A Novel Approach for Simulation-Based Power Estimation and Joint Modeling of Microbiome Counts

A Novel Approach for Simulation-Based Power Estimation and Joint Modeling of Microbiome Counts

By Michael Agronah,

A Thesis Submitted to the School of Graduate Studies in the Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy

McMaster University
Doctor of Philosophy (2025)
Hamilton, Ontario (School of Computational Science & Engineering)

TITLE: A Novel Approach for Simulation-Based Power Estimation and Joint Modeling of Microbiome Counts
AUTHOR: Michael Agronah (McMaster University)
SUPERVISOR: Dr. Benjamin Bolker
NUMBER OF PAGES: xiv, 98

Abstract

Advances in microbiome research have greatly enhanced our understanding of how microbial communities influence human health and disease. The advent of high-throughput sequencing technologies such as 16S rRNA amplicon and shotgun metagenomic sequencing has enabled researchers to generate microbiome abundance data for statistical analysis. These technological developments, together with the development of statistical methods have enabled researchers to detect differences in microbial composition across experimental conditions.

Despite these advancements, challenges remain in the areas of statistical power and sample size estimation, and the modeling of correlations between taxa in subject when analyzing associations between microbiome data and covariates. This PhD thesis addresses these challenges by developing new methods for power and sample size determination, and proposing methods for joint analysis of microbiome data while accounting for correlations among taxa in differential abundance studies.

We first developed two novel simulation methods (Chapter 2) designed to generate realistic microbiome count data for power and sample size estimation, and for evaluating the performance of the models we propose in this thesis. We then developed a new method for estimating statistical power in differential abundance studies. We apply this method to evaluate whether existing microbiome studies have sufficient power to detect differences in microbiome abundance (Chapter 3). Our findings suggest that differential abundance studies have low power to detect biologically meaningful differences. We extended our power estimation procedure to develop a novel method for sample size estimation for differential abundance studies (Chapter 4). Applying our sample size estimation procedure to real microbiome data sets suggests that the sample sizes seen in differential abundance microbiome literature may be too small to detect meaningful effects.

Most existing methods for differential abundance studies analyze individual taxa separately. We propose the Reduced Rank Multivariate Mixed Model (RRMM), which jointly models all taxa while accounting for correlations within subjects (Chapter 5). Due to the high dimensionality of microbiome data, modeling the correlations between taxa through a full variance-covariance matrix requires estimating thousands or even millions of parameters, making it computationally infeasible. RRMM reduces the number of parameters by applying rank reduction to the variance-covariance matrix. We show through simulation and using real microbiome data that RRMM improves precision in effect size estimates relative to standard methods such as the models implemented in the `DESeq2` and `NBZIMM` R packages. We extend RRMM to longitudinal microbiome design and developed the Longitudinal Reduced Rank Mixed Model (LRRMM) (Chapter 6). LRRMM jointly analyzes all taxa in a longitudinal study and models correlation and changes over time. Analyses of real and simulated data demonstrate that LRRMM improves precision in effect size (ie, a measure of the magnitude of the difference in taxon

abundance between experimental conditions) estimates than the models implemented in the `NBZIMM` package which model individual taxa separately.

Together, these contributions enhance the methodological foundation for microbiome research, offering methods for simulation, power analysis, and modeling that accounts for correlations between taxa within subjects in microbiome data.

Acknowledgements

I am deeply grateful to my wife, Mary, who has stood by my side through every challenge and triumph in our journey together. To my dear children–Selorm, Sena, and Selasi–your joy and laughter fill our home and continually inspire me to be the best father I can be.

My heartfelt thanks go to my supervisor, Dr. Benjamin Bolker, for his unwavering support, guidance, and mentorship over the years. I have learned so much from you–not only about statistics and computing, but also about kindness and intellectual curiosity. Your deep knowledge and generous spirit have been a constant source of inspiration. I could not have asked for a better supervisor.

I am also thankful to Dr. Jeganathan for her encouragement and support throughout my PhD journey, and to Dr. Michael Surette for his thoughtful feedback and guidance as a member of my committee. I sincerely thank you both as my supervisory committee members for helping me get this far.

Many others have also contributed to this work in meaningful ways. I would like to thank the members of the Theobio Lab at McMaster University for providing me with feedback that has helped me with my presentations and with this research. I am also thankful to all members of the Bio Data Lunch Lab at McMaster University for adding to my understanding of many statistical concepts. And to Dr. Ian Dworkin, thank you for generously sharing your ideas during the weekly Bio Data Lunch lab meetings.

I have received feedback on my write-ups and ideas from Dr Michael Li, Dr. Steve Cygu, Dr. Evan Mitchell, Dr Daniel Park (who all happen to be former students of either Dr Benjamin Bolker, Dr Jonathan Dushoff or Dr David Earn), Dr. Mary Edward (my beloved wife); Dr. Jennifer Stearns, Jake Szamosi (all from from McMaster University) and Dr Lord Kavi (Concordia University) and family for your feedback and support.

I am very thankful to the undergraduate team at the Department of Mathematics and Statistics at McMaster University for the many opportunities to develop my teaching skills both as a teaching assistant and sessional faculty member. These experiences have shaped me as an educator and enriched my academic journey. A big thanks to Dr. Erin Clements at McMaster University (Department of Mathematics and Statistics) for sharing with me her insights and knowledge about teaching.

Great thanks to Dr Erin Allard and Kris Knorr (Educational Developers at the Paul R. MacPherson Institute for Leadership, Innovation and Excellence in Teaching) for your valuable feedback on my teaching dossier and pedagogy

And to many others; I may not be able to list you all but I say a very big thank you.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Michael AGRONAH, declare that this thesis titled, "A Novel Approach for Simulation-Based Power Estimation and Joint Modeling of Microbiome Counts" and the work presented in it are my own. I confirm that:

- Chapter 2: I developed a novel simulation approach (MixGaussSim) to model the effect size and mean abundance distributions of taxa in microbiome studies. I also created a second simulator (RRSim) that jointly simulates microbiome count data while incorporating correlations between taxa within subjects. I implemented these simulation methods in R.
- Chapter 3: I investigated whether differential abundance microbiome studies are underpowered by developing a novel method for estimating statistical power at the level of individual taxa. I analyzed real microbiome data sets to assess statistical power in existing studies.
- Chapter 4: I introduced a new methodology for estimating sample size required in differential abundance microbiome studies, incorporating effect size, statistical power, and taxon mean abundance.
- Chapter 5: I developed the Reduced Rank Mixed Model (RRMM) to jointly model microbiome taxa while accounting for within-subject correlations. I implemented this modeling framework in R and used the RRSim simulator from Chapter 2 to evaluate its effectiveness. I also applied RRMM to real microbiome data sets to compare its performance with models that analyze individual taxa separately.
- Chapter 6: I extended the RRMM framework to longitudinal microbiome studies by developing the Longitudinal Reduced Rank Mixed Model (LRRMM). I formulated the model, implemented it in R, and evaluated its performance using both simulated data (generated by RRSim) and real microbiome data sets. I also compared LRRMM's performance to models that assume independence of taxa across time points.

I confirm that this work is original and has not been submitted for any other degree or professional qualification. Any external sources and collaborations have been properly acknowledged.

Chapter 1

Introduction

1.1 Background

The human microbiome is a community of microorganisms (bacteria, viruses, fungi, etc) that reside throughout the body. While the term "*microbiome*" encompasses all these microbial groups, most studies, particularly those using 16S rRNA gene sequencing, focus specifically on bacterial communities (Xia et al. 2018; Kotthapalli and Archer 2024). The microbiome is vital for human health, contributing to metabolic regulation, immune system support, and even neurological function (Lynch and Pedersen 2016).

High-throughput genomic sequencing technologies, such as 16S rRNA gene-targeted amplicon sequencing and shotgun metagenomic sequencing, have made it easier to profile and analyze microbiome composition (Kuczynski et al. 2012). Sequence reads are processed through bioinformatic pipelines like `DADA2` (Callahan et al. 2016) for 16S-targeted sequencing or `MetaPhlAn2` (Truong et al. 2015) for shotgun metagenomic data, resulting in an abundance table (also known as Amplicon Sequence Variants (ASV) table that records the frequencies of detected microbial taxa (Tab. 1.1). This abundance table, combined with metadata capturing sample-level characteristics, serves as the foundation for downstream statistical analyses. Advancements in sequencing technologies have significantly accelerated progress in microbiome research (Fig. 1.1). The Nature Navigator, for instance, reports that human microbiome studies have seen rapid growth with 118,970 publications and 298,117 researchers as of December 10, 2024.

Current microbiome studies focus on two main research goals (Xia et al. 2018): (1) investigating connections between the microbiome and biological, genetic, clinical, or experimental conditions of hosts, such as the relationship between dysbiosis and disease progression (Lewis et al. 2015), and (2) finding relationships between biological and environmental factors and microbiome composition, such as the impact of dietary interventions on the gut microbiota (Albenberg et al. 2012).

A common research goal of both kinds of analysis is identifying microbial taxa that differ in their abundance between experimental conditions, such as diseased versus healthy individuals, treated versus untreated subjects, or different environmental conditions (Hawinkel et al. 2019). Studies with such objectives, often referred to as differential abundance (DA) analysis, could aid in identifying microbial biomarkers associated with

TABLE 1.1: An example of Amplicon Sequence Variants (ASV) table from 16S rRNA sequencing. Each row represents a taxon and each column indicates the read count for a corresponding sample. Data Source: "*Statistical Analysis of Microbiome Data with R"-page 32* (Xia et al. 2018; Jin et al. 2015).

Species	**5_15_drySt-28F**	**20_12_CeSt-28F**	**1_11_drySt-28F**	**2_12_drySt-28F**
Tannerella sp.	474	66	543	569
Lactococcus lactis	326	737	2297	548
Lactobacillus murinus	11	42	114	28
Lactobacillus murinus:: Lactococcus lactis	1	12	25	5
Parasutterella excrementihominis	1	0	1	4
Helicobacter hepaticus	87	0	0	13
Prevotella sp.	116	5	237	59
Bacteroides sp.	174	31	945	353
Barnesiella intestinihominis	8	1	0	2
Lactobacillus murinus:: Lactobacillus sp.	1	9	7	4

health outcomes and environmental changes. Just as with microbiome research in general, small sample sizes are also common in differential abundance (DA) microbiome studies, raising concerns about low statistical power and the reproducibility of findings (Kers and Saccenti 2021; Kelly et al. 2015). In addition, modeling correlations between taxa is difficult due to the high dimensionality of microbiome data, making joint modeling approaches challenging. This thesis addresses these two challenges in differential abundance microbiome analysis.

