(\boldsymbol{\alpha}_{\cdot j}^{(l)}|\gamma_j^{(l)^*}\right)}{p\left(\boldsymbol{\alpha}_{\cdot j}^{(l)}|\gamma_j^{(l)}\right)} \frac{p\left(\gamma_j^{(l)^*}|\cdot\right)}{p\left(\gamma_j^{(l)}|\cdot\right)} \frac{J\left(\boldsymbol{\gamma}^{(l)}|\boldsymbol{\gamma}^{(l)^*}\right)}{J\left(\boldsymbol{\gamma}^{(l)^*}|\boldsymbol{\gamma}^{(l)}\right)}.$$

Note that the proposal density ratio equals 1. Here, we have two choices of $p\left(\gamma_j^{(l)}|\cdot\right)$, either independent Bernoulli prior or Markov random field prior (see Equation (7) in the main text). We should also notice that the feature selection and the abundance estimation are determined simultaneously in the MCMC algorithm. Therefore, to improve mixing, it is necessary to allow the selection to stabilize for any visited configurations of $\boldsymbol{A}^{(1)}$ and its induced $\boldsymbol{A}^{(l)}$'s. We suggest repeating the above Metropolis step multiple times within each iteration. In the simulations conducted for this Chapter, no improvement in the MCMC performance was noticed after repeating the step above 20 times.

Update of normalized abundance at upper levels $a_{ij}^{(l)}, l \geq 2$: For the DM model, the aggregation property can be used to derive the normalized abundance at upper levels

sequentially just from the one at the bottom level via $\alpha_{ij}^{(l)} = \sum_{\{j':g_{jj'}=1\}} \alpha_{ij'}^{(l-1)}$. For the ZINB model, the aggregation property does not hold. We assume that the size factor estimation should be irrelevant to the choices of microbiome count data at different taxonomic levels. Therefore, we update each $\alpha_{ij}^{(l)}, i = 1, \ldots, n, j = 1, \ldots, p^{(l)}, l = 2, \ldots, L$ by using a Metropolis-Hastings random walk algorithm conditional on the size factors estimated by $\boldsymbol{Y}^{(1)}$. We first propose a new $\alpha_{ij}^{(l)*}$ from $\mathsf{N}(\alpha_{ij}^{(l)}, \tau_\alpha^2)$, and then accept the proposed value with probability $\min(1, \mathsf{m}_{\mathsf{MH}})$, where

$$\mathsf{m}_{\mathsf{MH}} = \frac{f_{\mathsf{ZINB}}(y_{ij}^{(l)} | \alpha_{ij}^{(l)*}, \eta_{ij}, \phi_j^{(l)}, s_i)}{f_{\mathsf{ZINB}}(y_{ij}^{(l)} | \alpha_{ij}^{(l)}, \eta_{ij}, \phi_j^{(l)}, s_i)} \frac{p\left(\boldsymbol{\alpha}_{\cdot j}^{(l)*} | \gamma_j^{(l)}\right)}{p\left(\boldsymbol{\alpha}_{\cdot j}^{(l)} | \gamma_j^{(l)}\right)} \frac{J\left(\alpha_{ij}^{(l)}; \alpha_{ij}^{(l)*}\right)}{J\left(\alpha_{ij}^{(l)*}; \alpha_{ij}^{(l)}\right)}.$$

Note that the last term, which is the proposal density ratio, equals 1 for this random walk Metropolis update.

A.3. Sensitivity analysis

We examined the ZINB model sensitivity with respect to the choice of hyperparameters $b_0, \ldots, b_K$ and $h_0, \ldots, h_K$ in the top level as discussed in Section 2.2.2. The results in Table A.5 show that our approach is considerably insensitive to the hyperparameter settings.

The choice of b_k and h_k for $k = 0, \ldots, K$ are related to the variance terms in the Gaussian mixture model in the top level. Large values of h_k would achieve a noninformative prior on μ_{kj}'s. On the other hand, as we specify $\mathsf{IG}(a_k, b_k)$ prior for σ_{kj}^2 with $a_k = 2$ for all $k = 0, \ldots, K$, the resulting variance of inverse gamma distribution does not exist. We considered a range of (b_k, h_k) settings as $b_k \in \{0.1, 1, 2, 10\}$ and $h_k \in \{1, 10, 100\}$. Then we applied the ZINB-DPP model with different combinations of (b_k, h_k) to datasets simulated from the ZINB model discussed in Section 2.4.1. To fully assess the impact of hyperparameters under different scenarios, we considered $K = 2, 3$ and $n = 24, 108$ with a weak log effect size of $\sigma = 1$. We generated 50 independent datasets for each scenario and reported the averaged AUCs (in Table A.5). Clearly, the AUCs remain stable under

different choices of (b_k, h_k). We suggest to set $b_k = 1$ and h_k to be any value ranging from 10 to 100 for $k = 1, \ldots K$.

A.4. Additional results for the colorectal cancer study

A.4.1. Quality control

Before analyzing a given microbiome count dataset, we first implement a simple quality control step. It ensures that the dataset is of the best quality to perform the subsequent modeling. This step includes: 1) examining the total number of reads sequenced, and 2) verifying the richness of taxa discovered. In all, quality control is considered for both sample (patient) and feature (taxon) levels.

A.4.1.1. Sample-wise quality control

In sequencing data analysis, if the total number of reads for a sample falls above or below specific values (discussed below), then this may indicate poor sequence quality owing to duplicate reads or limited sampling bias. Specifically, let $y_i = \sum_{j=1}^{p} y_{ij}$ denote the total number of reads observed in sample i. A sample i will be removed if $y_i <$ Q1$-3($Q3$-$Q1$)$ or $>$ Q3$+3($Q3$-$Q1$)$, where Q1 and Q3 are the lower and upper quartiles (i.e. the 25th and 75th percentiles) of the total reads of all the samples (i.e. $\{y_1, \ldots, y_n\}$). Note that in the context of box-and-whisker plotting, a data point is defined as an extreme outlier if it stands outside these two limits. Second, in ecology, investigators find that the number of species increases as sampling effort increases. This species-abundance distribution can be depicted by *the collector's curve*, which is monotonically increasing and negatively accelerated. Hence, we assume that the logarithmic count of taxa discovered in one sample had a linear relationship with the total reads observed in the same sample. As suggested by [46], we fit the regression model to compute the Cook's distance for each patient, and remove the ones with distances above $4/(n-2)$ since they would be

considered the influential data points for a least-squares regression analysis.

A.4.1.2. Feature-wise quality control

Another common procedure in microbiome studies is to filter out the extremely low-abundant taxa. For example, [122] requires each genus in their model to be present in at least 5% of the samples. Similarly, [101] keeps the taxa with median compositional abundance greater than 0.01% of total abundance in either the healthy control group or the disease group. In our ZINB model, the estimation of the dispersion parameter of each feature (taxon) involves the calculation of the second moment, similar to computing the variance component in the Gaussian mixture model. Therefore, it requires at least two observed reads in each patient group to perform the analysis. In practice, we suggest removing a taxon if it has fewer than three nonzero reads in any patient group.

A.4.2. Result comparison with alternative methods

Along with the simulation study conducted in the Section 2.4, we compared the results given by our proposed models (DM, ZINB-DPP) on the case study data with those from alternative approaches, including ANOVA, Kruskal–Wallis test, `WaVE-DESeq2`, `WaVE-edgeR` and `metagenomeSeq`.

We adopted a 1% significance level threshold on the BH-adjusted p-values provided by the alternative methods. The choice of 1% was set to be consistent with the Bayesian false discovery rate (FDR) of the ZINB-DPP model. For the DM model, we kept the same hyperprior settings as for the ZINB-DPP model, i.e., we set $a_0 = a_1 = \ldots = a_k = 2$, $b_0 = b_1 = \ldots = b_k = 1$ for variance components σ^2_{0j} and σ^2_{kj}, and we let $h_0 = h_1 = \ldots = h_K = 50$. We further adopted the same Markov random field settings as $d = -2.2$ and $f = 0.5$. The differentially abundant taxa selected by the DM model were obtained by controlling the Bayesian FDR to be less than 1% on the corresponding PPIs. The clado-

grams in Fig A.3 compare the taxa selected by all different methods. First, ANOVA lacked statistical power when the data contained too many zeros, and it failed to identify any discriminating taxa in this case. Therefore, we excluded the result by ANOVA in Fig A.3. The Kruskal–Wallis test identified 30 discriminating taxa, 19 of which were also reported by the ZINB-DPP model. Although Kruskal–Wallis selected the branch of species *Fusobacterium nucleatum* as all the other methods did, it failed to detect the taxonomic branch from *Synergistaceae* to *Synergistetes*, which was reported by the ZINB-DPP model and *Synergistaceae* was found to be CRC-enriched in a previous study [25]. Next, under a stringent significance level of 1%, `WaVE-DESeq2` and `WaVE-edgeR` still led the selection of 238 and 77 discriminating taxa , respectively. The large number of detections might suggest a high FDR. Furthermore, `WaVE-edgeR` failed to detect the taxonomic branch from *Synergistaceae* to its phylum level. Lastly, we found that `metagenomeSeq` and the DM model performed conservatively, as they only reported 20 and 27 discriminating taxa, respectively. 14 out of 20 taxa detected by `metagenomeSeq` were consistent with the result by ZINB-DPP, while 15 out of 27 findings from the DM model overlapped with results by the ZINB-DPP model. Although both of these methods reported *Fusobacterium nucleatum* to be differentially abundant between two groups, neither of them detected the co-occurrence between *Fusobacterium nucleatum* and *Campylobacter*.

A.5. Additional tables and figures

Table A.1: List of commonly used normalization techniques for sequencing count data

	Definition	Constraint	Reference
TSS	$\hat{s}_i \propto Y_{i\cdot}$	$\sum_{i=1}^{n} \log s_i = 0$	
[1]Q75	$\hat{s}_i \propto q_i^{0.75p}$,	$\sum_{i=1}^{n} \log s_i = 0$	[14]
RLE	$\hat{s}_i \propto \text{median}_j \left\{ y_{ij} / \sqrt[n]{\prod_{i'=1}^{n} y_{i'j}} \right\}$	$\sum_{i=1}^{n} \log s_i = 0$	[5]
[2]TMM	$\hat{s}_i \propto \sum_{j=1}^{p} y_{ij} \cdot \exp\left(\frac{\sum_{j \in G^*} \psi_j(i,r) M_j(i,r)}{\sum_{j \in G^*} \psi_j(i,r)} \right)$	$\sum_{i=1}^{n} \log s_i = 0$	[105]
[1]CSS	$\hat{s}_i \propto \sum_{j=1}^{p} y_{ij} \cdot \mathsf{I}(y_{ij} \leq q_i^{0.5p})$	$\sum_{i=1}^{n} \log s_i = 0$	[97]
DPP	$p(\log s_i \vert \cdot) = \sum_{m=1}^{M} \psi_m \Big[t_m \mathsf{N}(\nu_m, \sigma_s^2) + (1 - t_m) \mathsf{N}\left(\frac{c_s - t_m \nu_m}{1 - t_m}, \sigma_s^2 \right) \Big]$	$\mathsf{E}(\log s_i) = 0$	[72]

Abbreviations: TSS is total sum scaling, Q75 is upper-quantile (i.e. 75%), RLE is relative log expression, TMM is trimmed mean by M-values, CSS is cumulative sum scaling, and DPP is Dirichlet process prior.

[1]Note for Q75 and CSS: q_i^l is defined as the l-th quantile of all the counts in sample i, i.e. there are l features in sample i whose values y_{ij}'s are less than q_i^l.

[2]Note for TMM: the M-value $M_j(i,r) = \log(y_{ij}/Y_{i\cdot})/\log(y_{rj}/Y_{r\cdot})$ is the log-ratio of scaled counts between sample i and the reference sample r, if not within the upper and lower 30% of all the M-values (as well as the upper and lower 5% of all the A-values, defined as $A_j(i,r) = \log \sqrt{y_{ij}/Y_{i\cdot} \cdot y_{rj}/Y_{r\cdot}}$, and the corresponding weight $\psi_{j'}(i,r)$ is the inverse of the approximate asymptotic variances, calculated as $\frac{Y_{i\cdot} - y_{ij'}}{y_{ij'} Y_{i\cdot}} + \frac{Y_{r\cdot}}{y_{rj'} Y_{r\cdot}}$ by the delta method.

Table A.2: DM and ZINB simulation: area under the curve (AUC)

Generative Model	Methods	Simulation Setting							
		$K = 2$				$K = 3$			
		$n = 24$		$n = 108$		$n = 24$		$n = 108$	
		$\sigma = 1$	$\sigma = 2$	$\sigma = 1$	$\sigma = 2$	$\sigma = 1$	$\sigma = 2$	$\sigma = 1$	$\sigma = 2$
DM	ZINB-DPP	**1.000** (0.0102)	**1.000** (0.0000)	**1.000** (0.0000)	**1.000** (0.0000)	**1.000** (0.0003)	**1.000** (0.0000)	**1.000** (0.0000)	**1.000** (0.0000)
	ZINB-TSS	0.977 (0.0111)	1.000 (0.0002)	1.000 (0.0002)	1.000 (0.0001)	0.999 (0.0006)	1.000 (0.0002)	1.000 (0.0001)	1.000 (0.0002)
	ZINB-Q75	0.977 (0.0112)	1.000 (0.0001)	1.000 (0.0001)	1.000 (0.0001)	0.999 (0.0005)	**1.000** (0.0000)	**1.000** (0.0000)	1.000 (0.0001)
	ZINB-RLE	0.978 (0.0100)	1.000 (0.0001)	1.000 (0.0001)	**1.000** (0.0000)	1.000 (0.0004)	**1.000** (0.0000)	1.000 (0.0001)	1.000 (0.0001)
	ZINB-TMM	0.978 (0.0105)	1.000 (0.0001)	1.000 (0.0001)	**1.000** (0.0000)	1.000 (0.0005)	**1.000** (0.0000)	1.000 (0.0001)	1.000 (0.0001)
	ZINB-CSS	0.976 (0.0115)	**1.000** (0.0000)	1.000 (0.0000)	1.000 (0.0001)	0.999 (0.0008)	**1.000** (0.0000)	1.000 (0.0001)	1.000 (0.0001)
	DM	0.972 (0.0111)	1.000 (0.0002)	1.000 (0.0002)	1.000 (0.0001)	0.999 (0.0006)	1.000 (0.0002)	1.000 (0.0001)	1.000 (0.0002)
	ANOVA	0.977 (0.0101)	1.000 (0.0003)	1.000 (0.0003)	1.000 (0.0004)	0.999 (0.0010)	1.000 (0.0004)	1.000 (0.0003)	1.000 (0.0004)
	Kruskal–Wallis	0.978 (0.0102)	1.000 (0.0002)	1.000 (0.0003)	1.000 (0.0004)	0.999 (0.0008)	1.000 (0.0001)	1.000 (0.0003)	1.000 (0.0003)
	`WaVE-DESeq2`	0.921 (0.0154)	0.946 (0.0004)	0.997 (0.0000)	0.999 (0.0000)	0.933 (0.0218)	0.944 (0.0249)	0.997 (0.0031)	0.999 (0.0013)
	`WaVE-edgeR`	0.951 (0.0141)	0.958 (0.0003)	0.998 (0.0000)	0.998 (0.0000)	0.942 (0.0203)	0.956 (0.0225)	0.997 (0.0030)	0.999 (0.0014)
	`metagenomeSeq`	0.957 (0.0146)	0.998 (0.0017)	0.895 (0.0239)	0.918 (0.0199)	0.972 (0.0122)	0.999 (0.0015)	0.992 (0.0043)	0.994 (0.0040)
ZINB	ZINB-DPP	**0.907** (0.0203)	**0.990** (0.0095)	**0.994** (0.0051)	**0.998** (0.0039)	**0.982** (0.0085)	0.998 (0.0072)	**0.998** (0.0026)	**1.000** (0.0003)
	ZINB-TSS	0.888 (0.0240)	0.988 (0.0097)	0.993 (0.0059)	0.997 (0.0047)	0.975 (0.0125)	**0.998** (0.0065)	0.997 (0.0044)	1.000 (0.0005)
	ZINB-Q75	0.888 (0.0225)	0.988 (0.0101)	0.991 (0.0070)	0.997 (0.0051)	0.975 (0.0106)	0.998 (0.0043)	0.997 (0.0043)	1.000 (0.0005)
	ZINB-RLE	NA (-)	NA (-)	NA (-)	NA (-)	NA (-)	NA (-)	NA (-)	NA (-)
	ZINB-TMM	0.887 (0.0247)	0.988 (0.0103)	0.992 (0.0064)	0.997 (0.0045)	0.974 (0.0116)	0.998 (0.0062)	0.997 (0.0040)	**1.000** (0.0003)
	ZINB-CSS	0.881 (0.0245)	0.988 (0.0094)	0.991 (0.0073)	0.997 (0.0049)	0.973 (0.0128)	0.998 (0.0046)	0.997 (0.0047)	**1.000** (0.0003)
	DM	0.659 (0.0378)	0.856 (0.0261)	0.929 (0.0173)	0.990 (0.0109)	0.759 (0.0499)	0.947 (0.0245)	0.968 (0.0175)	0.993 (0.0098)
	ANOVA	0.714 (0.0601)	0.908 (0.0333)	0.972 (0.0126)	0.995 (0.0052)	0.659 (0.0689)	0.788 (0.0608)	0.989 (0.0066)	0.997 (0.0031)
	Kruskal–Wallis	0.547 (0.0528)	0.634 (0.0834)	0.824 (0.0331)	0.942 (0.0197)	0.509 (0.0211)	0.512 (0.0216)	0.892 (0.0299)	0.982 (0.0098)
	`WaVE-DESeq2`	0.611 (0.0383)	0.532 (0.0300)	0.911 (0.0110)	0.962 (0.0061)	0.611 (0.0465)	0.540 (0.0405)	0.924 (0.0142)	0.970 (0.0080)
	`WaVE-edgeR`	0.801 (0.0282)	0.920 (0.0171)	0.933 (0.0090)	0.989 (0.0061)	0.855 (0.0403)	0.913 (0.0327)	0.982 (0.0123)	0.996 (0.0060)
	`metagenomeSeq`	0.609 (0.0410)	0.766 (0.0443)	0.750 (0.0401)	0.933 (0.0252)	0.599 (0.0555)	0.682 (0.0697)	0.608 (0.0493)	0.686 (0.0535)

Area under the curve (AUC) given by all methods on the simulated data. In each cell, the top number is the averaged AUC over 50 independent datasets, and the bottom number in parentheses is the standard error. The result from the model that achieved best performance under the associated scenario (each column) is marked in bold.