1.2 Motivation and Thesis Objectives

The objectives of this thesis are as follows:

- **Statistical power and sample size determination:** Microbiome research is often faced with low sample sizes due to the lack of resources and constraints which could lead to low statistical power. Many microbiome studies have argued that microbiome research, in general, lacks sufficient power to detect meaningful biological differences (Brüssow 2020; Kers and Saccenti 2021). However, this claim has not been specifically investigated for differential abundance (DA) microbiome studies. Existing methods for power calculation in microbiome studies are limited. To our knowledge, no methods exist to estimate statistical power in DA studies. Since the goal of DA analysis is to identify taxa with significant difference between experimental conditions, methods for sample size calculation tailored to differential abundance analysis need to account for the range of effect size and statistical

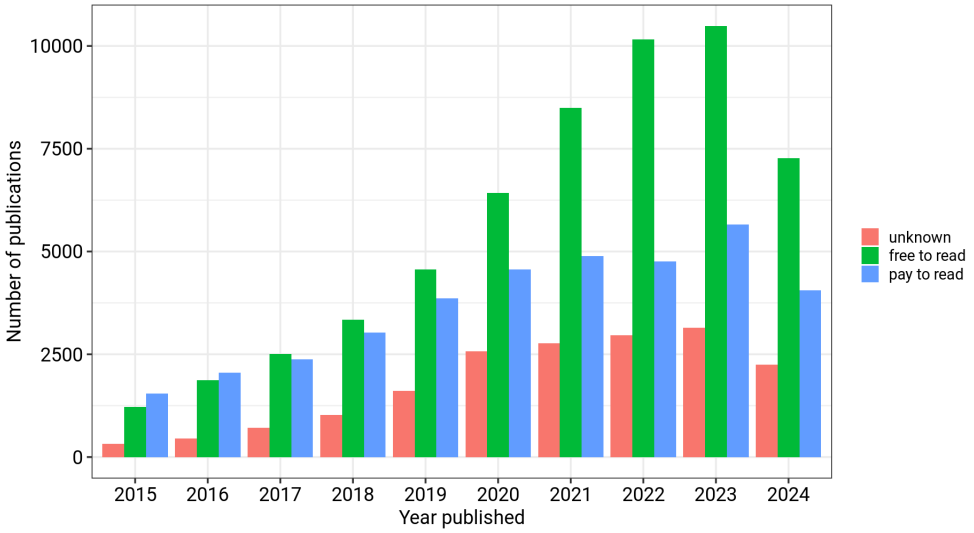

FIGURE 1.1: Growth in Human Microbiome Research. Data source: Nature Navigators (118,970 publications and 298,117 researchers as of December 10, 2024.)

power of individual taxa. This thesis will develop methods for statistical power and sample size determination for DA microbiome studies.

- **Correlation structures in microbiome data:** In microbiome data, abundances of taxa within a subject are often correlated (Hawinkel et al. 2019). Ignoring these correlations when modeling microbiome data could lead to imprecise effect size estimates. Most existing methods that analyze the association between taxa counts and covariates treat each taxon separately, assuming no correlations between taxa within subjects (Hawinkel et al. 2019). A major challenge in modelling these correlations is the large number of parameter estimates required for the correlation matrix caused by the high dimensionality of microbiome data. A typical microbiome data set with hundreds or thousands of taxa may require thousands or millions of correlation parameters. Fitting such correlations is neither computationally nor statistically feasible. This thesis will develop statistical models to jointly analyze the relationship between the count abundances of all taxa and covariates, while accounting for correlations among taxa. To address the challenge of estimating an excessive number of parameters in the correlation matrix, the model will incorporate a latent variable approach that significantly reduces the number of required parameter estimates.

1.3 Thesis Contributions

The main contributions of this thesis are as follows:

- a novel simulation framework (MixGaussSim) that improves power estimation in microbiome DA studies by explicitly modeling the relationship between taxon abundance and effect size.
- a novel method for estimating statistical power at the level of individual taxa, addressing a key gap in DA studies.
- a novel sample size determination method that accounts for taxon-specific effect sizes and statistical power.
- a statistical model; the Reduced Rank Mixed Model (RRMM), for jointly analyzing microbiome count data while accounting for correlations among taxa.
- an extension of RRMM to longitudinal data (LRRMM) that models correlation and changes over time and improves precision of estimates in longitudinal microbiome studies.

1.4 Thesis Structure

The remainder of this thesis is organized as follows:

Chapter 2: Statistical power calculation for microbiome studies cannot be performed analytically using theoretical distributions of test statistics (e.g., non-central t or chi-squared distributions), due to the complexity of microbiome data (Arnold et al. 2011). Instead, it requires simulating data that closely resemble real microbiomes to estimate statistical power. For differential abundance studies, statistical power should be calculated at the level of individual taxa since each taxon has its own effect size, leading to variations in power to detect these effects. Effect size of taxa are also related to by the mean abundances of taxa. A simulator that flexibly models the distribution of effect sizes and their relationship with abundances of taxa can facilitate the calculation of statistical power for individual taxa. Chapter 2 presents a novel simulation approach to model the effect size and mean abundance distributions of taxa and their relationships (MixGaussSim).

In order to demonstrate the value of accounting for correlations between taxa within subjects when modeling microbiome data, Chapter 2 also presents a simulator that allows the joint simulation of microbiome count while including correlations between taxa within subjects (RRSim). Unlike other models for simulating microbiome data, which simulate individual taxa separately, RRSim simulates counts of taxa jointly. The simulator RRSim is used in chapter 6 and 7 to account for correlations between taxa in differential abundance microbiome studies.

Chapter 3: In this chapter, we introduce a novel approach for estimating statistical power for individual taxa in differential abundance analyses. Using this power calculation framework, we investigate whether existing DA studies are underpowered. Our findings suggest that differential abundance studies may have low power to detect biologically meaningful differences.

Chapter 4: A key application of power analysis is determining the sample size required to detect an effect of a given size at a specified statistical power. Sample size determination depends on both effect size and statistical power. In differential abundance studies, these factors vary across taxa, as each taxon has its own effect size and corresponding statistical power. Effect sizes in differential abundance studies are also influenced by the mean abundances of taxa. Therefore, sample size estimation for differential abundance studies must account for taxon-specific effect sizes, statistical power, and mean abundances.

In this chapter, we introduce a novel method for determining sample size in differential abundance microbiome studies. Our approach estimates sample size as a function of effect size, statistical power, and the mean abundance of a given taxon. Using our sample size calculation framework along with the MixGaussSim simulator from Chapter 2, we demonstrate that differential abundance microbiome studies may require larger sample sizes than those commonly reported in the microbiome literature to achieve high statistical power (80% or greater).

Chapter 5: This chapter introduces the Reduced Rank Mixed Model (RRMM), a framework that jointly analyzes all taxa while accounting for correlations within subjects. Due to the high dimensionality of microbiome data, the correlation matrix can involve thousands or even millions of parameters, especially in data sets with hundreds or thousands of taxa. RRMM uses a latent variable model, specifically the rank reduction, to reduce the number of parameter estimates required for the correlation matrix.

Using the RRSim simulator introduced in Chapter 2, we generate simulated microbiome data to investigate whether incorporating taxon correlations leads to more precise effect size estimates compared to models that analyze individual taxa independently. We extend our analysis to real microbiome data sets.

Chapter 6: Longitudinal microbiome studies involve repeated sampling of subjects over time, allowing researchers to examine changes in microbial communities and their associations with covariates such as treatment effects, disease progression, and environmental factors.

This chapter extends the Reduced Rank Mixed Model (RRMM) introduced in Chapter 5 to a longitudinal microbiome study design. Our proposed Longitudinal Reduced Rank Mixed Model (LRRMM) jointly models all taxa across multiple time points while accounting for correlations within subjects over time. This extension addresses the complexity introduced by repeated measures and the need to model interactions between taxa and time-dependent effects. By using the reduced-rank approximation of the variance-covariance structure, LRRMM provides a computationally efficient approach to modeling high-dimensional microbiome data in a longitudinal framework.

We estimate how the rates of change in taxon abundance differ between groups (e.g., control vs. treatment or healthy vs. diseased subjects). Our approach is evaluated using simulated data generated by the RRSim simulator from Chapter 2 as well as with

real microbiome data sets. We investigated whether LRRMM leads to improvement in estimates in comparison to models that assume independence.

1.5 Summary

Differential abundance analysis are common in microbiome research, yet existing statistical methods face challenges related to power estimation, sample size determination, and taxon correlation structures. This thesis addresses these challenges by developing novel simulation methods, statistical power estimation frameworks, and correlation-aware modeling approaches. This work contributes methodological advancements that improve the reliability and reproducibility of microbiome DA studies.

1.6 Glossary of Terms

This section provides definitions for key terms and abbreviations used throughout the thesis.

TABLE 1.2: Glossary of Key Terms Used in the Thesis

Term	**Definition**
Power	The probability that a statistical test correctly detects a true effect (e.g., a truly differentially abundant taxon).
Average power	The arithmetic mean of the power estimates for all taxa
FDR (False Discovery Rate)	The expected proportion of false positives among the declared significant results, controlled using procedures like the Benjamini–Hochberg method.
fold change	A measure of effect size used in this thesis
ASV (Amplicon Sequence Variant)	A high-resolution method for distinguishing microbial taxa based on unique DNA sequences derived from marker gene sequencing, typically the 16S rRNA gene.
Differential Abundance	The process of identifying taxa whose abundance significantly differs across conditions, such as disease vs. healthy states.
Taxon (plural: taxa)	A classification unit in biological taxonomy, such as species, genus, or family.
`count` (used within `glmmTMB` syntax)	Denotes a long vector (ie; a concatenation of counts of all taxa across all subjects (and across all time points in the case of longitudinal data))
`taxon` (used within `glmmTMB` syntax)	refers to a factor label for a given taxa

1.7 Code Availability

The `R` code used for the analysis presented in this thesis can be found in the repository: https://doi.org/10.5281/zenodo.15556323

Chapter 2

Two Approaches for Simulating Microbiome Data

2.1 Abstract

Simulating microbiome count data offers significant advantages for microbiome research, including evaluating statistical methods, estimating statistical power, and studying the effects of parameters in controlled settings where true values are known. Most microbiome simulation approaches generate synthetic data, which are valuable for benchmarking statistical methods, testing computational tools, and exploring hypothetical scenarios. However, synthetic data often fail to fully capture the complex structure of real microbiome data sets, particularly taxon-taxon correlations, the relationship between mean abundance and effect sizes, zero inflation, and compositionality.

We introduce two novel simulation approaches: MixGaussSim and RRSim. MixGaussSim models the distribution of mean abundances and effect sizes of taxa as well as the relationship between mean abundance and effect size of taxa using a mixture of gaussian distributions. In contrast, RRSim jointly models all taxa and models the mean abundance of taxa with a mixed-effects model while accounting for correlations between taxa within subjects. RRSim uses the reduced-rank method to reduce the number of estimates required for the variance-covariance structure and provides flexibility in modeling zero inflation at both the taxon and group levels.

We compared MixGaussSim and RRSim against two existing simulators in the `HMP` and `metaSPARSim` R pacakges, using Kolmogorov-Smirnov (KS) goodness-of-fit tests. Results indicate that `metaSPARSim` best replicated the distribution of mean abundance of taxa, while MixGaussSim and RRSim ranked second or third across data sets. RRSim is the best simulator in terms of replicating the distributions of the proportions of zero counts across samples for most data sets we considered. RRSim and MixGaussSim are also effective in replicating the distribution of variance of taxa for most data sets.

2.2 Introduction

Data simulation approaches provide significant advantages for scientific research. Simulation studies help researchers gain insights about systems that may be difficult to obtain through observation and experiments alone. Across disciplines, data simulation is widely used to benchmark and validate statistical methods in controlled settings where the true parameter values are known. Simulation methods enable researchers to assess the reliability, accuracy and robustness of statistical methods before applying them to empirical data.

In microbiome research, simulations are particularly valuable for evaluating statistical power, as analytical power calculations based on distributional assumptions are rarely feasible (Arnold et al. 2011; Agronah and Bolker 2025). For example, Kelly et al. 2015 developed a simulation framework to aid the estimation of power for a community-wide microbiome analysis. Simulation studies also facilitate the evaluation of key study design factors, such as sequencing depth, sample size and differences in microbial communities between groups (e.g., control vs. treatment groups), helping researchers optimize experimental designs (Johnson et al. 2015).

Most microbiome simulation approaches generate synthetic data, which are useful for benchmarking statistical methods, testing computational tools, and exploring hypothetical scenarios (Kelly et al. 2015; Patuzzi et al. 2024). However, synthetic data often fail to capture the structure of real microbiome data sets, including correlations among taxa, the relationship between mean abundance and effect sizes, zero inflation, and compositionality. Simulating data that closely resemble real microbiome data is challenging for several reasons. First, microbiome data sets are high-dimensional, containing many taxa but relatively few samples (Xia et al. 2018). Moreover, taxa within individual subjects are correlated (Hawinkel et al. 2019). Modeling these correlations is computationally demanding and requires statistical methods to reduce the dimensionality of these correlations. Second, microbiome data sets are sparse, with many zero counts arising from taxa being absent or falling below detection thresholds (Hu et al. 2018; Fang et al. 2016). Standard count models may not fully capture this zero structure, requiring specialized approaches to model excess zeros appropriately. Third, microbiome data are typically compositional, meaning relative abundances are constrained to a fixed total, necessitating statistical models that account for these constraints (Xia et al. 2018; Chen and Li 2016). Overdispersion is also a key issue, with variance often exceeding the mean due to biological variability and technical noise (Xia et al. 2018).

In addition to the challenges posed by the properties of microbiome data, estimating the distribution of effect sizes (ie, a measure of the magnitude of the difference in taxon abundance between experimental conditions) for simulation is also challenging, as published microbiome studies often fail to report effect sizes, and those that do may present inflated values due to small sample sizes—a phenomenon known as the winner's curse (Button et al. 2013). Addressing these challenges is crucial for developing realistic microbiome simulations.

Several simulation approaches have been developed to generate microbiome data, each addressing different challenges in modeling community structure and sequencing biases. Synthetic data simulators (Patuzzi et al. 2024; Kelly et al. 2015) use predefined statistical distributions to create artificial data sets, allowing controlled benchmarking of statistical methods. Empirical data-driven simulators (Hawinkel et al. 2019) modify or resample real microbiome data sets to better capture key characteristics such as zero inflation, overdispersion, and compositionality. Mechanistic and ecological models incorporate ecological interactions and evolutionary dynamics to simulate microbiome communities (Pasqualini et al. 2024).

We propose two novel approaches for simulating microbiome data: a discrete mixture of Gaussians simulation method (MixGaussSim) and a Reduced Rank simulation method (RRSim). An advantage of MixGaussSim is its ability to model the distribution of mean abundance and effect size of taxa as well as modeling the relationship between effect size and mean abundance of taxa. This enables researchers to quantify the distribution of effect sizes given the mean abundance of taxa based on existing data. In contrast, RRSim jointly models all taxa and uses a dimension reduction technique to model correlations between taxa. RRSim can also model zero inflation in microbiome data.

2.3 Data Collection and Processing

In order to determine realistic parameter values for the simulation models, we estimated these parameters from real microbiome data sets for use in subsequent data simulations. We obtained seven microbiome data sets from the European Nucleotide Archive (EBA) (Leinonen et al. 2010) and the National Center for Biotechnology Information (NCBI) (Sayers et al. 2021). The selection of autism-related microbiome data sets was motivated by prior experience working with children with autism spectrum, which provided motivation and contextual insight. We performed a search using the query terms *"autism[All Fields] AND 16S[All Fields]"* and *"autism[All Fields] AND 16S[All Fields] AND Fecal[All Fields]"* on November 6, 2021. The search resulted in 10 data sets with accession numbers PRJNA168470, PRJNA355023, PRJNA453621, PRJEB45948, PRJNA644763, PRJNA589343, PRJNA687773, PRJNA578223, PRJNA624252 and PRJNA642975. Each data included a "treatment" group of children with autism spectrum disorder and a "control" group of neurotypical children.

To prepare the data sets for downstream analysis, we removed adaptors and primer sequences using the `cutadapt` function. We then processed the trimmed sequences into Amplicon Sequence Variant (ASV) data using the `Dada2` (Callahan et al. 2016) pipeline, which involved various steps such as filtering and trimming, error estimation, denoising, merging paired reads, and removing chimeras (Chen et al. 2020). Three of the data sets (PRJNA578223, PRJNA624252, and PRJNA642975) had very low count abundances. Pre-filtering performed to remove low mean abundances resulted in the exclusion of the majority of taxa from these data sets. Consequently, these data sets were excluded, leaving seven data sets for our analysis.

2.3.1 Pre-filtering low abundant taxa

Taxa with low abundance exhibit high variability, which can pose challenges in detecting significant differences between groups. As is routinely done in differential abundance analysis, we filtered rare taxa in each of the seven autism data sets, retaining only those taxa that had an abundance of five or more reads in at least three samples (Xia et al. 2018; Love et al. 2014). Table 2.1 provides a summary of the seven autism data sets used in this study.

TABLE 2.1: Summary of autism microbiome data sets used in the analyses presented in Chapter 2 -4. The data sets were obtained from the European Nucleotide Archive (EBA) and the National Center for Biotechnology Information (NCBI) repositories and consists of fecal samples. For all data sets, rare taxa were filtered out by retaining only those with an abundance of at least five reads in three or more samples.

Accession Number	**DNA Region**	**Platform**	**# Samples**	**# ASV after processing**	**# ASV after prefiltering**
PRJNA168470	V2-V3	454 GS FLX Titanium	40 (20 ASD, 20 NT)	3264	680
PRJNA355023	V3	NextSeq 500	54 (30 ASD, 24 NT)	4401	1208
PRJNA453621	V4	Illumina HiSeq 2500	286 (143 ASD, 143 NT)	2522	1233
PRJEB45948	V3-V4	Illumina MiSeq	92 (54 ASD, 38 NT)	20356	5039
PRJNA644763	V3-V4	Illumina MiSeq	123 (76 NT, 47 ASD)	8477	1053
PRJNA589343	V4	Illumina HiSeq 4000	127 (77 ASD, 50 NT)	1656	599
PRJNA687773	V4-V5	Illumina MiSeq	83 (44 ASD, 39 NT)	4864	897

2.4 The Mixture of Gaussian Simulation Approach (MixGaussSim)

The descriptions presented in this method section are adapted from a published paper with the PLOS ONE journal, as cited in (Agronah and Bolker 2025).

MixGaussSim uses the negative binomial model, a standard approach for analyzing microbiome count data. For example, the negative binomial model is implemented in the `DEseq2` (Love et al. 2014) and `edgeR` (Robinson et al. 2010) `R` packages, both originally developed for transcriptomics data and now widely used in microbiome analysis. MixGaussSim uses a mixture of Gaussian distributions to model distributions of fold change and mean abundance of taxa and generates microbiome data from the negative binomial model.

2.4.1 The Negative Binomial Model

Let K_{ij} denote the count data for the i^{th} taxon in the j^{th} sample. Then K_{ij} follows a negative binomial distribution:

$$\begin{aligned} K_{ij} &\sim \text{NB}(\text{mean} = \mu_{ij}, \text{dispersion} = \alpha_{ij}), \\ \mu_{ij} &= s_j q_{ij} \\ \log q_{ij} &= \sum_r x_{jr}\beta_{ir}, \end{aligned} \tag{2.1}$$

where μ_{ij} and s_j are the mean abundance and normalization constants is a sample-specific scaling factor that accounts for differences in sequencing depth or other technical variation across samples respectively. q_{ij} is the expected mean abundance of a given taxon in a sample prior to normalization. We assume the dispersion parameter is constant for a given taxon. Thus, $\alpha_{ij} = \alpha_i$. The coefficients $\hat{\beta}_{ir}$ are estimates of the effect sizes and x_{jr} are the covariates. The relationship between the variance of counts and the dispersion is defined by $\text{var}(K_{ij}) = \mu_i + \alpha_i \mu_i^2$. In this study, the estimating procedure implemented in the `DESeq2` package (Love et al. 2014; Anders and Huber 2010) in `R` is used for estimating $\hat{\beta}_{ir}$ and $\hat{\alpha}_i$.