Table A.3: DM and ZINB simulation: Matthews correlation coefficient (MCC)

Generative Model	Methods	Simulation Setting							
		$K = 2$				$K = 3$			
		$n = 24$		$n = 108$		$n = 24$		$n = 108$	
		$\sigma = 1$	$\sigma = 2$	$\sigma = 1$	$\sigma = 2$	$\sigma = 1$	$\sigma = 2$	$\sigma = 1$	$\sigma = 2$
DM	ZINB-DPP	**0.800** (0.0379)	0.997 (0.0074)	0.999 (0.0042)	**1.000** (0.0000)	**0.966** (0.0180)	**1.000** (0.0000)	**1.000** (0.0000)	**1.000** (0.0000)
	ZINB-TSS	0.779 (0.0474)	0.996 (0.0082)	0.999 (0.0042)	1.000 (0.0030)	0.954 (0.0260)	**1.000** (0.0000)	1.000 (0.0030)	**1.000** (0.0000)
	ZINB-Q75	0.777 (0.0479)	0.996 (0.0085)	0.999 (0.0051)	0.998 (0.0058)	0.957 (0.0267)	1.000 (0.0000)	1.000 (0.0030)	1.000 (0.0030)
	ZINB-RLE	0.782 (0.0495)	0.995 (0.0091)	1.000 (0.0030)	1.000 (0.0000)	0.960 (0.0220)	1.000 (0.0000)	0.999 (0.0042)	1.000 (0.0030)
	ZINB-TMM	0.776 (0.0451)	0.995 (0.0091)	0.999 (0.0042)	0.999 (0.0042)	0.957 (0.0224)	**1.000** (0.0000)	1.000 (0.0030)	**1.000** (0.0000)
	ZINB-CSS	0.774 (0.0494)	0.995 (0.0093)	1.000 (0.0030)	0.999 (0.0042)	0.955 (0.0266)	1.000 (0.0035)	0.999 (0.0042)	1.000 (0.0030)
	DM	0.751 (0.0515)	0.997 (0.0074)	0.999 (0.0042)	0.999 (0.0051)	0.951 (0.0231)	0.999 (0.0037)	0.999 (0.0051)	0.999 (0.0042)
	ANOVA	0.739 (0.0529)	0.982 (0.0153)	**1.000** (0.0000)	**1.000** (0.0000)	0.935 (0.0280)	0.999 (0.0052)	**1.000** (0.0000)	**1.000** (0.0000)
	Kruskal–Wallis	0.741 (0.0579)	0.984 (0.0133)	**1.000** (0.0000)	**1.000** (0.0000)	0.938 (0.0240)	0.999 (0.0037)	**1.000** (0.0000)	**1.000** (0.0000)
	`WaVE-DESeq2`	0.783 (0.0525)	**0.999** (0.0175)	**1.000** (0.0074)	**1.000** (0.0000)	0.731 (0.0539)	0.916 (0.0548)	0.980 (0.0313)	**1.000** (0.0269)
	`WaVE-edgeR`	0.717 (0.0577)	0.987 (0.0187)	0.999 (0.0074)	**1.000** (0.0000)	0.699 (0.0552)	0.930 (0.0476)	0.977 (0.0313)	**1.000** (0.0248)
	`metagenomeSeq`	0.659 (0.0666)	0.996 (0.0082)	0.982 (0.0128)	**1.000** (0.0000)	0.746 (0.0528)	0.970 (0.0238)	0.995 (0.0093)	**1.000** (0.0000)
ZINB	ZINB-DPP	**0.459** (0.0582)	**0.845** (0.0467)	**0.912** (0.0267)	0.971 (0.0186)	**0.716** (0.0514)	**0.705** (0.0532)	**0.956** (0.0223)	0.986 (0.0142)
	ZINB-TSS	0.403 (0.0651)	0.835 (0.0402)	0.906 (0.0289)	0.969 (0.0167)	0.681 (0.0487)	0.704 (0.0520)	0.954 (0.0216)	0.986 (0.0144)
	ZINB-Q75	0.407 (0.0598)	0.837 (0.0428)	0.906 (0.0314)	**0.972** (0.0187)	0.676 (0.0559)	0.704 (0.0520)	0.952 (0.0210)	0.985 (0.0157)
	ZINB-RLE	NA (-)	NA (-)	NA (-)	NA (-)	NA (-)	NA (-)	NA (-)	NA (-)
	ZINB-TMM	0.399 (0.0548)	0.832 (0.0448)	0.905 (0.0315)	0.967 (0.0185)	0.674 (0.0551)	0.703 (0.0520)	0.955 (0.0223)	**0.987** (0.0138)
	ZINB-CSS	0.399 (0.0580)	0.832 (0.0433)	0.904 (0.0292)	0.969 (0.0190)	0.663 (0.0540)	0.703 (0.0522)	0.954 (0.0212)	0.986 (0.0125)
	DM	0.077 (0.0471)	0.327 (0.0553)	0.472 (0.0594)	0.923 (0.0277)	0.217 (0.0689)	0.518 (0.0593)	0.767 (0.0466)	0.957 (0.0243)
	ANOVA	0.099 (0.0720)	0.412 (0.0847)	0.741 (0.0511)	0.930 (0.0234)	0.217 (0.1070)	0.374 (0.0917)	0.870 (0.0341)	0.954 (0.0271)
	Kruskal–Wallis	0.031 (0.0583)	0.122 (0.0862)	0.372 (0.0689)	0.634 (0.0557)	0.007 (0.0347)	0.010 (0.0392)	0.488 (0.0726)	0.795 (0.0435)
	`WaVE-DESeq2`	0.233 (0.0944)	0.247 (0.1276)	0.717 (0.0688)	0.825 (0.0324)	0.254 (0.0858)	0.245 (0.0945)	0.753 (0.0929)	0.854 (0.0620)
	`WaVE-edgeR`	0.432 (0.0597)	0.602 (0.0459)	0.847 (0.0320)	0.950 (0.0226)	0.448 (0.0720)	0.631 (0.0764)	0.850 (0.0345)	0.942 (0.0280)
	`metagenomeSeq`	0.072 (0.0518)	0.232 (0.0598)	0.223 (0.0603)	0.672 (0.0675)	0.094 (0.0506)	0.154 (0.0743)	0.186 (0.0560)	0.360 (0.0625)

Matthews correlation coefficient (MCC) given by all methods on the simulated data. In each cell, the top number is the averaged MCC over 50 independent datasets, and the bottom number in parentheses is the standard error. The result from the model that achieved best performance under the associated scenario (each column) is marked in bold.

Table A.4: Synthetic data: area under the curve (AUC) and Matthews correlation coefficient (MCC)

Real Data Sample Type	Methods	Synthetic Setting							
		AUC				MCC			
		$\log(\sigma) = 1$		$\log(\sigma) = 2$		$\log(\sigma) = 1$		$\log(\sigma) = 2$	
		$n = 24$	$n = 108$	$n = 24$	$n = 108$	$n = 24$	$n = 108$	$n = 24$	$n = 108$
Skin	ZINB-DPP	0.938 (0.0216)	0.978 (0.0154)	**0.994** (0.0061)	0.999 (0.0016)	**0.666** (0.0639)	**0.845** (0.0541)	**0.928** (0.0442)	**0.988** (0.0178)
	ZINB-TSS	0.923 (0.0293)	0.957 (0.0259)	0.991 (0.0110)	0.999 (0.0041)	0.670 (0.0727)	0.825 (0.0591)	0.920 (0.0469)	0.983 (0.0214)
	ZINB-Q75	0.920 (0.0335)	0.959 (0.0247)	0.991 (0.0109)	0.998 (0.0036)	0.658 (0.0704)	0.813 (0.0756)	0.920 (0.0450)	0.986 (0.0190)
	ZINB-RLE	0.923 (0.0249)	0.952 (0.0277)	0.990 (0.0117)	0.998 (0.0042)	0.658 (0.0723)	0.825 (0.0609)	0.921 (0.0436)	0.982 (0.0215)
	ZINB-TMM	0.925 (0.0271)	0.952 (0.0274)	0.990 (0.0099)	0.998 (0.0053)	0.658 (0.0786)	0.825 (0.0608)	0.921 (0.0467)	0.986 (0.0203)
	ZINB-CSS	0.909 (0.0348)	0.956 (0.0233)	0.988 (0.0116)	0.998 (0.0047)	0.640 (0.0857)	0.822 (0.0752)	0.914 (0.0491)	0.986 (0.0220)
	DM	**0.929** (0.0246)	**0.978** (0.0124)	**0.994** (0.0060)	1.000 (0.0011)	0.639 (0.0684)	0.819 (0.0480)	**0.928** (0.0418)	0.985 (0.0172)
	ANOVA	0.851 (0.0528)	0.946 (0.0263)	0.976 (0.0179)	0.998 (0.0046)	0.579 (0.0884)	0.744 (0.0753)	0.831 (0.0635)	0.960 (0.0277)
	Kruskal–Wallis	0.846 (0.0557)	0.966 (0.0215)	0.979 (0.0154)	**1.000** (0.0005)	0.572 (0.0968)	0.787 (0.0692)	0.844 (0.0635)	0.983 (0.0182)
	`WaVE-DESeq2`	0.539 (0.0509)	0.966 (0.0193)	0.914 (0.0325)	0.999 (0.0053)	0.575 (0.0593)	0.827 (0.0644)	0.822 (0.0626)	0.983 (0.0187)
	`WaVE-edgeR`	0.834 (0.0943)	0.969 (0.0364)	0.978 (0.0225)	0.999 (0.0027)	0.589 (0.0804)	0.833 (0.0431)	0.874 (0.0431)	0.990 (0.0146)
	`metagenomeSeq`	0.637 (0.0783)	0.953 (0.0242)	0.971 (0.0195)	1.000 (0.0006)	0.558 (0.0863)	0.704 (0.0592)	0.813 (0.0703)	0.936 (0.0346)
Feces	ZINB-DPP	**0.917** (0.0320)	0.891 (0.0390)	**0.987** (0.0117)	0.979 (0.0121)	0.619 (0.0995)	**0.658** (0.0784)	**0.884** (0.0832)	0.900 (0.0504)
	ZINB-TSS	0.900 (0.0370)	0.857 (0.0499)	0.975 (0.0222)	0.968 (0.0237)	0.620 (0.1003)	0.627 (0.0977)	0.872 (0.0810)	0.883 (0.0552)
	ZINB-Q75	0.874 (0.0691)	0.858 (0.0482)	0.970 (0.0254)	0.966 (0.0247)	0.553 (0.1310)	0.625 (0.1009)	0.861 (0.0818)	0.879 (0.0634)
	ZINB-RLE	0.909 (0.0346)	0.863 (0.0462)	0.975 (0.0235)	0.962 (0.0238)	**0.630** (0.1028)	0.630 (0.0848)	0.868 (0.0826)	0.876 (0.0650)
	ZINB-TMM	0.912 (0.0333)	0.869 (0.0435)	0.976 (0.0218)	0.966 (0.0235)	0.623 (0.0981)	0.638 (0.0921)	0.873 (0.0895)	0.878 (0.0574)
	ZINB-CSS	0.887 (0.0493)	0.858 (0.0564)	0.978 (0.0202)	0.968 (0.0262)	0.593 (0.1096)	0.639 (0.0904)	0.870 (0.0657)	0.887 (0.0594)
	DM	**0.917** (0.0295)	**0.929** (0.0293)	**0.987** (0.0137)	**0.993** (0.0080)	0.594 (0.0982)	0.655 (0.0811)	**0.884** (0.0764)	**0.928** (0.0390)
	ANOVA	0.822 (0.0636)	0.867 (0.0491)	0.955 (0.0339)	0.981 (0.0177)	0.560 (0.1116)	0.610 (0.0723)	0.801 (0.0865)	0.881 (0.0542)
	Kruskal–Wallis	0.819 (0.0611)	0.886 (0.0436)	0.957 (0.0354)	0.990 (0.0112)	0.553 (0.1077)	0.642 (0.0860)	0.810 (0.0980)	0.911 (0.0508)
	`WaVE-DESeq2`	0.733 (0.0675)	0.752 (0.0621)	0.917 (0.0562)	0.916 (0.0428)	0.424 (0.0711)	0.584 (0.0958)	0.859 (0.0633)	0.857 (0.0760)
	`WaVE-edgeR`	0.738 (0.1034)	0.832 (0.1054)	0.925 (0.0988)	0.966 (0.0163)	0.407 (0.0816)	0.537 (0.0929)	0.754 (0.0976)	0.851 (0.0527)
	`metagenomeSeq`	0.621 (0.0892)	0.816 (0.0623)	0.943 (0.0426)	0.985 (0.0168)	0.559 (0.1013)	0.584 (0.0937)	0.783 (0.0860)	0.881 (0.0557)

Area under the curve (AUC) and Matthews correlation coefficient (MCC) given by all methods on the synthetic data. In each cell, the top number is the averaged AUC (or MCC) over 50 independent datasets, and the bottom number in parentheses is the standard error. The result from the model that achieved best performance under the associated scenario (each column) is marked in bold.

Table A.5: Sensitivity Analysis: AUCs and the corresponding standard error (in parenthesis) for different choice of hyperparameters

b_k	0.1			1			2			10		
h_k	1	10	100	1	10	100	1	10	100	1	10	100
$K = 2$	0.887	0.888	0.879	0.888	0.878	0.869	0.871	0.858	0.846	0.739	0.730	0.711
$n = 24$	(0.0309)	(0.0253)	(0.0239)	(0.0244)	(0.0258)	(0.0243)	(0.0261)	(0.0290)	(0.0268)	(0.0354)	(0.0336)	(0.0421)
$K = 2$	0.987	0.986	0.981	0.996	0.994	0.993	0.996	0.996	0.995	0.967	0.964	0.946
$n = 108$	(0.0127)	(0.0136)	(0.0184)	(0.0057)	(0.0064)	(0.0069)	(0.0046)	(0.0051)	(0.0052)	(0.0098)	(0.0148)	(0.0193)
$K = 3$	0.780	0.785	0.783	0.787	0.784	0.792	0.785	0.783	0.793	0.722	0.706	0.675
$n = 24$	(0.0562)	(0.0577)	(0.0589)	(0.0581)	(0.0569)	(0.0542)	(0.0576)	(0.0563)	(0.0503)	(0.0549)	(0.0511)	(0.0438)
$K = 3$	0.992	0.991	0.983	0.998	0.997	0.996	0.998	0.998	0.997	0.994	0.986	0.936
$n = 108$	(0.0104)	(0.0110)	(0.0163)	(0.0062)	(0.0067)	(0.0063)	(0.0053)	(0.0047)	(0.0057)	(0.0058)	(0.0103)	(0.0279)

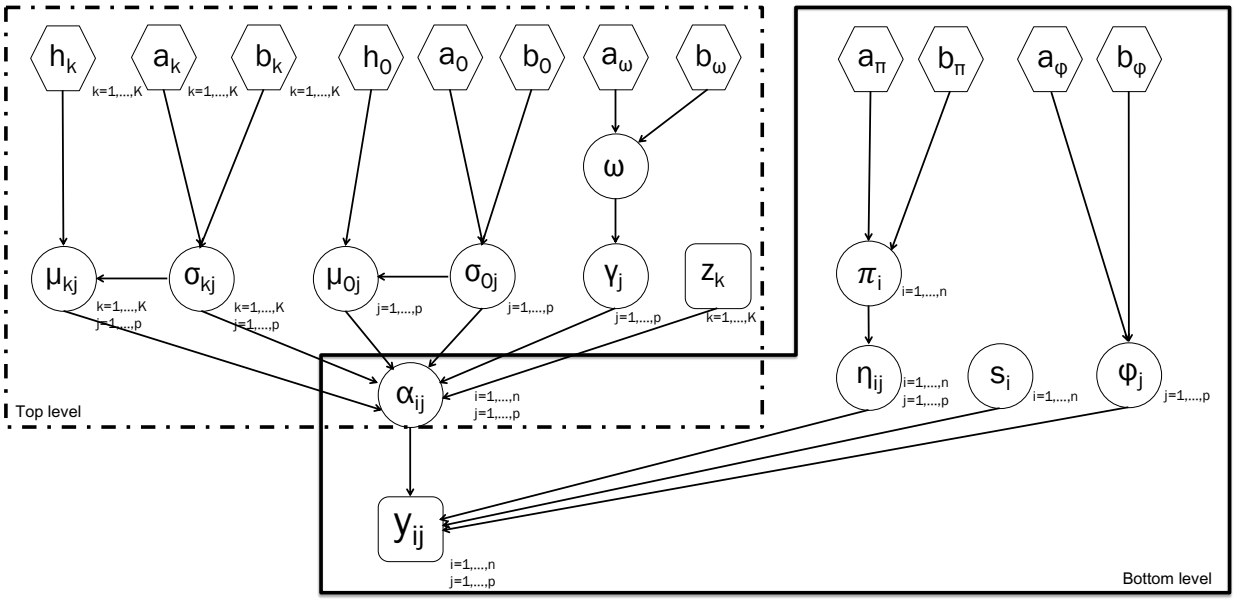


Figure A.1: A graphical representation of the proposed bi-level Bayesian framework for microbial differential abundance analysis, with the bottom level (within the solid border) of zero-inflated negative binomial (ZINB) model. Each node in a circle/hexagon/square refers to a model parameter/a fixed hyperparameter/observable data. The link between two nodes represents a direct probabilistic dependence. Note that both Fig A.1 and A.2 share the same top level (within the dashed border).

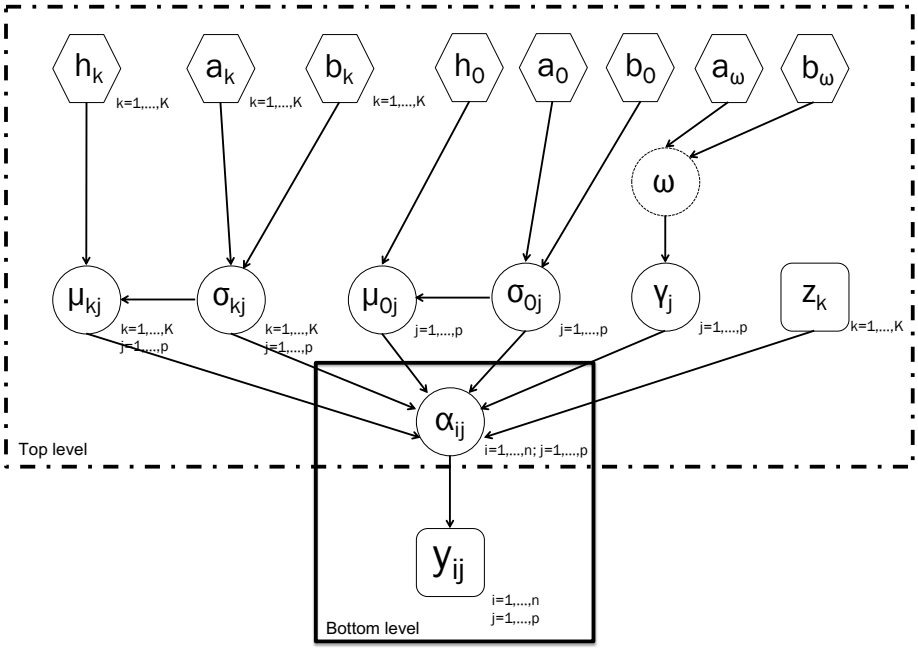


Figure A.2: A graphical representation of the proposed bi-level Bayesian framework for microbial differential abundance analysis, with the bottom level (within the solid border) of Dirichlet-multinomial (DM) model. Each node in a circle/hexagon/square refers to a model parameter/a fixed hyperparameter/observable data. The link between two nodes represents a direct probabilistic dependence. Note that both Fig A.1 and A.2 share the same top level (within the dashed border).

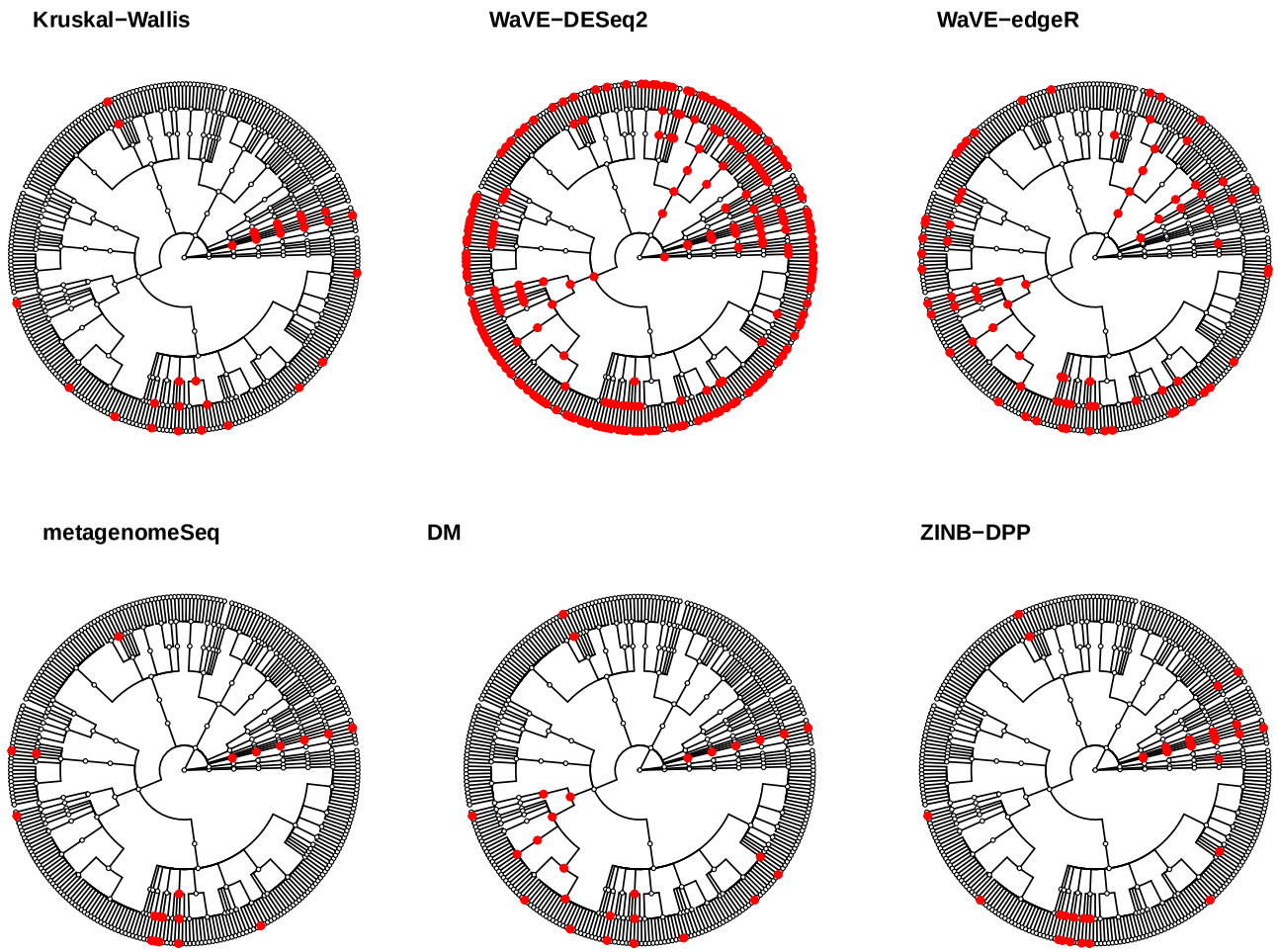


Figure A.3: Colorectal cancer study: the discriminating taxa identified by different methods. The red dots in each of the first four cases represent the taxa with Benjamini-Hochberg adjusted p-values below the significance level of 1%. The red dots in the DM and ZINB-DPP are taxa detected by controlling the Bayesian FDR to be less than 1%.

APPENDIX B

APPENDIX of CHAPTER 3

B.1. Details of the MCMC algorithms

First, we write the likelihood function as follows:

$$\prod_{k=1}^{K}\prod_{i:z_i=k}\prod_{j:\gamma_j=1,r_{ij}=0}\frac{\Gamma(y_{ij}+\phi_j)}{y_{ij!}\Gamma(\phi_j)}\left(\frac{\phi_j}{s_i e^{\mu_{0j}+\mu_{kj}+\boldsymbol{x}_i\boldsymbol{\beta}_j^T}+\phi_j}\right)^{\phi_j}\left(\frac{s_i e^{\mu_{0j}+\mu_{kj}+\boldsymbol{x}_i\boldsymbol{\beta}_j^T}}{s_i e^{\mu_{0j}+\mu_{kj}+\boldsymbol{x}_i\boldsymbol{\beta}_j^T}+\phi_j}\right)^{y_{ij}}\times$$
$$\prod_{i}\prod_{j:\gamma_j=0,r_{ij}=0}\frac{\Gamma(y_{ij}+\phi_j)}{y_{ij!}\Gamma(\phi_j)}\left(\frac{\phi_j}{s_i e^{\mu_{0j}+\boldsymbol{x}_i\boldsymbol{\beta}_j^T}+\phi_j}\right)^{\phi_j}\left(\frac{s_i e^{\mu_{0j}+\boldsymbol{x}_i\boldsymbol{\beta}_j^T}}{s_i e^{\mu_{0j}+\boldsymbol{x}_i\boldsymbol{\beta}_j^T}+\phi_j}\right)^{y_{ij}}.$$

Then, we update the parameters in each iteration following the steps below:

1. **Update of zero-inflation latent indicator** r_{ij}: Notice that we only need to update the r_{ij}'s that correspond to $y_{ij}=0$. We write the posterior as:

$$\begin{aligned}&p(r_{ij}|y_{ij}=0,\phi_j,z_i=k,s_i,\mu_{0j},\mu_{kj},\gamma_j)\\&\propto\int L(r_{ij}|y_{ij}=0,\phi_j,z_i=k,\mu_{0j},\mu_{kj},\gamma_j)\times p(r_{ij}|\pi)\times p(\pi)d\pi\end{aligned}$$

   Then it follows that

$$p(r_{ij}|\cdot)\propto\begin{cases}\left(\frac{\phi_j}{s_i e^{\mu_{0j}+\mu_{kj}+\boldsymbol{x}_i\boldsymbol{\beta}_j^T}+\phi_j}\right)^{\phi_j(1-r_{ij})}\times\frac{Be(a_\pi+r_{ij},b_\pi-r_{ij}+1)}{Be(a_\pi,b_\pi)} & if\ \gamma_j=1\\ \left(\frac{\phi_j}{s_i e^{\mu_{0j}+\boldsymbol{x}_i\boldsymbol{\beta}_j^T}+\phi_j}\right)^{\phi_j(1-r_{ij})}\times\frac{Be(a_\pi+r_{ij},b_\pi-r_{ij}+1)}{Be(a_\pi,b_\pi)} & if\ \gamma_j=0\end{cases}$$

2. **Update of** $\boldsymbol{\mu}_0$: We update each $\mu_{0j},\ j=1,2,...p$ sequentially using an independent Metropolis-Hasting algorithm. We first propose a new μ_{0j}^* from $N(\mu_{0j},\tau_0^2)$ and then

accept the proposed value with probability min$(1,\ m_{MH})$, where

$$m_{MH} = \frac{\prod_{i=1}^{n} f(y_{ij}|\mu_{0j}^*, \mu_{kj}, \phi_j, s_i, \gamma_j, \boldsymbol{R}, \boldsymbol{X}, \boldsymbol{B}) \times p(\mu_{0j}^*) \times J(\mu_{0j}; \mu_{0j}^*)}{\prod_{i=1}^{n} f(y_{ij}|\mu_{0j}, \mu_{kj}, \phi_j, s_i, \gamma_j, \boldsymbol{R}, \boldsymbol{X}, \boldsymbol{B}) \times p(\mu_{0j}) \times J(\mu_{0j}^*; \mu_{0j})}$$

3. **Joint Update of $\mu_{k\cdot}$ and $\boldsymbol{\gamma}$**: A between-model step is implemented first to jointly update $\mu_{k\cdot}$ and $\boldsymbol{\gamma}$. We use an *add-delete* algorithm, where we select a $j \in \{1, \ldots, p\}$ at random and change the value of γ_j. For the *add* case, i.e. $\gamma_j = 0 \to \gamma_j = 1$, we propose μ_{kj}^* for each $k = 2, \ldots, n$ from $N(0, \tau_{\mu j}^2)$. For the *delete* case, i.e. $\gamma_j = 1 \to \gamma_j = 0$, we set $\mu_{kj}^* = 0$ for all k. We finally accept the proposed values with probability min$(1,\ m_{MH})$, where

$$\begin{aligned} m_{MH} =& \frac{\prod_{i=1}^{n} f(y_{ij}|\mu_{kj}^*, \mu_{0j}, \phi_j, s_i, \gamma_j^*, \boldsymbol{R}, \boldsymbol{X}, \boldsymbol{B}) \times p(\mu_{kj}^*|\gamma_j^*) \times p(\boldsymbol{\gamma}^*)}{\prod_{i=1}^{n} f(y_{ij}|\mu_{kj}, \mu_{0j}, \phi_j, s_i, \gamma_j, \boldsymbol{R}, \boldsymbol{X}, \boldsymbol{B}) \times p(\mu_{kj}|\gamma_j) \times p(\boldsymbol{\gamma})} \\ & \times \frac{J(\mu_{kj}; \mu_{kj}^*|\gamma_j; \gamma_j^*) \times J(\boldsymbol{\gamma}; \boldsymbol{\gamma}^*)}{J(\mu_{kj}^*; \mu_{kj}|\gamma_j^*; \gamma_j) \times J(\boldsymbol{\gamma}^*; \boldsymbol{\gamma})} \end{aligned}$$

Further update of μ_{kj} when $\gamma_j^* = 1$: A within-model step is followed to further update each $\mu_{kj},\ k = 2, \ldots, K$ that corresponds to $\gamma_j^* = 1$ in the current iteration. We first propose a new μ_{kj}^* from $N(\mu_{kj}, (\tau_{\mu j}/2)^2)$ and then accept the proposed value with probability min$(1, m_{MH})$, where

$$m_{MH} = \frac{\prod_{i=1}^{n} f(y_{ij}|\mu_{kj}^*, \mu_{0j}, \phi_j, s_i, \gamma_j, \boldsymbol{R}, \boldsymbol{X}, \boldsymbol{B}) \times p(\mu_{kj}^*) \times J(\mu_{kj}; \mu_{kj}^*)}{\prod_{i=1}^{n} f(y_{ij}|\mu_{kj}, \mu_{0j}, \phi_j, s_i, \gamma_j, \boldsymbol{R}, \boldsymbol{X}, \boldsymbol{B}) \times p(\mu_{kj}) \times J(\mu_{kj}^*; \mu_{kj})}$$

4. **Joint update of $\beta_{\cdot j}$ and $\delta_{\cdot j}$**: Very similar to the above, we perform a between-model step first using an *add-delete* algorithm. For each $j = 1, \ldots, p$, we first select an $r \in \{1, \ldots, R\}$ at random and change the value of δ_{rj}. For the *add* case, i.e. $\delta_{rj} = 0 \to \delta_{rj} = 1$, we propose β_{rj}^* from $N(0, \tau_{\beta j}^2)$. For the *delete* case, i.e. $\delta_{rj} = 1 \to \delta_{rj} = 0$, we set $\beta_{rj}^* = 0$. Then finally we accept the proposed values with probability

min$(1, m_{MH})$, where

$$m_{MH} = \frac{\prod_{i=1}^{n} f(y_{ij}|\beta^*_{rj}, \delta^*_{rj}, \mu_{0j}, \mu_{\cdot j}, s_i, \gamma_j, \boldsymbol{R}, \boldsymbol{X}) \times p(\beta^*_{\cdot j}|\delta^*_{\cdot j}) \times p(\delta^*_{\cdot j})}{\prod_{i=1}^{n} f(y_{ij}|\beta_{rj}, \delta_{rj}, \mu_{0j}, \mu_{\cdot j}, s_i, \gamma_j, \boldsymbol{R}, \boldsymbol{X}) \times p(\beta_{\cdot j}|\delta_{\cdot j}) \times p(\delta_{\cdot j})} \times \frac{J(\beta_{\cdot j}; \beta^*_{\cdot j}|\delta_{\cdot j}; \delta^*_{\cdot j}) \times J(\delta_{\cdot j}; \delta^*_{\cdot j})}{J(\beta^*_{\cdot j}; \beta_{\cdot j}|\delta^*_{\cdot j}; \delta_{\cdot j}) \times J(\delta^*_{\cdot j}; \delta_{\cdot j})}$$

Further update of β_{rj} when $\delta^*_{rj} = 1$: A within-model step is followed to further update each $\beta_{rj},\ r = 1, \ldots, R$ that corresponds to $\delta^*_{rj} = 1$. We first propose a new β^*_{rj} from $N(\beta_{rj}, (\sigma_{\beta j}/2)^2)$ and then accept the proposed value with probability min$(1, m_{MH})$, where

$$m_{MH} = \frac{\prod_{i=1}^{n} f(y_{ij}|\beta^*_{rj}, \delta_{rj}, \mu_{0j}, \mu_{\cdot j}, s_i, \phi_j, \gamma_j, \boldsymbol{R}, \boldsymbol{X}) \times p(\beta^*_{rj}) \times J(\beta_{rj}; \beta^*_{rj})}{\prod_{i=1}^{n} f(y_{ij}|\beta_{rj}, \delta_{rj}, \mu_{0j}, \mu_{\cdot j}, s_i, \phi_j, \gamma_j, \boldsymbol{R}, \boldsymbol{X}) \times p(\beta_{rj}) \times J(\beta^*_{rj}; \beta_{rj})}$$

5. **Update of $\phi_{\cdot}$:** We update each $\phi_j\ j = 1, \ldots, p$ sequentially by using an independent Metropolis-Hasting algorithm. We first propose a new ϕ^*_j from the normal distribution $N(\phi_j, \tau^2_\phi)$ that truncated at 0, and accept the proposed value with probability min(1,m_{MH}), where

$$m_{MH} = \frac{\prod_{i=1}^{n} f(y_{ij}|\phi^*_j, \beta_{rj}, \delta_{rj}, \mu_{0j}, \mu_{\cdot j}, s_i, \gamma_j, \boldsymbol{R}, \boldsymbol{X}, \boldsymbol{B}) \times p(\phi^*_j) \times J(\phi_j; \phi^*_j)}{\prod_{i=1}^{n} f(y_{ij}|\phi_j, \beta_{rj}, \delta_{rj}, \mu_{0j}, \mu_{\cdot j}, s_i, \gamma_j, \boldsymbol{R}, \boldsymbol{X}, \boldsymbol{B}) \times p(\phi_j) \times J(\phi^*_j; \phi_j)}$$

B.2. Additional results of simulation study

We performed a comprehensive simulation study for model comparison. First, we introduce the following reference setting in the simulation study, that is, 1) $n = 60$ samples split into $K = 2$ equally sized groups; 2) $p = 300$ features, 20 of which were truly discriminating ones; 3) $\pi_0 = 40\%$ false zeros (i.e. structural zeros) randomly assigned among all counts; 4) $R = 7$ covariates, four of which true coefficients were nonzero; 5) noise level $\epsilon^2_e = 1$. Furthermore, we varied the following settings to comprehensively examine the model performance, including the choices for sample size per group ($n/2 = 10$ or 30), the three log-scale noise levels ($\sigma_e = 0.5$, 1.0, or 1.5) and the extra zero proportions ($\pi_0 = 30\%$, 40%, or 70%). In all cases, we randomly set four out of seven nonzero β_{rj} for each taxon

j.