In order to simulate microbiome data using MixGaussSim, we need to estimate realistic distributions for effect size and mean abundance of taxa. We estimate these distributions from the seven autism data sets presented in Section 2.3.

The following sections describe our methods for fitting distributions to the mean abundance and fold change, and for simulating microbiome count data.

2.4.2 Effect size shrinkage

Rare taxa also often lead to implausibly large fold change estimates which can distort the accuracy of estimating the true distribution of effect sizes. To tackle the problem of exaggerated effect sizes from rare taxa, we used a shrinkage functionality (ie, shrinkage `type = normal`) in the `DESeq2` package, which shrinks large fold change estimates for low-abundance taxa towards zero.

2.4.3 Estimating Distributions for Mean Abundance and Effect Size

We modeled log mean abundance (that is, log of the arithmetic mean abundance from both control and treatment groups) as a finite mixture of Gaussian distributions. To determine the optimal number of components (that is, number of distinct Gaussian distributions), we used a parametric bootstrap approach to sequentially test mixtures with 1 to 5 components. We used the implementation of the parametric bootstrap in the `mixtool` R package (Benaglia et al. 2010) For each successive pair of components (k and $k+1$ components), we conducted a parametric bootstrap by generating 100 bootstrap samples from the null model (the model with k components) and fitted both the null and

alternate model (i.e., the model with $k+1$ components) for each bootstrap sample to calculate a distribution of the likelihood ratio statistic under the null hypothesis. This statistic is used to test the null hypothesis of a k component fit against the alternative hypothesis of a $k+1$ component fit across different mixture models. A p-value (with a 5% significance threshold) is used as a decision rule for selecting the optimal number of components. Once the p-value exceeds the significance threshold, the testing terminates and the null model for the test where the procedure terminates is chosen as the number of components (Benaglia et al. 2010).

We also modelled log fold change; representing our measure of effect size, as a finite mixture of Gaussian distributions. Because fold change is typically related to mean abundance (Love et al. 2014), we model the dependency of mean abundance and effect sizes. We modelled the mean and standard deviation parameters of the individual Gaussian components as functions of log mean abundance. In order to determine an appropriate way to model log fold change as a function of log mean abundance, we examined the relationship between log mean abundance and log fold change for each data. Fig 2.1 shows the relationship between log mean abundance and log fold change for three of the microbiome data sets.

The smooth line representing the mean of log fold change as a function of log mean abundance (a loess curve, i.e. a locally quadratic regression) appears to follow a linear trend in some cases, such as in the rightmost panel of Fig 2.1. To allow for this possibility (even though it may be relevant only in some cases), we modelled the mean parameter for each Gaussian component as a linear function of log mean abundance. Consequently, the overall mean of the mixture distribution is also a linear function of log mean abundance.

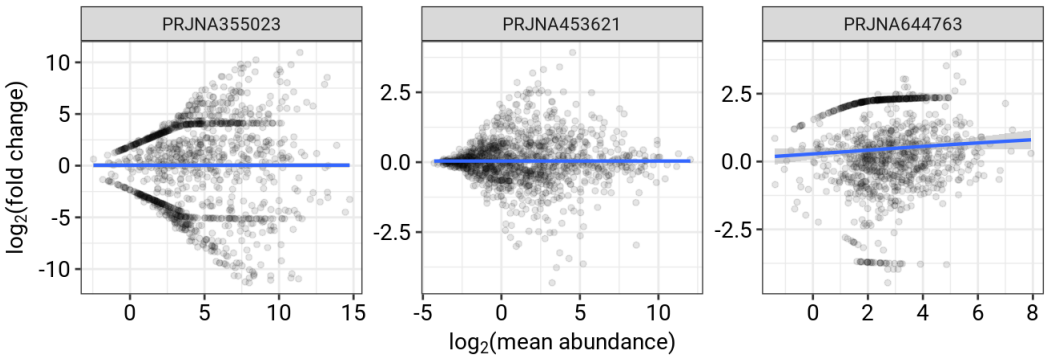

FIGURE 2.1: Relationship between log fold changes and log mean abundance for three typical data sets. The unusual features in the plot (concentrations of points along symmetric curves above and below zero) in the first two panels correspond to taxa with zero counts across all subjects in either the control or the treatment group.

Upon examining variations of log fold change around the smooth line, we observed either a linear or quadratic trend (refer to Fig. 2.8 in the scale-location plot (see *Sup.* 2.9). We therefore modelled the variance of each Gaussian component as both linear

and quadratic functions of log mean abundance. We compared Gaussian mixtures with 1-5 components. For a given model (Gaussian mixture model with a specified number of components), we modelled the variance parameter of all components either by a linear or quadratic function. We selected the model that yielded the minimum Akaike Information Criteria (AIC) value across all the fitted components. The model of log fold change as a function of log mean abundance is:

$$\begin{aligned} y_j &\sim \sum_{i=1}^{K} \pi_i \mathcal{N}(y_j \mid \mu_i(x_j), \sigma_i(x_j)) \\ \mu_i(x_j) &= m_i^0 + m_i^1 x_j \\ \sigma_i(x_j) &= \exp(f(x_j)) \\ \pi_i &= \frac{\exp(\lambda_i)}{1 + \sum_i \exp(\lambda_i)}; \quad \sum_i^K \pi_i = 1, \lambda_1 = 0, \end{aligned} \tag{2.2}$$

where K is the number of Gaussian components. μ_i and σ_i are the mean and standard deviation of the i^{th} component, conditional on the log mean abundance of the j^{th} taxa (x_j). y_j is the log fold change of the j^{th} taxa. The function f denotes a linear or quadratic function of log mean abundance used to model the standard deviation parameter and π_i is the mixture probability with parameter λ_i.

2.4.4 Modelling Dispersion

We used the `DESeq2` package to estimate dispersion for the negative binomial model. Dispersion typically varies based on count abundance, with rarer taxa exhibiting higher dispersion (Love et al. 2014). To accommodate this variability and to simulate dispersion for subsequent power analyses, we used a nonlinear function of mean abundance to model the dispersion estimates, as implemented in the `DESeq2` package:

$$d = c_0 + \frac{c_1}{m}, \tag{2.3}$$

where d and m denote the dispersion and mean abundance respectively. The term c_0 represents the asymptotic dispersion level for high abundance taxa, and c_1 captures additional dispersion variability.

The dispersion estimates from the `DESeq2` package were unrealistically high, for example, ranging from 150 to 200 (see Fig. 2.9 under Supplementary Materials- *Sup.* 2.9). Using these dispersion estimates, we simulated count data from a negative binomial model with mean abundance from the microbiome data set and log fold change estimates from the `DESeq2` package. The variability in the coefficients of variation of taxa abundance computed from the dispersion estimate was notably greater than observed in the actual data set (see Fig 2.10 under supplementary materials). We therefore scaled

the dispersion to align the coefficients of variation from the simulated data more closely with those from the observed data sets. We experimented with different scaling values in the interval $(0, 1)$ to make the distribution of the coefficient of variation from the observed data roughly match those of the simulated data. We found that a scale parameter of 0.3 made the coefficient of variation align much better with the true coefficient of variation. With a scale factor of 0.3, the observed distribution of mean counts and variances from the simulations closely matched the true distributions of mean counts and taxa variances (refer to Figs 2.11 and 2.12 under Supplementary Materials- *Sup.* 2.9)).

2.4.5 Procedure for Simulating Microbiome Count Data

We simulated count data to assess how well MixGaussSim replicates each of the autism data sets described in Section 2.3. The simulation parameters for each data set were estimated by fitting the MixGaussSim model to the corresponding real autism data.

The following steps outline our procedure for simulating microbiome count data.

- **Simulate overall log mean abundance:** For each data set, we simulated log mean abundance from the fitted Gaussian mixture distributions.
- **Simulate log fold changes:** Using the simulated log mean abundance, we simulated log fold change from the fitted Gaussian mixture distributions (equation (2.2)).
- **Predict dispersion values:** Next, we predicted dispersion values as a function of the simulated mean abundance from the fitted non-linear function (equation (2.3)).
- **Calculate per-group mean abundance:** We then calculated mean abundance for control and treatment groups using the simulated mean abundances and the simulated log fold changes.
- **Simulate count data:** Using the calculated mean abundance for control and treatment groups and the predicted dispersion values, we simulated count abundances from the negative binomial distribution (equation (2.1)).

2.5 The Reduced Rank Simulation Approach (RRSim)

2.5.1 General Model Description

The Negative Binomial Mixed Model

Let $\boldsymbol{Y}$ denote a $m \times n$ ASV table, with rows $i = 1, \ldots, m$ and columns $j = 1, \ldots, n$, where m and n represent number of subjects and taxa respectively. Let $\boldsymbol{x}_i$ denote a d-dimensional vector of covariates for each subject, and $\boldsymbol{y_i} = (y_{i1}, \ldots, y_{in})'$ represent a $1 \times n$ vector of count abundance for subject i. Define $\bar{\boldsymbol{y}} = (\boldsymbol{y_1}, \boldsymbol{y_2}, \ldots, \boldsymbol{y_n})'$ as the long-format concatenation of all the individual count vectors across the m subjects. Then

$\bar{y}_k \in \bar{\boldsymbol{y}}$, $\{k = 1, ..., mn\}$ follows a negative binomial distribution:

$$\bar{y}_k \sim \text{NegBin}(\mu_k, \theta_k),$$

where μ_k and θ_k are the expected mean counts and dispersion for $\bar{y}_k$, respectively.