B.2.1. Evaluation for sample size

Figure B.3 compares the model performance under two choices of group sizes ($n/2 = 10$ or 30) with fixed log-scale noise level at 1.0 and 40% of extra zeros as in the reference setting. Different methods are compared using the receiving operating characteristic (ROC) curve and area under the curve (AUC). The left part in Figure B.3 shows the results of identifying the differentially abundant taxa ($\boldsymbol{\gamma}$) and the right part is the results of detecting significant covariate-taxa associations ($\boldsymbol{\Delta}$). Clearly, decreasing the sample size hampers the performance of all the methods, but the proposed ZINB model maintains the highest AUC in both cases, and achieves the highest true positive rate under a fixed small false positive rate.

B.2.2. Evaluation for log-scale noise level

Figure B.4 compares the model performance under three choices of log-scale noise level ($\sigma_e = 0.5$, 1.0, or 1.5) with fixed group size of $n/2 = 30$ and 40% extra zeros as in the reference setting. The ZINB model maintains the highest AUC across all settings of identifying the differentially abundant taxa (AUC > 0.9), and detecting significant covariate-taxa associations (AUC > 0.8). Notice that the true log-scale signal level is set to be 2, and the ZINB model still shows an obvious advantage over the alternative methods under a large log-scale noise level of 1.5.

B.2.3. Evaluation for extra zero proportion

Figure B.5 compares the model performance under three scenarios of extra zero proportions ($\pi_0 = 30\%$, 40%, and 70%) with a group size of $n/2 = 30$ and a noise level of $\epsilon_e^2 = 1$ as in the reference setting. Although a higher proportion of zeros like 70% dose down-

grade the performance of all methods, the proposed ZINB model is the best under any circumstance. Particularly, our methods always exhibits a considerably advantage over the others in terms of the identification of feature-covariate association ($\mathbf{\Delta}$) as shown in the right column of Figure B.5.

B.3. Sensitivity analysis

To assess model robustness with respect to the choice of size factor estimation methods, we compared the model performance under five typical normalization methods for the analysis of high-dimensional count data. They are: 1) geometric mean of pairwise ratios (GMPR) proposed by [22]; 2) cumulative sum scaling (CSS) proposed by [97]; 3) The 0.75-th quantile (Q75) proposed by [14]; 4) trimmed mean of M values (TMM) proposed by [106]; 5) relative log expression (RLE) proposed by [5]. The first two, designed for normalizing the microbiome count data, have been described in Section 3.2.3, while Q75, TMM and RLE are commonly used in RNA-seq data studies. In particular, Q75 calculates the size factor based on the upper-quantile (75%) of the count distribution of a sample. TMM first sets a reference sample, and calculates the trimmed mean of the log ratios between all other samples with the selected reference to estimate size factors. RLE, on the other hand, computes a reference value of each feature (taxon) as the geometric mean across all samples, and then obtains ratios by dividing all features by the reference. The size factor for a sample by RLE is set to be the median of the ratios. Due to the high sparsity observed in the microbiome data, it is needed to add a pseudo-count such as 1 to the count matrix when using RLE to estimate size factors.

To test if the performance of our model is robust to the choice of different normalization techniques, we used the simulate datasets generated by the reference setting described in Section B.2. The resulting AUCs for the discriminating feature indicator γ and the feature-covariate association indicator $\mathbf{\Delta}$ over 100 data replicates are summarized in Figure S6. First, the result suggests that the proposed ZINB model is robust with respect

to plug-in size factor estimations. Next, the ZINB-CSS and ZINB-GMPR show better performances due to the smaller variation and slightly higher average AUCs, since both are based on the normalization methods that better account for the characteristics of the microbiome data. Notice that RLE is less stable compared to the other methods for such sparse count data, which is also mentioned in [22].

Next, we assess impacts of setting priors via sensitivity analysis. In our model, the choice of a and b in the IG(a, b) prior for $\sigma^2_{\mu j}$ has an impact on the posterior probabilities of inclusion of γ. To investigate model performance with respect to the choice of these hyperparameters, we simulated 30 datasets under the reference setting described in Section B.2, and benchmarked our model with varying values of a from 0.5 to 6 and b from 0.5 to 25. The choices of a and b are illustrated in Figure B.7.

The results given by different values of (a, b) were compared based on the Matthews correlation coefficient (MCC) [87] across 30 replicated datasets. In each replicate, we controlled a 5% Bayesian false discovery rate and selected discriminating features. We then calculated the number of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) and MCC. Here MCC is defined as

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}.$$

MCC ranges from -1 to 1, and larger values represents favorable prediction results. It is also demonstrated in the above formula that the MCC-based evaluation is suitable for classes with very different sizes, since it strikes a balance between TP and FP counts. In our scenario, the size of truly discriminating features are relatively small compared to the total number. Therefore, we adopt MCC as an appropriate performance metric to handle the imbalanced setting. As can be seen in Figure B.7, given a small value of b ($b \leq 2$), the MCC is undesirable with any value of a displayed here. As shown by [42], the IG(a, b) prior with small a and b would distort the posterior inferences. On the other hand, if we increase

a to have $a > 2$ while fixing b to be small, the corresponding prior distribution is strongly informative since IG(a, b) has the mean of $b/(a-2)$ and the variance of $b^2/(a-2)^2(a-1)$. Therefore, we choose $a = 2$ and $b = 10$ to be the default setting, since this ensures a flat prior and yields a beneficial variable selection result as shown in Figure B.7.

B.4. Additional tables and figures

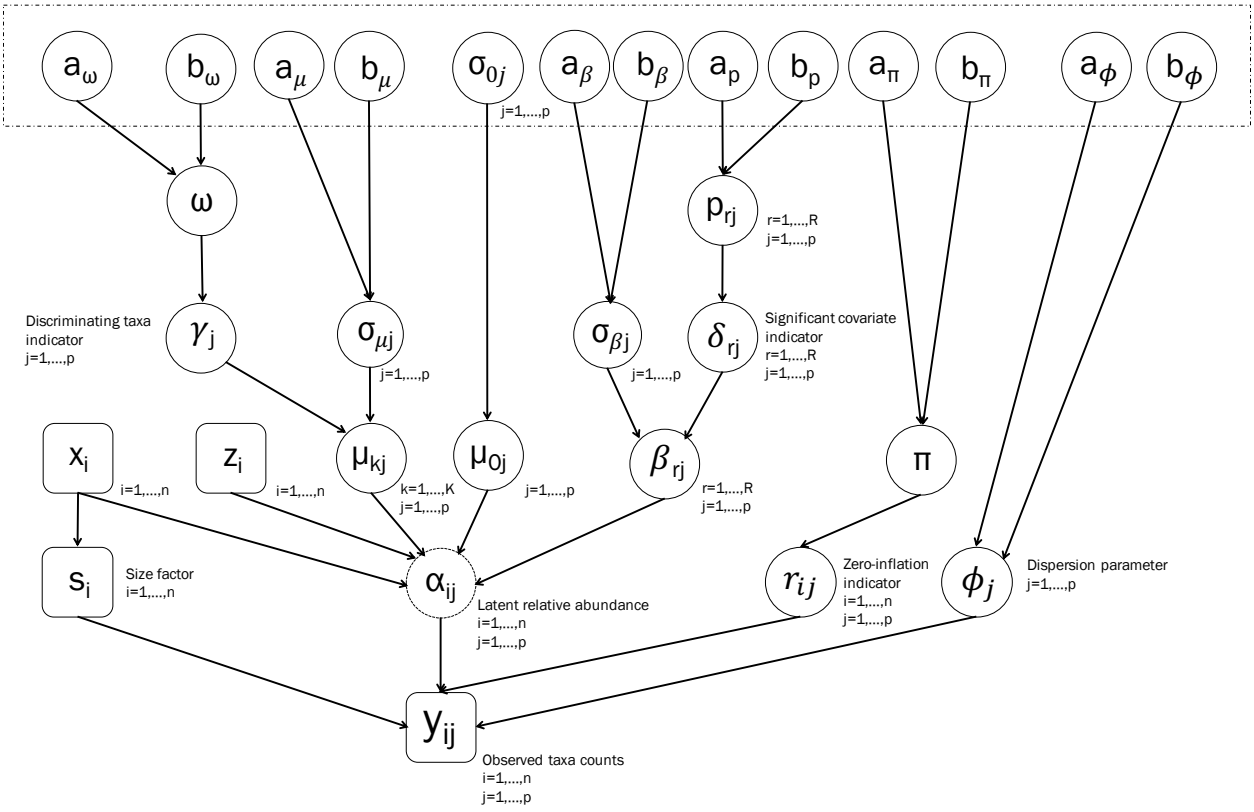


Figure B.1: The graphical formulation of the proposed Bayesian zero-inflated negative binomial regression model. Node in a circle refers to a parameter of the model. Node in a rectangle is observable data. Circle nodes in the dashed block are fixed hyperparameters. The link between two nodes represents a direct probabilistic dependence.

Parameters:

$$
\begin{aligned}
\boldsymbol{R} &= [\boldsymbol{r}_{\cdot 1}, \ldots, \boldsymbol{r}_{\cdot j}, \ldots, \boldsymbol{r}_{\cdot p}] \\
\boldsymbol{\phi} &= [\phi_1, \ldots, \phi_j, \ldots, \phi_p] \\
\boldsymbol{\mu}_0 &= [\mu_{01}, , \ldots, \mu_{0j}, \ldots, \mu_{0p}] \\
\boldsymbol{M} &= [\boldsymbol{\mu}_{\cdot 1}, \ldots, \boldsymbol{\mu}_{\cdot j}, \ldots, \boldsymbol{\mu}_{\cdot p}] \\
\boldsymbol{B} &= [\boldsymbol{\beta}_{\cdot 1}, \ldots, \boldsymbol{\beta}_{\cdot j}, \ldots, \boldsymbol{\beta}_{\cdot p}] \\
\boldsymbol{\gamma} &= [\gamma_1, \ldots, \gamma_j, \ldots, \gamma_p] \\
\boldsymbol{\Delta} &= [\boldsymbol{\delta}_{\cdot 1}, \ldots, \boldsymbol{\delta}_{\cdot j}, \ldots, \boldsymbol{\delta}_{\cdot p}]
\end{aligned}
$$

Mixture model likelihood:

$$
\begin{aligned}
y_{ij}|r_{ij} = 0, z_i = k, \gamma_j = 1 &\overset{ind}{\sim} \mathsf{NB}(y_{ij}; s_i\alpha_{ijk}, \phi_j) \quad \text{with } \log(\alpha_{ijk}) = \mu_{0j} + \mu_{kj} + \boldsymbol{x}_{i\cdot}^T\boldsymbol{\beta}_{\cdot j} \\
y_{ij}|r_{ij} = 0, \gamma_j = 0 &\overset{ind}{\sim} \mathsf{NB}(y_{ij}; s_i\alpha_{ij0}, \phi_j) \quad \text{with } \log(\alpha_{ij0}) = \mu_{0j} + \boldsymbol{x}_{i\cdot}^T\boldsymbol{\beta}_{\cdot j} \\
y_{ij}|r_{ij} = 1 &\equiv 0
\end{aligned}
$$

Zero-inflation prior:

$$r_{ij}|\pi \sim \mathsf{Bernoulli}(\pi), \quad \pi \sim \mathsf{Beta}(a_\pi, b_\pi) \quad \Rightarrow \quad r_{ij}|a_\pi, b_\pi \sim \text{Beta-Bernoulli}(r_{ij}; a_\pi, b_\pi)$$

Feature selection prior:

$$\gamma_j|\omega \sim \mathsf{Bernoulli}(\omega), \quad \omega \sim \mathsf{Beta}(a_\omega, b_\omega) \quad \Rightarrow \quad \gamma_j|a_\omega, b_\omega \sim \text{Beta-Bernoulli}(\gamma_j; a_\omega, b_\omega)$$

Dispersion prior:

$$\phi_j \sim \mathsf{Ga}(a_\phi, b_\phi)$$

Feature / Covariate characterization priors:

$$
\begin{aligned}
\mu_{0j}|\sigma_{0j} &\sim \mathsf{N}(0, \sigma_{0j}^2) \\
\mu_{kj}|\gamma_j, \sigma_{kj} \sim (1-\gamma_j)\mathsf{I}(\mu_{kj} = 0) + \gamma_j\mathsf{N}(0, \sigma_{\mu j}^2) \quad &, \quad \sigma_{\mu j}^2 \sim \mathsf{IG}(a_\mu, b_\mu) \Rightarrow \\
&\quad \mu_{kj}|\gamma_j \sim (1-\gamma_j)\mathsf{I}(\mu_{kj} = 0) + \gamma_j t_{2a_\mu}(0, b_\mu/a_\mu) \\
\beta_{rj}|\delta_{rj}, \sigma_{\beta j} \sim (1-\delta_{rj})\mathsf{I}(\beta_{rj} = 0) + \delta_{rj}\mathsf{N}(0, \sigma_{\beta j}^2) \quad &, \quad \sigma_{\beta j}^2 \sim \mathsf{IG}(a_\beta, b_\beta) \Rightarrow \\
&\quad \beta_{rj}|\delta_{rj} \sim (1-\delta_{rj})\mathsf{I}(\beta_{rj} = 0) + \delta_{rj} t_{2a_\beta}(0, b_\beta/a_\beta)
\end{aligned}
$$

Covariate selection prior:

$$\delta_{rj}|p_{rj} \sim \mathsf{Bernoulli}(p_{rj}), \quad p_{rj} \sim \mathsf{Beta}(a_p, b_p) \quad \Rightarrow \quad \delta_{rj}|a_p, b_p \sim \text{Beta-Bernoulli}(\delta_{rj}; a_p, b_p)$$

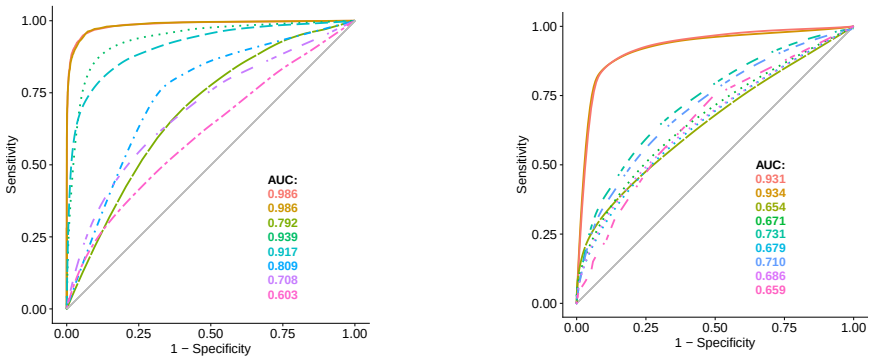
Fixed hyperparameters:

$$\sigma_{0j}, a_\pi, b_\pi, a_\omega, b_\omega, a_\phi, b_\phi, a_\mu, b_\mu, a_\beta, b_\beta$$