We model $\boldsymbol{\mu} = (\mu_1, ..., \mu_{mn})$, the vector of expected mean counts, by a mixed model defined as follows:

$$\begin{aligned} \boldsymbol{g}(\boldsymbol{\mu}) &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}, \\ \boldsymbol{b} &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}), \end{aligned}$$

where $\boldsymbol{X} \in \mathbb{R}^{mn \times d}$ and $\boldsymbol{Z} \in \mathbb{R}^{mn \times q}$ are the fixed and random effect model matrices respectively, and $\boldsymbol{\beta}$ is a d-dimensional vector of fixed effect coefficients. $\boldsymbol{b}$ is a q-dimensional random effect vector. The unconditional distribution of $\boldsymbol{b}$ is assumed to come from a multivariate normal distribution with variance-covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{mn \times mn}$. $\boldsymbol{g}$ is the log link function.

2.5.2 Specific Model Description

The Reduced Rank Mixed Effect Model (RRMM)

In microbiome data, taxa exhibit varying levels of abundance. We account for this variation by including a taxon-specific random intercept, which captures differences in baseline counts across taxa. The effect of group conditions (e.g., treatment vs. control) on abundance varies across taxa. For example, one taxon may show a significant increase in counts in the treatment group compared to the control, while another taxon might exhibit a smaller increase or even a decrease. We therefore included a taxon-specific random slope for the group effect to capture differences in how each taxon responds to the group condition. We allow the random intercept and the random group effect terms to be correlated to account for potential correlations between random intercept (baseline count) and group (treatment) effects.

While incorporating all taxa into a single mixed model increases computational demands, it enables the modeling of correlations and information sharing between taxa. Including the random intercept and slope terms allows the model to shrink estimates for individual taxa towards the overall population trends, balancing individual variability with shared patterns across the taxa. To account for correlations between taxa within subjects, we include a taxon-specific random effect that varies within subjects. This formulation allows each taxon to have a subject-specific deviation, capturing within-subject variability while also modeling potential correlations between taxa within the same subject.

We do not include fixed effect terms for taxon or group (control vs. treatment). Estimating taxon-specific fixed effects could lead to overfitting, especially when the number

of taxa is large. Instead, we model the effect of taxon on count as a random effect, which captures variability across taxa while avoiding the need to estimate an excessive number of parameters. Including a fixed effect for group (that is, a main effect for group) would measure the degree to which all taxa consistently increase or decrease across all subjects in one group compared to another group. However, this assumption does not accurately reflect the nature of microbiome data. Researchers typically treat microbiome data as compositional, meaning the changes in abundance of each taxon are measured in relation to the abundance of others. If one taxon increases, others must decrease. In practice, this compositionality is typically handled by normalizing the data to adjust for differences in overall abundance, as caused for example by differences in sequencing depth between samples. Thus, the normalization and compositional nature of microbiome data make the assumption of a consistent increase or decrease in all taxa abundance across all subjects in a group inappropriate for microbiome data.

Consider a study involving two groups, in this case "control" and "treatment". We model $\boldsymbol{\mu}$ as follows:

$$\begin{cases} \log(\boldsymbol{\mu}) = \boldsymbol{o} + \boldsymbol{Z_1 b_1} + \boldsymbol{Z_2 b_2} \\ \quad \boldsymbol{b_1} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma_1}), \\ \quad \boldsymbol{b_2} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma_2}) \end{cases} \tag{2.4}$$

where $\boldsymbol{o}$ is the offset term to account for differences in sequencing depth (read depth) between samples. $\boldsymbol{Z_1} \in \mathbb{R}^{mn \times 2n}$ and $\boldsymbol{Z_2} \in \mathbb{R}^{mn \times mn}$ are the model matrices for the random effects. The variance-covariance matrices $\boldsymbol{\Sigma_1} \in \mathbb{R}^{2n \times 2n}$ and $\boldsymbol{\Sigma_2} \in \mathbb{R}^{mn \times mn}$ are block diagonal matrices. Each block matrix in $\boldsymbol{\Sigma_1}$ models correlation between the random intercept and random group effect for each taxon as well as the variances of intercept and slope. The block matrices of $\boldsymbol{\Sigma_2}$ model correlations among taxa within each subject. Consequently, we define the variance-covariance matrices $\boldsymbol{\Sigma_1}$ and $\boldsymbol{\Sigma_2}$ as follows:

$$\begin{cases} \boldsymbol{\Sigma_1} = \boldsymbol{\sigma_1^*} \otimes \boldsymbol{I}_{2n}, \\ \boldsymbol{\Sigma_2} = \boldsymbol{\sigma_2^*} \otimes \boldsymbol{I}_{mn}, \end{cases} \tag{2.5}$$

where $\boldsymbol{\sigma_1^*} \in \mathbb{R}^{2\times 2}$ models correlations between the random intercept and group effects for each taxon, $\boldsymbol{\sigma_2^*} \in \mathbb{R}^{n\times n}$ models correlations between taxa within each subject, and $\otimes$ denotes the Kronecker product. $\boldsymbol{I}_{2n}$ and $\boldsymbol{I}_{mn}$ are identity matrices. From equation (2.5), $\boldsymbol{\Sigma_1}$ and $\boldsymbol{\Sigma_2}$ are block-diagonal covariance matrices with homogeneous blocks.

Since microbiome data sets typically contain hundreds or thousands of taxa, estimating $\boldsymbol{\sigma_2^*}$ is impractical as it requires estimating $n(n+1)/2$ parameters—potentially hundreds or even millions of parameters, unless we impose constraints. To overcome this challenge, we apply a rank reduction approach (also known as the factor analytic method) (McGillycuddy et al. 2025), which reduces the rank of $\boldsymbol{\sigma_2^*}$ to $d \ll n$. The vectors $\boldsymbol{b_1}$ and $\boldsymbol{b_2}$ are random effects with unconditional distributions assumed to be

multivariate Gaussian with zero mean and variance covariance matrices defined by $\boldsymbol{\Sigma_1}$ and $\boldsymbol{\Sigma_2}$, respectively. $\boldsymbol{b_1}$ and $\boldsymbol{b_2}$ are block vectors, each block representing the random effects associated with a specific grouping level. Thus, we define $\boldsymbol{b_1}$ and $\boldsymbol{b_2}$ as follows:

$$\begin{cases} \boldsymbol{b_1} = \begin{bmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \\ \vdots \\ \boldsymbol{\alpha}_n \end{bmatrix}, \quad b_2 = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_m \end{bmatrix} \\ \boldsymbol{\alpha}_j \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma_1^*}), \quad \boldsymbol{\beta}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma_2^*}), \end{cases} \tag{2.6}$$

where $\boldsymbol{\alpha}_j$ (for $j = 1, \ldots, n$) is a vector of length 2 whose entries are the random intercept and random group effect for taxon j, and $\boldsymbol{\beta}_i$ (for $i = 1, \ldots, m$) is a vector of length n whose entries are the taxon-specific effects for subject j. Both $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_i$ are assumed to follow the standard normal distributions with variance-covariance matrices $\boldsymbol{\sigma_1^*}$ and $\boldsymbol{\sigma_2^*}$ respectively.

The Reduced Rank Method

The reduced rank model of dimension d expresses $\boldsymbol{\beta}_i$ as a linear combination of latent variables:

$$\boldsymbol{\beta}_i = \boldsymbol{\Lambda} \boldsymbol{u}_i \tag{2.7}$$

where $\boldsymbol{u}_i$ is a vector of d latent variables, and $\boldsymbol{\Lambda}$ is an $n \times d$ matrix of factor loadings. The latent variables, also referred to as spherical latent variables (Bates 2014), are assumed to come from a multivariate standard normal distribution:

$$\boldsymbol{u}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I_d}), \tag{2.8}$$

where $\boldsymbol{I_d}$ is the identity matrix. Thus, we have:

$$\boldsymbol{\beta}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top) \tag{2.9}$$

$\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top$ is the reduced rank approximation of $\boldsymbol{\sigma_2^*}$. The estimation procedure for both $\boldsymbol{\Lambda}$ and $\boldsymbol{u}_i$, as implemented in the `glmmTMB` R package, is described in detail by McGillycuddy et al. 2025.

Thus the specific model used in this paper is defined as follows:

$$\begin{cases} \log(\boldsymbol{\mu}) = \boldsymbol{o} + \mathbf{1} + \boldsymbol{Z_1 b_1} + \boldsymbol{Z_2 b_2} \\ \quad \boldsymbol{b_1} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma_1}), \ \boldsymbol{b_2} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma_2}) \\ \quad \boldsymbol{\Sigma_1} = \boldsymbol{\sigma_1^*} \otimes \boldsymbol{I}_{2n}, \\ \quad \boldsymbol{\Sigma_2} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top \otimes \boldsymbol{I}_{mn}, \end{cases} \tag{2.10}$$

where $\boldsymbol{o}$ is an offset accounting for library size normalization, and $\mathbf{1}$ represents an intercept-only fixed effect. The term $\boldsymbol{Z_1 b_1}$ includes taxon-specific random intercepts and slopes for the group-level covariate (e.g., treatment vs. control), with $\boldsymbol{Z_1}$ indicating the design matrix linking observations to these random effects. The term $\boldsymbol{Z_2 b_2}$ models subject-specific latent effects with a low-rank covariance structure. The Kronecker product $\boldsymbol{\Sigma_2} = \boldsymbol{\Lambda\Lambda}^\top \otimes \boldsymbol{I}_n$ captures correlation among taxa within each subject.

Since microbiome data exhibits zero-inflation (Xia et al. 2018; Zhang et al. 2020), we incorporate a zero-inflation component in our model to account for excess zeros in each taxon. Zero-inflation could be modelled by incorporating both an overall zero-inflation term and taxon-specific deviations to account for varying excess zeros across taxa. However, we assume a single zero-inflation probability shared across all taxa for simplicity.

2.5.3 Microbiome count simulation procedure

To simulate count data from the model described in equation (2.10), we used the `simulate_new` function in the `glmmTMB` *R* package. The `simulate_new` function allows us to set up a simulation directly, specifying the right hand side of the model formula described in Listing 2.1, a data frame describing the experimental setup (e.g., number of taxa, number of subjects, group names of subjects), and a list of simulation parameters. The function requires input values for the standard deviations and correlations of each random effect, as well as intercept terms (that is, the overall average count, specific-taxon effect and average zero-inflation probability). To ensure that the input parameter values are realistic for microbiome data, we estimated them using actual microbiome data sets described in Section 2.3.