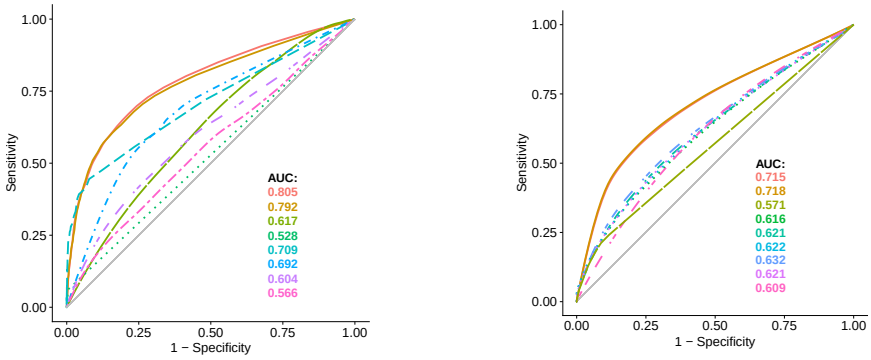
Posterior:

$$
\begin{aligned}
&p(\boldsymbol{R}, \boldsymbol{\phi}, \boldsymbol{\mu}_0, \boldsymbol{M}, \boldsymbol{B}, \boldsymbol{\gamma}, \boldsymbol{\Delta}|\boldsymbol{Y}, \boldsymbol{X}) \quad \propto \\
&\prod_{k=1}^{K} \prod_{i:z_i=k} \prod_{\substack{j:\gamma_j=1 \\ r_{ij}=0}} \mathsf{NB}(y_{ij}; s_i\alpha_{ijk}, \phi_j) \times \prod_i \prod_{\substack{j:\gamma_j=0 \\ r_{ij}=0}} \mathsf{NB}(y_{ij}; s_i\alpha_{ij0}, \phi_j) \\
&\times \prod_{i,j} \text{Beta-Bernoulli}(r_{ij}; a_\pi, b_\pi) \times \prod_j \mathsf{Ga}(\phi_j; a_\phi, b_\phi) \times \prod_j \mathsf{N}(\mu_{0j}; 0, \sigma_{0j}^2) \\
&\times \prod_{j,k} [(1-\gamma_j)\mathsf{I}(\mu_{kj} = 0) + \gamma_j t_{2a_\mu}(\mu_{kj}; 0, b_\mu/a_\mu)] \times \prod_j \text{Beta-Bernoulli}(\gamma_j; a_\omega, b_\omega) \\
&\times \prod_{r,j} [(1-\delta_{rj})\mathsf{I}(\beta_{rj} = 0) + \delta_{rj} t_{2a_\beta}(\beta_{rj}; 0, b_\beta/a_\beta)] \times \prod_{r,j} \text{Beta-Bernoulli}(\delta_{rj}; a_p, b_p)
\end{aligned}
$$

Figure B.2: Hierarchical formulation of the proposed hierarchical mixture model



(a) ROC curves for γ (left) and $\mathbf{\Delta}$ (right) with a sample size per group of $n/2 = 30$



(b) ROC curves for γ (left) and $\mathbf{\Delta}$ (right) with a sample size per group of $n/2 = 10$

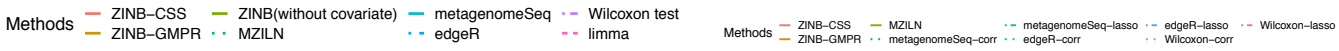
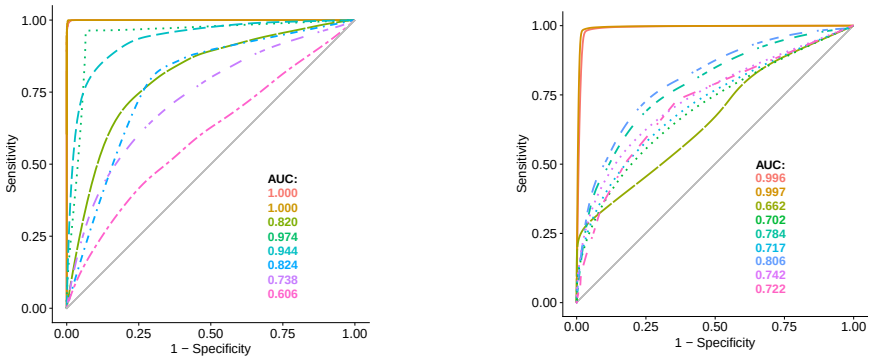
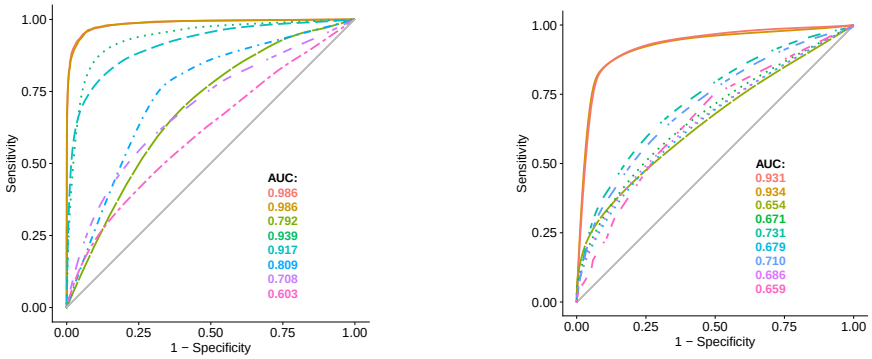


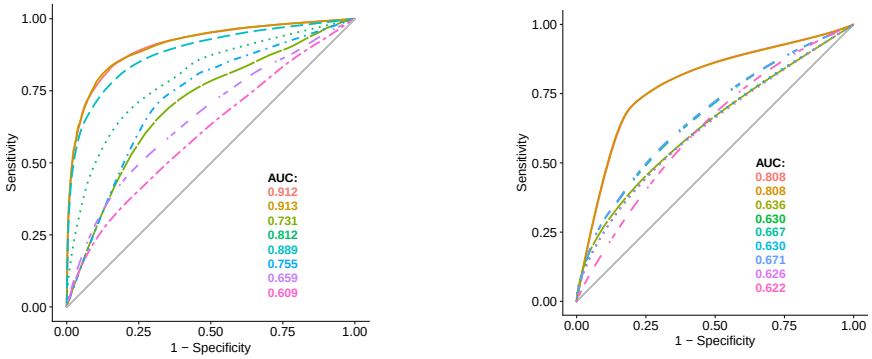
Figure B.3: Averaged ROC curves for the discriminating feature indicator γ (left) and the feature-covariate association indicator $\mathbf{\Delta}$ (right) with respect to different sample sizes per group (a) $n/2 = 30$ and (b) 10, over 100 replicates in each scenario.



(a) ROC curves for γ (left) and $\mathbf{\Delta}$ (right) with a log-scale noise level of $\sigma_e = 0.5$



(b) ROC curves for γ (left) and $\mathbf{\Delta}$ (right) with a log-scale noise level of $\sigma_e = 1.0$



(c) ROC curves for γ (left) and $\mathbf{\Delta}$ (right) with a log-scale noise level of $\sigma_e = 1.5$

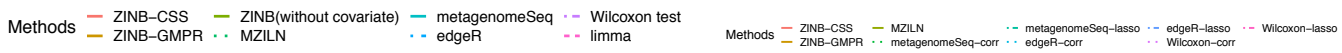
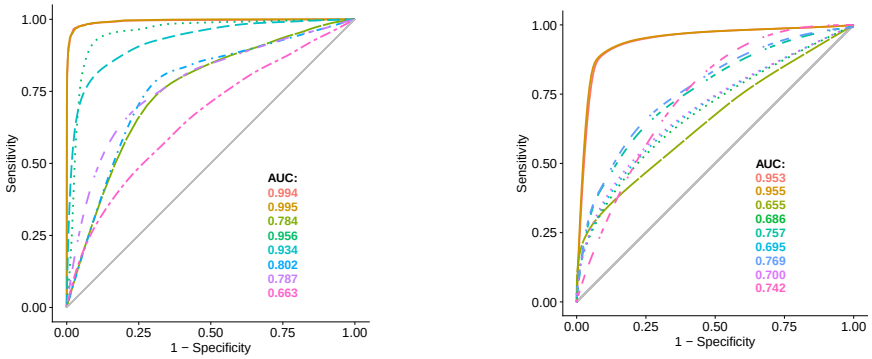
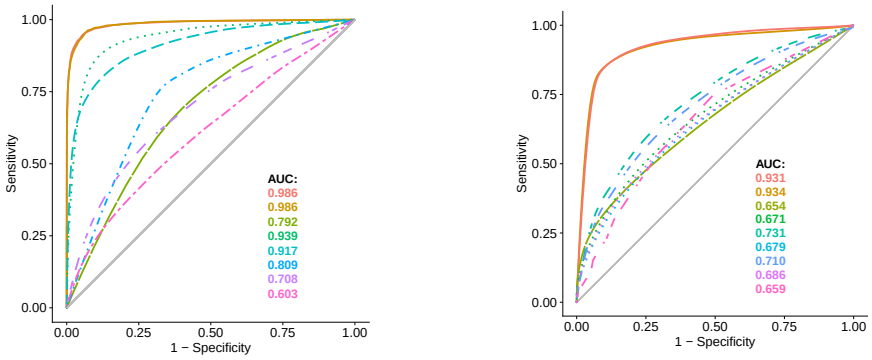


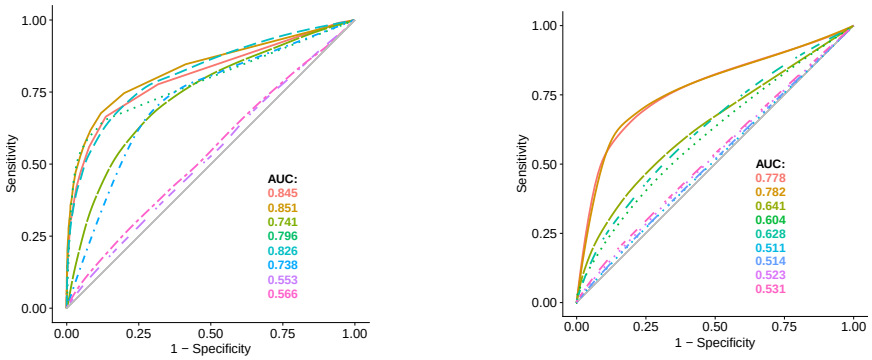
Figure B.4: Averaged ROC curves for the discriminating feature indicator γ (left) and the feature-covariate association indicator $\mathbf{\Delta}$ (right) with respect to different noise levels (a) $\sigma_e = 0.5$, (b) $\sigma_e = 1.0$, and (c) $\sigma_e = 1.5$, over 100 replicates in each scenario.



(a) ROC curves for γ (left) and $\boldsymbol{\Delta}$ (right) with a false zero proportion of $\pi_0 = 30\%$



(b) ROC curves for γ (left) and $\boldsymbol{\Delta}$ (right) with a false zero proportion of $\pi_0 = 40\%$



(c) ROC curves for γ (left) and $\boldsymbol{\Delta}$ (right) with a false zero proportion of $\pi_0 = 70\%$

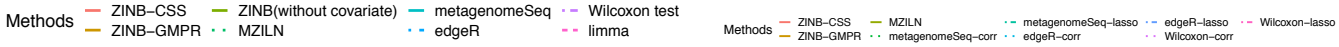


Figure B.5: Averaged ROC curves for the discriminating feature indicator γ (left) and the feature-covariate association indicator $\boldsymbol{\Delta}$ (right) with respect to different false zero proportions (a) $\pi_0 = 30\%$, (b) $\pi_0 = 40\%$, and (c) $\pi_0 = 70\%$, over 100 replicates in each scenario.

Liver Cirrhosis Study	Major Parameter Estimation		Covariate Effect by Group		Estimation for Health		Estimation for Liver Cirrhosis	
Discriminatory Taxa (ordered by posterior mean of μ_{2j})	μ_{0j} (CI for μ_{0j})	μ_{2j} (CI for μ_{2j})	Ave($\boldsymbol{x}\boldsymbol{\beta}^T$) Health	Ave($\boldsymbol{x}\boldsymbol{\beta}^T$) Liver Cirrhosis	Normalized $\log(y_{\cdot j})$	Estimated $\alpha_{\cdot j}$	Normalized $\log(y_{\cdot j})$	Estimated $\alpha_{\cdot j}$
Bacteroides eggerthii	12.93 (12.77, 13.11)	-0.23 (-0.38, -0.08)	0.47	-0.37	10.79	13.40	8.82	12.32
Gammaproteobacteria	14.49 (14.31, 14.70)	0.22 (0.07, 0.37)	-0.54	0.50	12.52	13.95	14.24	15.22
Veillonella dispar	11.18 (10.96, 11.42)	0.22 (0.06, 0.38)	-0.84	0.75	8.14	10.34	10.79	12.16
Bacteroides caccae	13.37 (13.19, 13.56)	0.23 (0.08, 0.38)	-0.11	0.12	11.48	13.26	11.88	13.72
Streptococcus salivarius	12.28 (12.10, 12.47)	0.26 (0.11, 0.41)	-0.67	0.63	10.22	11.61	12.25	13.17
Haemophilus parainfluenzae	12.98 (12.80, 13.18)	0.26 (0.12, 0.40)	-0.57	0.52	10.12	12.40	12.51	13.76
Haemophilus	13.00 (12.83, 13.19)	0.27 (0.12, 0.41)	-0.57	0.52	10.13	12.43	12.53	13.79
Streptococcus parasanguinis	11.48 (11.29, 11.70)	0.28 (0.12, 0.42)	-0.82	0.76	8.70	10.66	11.38	12.53
Pasteurellales	13.04 (12.86, 13.22)	0.29 (0.13, 0.42)	-0.57	0.52	10.14	12.47	12.56	13.84
Pasteurellaceae	13.03 (12.86, 13.2)	0.29 (0.13, 0.43)	-0.57	0.52	10.14	12.47	12.56	13.84
Veillonella parvula	12.01 (11.83, 12.21)	0.29 (0.13, 0.43)	-0.63	0.63	9.39	11.38	11.96	12.93
Streptococcus	12.88 (12.71, 13.05)	0.32 (0.17, 0.45)	-0.66	0.61	10.61	12.22	13.17	13.81
Streptococcaceae	12.89 (12.72, 13.07)	0.32 (0.17, 0.46)	-0.66	0.61	10.61	12.22	13.19	13.82
Klebsiella pneumoniae	13.09 (12.91, 13.28)	0.32 (0.17, 0.46)	-0.39	0.42	10.38	12.71	11.78	13.84
Bacilli	13.36 (13.19, 13.53)	0.34 (0.19, 0.47)	-0.71	0.66	11.03	12.64	13.59	14.36
Klebsiella	13.19 (13.00, 13.37)	0.34 (0.20, 0.48)	-0.45	0.48	10.32	12.74	11.69	14.01
Lactobacillales	13.33 (13.16, 13.51)	0.35 (0.20, 0.49)	-0.71	0.66	10.88	12.62	13.58	14.34
Veillonella	13.47 (13.30, 13.65)	0.39 (0.25, 0.52)	-0.81	0.75	10.27	12.66	13.88	14.61
Veillonella unclassified	12.93 (12.75, 13.13)	0.40 (0.26, 0.55)	-0.88	0.83	9.28	12.05	13.27	14.17

Table B.1: Liver cirrhosis dataset: parameter estimation for the identified discriminating taxa from the liver cirrhosis study. Posterior mean and 95% Credible Interval (CI) are reported for the estimated μ_{0j} (feature-specific baseline parameter) and μ_{2j}(group-specific parameter); Covariate effect represents the mean of $\boldsymbol{x}\hat{\boldsymbol{\beta}}^T$ of all samples in the corresponding patient group; Normalized $\log(y_{\cdot j})$ is the mean of log scaled observations after accounting for the sample heterogeneity factor (i.e. size factor) s_i. Estimated $\alpha_{\cdot j}$ is the mean of α_{ij} for all sample i from the same patient group.

Liver Cirrhosis Study	Major Parameter Estimation		Covariate Effect by Group		Estimation for Health		Estimation for Liver Cirrhosis	
Discriminatory Taxa (ordered by posterior mean of μ_{2j})	μ_{0j} (CI for μ_{0j})	μ_{2j} (CI for μ_{2j})	Ave($\boldsymbol{x}\boldsymbol{\beta}^T$) Health	Ave($\boldsymbol{x}\boldsymbol{\beta}^T$) Liver Cirrhosis	Normalized $\log(y_{.j})$	Estimated $\alpha_{.j}$	Normalized $\log(y_{.j})$	Estimated $\alpha_{.j}$
Bifidobacterium	12.50 (12.16, 12.86)	-1.34 (-2.08, -0.62)	-0.19	0.62	10.78	12.31	10.97	11.79
Bifidobacteriaceae	12.50 (12.16, 12.86)	-1.34 (-2.07, -0.62)	-0.19	0.62	10.78	12.31	10.97	11.78
Bifidobacteriales	12.50 (12.16, 12.86)	-1.33 (-2.06, -0.60)	-0.19	0.61	10.78	12.31	10.97	11.78
Clostridium methylpentosum	8.13 (5.95, 6.90)	0.94 (0.54, 1.34)	0.18	-0.39	8.21	8.31	8.69	8.68
Carnobacteriaceae	6.40 (10.13, 11.22)	1.03 (0.56, 1.45)	0.31	-0.24	6.56	6.71	7.12	7.18
Clostridium bartlettii	10.60 (7.88, 8.54)	1.40 (0.65, 2.15)	-0.28	0.46	9.35	10.32	11.19	12.46
Eubacterium siraeum	11.84 (11.35, 12.42)	1.42 (0.64, 2.18)	0.30	-0.35	10.70	12.14	12.41	12.92

Table B.2: Metastatic melanoma dataset: parameter estimation for the identified discriminating taxa from the metastatic melanoma study. Posterior mean and 95% Credible Interval (CI) are reported for the estimated μ_{0j} (feature-specific baseline parameter) and μ_{2j}(group-specific parameter); Covariate effect represents the mean of $\boldsymbol{x}\hat{\boldsymbol{\beta}}^T$ of all samples in the corresponding patient group; Normalized $\log(y_{.j})$ is the mean of log scaled observations after accounting for the sample heterogeneity factor (i.e. size factor) s_i. Estimated $\alpha_{.j}$ is the mean of α_{ij} for all sample i from the same patient group.