The simulation model (equation (2.10)) is specified in `glmmTMB` as follows:

```
simulate_new( ~ 1 + us(1 + group | taxon) +
                    rr(0 + taxon | subject, d),
                    ziformula = ~1,
                    data = data,
                    family = nbinom2)
```

LISTING 2.1: Example code for glmmTMB

where `us()` and `rr()` are functions for the unstructured and reduced-rank variance-covariance matrices for the random effects respectively, and `d` specifies the rank of the reduced rank matrix. `nbinom2` denotes a negative binomial conditional distribution.

2.6 Goodness of Fit Test

We assess the performance of MixGaussSim, RRSim, and two simulation models: `metaSPARSim` (Patuzzi et al. 2024) and `HMP` (La Rosa et al. 2012), by examining how well these simulation models reproduce the distribution of mean abundance of taxa, the distribution

of variance of taxa, and the distribution of the proportion of zero counts across samples for the microbiome data sets described in Section 2.3. For each of these criteria we conducted a Kolmogorov-Smirnov (KS) goodness-of-fit test (Lopes et al. 2007; Kotz et al. 2005) to identify the best simulation method. The Kolmogorov-Smirnov (KS) test is a non-parametric method that compares empirical cumulative distribution functions (ECDFs). The KS test measures the maximum absolute difference between two ECDFs:

$$D_n = \sup_x |F_1(x) - F_2(x)| \tag{2.11}$$

where $F_1(x)$ and $F_2(x)$ are the ECDFs of the real and simulated data sets, respectively. The KS statistic (D_n) quantifies this difference, and is associated with a p-value for determining whether the observed divergence is statistically significant.

A small KS statistic and a high p-value suggest that the simulated and real distributions match closely, while a large KS statistic and a low p-value indicate significant differences. This approach allows us to systematically compare simulation methods and determine which best reproduces real microbiome data.

2.7 Results and Discussion

We compare our simulation approaches (MixGaussSim described in Section 2.4 and RRSim described in section 2.5) with two existing simulation approaches implemented in the `metaSPARSim` (Patuzzi et al. 2024) and `HMP` (La Rosa et al. 2012) `R` packages.

The `HMP` package simulates microbiome data from a Dirichlet multinomial model and the `metaSPARSim` package simulates microbiome data from a Multivariate hypergeometric model. `metaSPARSim` models variation in taxa abundances between biological samples using a Gamma distribution and models technical variability introduced by the sequencing process using a multivariate hypergeometric model. The `HMP` and `metaSPARSim` packages both model taxa jointly, capturing correlations between them. The `HMP` package introduces negative correlations and accounts for compositionality through the Dirichlet distribution. In contrast, `metaSPARSim` accounts for the compositional nature of microbiome data using the multivariate hypergeometric distribution and includes a parameter specifically designed to introduce sparsity into the data. Both models also capture overdispersion in microbiome data.

We show results of the comparison for four of the autism microbiome data sets described in Section 2.3. To ensure a fair comparison, the simulation parameters for each model were estimated from the corresponding real microbiome data set before data generation.

For each simulation method, we generated count data using parameters estimated from the real microbiome data sets described in Section 2.3. The simulations were performed using the same number of taxa and sample sizes as in the corresponding real

data sets. Figures 2.4 and 2.3 present comparisons of the distributions of mean abundance and variance of taxa from four of the microbiome data sets against those obtained from simulations using `HMP`, `metaSPARSim`, and our proposed methods, `MixGaussSim` and `RRSim`.

The distribution of mean abundance of taxa from simulations using `metaSPARSim`, MixGaussSim, and RRSim matches that of actual microbiome data sets (Fig. 2.4). In contrast, simulations generated with the `HMP` package fail to accurately replicate the distribution of mean abundance of taxa. Goodness-of-fit test results indicate that `metaSPARSim` is the most effective simulator, consistently yielding the lowest KS statistics and highest p-values for all four data sets (Fig. 2.5). MixGaussSim ranks as the second best simulator for all data sets except for data set PRJNA589343, where RRSim ranks as the second best simulator having the second lowest KS statistic and the second highest p-value. In contrast, `HMP` is the worst performing simulator, showing the highest KS statistics and the lowest p values in all data sets.

A comparison of the distribution of variance of taxa shows that RRSim has the best performance in replicating the distribution of variance of taxa for two of the data sets (ie. PRJNA589343, PRJNA687773), with the lowest KS statistics and highest p-value for these data sets (Fig. 2.6). On the other hand, the KS goodness of test shows that MixGaussSim and `metaSPARSim` are the best performing simulators in replicating the distribution of variance of taxa for data sets PRJNA168470 and PRJNA355023 respectively (Fig. 2.6). In contrast, the `HMP` simulator fails to replicate the distribution of the variance of taxa compared with the other simulators (Fig. 2.3).

A comparison of the distribution of the proportion of zeros across samples indicates that RRSim performs best in replicating this distribution across all data sets, except for the PRJNA589343 data set, where `HMP` achieves the lowest KS statistic and highest p-value (Fig. 2.7). Across all data sets, `HMP` consistently ranks as the second-best simulator, as evidenced by its second-lowest KS statistic and second-highest p-value (Fig. 2.7). In contrast, `metaSPARSim` has the worst performance across all four data sets (Figs. 4.4 and 2.7).

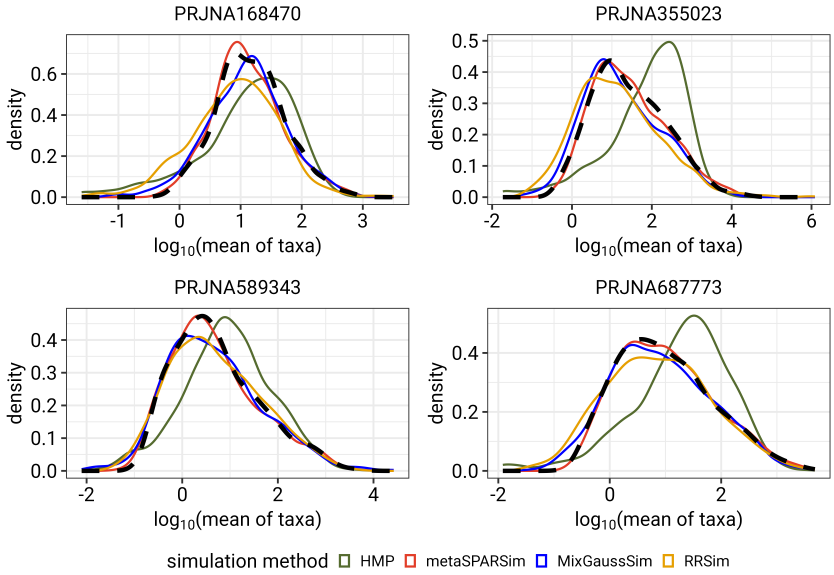

FIGURE 2.2: Comparison of distributions of mean abundance of taxa between observed data and simulations generated from `HMP`, `metaSPARSim`, `MixGaussSim` and `RRSim`. Black dashed lines represent the distribution of mean abundance of taxa for the microbiome data set.

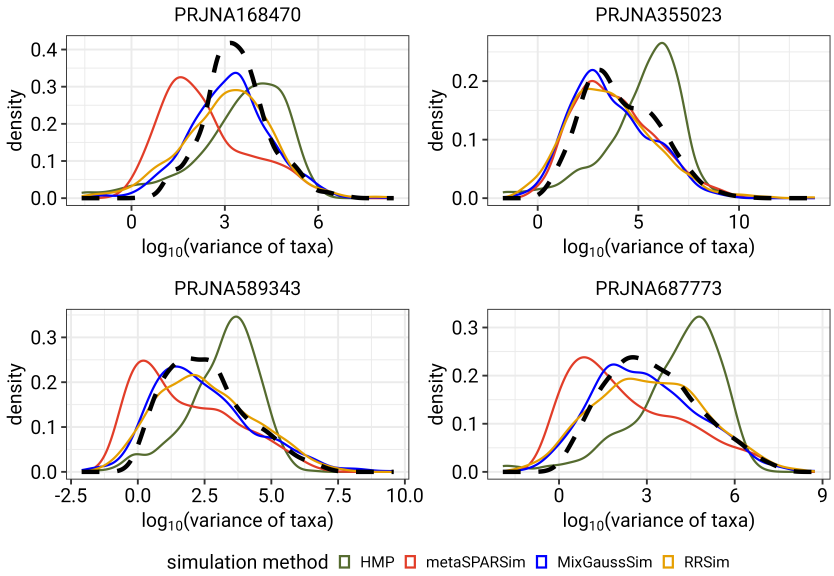

FIGURE 2.3: Comparison of the distributions of variance of taxa between observed data and simulation generated from `HMP`, `metaSPARSim`, `MixGaussSim` and `RRSim`. Black dashed lines represent the distribution of variance of taxa for the microbiome data set.

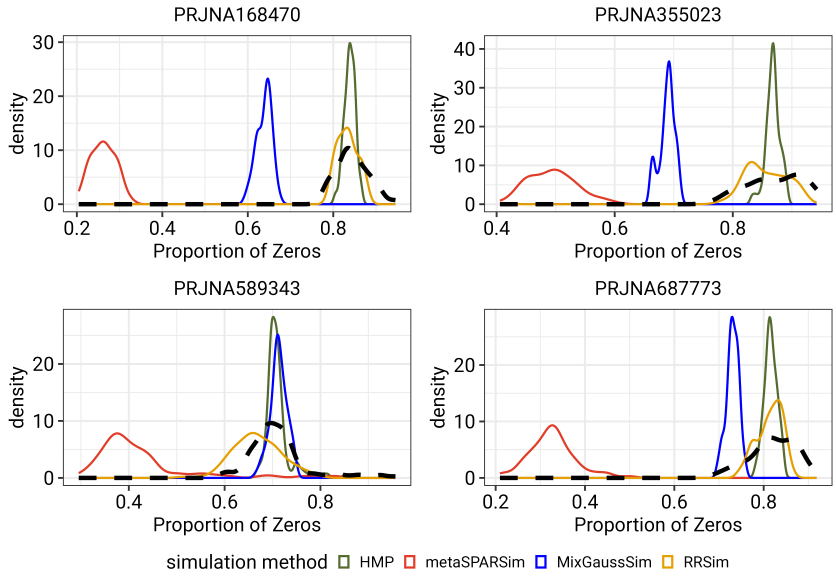

FIGURE 2.4: Comparison of distribution of proportion of zeros across samples between observed data and simulations generated from `HMP`, `metaSPARSim`, `MixGaussSim` and `RRSim`. Black dashed lines represent the distribution of proportion of zeros across sample for the microbiome data set.