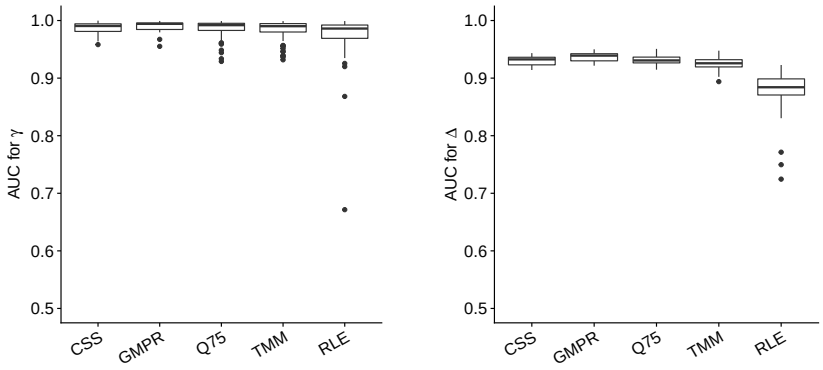


Figure B.6: Side-by-side box plots of AUCs for the discriminating feature indicator $\boldsymbol{\gamma}$ (left) and the feature-covariate association indicator $\boldsymbol{\Delta}$ (right) with respect to different normalization techniques, over 100 reference simulated datasets. CSS for cumulative sum scaling. GMPR for geometric mean of pairwise ratios. Q75 for 0.75-th quantile. TMM for trimmed mean of M values. RLE for relative log expression.

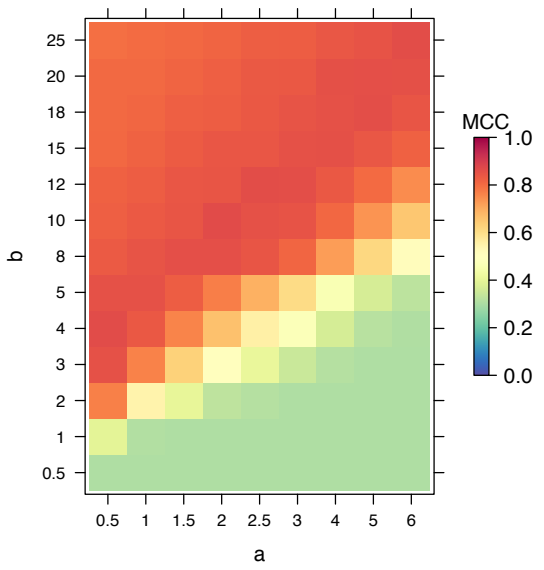


Figure B.7: Heatmap of Matthews correlation coefficients (MCC) for the discriminating feature indicator γ with the choice of (a, b) from the inverse-gamma prior on the variance terms $\sigma^2_{\mu j}$ and $\sigma^2_{\beta j}$. Each value of MCC represents the averaged result of 30 replicates.

APPENDIX C

APPENDIX of CHAPTER 4

C.1. Details of the MCMC algorithms

We start by writing the likelihood for each sample $i, i = 1, \ldots, n$ as

$$f_{\text{ZINB}}(\boldsymbol{y}_{i\cdot}|\boldsymbol{\alpha}_{i\cdot}, \boldsymbol{\eta}_{i\cdot}, \boldsymbol{\phi}, s_i) = \prod_{j=1}^{p} f_{\text{ZINB}}(y_{ij}|\alpha_{ij}, \eta_{ij}, \phi_j, s_i),$$

where

$$\begin{aligned} & f_{\text{ZINB}}(y_{ij}|\alpha_{ij}, \eta_{ij}, \phi_j, s_i) \\ =& \mathsf{I}(y_{ij} = 0)^{\eta_{ij}} \left(\frac{\Gamma(y_{ij} + \phi_j)}{y_{ij}!\,\Gamma(\phi_j)} \left(\frac{\phi_j}{s_i\alpha_{ij} + \phi_j} \right)^{\phi_j} \left(\frac{s_i\alpha_{ij}}{s_i\alpha_{ij} + \phi_j} \right)^{y_{ij}} \right)^{1-\eta_{ij}}. \end{aligned}$$

Update of zero-inflation indicator η_{ij}: We update each $\eta_{ij}, i = 1, \ldots, n, j = 1, \ldots, p$ that corresponds to $y_{ij} = 0$ by sampling from the normalized version of the following conditional:

$$p(\eta_{ij}|\cdot) \propto f_{\text{ZINB}}(y_{ij}|\alpha_{ij}, \eta_{ij}, \phi_j, s_i) \cdot \text{Bern}(\eta_{ij}; \pi_i).$$

After the Metropolis-Hasting steps for all η_{ij}, we use a Gibbs sampler to update each $\pi_i, i = 1, \ldots, n$:

$$\pi_i|\cdot \sim \text{Be}(a_\pi + \sum_{j=1}^{p} \eta_{ij}, b_\pi + p - \sum_{j=1}^{p} \eta_{ij}).$$

Update of dispersion parameter ϕ_j: We update each $\phi_j, j = 1, \ldots, p$ by using a random walk Metropolis-Hastings algorithm. We first propose a new ${\phi_j}^*$ from $\text{Ga}({\phi_j}^2/\tau_\phi, \phi_j/\tau_\phi)$

and then accept the proposed value ${\phi_j}^*$ with probability $\min(1, m_{\mathsf{MH}})$, where

$$m_{\mathsf{MH}} = \frac{\prod_{i=1}^{n} f_{\mathsf{ZINB}}(y_{ij}|\alpha_{ij}, \eta_{ij}, \phi_j, s_i)}{\prod_{i=1}^{n} f_{\mathsf{ZINB}}(y_{ij}|\alpha_{ij}, \eta_{ij}, \phi_j, s_i)} \frac{\mathsf{Ga}({\phi_j}^*; a_\phi, b_\phi)}{\mathsf{Ga}(\phi_j; a_\phi, b_\phi)} \frac{J(\phi_j; {\phi_j}^*)}{J({\phi_j}^*; \phi_j)}.$$

Here we use $J(\cdot|\cdot)$ to denote the proposal probability distribution for the selected move. Note that the last term, which is the proposal density ratio, can be canceled out for this random walk Metropolis update.

Update of size factor s_i: We can rewrite Equation (2) in the main text, i.e.

$$\log s_i \sim \sum_{m=1}^{M} \psi_m \left[t_m \, \mathsf{N}(\nu_m, \sigma_s^2) + (1 - t_m) \, \mathsf{N}\left(-\frac{t_m \nu_m}{1 - t_m}, \sigma_s^2 \right) \right]$$

by introducing latent auxiliary variables to specify how each sample (in terms of $\log s_i$) is assigned to any of the inner and outer mixture components. More specifically, we can introduce an $n \times 1$ vector of assignment indicators $\boldsymbol{g}$, with $g_i = m$ indicating that $\log s_i$ is a sample from the m-th component of the outer mixture. The weight ψ_m determines the probability of each value $g_i = m$, with $m = 1, \ldots, M$. Similarly, we can consider an $n \times 1$ vector $\boldsymbol{\epsilon}$ of binary elements ϵ_i, where $\epsilon_i = 1$ indicates that, given $g_i = m$, $\log s_i$ is drawn from the first component of the inner mixture, i.e. $\mathsf{N}(\nu_m, \sigma_s^2)$ with probability t_m, and $\epsilon_i = 0$ indicates that $\log s_i$ is drawn from the second component of the inner mixture, i.e. $\mathsf{N}\left(-\frac{t_m \nu_m}{1 - t_m}, \sigma_s^2 \right)$, with probability $1 - t_m$. Thus, the Dirichlet process prior (DPP) model can be rewritten as

$$\log s_i | g_i, \epsilon_i, \boldsymbol{t}, \boldsymbol{\nu} \sim \mathsf{N}\left(\epsilon_i \nu_{g_i} + (1 - \epsilon_i) \frac{-t_{g_i} \nu_{g_i}}{1 - t_{g_i}}, \sigma_s^2 \right),$$

where $\boldsymbol{t}$ and $\boldsymbol{\nu}$ denote the collections of t_m and ν_m, respectively. Therefore, the update of the size factor $s_i, i = 1, \ldots, n$ can proceed by using a random walk Metropolis-Hastings algorithm. We propose a new $\log s_i^*$ from $\mathsf{N}(\log s_i, \tau_s^2)$ and accept it with proba-

bility $\min(1, \mathsf{m}_{\mathsf{MH}})$, where

$$\mathsf{m}_{\mathsf{MH}} = \frac{\prod_{j=1}^{p} f_{\mathsf{ZINB}}(y_{ij}|\alpha_{ij}, \eta_{ij}, \phi_j, s_i^*)}{\prod_{j=1}^{p} f_{\mathsf{ZINB}}(y_{ij}|\alpha_{ij}, \eta_{ij}, \phi_j, s_i)} \frac{\mathsf{N}(\log s_i^*; \epsilon_i \nu_{g_i} + (1-\epsilon_i)\frac{-t_{g_i}\nu_{g_i}}{1-t_{g_i}}, \sigma_s^2)}{\mathsf{N}(\log s_i; \epsilon_i \nu_{g_i} + (1-\epsilon_i)\frac{-t_{g_i}\nu_{g_i}}{1-t_{g_i}}, \sigma_s^2)} \times \frac{J(\log s_i; \log s_i^*)}{J(\log s_i^*; \log s_i)}.$$

Note that the last term, which is the proposal density ratio, equals 1 for this random walk Metropolis update. Since $\boldsymbol{g}$, $\boldsymbol{\epsilon}$, $\boldsymbol{t}$, and $\boldsymbol{\nu}$ have conjugate full conditionals, we use Gibbs samplers to update them one after another:

- Gibbs sampler for updating $g_i, i = 1, \ldots, n$, by sampling from the normalized version of the following conditional:

$$p(g_i = m|\cdot) \propto \psi_m \mathsf{N}\left(\log s_i; \epsilon_i \nu_m + (1-\epsilon_i)\frac{-t_m \nu_m}{1-t_m}, \sigma_s^2\right).$$

- Gibbs sampler for updating $\epsilon_i, i = 1, \ldots, n$, by sampling from the normalized version of the following conditional:

$$p(\epsilon_i|\cdot) \propto \begin{cases} (1-t_m)\mathsf{N}\left(\log s_i; -\frac{t_m\nu_m}{1-t_m}, \sigma_s^2\right) & \textit{if } \epsilon_i = 0 \\ t_m \mathsf{N}\left(\log s_i; \nu_m, \sigma_s^2\right) & \textit{if } \epsilon_i = 1 \end{cases}.$$

- Gibbs sampler for updating $t_m, m = 1, \ldots, M$:

$$t_m|\cdot \sim \mathsf{Be}(a_t + \sum_{i=1}^{n} \mathsf{I}(g_i = m)\mathsf{I}(\epsilon_i = 1), b_t + \sum_{i=1}^{n} \mathsf{I}(g_i = m)\mathsf{I}(\epsilon_i = 0)).$$

- Gibbs sampler for updating $\nu_m, m = 1, \ldots, M$:

$$\nu_m|\cdot \sim \mathsf{N}\left(\frac{c_m/\sigma_s^2}{e_m/\sigma_s^2 + 1/\tau_\nu^2}, \frac{1}{e_m/\sigma_s^2 + 1/\tau_\nu^2}\right),$$

where $c_m = \sum_{\{i:g_i=m, \epsilon_i=1\}} \log s_i - \frac{t_m}{1-t_m}\sum_{\{i:g_i=m,\epsilon_i=0\}} \log s_i$ and $e_m = \sum_{i=1}^{n} \mathsf{I}(g_i =$

$m)\mathsf{I}(\epsilon_i = 1) + \sum_{\{i:g_i=m,\epsilon_i=0\}} \left(\frac{t_m}{1-t_m}\right)^2$.

- Gibbs sampler for updating $\psi_m, m = 1, \ldots, M$ by stick-breaking process:

$$\begin{aligned}
\psi_1 &= v_1, \\
\psi_2 &= (1 - v_1)v_2, \\
&\vdots \\
\psi_M &= (1 - v_1) \cdots (1 - v_{M-1})v_M,
\end{aligned}$$

where $v_m|\boldsymbol{\nu} \sim \mathsf{Be}\left(a_m + \sum_{i=1}^n \mathsf{I}(g_i = m), b_m + \sum_{i=1}^n \mathsf{I}(g_i > m)\right)$.

For the sake of convenience, we have copied Equation (3) in the main text here,

$$p(\boldsymbol{\alpha}_{\cdot j}) = (nh_0 + 1)^{-\frac{1}{2}} \frac{\Gamma\left(a_0 + \frac{n}{2}\right)}{\Gamma(a_0)} \frac{b_0^{a_0}}{\left\{b_0 + \frac{1}{2}\left[\sum_{i=1}^n \log \alpha_{ij}^2 - \frac{\left(\sum_{i=1}^n \log \alpha_{ij}\right)^2}{n + \frac{1}{h_0}}\right]\right\}^{a_0 + \frac{n}{2}}}.$$

Update of normalized abundance: We update each $\alpha_{ij}, i = 1, \ldots, n, j = 1, \ldots, p$ by using a Metropolis-Hastings random walk algorithm. We first propose a new ${\alpha_{ij}}^*$ from $\mathsf{N}(\alpha_{ij}, \tau_\alpha^2)$, and then accept the proposed value with probability $\min(1, \mathsf{m_{MH}})$, where

$$\mathsf{m_{MH}} = \frac{f_{\mathsf{ZINB}}(\boldsymbol{y}_{i\cdot}|{\boldsymbol{\alpha}_{i\cdot}}^*, \cdot)}{f_{\mathsf{ZINB}}(\boldsymbol{y}_{i\cdot}|\boldsymbol{\alpha}_{i\cdot}, \cdot)} \frac{p\left({\boldsymbol{\alpha}_{\cdot j}}^*|\gamma_j\right)}{p\left(\boldsymbol{\alpha}_{\cdot j}|\gamma_j\right)} \frac{J\left(\alpha_{ij}; {\alpha_{ij}}^*\right)}{J\left({\alpha_{ij}}^*; \alpha_{ij}\right)}.$$

Note that the last term, which is the proposal density ratio, equals 1 for this random walk Metropolis update.

C.2. Infer the normalized abundances for multiple groups

In practice, when there are two groups of subjects in a microbiome study (e.g., subjects with two distinct phenotypes), the sequencing data usually include measurements on the same taxonomic features for all the subjects. Then, if the abundance of a taxon j does not differ between two groups, we can improve the posterior influence of $\log \boldsymbol{\alpha_{\cdot j}}$

by merging two groups to increase the sample size. On the other hand, if the taxon is associated with subject's condition, i.e., a taxon that changes its abundance between two groups in the study, the inference of $\log \boldsymbol{\alpha_{\cdot j}}$ should rely on each subject group.

With the goal of borrowing information to improve the posterior inference for certain taxa, we combine the original count matrix from two different groups, to generate the count matrix $\boldsymbol{Y}_{n\times p}$. Here, the sample size is $n = n_1 + n_2$, with n_1, n_2 representing the number of subjects in the first and the second groups, respectively. Meanwhile, we let $\boldsymbol{z} = (z_1, \ldots, z_n)^T$ to allocate the n subjects into two groups, with $z_i = 1$ or 2 indicating the group label of subject i. In practice, if taxon j is irrelevant to the subject's phenotype, its abundances should not be differentiating between two groups. However, if taxon j is associated with the disease, its abundance could either increase or decrease from healthy subjects to patients. Therefore, we model the normalized abundance α_{ij} as following:

$$\log \alpha_{ij} | \gamma_j \sim \begin{cases} \mathsf{N}(\mu_{0j}, \sigma_{0j}^2) & \text{if } \gamma_j = 0 \\ \mathsf{N}(\mu_{1j}, \sigma_{1j}^2) & \text{if } \gamma_j = 1 \text{ and } z_i = 1 \\ \mathsf{N}(\mu_{2j}, \sigma_{2j}^2) & \text{if } \gamma_j = 1 \text{ and } z_i = 2 \end{cases} . \tag{C.1}$$

Here, γ_j is a latent binary variable, with $\gamma_j = 1$ if taxon j is differentially abundant between two groups, and $\gamma_j = 0$ otherwise. For the taxa with $\gamma_j = 0$, we can borrow information between groups to increase the sample size in estimating the corresponding posterior of $\log \boldsymbol{\alpha_{\cdot j}}$. As an extension to Section 2.1 where we assume $\log \alpha_{ij} \sim \mathsf{N}(\mu_j, \sigma_j^2)$, the current model includes μ_{0j}, μ_{1j}, and μ_{2j} as the mean parameters for the normal mixture model. Again, we take the conjugate Bayesian approach and impose the following priors for the parameters in the normal mixture model: $\mu_{0j} \sim \mathsf{N}(0, h_0\sigma_0^2)$, $\sigma_{0j}^2 \sim \mathsf{IG}(a_0, b_0)$, $\mu_{kj} \sim \mathsf{N}(0, h_k\sigma_k^2)$ and $\sigma_{kj}^2 \sim \mathsf{IG}(a_k, b_k)$ for $k = 1, 2$.

The estimation of γ_j's determines the resulted normalized abundance matrix. Specifically, for taxon j with $\gamma_j = 0$, we can impute the zeros due to missing by the posterior mean

of $\log \boldsymbol{\alpha_{\cdot j}}$ calculated using information from both groups. As an extension to Equation (3) in the main text , the posterior of $\alpha_{\cdot j}|\gamma_j$ is as following:

$$p(\boldsymbol{\alpha}_{\cdot j}|\gamma_j) = (2\pi)^{-\frac{n}{2}} \times$$

$$\begin{cases} \prod_{k=1}^{K}(n_k h_k + 1)^{-\frac{1}{2}} \frac{\Gamma\left(a_k + \frac{n_k}{2}\right)}{\Gamma(a_k)} \frac{b_k^{a_k}}{\left\{ b_k + \frac{1}{2}\left[\sum_{\{i:z_i=k\}} \log \alpha_{ij}^2 - \frac{\left(\sum_{\{i:z_i=k\}} \log \alpha_{ij}\right)^2}{n_k + \frac{1}{h_k}} \right] \right\}^{a_k + \frac{n_k}{2}}} \\ \qquad \text{if } \gamma_j = 1 \\ (nh_0 + 1)^{-\frac{1}{2}} \frac{\Gamma\left(a_0 + \frac{n}{2}\right)}{\Gamma(a_0)} \frac{b_0^{a_0}}{\left\{ b_0 + \frac{1}{2}\left[\sum_{i=1}^{n} \log \alpha_{ij}^2 - \frac{\left(\sum_{i=1}^{n} \log \alpha_{ij}\right)^2}{n + \frac{1}{h_0}} \right] \right\}^{a_0 + \frac{n}{2}}} \\ \qquad \text{if } \gamma_j = 0 \end{cases}, \tag{C.2}$$

Therefore, we can obtain the posterior mean of $\log \boldsymbol{\alpha_{\cdot j}}$ by averaging over the log-transformed MCMC samples of $\boldsymbol{\alpha}_{\cdot j}$ after burn-in.

C.3. Additional tables and figures

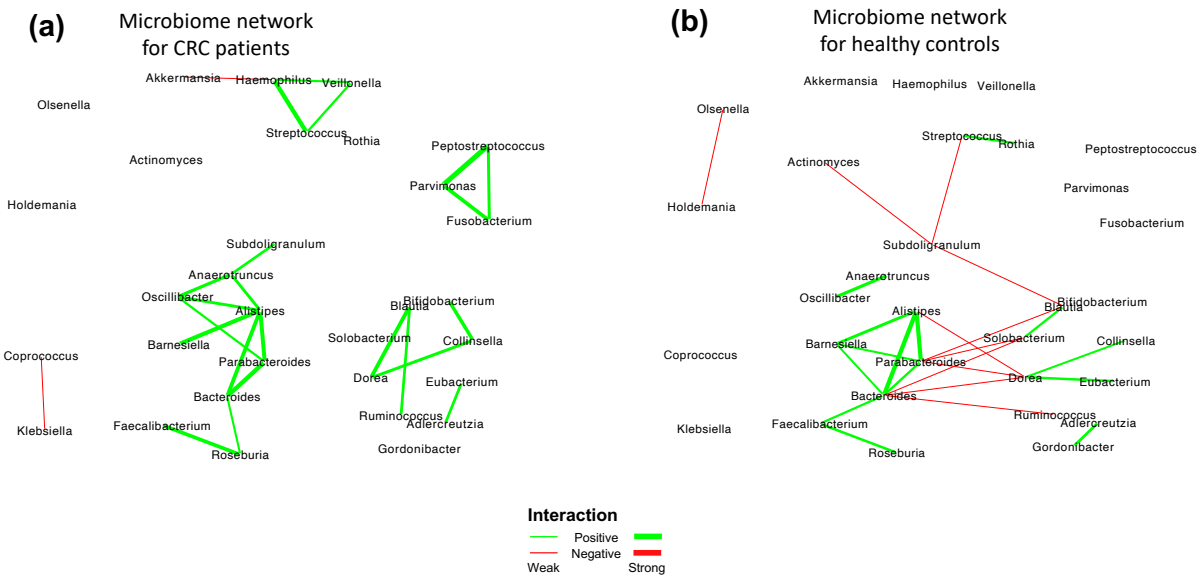


Figure C.1: CRC case study: The estimated networks by HARMONIES for (a) CRC patients and (b) healthy controls. All nodes are labeled in their genus names.

BIBLIOGRAPHY

[1] II Abubakar, T Tillmann, and A Banerjee. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease study 2013. *Lancet*, 385(9963):117–171, 2015.

[2] Edoardo M Airoldi and Jonathan M Bischof. Improving and evaluating topic models and other models of text. *Journal of the American Statistical Association*, 111(516):1381–1403, 2016.

[3] John Aitchison. Principal component analysis of compositional data. *Biometrika*, 70(1):57–65, 1983.

[4] Emma Allen-Vercoe and Christian Jobin. Fusobacterium and Enterobacteriaceae: important players for CRC? *Immunol Lett.*, 162(2):54–61, 2014.

[5] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010.

[6] Melina Arnold, Mónica S Sierra, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global patterns and trends in colorectal cancer incidence and mortality. *Gut*, 66(4):683–691, 2017.

[7] Fredrik Bäckhed, Ruth E Ley, Justin L Sonnenburg, Daniel A Peterson, and Jeffrey I Gordon. Host-bacterial mutualism in the human intestine. *Science*, 307(5717):1915–1920, 2005.

[8] Yuguang Ban, Lingling An, and Hongmei Jiang. Investigating microbial co-occurrence patterns based on metagenomic compositional data. *Bioinformatics*, 31(20):3322–3329, 2015.

[9] Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. *Hierarchical modeling and analysis for spatial data*. CRC Press, 2014.

[10] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, pages 289–300, 1995.

[11] Daniel Benten and Reiner Wiest. Gut microbiome and intestinal barrier failure–The "Achilles heel" in hepatology? *Journal of Hepatology*, 56(6):1221–1223, 2012.

[12] Christopher T Brown, Austin G Davis-Richardson, Adriana Giongo, Kelsey A Gano, David B Crabb, Nabanita Mukherjee, George Casella, Jennifer C Drew, Jorma Ilonen, Mikael Knip, et al. Gut microbiome metagenomics analysis suggests a functional model for the development of autoimmunity for type 1 diabetes. *PLoS One*, 6(10):e25792, 2011.

[13] Philip J Brown, Marina Vannucci, and Tom Fearn. Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 60(3):627–641, 1998.

[14] James H Bullard, Elizabeth Purdom, Kasper D Hansen, and Sandrine Dudoit. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11(1):94, 2010.

[15] A Colin Cameron and Pravin K Trivedi. *Regression analysis of count data*, volume 53. Cambridge University Press, 2013.

[16] J Gregory Caporaso, Christian L Lauber, William A Walters, Donna Berg-Lyons, Catherine A Lozupone, Peter J Turnbaugh, Noah Fierer, and Rob Knight. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. U S A.*, 108(Supplement 1):4516–4522, 2011.

[17] Marne C Cario and Barry L Nelson. Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Technical report, Citeseer, 1997.

[18] Ron Caspi, Hartmut Foerster, Carol A Fulcher, Pallavi Kaipa, Markus Krummenacker, Mario Latendresse, Suzanne Paley, Seung Y Rhee, Alexander G Shearer, Christophe Tissier, et al. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research*, 36(suppl_1):D623–D631, 2007.

[19] Mauro Castellarin, René L Warren, J Douglas Freeman, Lisa Dreolini, Martin Krzywinski, Jaclyn Strauss, Rebecca Barnes, Peter Watson, Emma Allen-Vercoe, Richard A Moore, et al. *Fusobacterium nucleatum* infection is prevalent in human colorectal carcinoma. *Genome Research*, 22(2):299–306, 2012.

[20] Doris SM Chan, Rosa Lau, Dagfinn Aune, Rui Vieira, Darren C Greenwood, Ellen Kampman, and Teresa Norat. Red and processed meat and colorectal cancer incidence: meta-analysis of prospective studies. *PloS One*, 6(6):e20456, 2011.

[21] Jun Chen and Hongzhe Li. Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *Annals of Applied Statistics*, 7(1), 2013.

[22] Li Chen, James Reeve, Lujun Zhang, Shengbing Huang, Xuefeng Wang, and Jun Chen. GMPR: A robust normalization method for zero-inflated count data with application to microbiome sequencing data. *PeerJ*, 6:e4600, 2018.

[23] Weiguang Chen, Fanlong Liu, Zongxin Ling, Xiaojuan Tong, and Charlie Xiang. Human intestinal lumen and mucosa-associated microbiota in patients with colorectal cancer. *PloS One*, 7(6):e39743, 2012.

[24] Yin Bin Cheung. Zero-inflated models for regression analysis of count data: a study of growth and development. *Statistics in Medicine*, 21(10):1461–1469, 2002.

[25] Olabisi Oluwabukola Coker, Geicho Nakatsu, Rudin Zhenwei Dai, William Ka Kei Wu, Sunny Hei Wong, Siew Chien Ng, Francis Ka Leung Chan, Joseph Jao Yiu Sung, and Jun Yu. Enteric fungal microbiota dysbiosis and ecological alterations in colorectal cancer. *Gut*, 68(4):654–662, 2019.

[26] Zhenwei Dai, Olabisi Oluwabukola Coker, Geicho Nakatsu, William KK Wu, Liuyang Zhao, Zigui Chen, Francis KL Chan, Karsten Kristiansen, Joseph JY Sung, Sunny Hei Wong, et al. Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers. *Microbiome*, 6(1):70, 2018.

[27] Marie-Agnès Dillies, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, Guillemette Marot, David Castel, Jordi Estelle, et al. A comprehensive evaluation of normalization methods for illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, 14(6):671–683, 2013.

[28] Julia L Drewes, Franck Housseau, and Cynthia L Sears. Sporadic colorectal cancer: microbial contributors to disease prevention, development and therapy. *British Journal of Cancer*, 115(3):273, 2016.

[29] Julia L Drewes, James R White, Christine M Dejea, Payam Fathi, Thevambiga Iyadorai, Jamuna Vadivelu, April C Roslani, Elizabeth C Wick, Emmanuel F Mongodin, Mun Fai Loke, et al. High-resolution bacterial 16S rRNA gene profile meta-analysis and biofilm status reveal common colorectal cancer consortia. *npj Biofilms Microbiomes*, 3(1):34, 2017.

[30] Huaying Fang, Chengcheng Huang, Hongyu Zhao, and Minghua Deng. CClasso: correlation inference for compositional data through lasso. *Bioinformatics*, 31(19):3172–3180, 2015.

[31] R Fang, BD Wagner, JK Harris, and SA Fillon. Zero-inflated negative binomial mixed model: an application to two microbial organisms important in oesophagitis. *Epidemiol. Infect.*, 144(11):2447–2455, 2016.

[32] Karoline Faust and Jeroen Raes. CoNet app: inference of biological association networks using Cytoscape. *F1000Research*, 5, 2016.

[33] Qiang Feng, Suisha Liang, Huijue Jia, Andreas Stadlmayr, Longqing Tang, Zhou Lan, Dongya Zhang, Huihua Xia, Xiaoying Xu, Zhuye Jie, et al. Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nat. Commun.*, 6:6528, 2015.

[34] Ronald A Fisher. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521, 1915.

[35] Burkhardt Flemer, Denise B Lynch, Jillian MR Brown, Ian B Jeffery, Feargal J Ryan, Marcus J Claesson, Micheal O'riordain, Fergus Shanahan, and Paul W O'toole. Tumour-associated and non-tumour-associated microbiota in colorectal cancer. *Gut*, 66(4):633–643, 2017.

[36] Kaitlin J Flynn, Nielson T Baxter, and Patrick D Schloss. Metabolic and community synergy of oral bacteria in colorectal cancer. *mSphere*, 1(3):e00102–16, 2016.

[37] Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

[38] Arthur E Frankel, Laura A Coughlin, Jiwoong Kim, Thomas W Froehlich, Yang Xie, Eugene P Frenkel, and Andrew Y Koh. Metagenomic shotgun sequencing and unbiased metabolomic profiling identify specific human gut microbiota and metabolites associated with immune checkpoint therapy efficacy in melanoma patients. *Neoplasia*, 19(10):848–855, 2017.

[39] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[40] Zhiguang Gao, Bomin Guo, Renyuan Gao, Qingchao Zhu, and Huanlong Qin. Microbiota disbiosis is associated with colorectal cancer. *Frontiers in Microbiology*, 6:20, 2015.

[41] Guadalupe Garcia-Tsao and Reiner Wiest. Gut microflora in the pathogenesis of the complications of cirrhosis. *Best Practice & Research: Clinical Gastroenterology*, 18(2):353–372, 2004.

[42] Andrew Gelman et al. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515–534, 2006.

[43] Jiawei Geng, Qingfang Song, Xiaodan Tang, Xiao Liang, Hong Fan, Hailing Peng, Qiang Guo, and Zhigang Zhang. Co-occurrence of driver and passenger bacteria in human colorectal cancer. *Gut Pathogens*, 6(1):26, 2014.

[44] Edward I George and Robert E McCulloch. Approaches for Bayesian variable selection. *Statistica Sinica*, pages 339–373, 1997.

[45] Dirk Gevers, Subra Kugathasan, Lee A Denson, Yoshiki Vázquez-Baeza, Will Van Treuren, Boyu Ren, Emma Schwager, Dan Knights, Se Jin Song, Moran Yassour, et al. The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host & Microbe*, 15(3):382–392, 2014.

[46] Joseph F Hair, William C Black, Barry J Babin, Rolph E Anderson, Ronald L Tatham, et al. Multivariate data analysis (vol. 6), 2006.

[47] Jonas Halfvarson, Colin J Brislawn, Regina Lamendella, Yoshiki Vázquez-Baeza, William A Walters, Lisa M Bramer, Mauro D'Amato, Ferdinando Bonfiglio, Daniel McDonald, Antonio Gonzalez, et al. Dynamics of the human gut microbiome in inflammatory bowel disease. *Nature Microbiology*, 2(5):17004, 2017.

[48] Ashley A Hibberd, Anna Lyra, Arthur C Ouwehand, Peter Rolny, Helena Lindegren, Lennart Cedgård, and Yvonne Wettergren. Intestinal microbiota is altered in patients with colon cancer and modified by probiotic intervention. *BMJ Open Gastroenterology*, 4(1):e000145, 2017.

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

[50] Kenya Honda and Dan R Littman. The microbiome in infectious disease and inflammation. *Annual Review of Immunology*, 30:759–795, 2012.

[51] HMP Integrative. The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell Host & Microbe*, 16(3):276, 2014.

[52] Hemant Ishwaran and Lancelot F James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

[53] Hemant Ishwaran, J Sunil Rao, et al. Spike and slab variable selection: frequentist and Bayesian strategies. *Annals of Statistics*, 33(2):730–773, 2005.

[54] Fredrik H Karlsson, Valentina Tremaroli, Intawat Nookaew, Göran Bergström, Carl Johan Behre, Björn Fagerberg, Jens Nielsen, and Fredrik Bäckhed. Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature*, 498(7452):99, 2013.

[55] James M Kinross, Ara W Darzi, and Jeremy K Nicholson. Gut microbiome-host interactions in health and disease. *Genome Medicine*, 3(3):14, 2011.

[56] Ioannis Koliarakis, Ippokratis Messaritakis, Taxiarchis Konstantinos Nikolouzakis, George Hamilos, John Souglakos, and John Tsiaoussis. Oral bacteria and intestinal dysbiosis in colorectal cancer. *International Journal of Molecular Sciences*, 20(17):4146, 2019.

[57] Matthew D Koslovsky, Kristi L Hoffman, Carrie R Daniel, Marina Vannucci, et al. A Bayesian model of microbiome data for simultaneous identification of covariate associations and prediction of phenotypic outcomes. *Annals of Applied Statistics*, 14(3):1471–1492, 2020.

[58] Aleksandar D Kostic, Eunyoung Chun, Lauren Robertson, Jonathan N Glickman, Carey Ann Gallini, Monia Michaud, Thomas E Clancy, Daniel C Chung, Paul Lochhead, Georgina L Hold, et al. *Fusobacterium nucleatum* potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host & Microbe*, 14(2):207–215, 2013.

[59] Aleksandar D Kostic, Dirk Gevers, Heli Siljander, Tommi Vatanen, Tuulia Hyötyläinen, Anu-Maaria Hämäläinen, Aleksandr Peet, Vallo Tillmann, Päivi Pöhö, Ismo Mattila, et al. The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell Host & Microbe*, 17(2):260–273, 2015.

[60] Ritesh Kumar, Jennifer L Herold, Deborah Schady, Jennifer Davis, Scott Kopetz, Margarita Martinez-Moczygemba, Barbara E Murray, Fang Han, Yu Li, Evelyn Callaway, et al. *Streptococcus gallolyticus* subsp. *gallolyticus* promotes colorectal tumor development. *PLoS Pathogens*, 13(7), 2017.

[61] Zachary D Kurtz, Christian L Müller, Emily R Miraldi, Dan R Littman, Martin J Blaser, and Richard A Bonneau. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Computational Biology*, 11(5):e1004226, 2015.

[62] Thomas NY Kwong, Xiansong Wang, Geicho Nakatsu, Tai Cheong Chow, Timothy Tipoe, Rudin ZW Dai, Kelvin KK Tsoi, Martin CS Wong, Gary Tse, Matthew TV Chan, et al. Association between bacteremia from specific microbes and subsequent diagnosis of colorectal cancer. *Gastroenterology*, 155(2):383–390, 2018.

[63] Minjung Kyung, Jeff Gill, and George Casella. Sampling schemes for generalized linear Dirichlet process random effects models. *Statistical Methods & Applications*, 20(3):259–290, 2011.