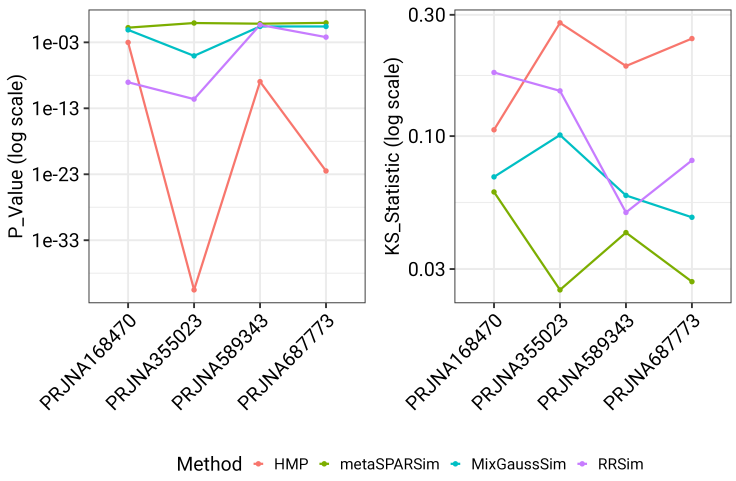

FIGURE 2.5: Comparison of KS Statistics and p-values estimates from goodness of fit test performed on the distribution of mean abundances of taxa for four data sets. metaSPARSim is the best model (with the owest KS statistic and the largest p-value)

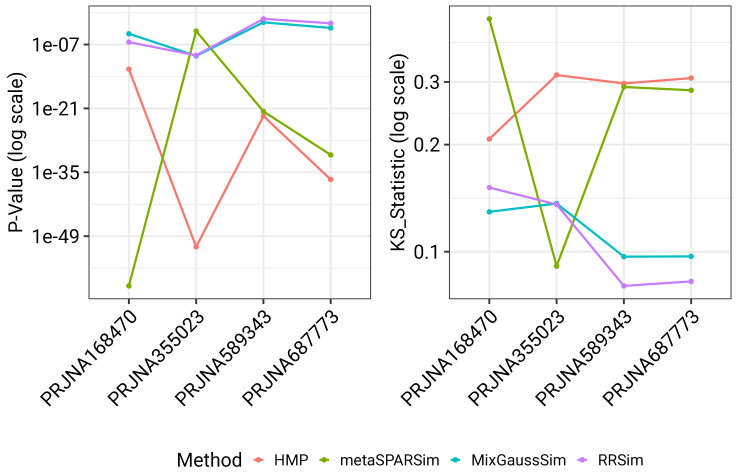

FIGURE 2.6: Comparison of KS Statistics and p-value estimates from goodness of fit test performed on the distribution of variance of taxa across four data sets.

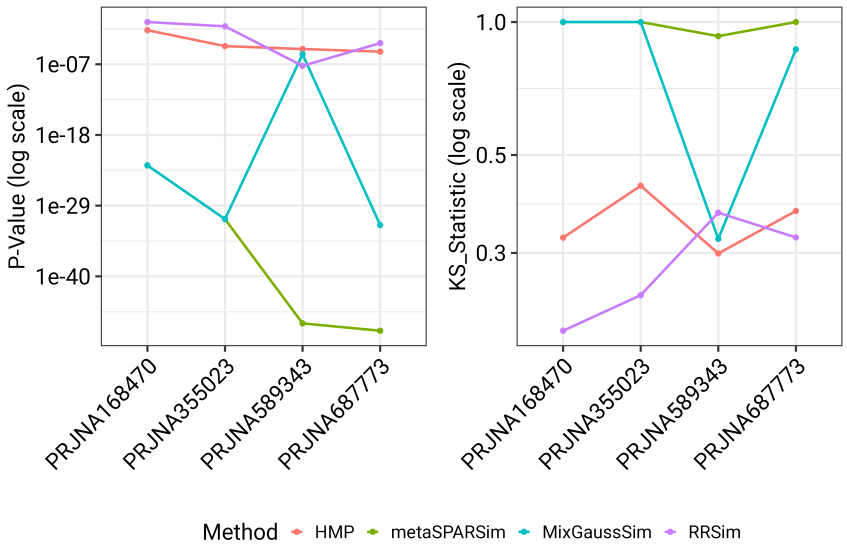

FIGURE 2.7: Comparison of KS Statistics and p-value from goodness of fit test performed on the distribution of the proportion of zero counts across samples for four data sets.

2.8 Conclusion

This chapter introduces two novel simulation approaches for microbiome count data: MixGaussSim and RRSim. MixGaussSim simulates microbiome data using a negative binomial distribution and captures both the distribution of mean abundances and effect sizes of taxa. A key advantage of this approach is its ability to estimate effect size distributions, which are crucial for statistical power calculations at the taxon level. Since microbiome studies rarely report effect sizes, MixGaussSim provides a practical framework for inferring these distributions using parameter estimates from real microbiome data.

In contrast, RRSim generates microbiome count data using a negative binomial distribution while modeling taxa mean abundances as a mixed-effects model. RRSim jointly models all taxa and accounts for correlations between taxa within individuals. It achieves this by using reduced-rank dimension reduction, which decreases the number of parameters required to quantify the variance-covariance matrix. RRSim offers flexibility in modeling zero inflation, allowing for more complex structures. For example, it can estimate zero inflation at the level of individual taxa as a random effect or account for zero inflation within specific groups of taxa.

We evaluated the performance of MixGaussSim and RRSim against two existing simulators from the `HMP` and `metaSPARSim` R packages. Results from the Kolmogorov-Smirnov (KS) goodness-of-fit test showed that `metaSPARSim` best captured the distribution of mean abundance of taxa, while MixGaussSim and RRSim ranked second or third across data sets. However, RRSim outperformed all other methods in replicating the variance distribution for two out of four data sets, while MixGaussSim and `metaSPARSim` were best for the remaining two. RRSim ranked as the best simulator in replicating the distribution of proportions of zeros counts across samples for three of four microbiome data sets. In contrast, `HMP` consistently failed to accurately reproduce both mean abundance and variance distributions but is effective in reproducing the distributions of the proportions of zero counts across samples. Overall, MixGaussSim and RRSim provide researchers with flexible and effective tools for simulating microbiome count data, each offering distinct advantages.

2.9 Supplementary Materials

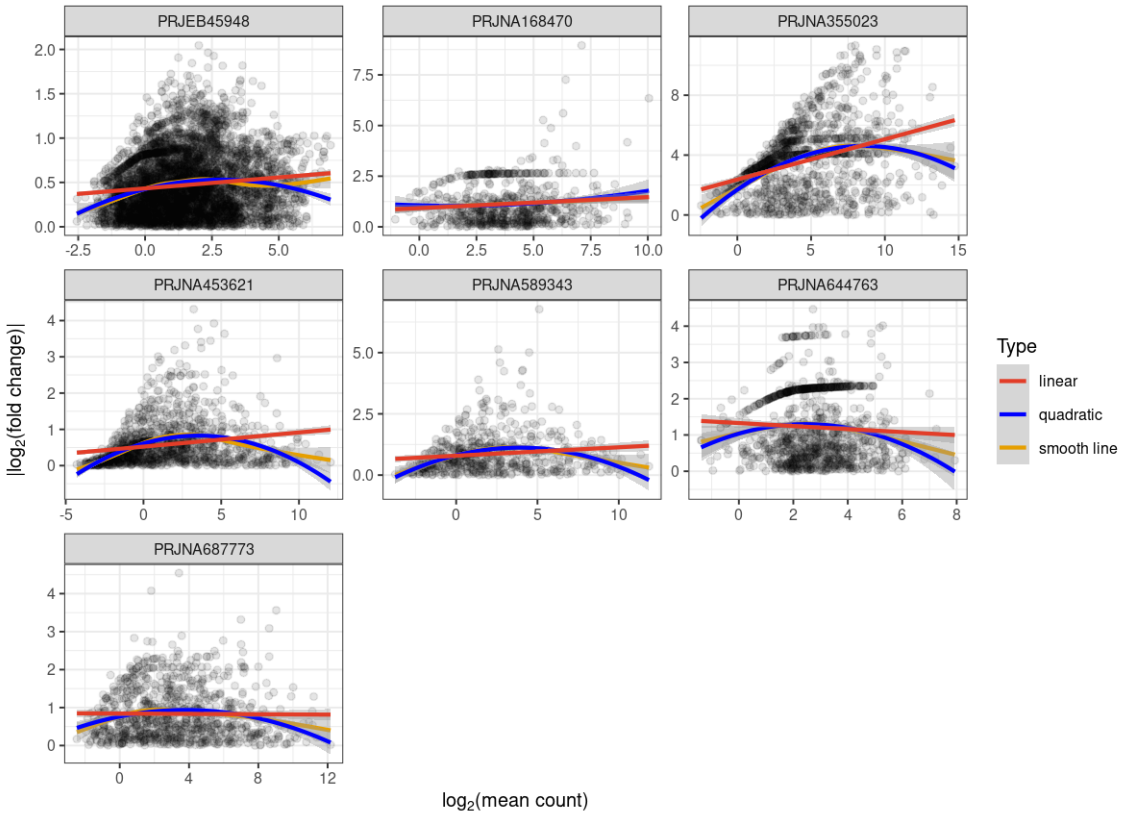

FIGURE 2.8: Scale-location plots to determine functions to model the standard deviation parameter