[64] Patricio S La Rosa, J Paul Brooks, Elena Deych, Edward L Boone, David J Edwards, Qin Wang, Erica Sodergren, George Weinstock, and William D Shannon. Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PloS One*, 7(12):e52078, 2012.

[65] Patricio S La Rosa, Yanjiao Zhou, Erica Sodergren, George Weinstock, and William D Shannon. Hypothesis testing of metagenomic data. In *Metagenomics for Microbiology*, pages 81–96. Elsevier, 2015.

[66] Clifford Lam and Jianqing Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics*, 37(6B):4254, 2009.

[67] Juhee Lee and Marilou Sison-Mangus. A bayesian semiparametric regression model for joint analysis of microbiome data. *Frontiers in Microbiology*, 9:522, 2018.

[68] Hongzhe Li. Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application*, 2:73–94, 2015.

[69] Jun Li, Daniela M Witten, Iain M Johnstone, and Robert Tibshirani. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics*, 13(3):523–538, 2012.

[70] Min Li, Baohong Wang, Menghui Zhang, Mattias Rantalainen, Shengyue Wang, Haokui Zhou, Yan Zhang, Jian Shen, Xiaoyan Pang, Meiling Zhang, et al. Symbiotic gut microbes modulate human metabolic phenotypes. *Proceedings of the National Academy of Sciences of the United States of America*, 105(6):2117–2122, 2008.

[71] Qiwei Li, Alberto Cassese, Michele Guindani, and Marina Vannucci. Bayesian negative binomial mixture regression models for the analysis of sequence count and methylation data. *Biometrics*, 75(1):183–192, 2019.

[72] Qiwei Li, Michele Guindani, Brian J Reich, Howard D Bondell, and Marina Vannucci. A Bayesian mixture model for clustering and selection of feature occurrence rates under mean constraints. *Statistical Analysis and Data Mining*, 10(6):393–409, 2017.

[73] Zhigang Li, Katherine Lee, Margaret R Karagas, Juliette C Madan, Anne G Hoen, A James O'malley, and Hongzhe Li. Conditional regression based on a multivariate zero-inflated logistic-normal model for microbiome relative abundance data. *Statistics in Biosciences*, 10(3):587–608, 2018.

[74] Qiaoyi Liang, Jonathan Chiu, Yingxuan Chen, Yanqin Huang, Akira Higashimori, Jingyuan Fang, Hassan Brim, Hassan Ashktorab, Siew Chien Ng, Simon Siu Man Ng, et al. Fecal bacteria act as novel biomarkers for noninvasive diagnosis of colorectal cancer. *Clinical Cancer Research*, 23(8):2061–2070, 2017.

[75] Han Liu, Kathryn Roeder, and Larry Wasserman. Stability approach to regularization selection (StARS) for high dimensional graphical models. In *Advances in Neural Information Processing Systems*, pages 1432–1440, 2010.

[76] Jason Lloyd-Price, Cesar Arze, Ashwin N Ananthakrishnan, Melanie Schirmer, Julian Avila-Pacheco, Tiffany W Poon, Elizabeth Andrews, Nadim J Ajami, Kevin S Bonham, Colin J Brislawn, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, 569(7758):655, 2019.

[77] Chieh Lo and Radu Marculescu. MPLasso: Inferring microbial association networks using prior microbial knowledge. *PLoS Computational Biology*, 13(12):e1005915, 2017.

[78] Xiaohang Long, Chi Chun Wong, Li Tong, Eagle SH Chu, Chun Ho Szeto, Minne YY Go, Olabisi Oluwabukola Coker, Anthony WH Chan, Francis KL

Chan, Joseph JY Sung, et al. Peptostreptococcus anaerobius promotes colorectal carcinogenesis and modulates tumour immunity. *Nature Microbiology*, pages 1–12, 2019.

[79] Sharon M Louie, Lindsay S Roberts, Melinda M Mulvihill, Kunxin Luo, and Daniel K Nomura. Cancer cells incorporate and remodel exogenous palmitate into structural and oncogenic signaling lipids. *Biochimica et Biophysica Acta*, 1831(10):1566–1572, 2013.

[80] Petra Louis, Georgina L Hold, and Harry J Flint. The gut microbiota, bacterial metabolites and colorectal cancer. *Nature Reviews Microbiology*, 12(10):661, 2014.

[81] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014.

[82] Lisa Maier, Mihaela Pruteanu, Michael Kuhn, Georg Zeller, Anja Telzerow, Exene Erin Anderson, Ana Rita Brochado, Keith Conrad Fernandez, Hitomi Dose, Hirotada Mori, et al. Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature*, 555(7698):623, 2018.

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

[84] Julian R Marchesi, Bas E Dutilh, Neil Hall, Wilbert HM Peters, Rian Roelofs, Annemarie Boleij, and Harold Tjalsma. Towards the human colorectal cancer microbiome. *PloS One*, 6(5):e20447, 2011.

[85] Vyara Matson, Jessica Fessler, Riyue Bao, Tara Chongsuwat, Yuanyuan Zha, Maria-Luisa Alegre, Jason J Luke, and Thomas F Gajewski. The commensal microbiome is associated with anti–PD-1 efficacy in metastatic melanoma patients. *Science*, 359(6371):104–108, 2018.

[86] Brian W Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta*, 405(2):442–451, 1975.

[87] DR Matthews, JP Hosker, AS Rudenski, BA Naylor, DF Treacher, and RC Turner. Homeostasis model assessment: insulin resistance and β-cell function from fasting plasma glucose and insulin concentrations in man. *Diabetologia*, 28(7):412–419, 1985.

[88] Paul J McMurdie and Susan Holmes. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.*, 10(4):e1003531, 2014.

[89] Nicolai Meinshausen, Peter Bühlmann, et al. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

[90] Patricia Menéndez, Yiannis AI Kourmpetis, Cajo JF ter Braak, and Fred A van Eeuwijk. Gene regulatory networks from multifactorial perturbations using Graphical Lasso: application to the DREAM4 challenge. *PloS One*, 5(12):e14147, 2010.

[91] Michael L Metzker. Sequencing technologies—the next generation. *Nature Reviews Genetics*, 11(1):31, 2010.

[92] Kosuke Mima, Reiko Nishihara, Zhi Rong Qian, Yin Cao, Yasutaka Sukawa, Jonathan A Nowak, Juhong Yang, Ruoxu Dou, Yohei Masugi, Mingyang Song, et al. *Fusobacterium nucleatum* in colorectal carcinoma tissue and patient prognosis. *Gut*, 65(12):1973–1980, 2016.

[93] Giorgia Mori, Simone Rampelli, Beatrice Silvia Orena, Claudia Rengucci, Giulia De Maio, Giulia Barbieri, Alessandro Passardi, Andrea Casadei Gardini, Giovanni Luca Frassineti, Stefano Gaiarsa, et al. Shifts of faecal microbiota during sporadic colorectal carcinogenesis. *Scientific Reports*, 8, 2018.

[94] Michael A Newton, Amine Noueiry, Deepayan Sarkar, and Paul Ahlquist. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5(2):155–176, 2004.

[95] Jung Hun Oh and Joseph O Deasy. Inference of radio-responsive gene regulatory networks using the graphical lasso algorithm. *BMC Bioinformatics*, 15(7):S5, 2014.

[96] Edoardo Pasolli, Lucas Schiffer, Paolo Manghi, Audrey Renson, Valerie Obenchain, Duy Tin Truong, Francesco Beghini, Faizan Malik, Marcel Ramos, Jennifer B Dowd, et al. Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods.*, 14(11):1023, 2017.

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

[98] Jie Peng, Pei Wang, Nengfeng Zhou, and Ji Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009.

[99] Xiaoling Peng, Gang Li, and Zhenqiu Liu. Zero-inflated beta regression for differential abundance analysis with metagenomics data. *Journal of Computational Biology*, 23(2):102–110, 2016.

[100] Rachel V Purcell, Martina Visnovska, Patrick J Biggs, Sebastian Schmeier, and Frank A Frizelle. Distinct gut microbiome patterns associate with consensus molecular subtypes of colorectal cancer. *Sci. Rep.*, 7(1):11590, 2017.

[101] Nan Qin, Fengling Yang, Ang Li, Edi Prifti, Yanfei Chen, Li Shao, Jing Guo, Emmanuelle Le Chatelier, Jian Yao, Lingjiao Wu, et al. Alterations of the human gut microbiome in liver cirrhosis. *Nature*, 513(7516):59, 2014.

[102] Allison L Richards, Amanda L Muehlbauer, Adnan Alazizi, Michael B Burns, Anthony Findley, Francesco Messina, Trevor J Gould, Camilla Cascardo, Roger Pique-Regi, Ran Blekhman, et al. Gut microbiota has a widespread and modifiable effect on host gene regulation. *mSystems*, 4(5):e00323–18, 2019.

[103] Jason M Ridlon, Joao Marcelo Alves, Phillip B Hylemon, and Jasmohan S Bajaj. Cirrhosis, bile acids and gut microbiota: unraveling a complex relationship. *Gut Microbes*, 4(5):382–387, 2013.

[104] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, 2015.

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

[106] Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25, 2010.

[107] Mara Roxana Rubinstein, Xiaowei Wang, Wendy Liu, Yujun Hao, Guifang Cai, and Yiping W Han. *Fusobacterium nucleatum* promotes colorectal carcinogenesis by modulating E-cadherin β-catenin signaling via its FadA adhesin. *Cell Host & Microbe*, 14(2):195–206, 2013.

[108] Terrance Savitsky and Marina Vannucci. Spiked Dirichlet process priors for Gaussian process models. *Journal of Probability and Statistics*, 2010, 2010.

[109] Mark A Schell, Maria Karmirantzou, Berend Snel, David Vilanova, Bernard Berger, Gabriella Pessi, Marie-Camille Zwahlen, Frank Desiere, Peer Bork, Michele Delley, et al. The genome sequence of Bifidobacterium longum reflects its adaptation to the human gastrointestinal tract. *Proceedings of the National Academy of Sciences, India Section A*, 99(22):14422–14427, 2002.

[110] Cynthia L Sears and Wendy S Garrett. Microbes, microbiota, and colon cancer. *Cell Host & Microbe*, 15(3):317–328, 2014.

[111] Nicola Segata, Levi Waldron, Annalisa Ballarini, Vagheesh Narasimhan, Olivier Jousson, and Curtis Huttenhower. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, 9(8):811, 2012.

[112] Ayelet Sivan, Leticia Corrales, Nathaniel Hubert, Jason B Williams, Keston Aquino-Michaels, Zachary M Earley, Franco W Benyamin, Yuk Man Lei, Bana Jabri,

Maria-Luisa Alegre, et al. Commensal Bifidobacterium promotes antitumor immunity and facilitates anti–PD-L1 efficacy. *Science*, page aac4255, 2015.

[113] Francesco C Stingo, Michele Guindani, Marina Vannucci, and Vince D Calhoun. An integrative Bayesian modeling approach to imaging genetics. *Journal of the American Statistical Association*, 108(503):876–891, 2013.

[114] Matthew A Taddy, Athanasios Kottas, et al. Mixture modeling for marked Poisson processes. *Bayesian Analysis*, 7(2):335–362, 2012.

[115] Mahlet G Tadesse, Naijun Sha, and Marina Vannucci. Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, 100(470):602–617, 2005.

[116] Ruqi Tang, Yiran Wei, Yanmei Li, Weihua Chen, Haoyan Chen, Qixia Wang, Fan Yang, Qi Miao, Xiao Xiao, Haiyan Zhang, et al. Gut microbial profile is altered in primary biliary cholangitis and partially restored after UDCA therapy. *Gut*, pages gutjnl–2016, 2017.

[117] Laura Tipton, Christian L Müller, Zachary D Kurtz, Laurence Huang, Eric Kleerup, Alison Morris, Richard Bonneau, and Elodie Ghedin. Fungi stabilize connectivity in the lung and skin microbial ecosystems. *Microbiome*, 6(1):12, 2018.

[118] Lorenzo Trippa and Giovanni Parmigiani. False discovery rates in somatic mutation studies of cancer. *Annals of Applied Statistics*, pages 1360–1378, 2011.

[119] Chu *et. al.* Tsoi, Ho. *Peptostreptococcus anaerobius* induces intracellular cholesterol biosynthesis in colon cells to induce proliferation and causes dysplasia in mice. *Gastroenterology*, 152(6):1419–1433, 2017.

[120] Tomotaka Ugai, Masataro Norizuki, Takahiro Mikawa, Goh Ohji, and Makito Yaegashi. Necrotizing fasciitis caused by haemophilus influenzae type b in a patient with rectal cancer treated with combined bevacizumab and chemotherapy: a case report. *BMC Infectious Diseases*, 14(1):198, 2014.

[121] Luke K Ursell, Jessica L Metcalf, Laura Wegener Parfrey, and Rob Knight. Defining the human microbiome. *Nutrition Reviews*, 70(suppl_1):S38–S44, 2012.

[122] W Duncan Wadsworth, Raffaele Argiento, Michele Guindani, Jessica Galloway-Pena, Samuel A Shelburne, and Marina Vannucci. An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC Bioinformatics*, 18(1):94, 2017.

[123] Tingting Wang, Guoxiang Cai, Yunping Qiu, Na Fei, Menghui Zhang, Xiaoyan Pang, Wei Jia, Sanjun Cai, and Liping Zhao. Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. *The ISME Journal*, 6(2):320–329, 2012.

[124] René L Warren, Douglas J Freeman, Stephen Pleasance, Peter Watson, Richard A Moore, Kyla Cochrane, Emma Allen-Vercoe, and Robert A Holt. Co-occurrence of anaerobic bacteria in colorectal carcinomas. *Microbiome*, 1(1):16, 2013.

[125] Sophie Weiss, Will Van Treuren, Catherine Lozupone, Karoline Faust, Jonathan Friedman, Ye Deng, Li Charlie Xia, Zhenjiang Zech Xu, Luke Ursell, Eric J Alm, et al. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *The ISME journal*, 10(7):1669, 2016.

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

[127] Daniela M Witten. Classification and clustering of sequencing data using a Poisson model. *Annals of Applied Statistics*, pages 2493–2518, 2011.

[128] Xiaoxuan Xia, William Ka Kei Wu, Sunny Hei Wong, Dabin Liu, Thomas Ngai Yeung Kwong, Geicho Nakatsu, Pearlly S Yan, Yu-Ming Chuang, Michael Wing-Yan Chan, Olabisi Oluwabukola Coker, et al. Bacteria pathogens drive host colonic epithelial cell promoter hypermethylation of tumor suppressor genes in colorectal cancer. *Microbiome*, 8(1):1–13, 2020.

[129] Yinglin Xia and Jun Sun. Hypothesis testing and statistical analysis of microbiome. *Genes and Diseases*, 4(3):138–148, 2017.

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

[131] Arthur W Yan, Derrick E Fouts, Johannes Brandl, Peter Stärkel, Manolito Torralba, Eckart Schott, Hide Tsukamoto, Karen E Nelson, David A Brenner, and Bernd Schnabl. Enteric dysbiosis associated with a mouse model of alcoholic liver disease. *Hepatology*, 53(1):96–105, 2011.

[132] Bahtiyar Yilmaz, Pascal Juillerat, Ove Øyås, Charlotte Ramon, Francisco Damian Bravo, Yannick Franc, Nicolas Fournier, Pierre Michetti, Christoph Mueller, Markus Geuking, et al. Microbial network disturbances in relapsing refractory Crohn's disease. *Nature Medicine*, 25(2):323–336, 2019.

[133] Dima Youssef, Ibrahim Youssef, Tariq S Marroush, and Mamta Sharma. Gemella endocarditis: A case report and a review of the literature. *Avicenna Journal of Medicine*, 9(4):164, 2019.

[134] Jun Yu, Qiang Feng, Sunny Hei Wong, Dongya Zhang, Qiao yi Liang, Youwen Qin, Longqing Tang, Hui Zhao, Jan Stenvang, Yanli Li, et al. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut*, 66(1):70–78, 2017.

[135] Georg Zeller, Julien Tap, Anita Y Voigt, Shinichi Sunagawa, Jens Roat Kultima, Paul I Costea, Aurélien Amiot, Jürgen Böhm, Francesco Brunetti, Nina Habermann, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular Systems Biology*, 10(11):766, 2014.

[136] Liangliang Zhang, Yushu Shi, Robert R Jenq, Kim-Anh Do, and Christine B Peterson. Bayesian compositional regression with structured priors for microbiome feature selection. *Biometrics*, 2020.

[137] Xinyan Zhang, Himel Mallick, Zaixiang Tang, Lei Zhang, Xiangqin Cui, Andrew K Benson, and Nengjun Yi. Negative binomial mixed models for analyzing microbiome count data. *BMC Bioinformatics*, 18(1):4, 2017.

[138] Xinyan Zhang, Himel Mallick, and Nengjun Yi. Zero-inflated negative binomial regression for differential abundance testing in microbiome studies. *Journal of Bioinformatics and Genomics*, (2 (2)), 2016.

[139] Haitao Zhao and Zhong-Hui Duan. Cancer Genetic Network Inference Using Gaussian Graphical models. *Bioinformatics and Biology Insights*, 13:1177932219839402, 2019.

[140] Wenhan Zhu, Maria G Winter, Mariana X Byndloss, Luisella Spiga, Breck A Duerkop, Elizabeth R Hughes, Lisa Büttner, Everton de Lima Romão, Cassie L Behrendt, Christopher A Lopez, et al. Precision editing of the gut microbiota ameliorates colitis. *Nature*, 553(7687):208, 2018.