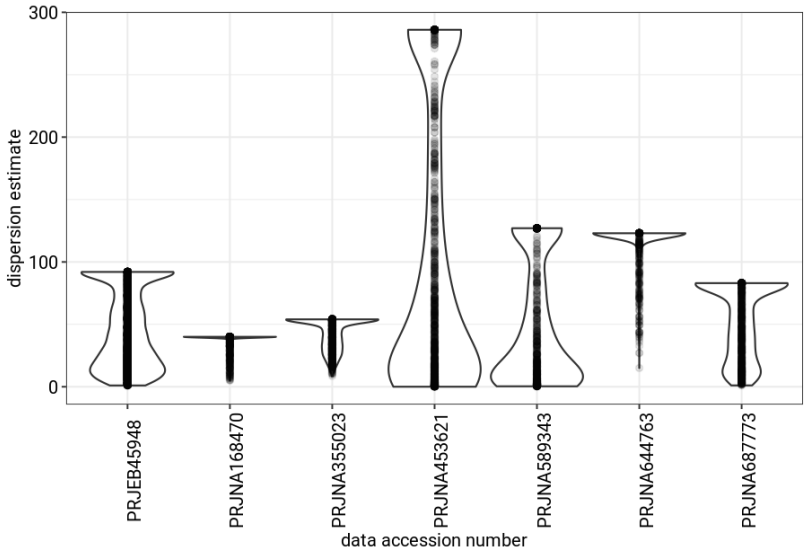

FIGURE 2.9: Distribution of dispersion estimates from `DESeq2` R package

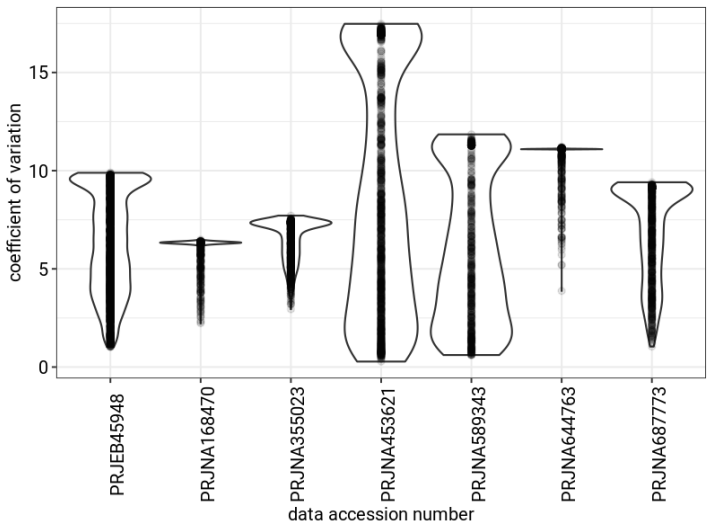

FIGURE 2.10: Coefficient of variation of taxa abundance using dispersion estimates from `DESeq2` R package

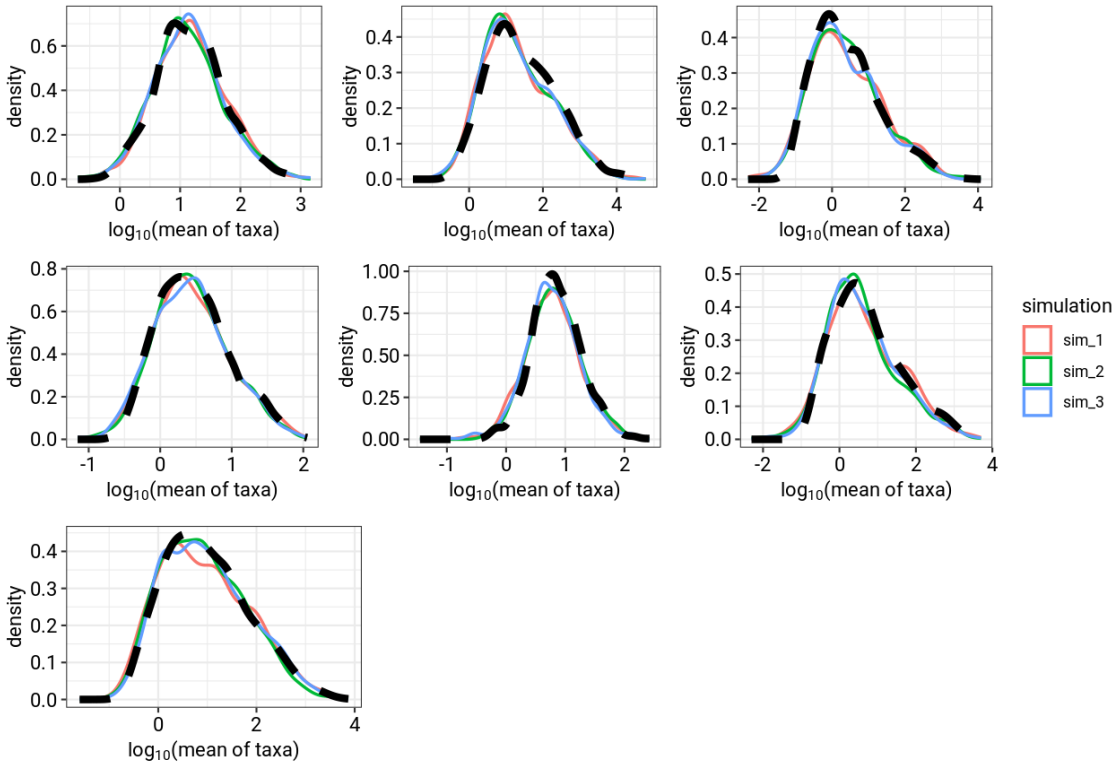

FIGURE 2.11: Comparison of mean abundance distributions between simulated and observed taxa data. The distribution of observed data is shown with a black dashed line. Simulation parameters: 1000 OTUs, 100 samples per group, and a dispersion scale value of 0.3.

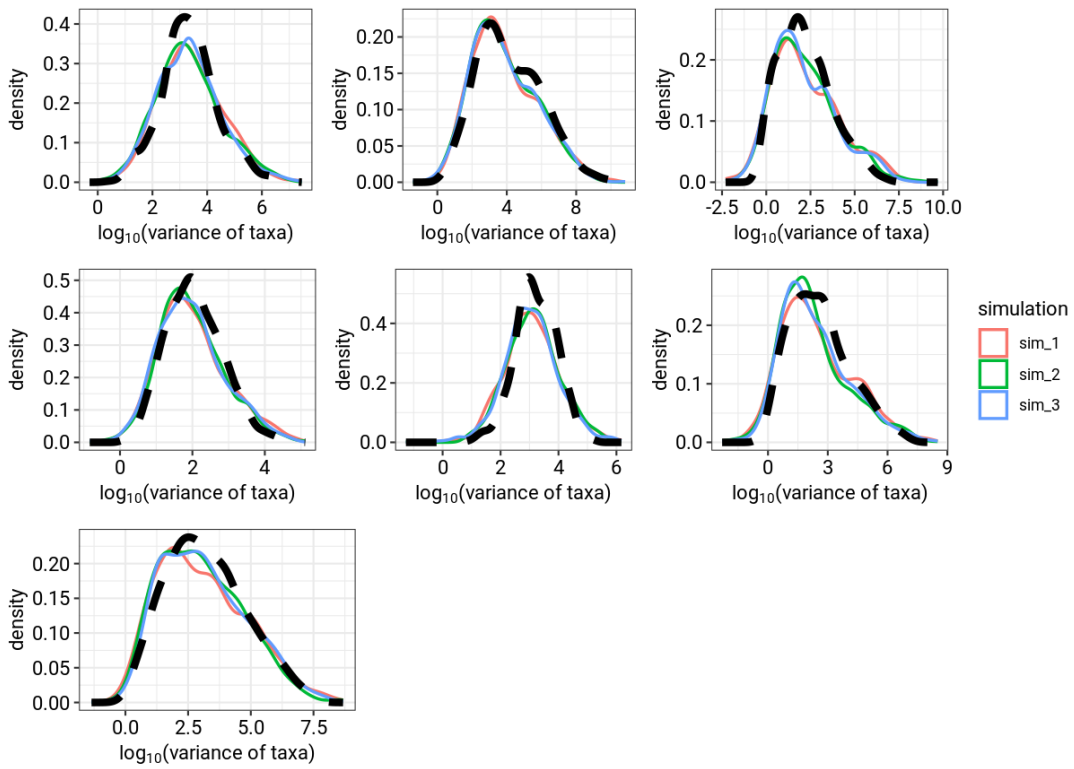

FIGURE 2.12: Comparison of distributions of variance of taxa between simulated and observed taxa data. The distribution of observed data is shown with a black dashed line. Simulation parameters: 1000 OTUs, 100 samples per group, and a dispersion scale value of 0.3.

Chapter 3

Investigating Statistical Power of Differential Abundance Studies

The descriptions in this chapter are adapted from a paper published in the PLOS ONE journal (Agronah and Bolker 2025) based on the research presented in this chapter.

3.1 Abstract

Identifying microbial taxa that differ in abundance between groups (control/treatment, healthy/diseased, etc.) is important for both basic and applied science. As in all scientific research, microbiome studies must have good statistical power to detect taxa with substantially different abundance between treatments; low power leads to poor precision and biased effect size estimates (Hawinkel et al. 2019). Several studies have raised concerns about low power in microbiome studies (Kers and Saccenti 2021). In this study, we investigate statistical power in differential abundance analysis. In particular, we present a novel approach for estimating the statistical power to detect effects at the level of individual taxa as a function of effect size (fold change) and mean abundance. We illustrate how power varies with effect size and mean abundance; our results suggest that typical differential abundance studies are underpowered for detecting changes in individual taxa.

3.2 Introduction

Identifying taxa that show differential abundance between groups holds great potential for clinical applications (Nearing et al. 2022). For example, a study aimed at assessing the effects of a dietary intervention on microbial composition might analyze the abundance of different microbial taxa between a control group on a standard diet and a treatment group on a gut-health-promoting regimen.

Power analysis allows researchers to determine whether they have a sufficient sample size to detect meaningful effects in their studies. The power of a statistical test is the probability of successfully rejecting the null hypothesis given a particular effect size (Cohen 2013). Power is determined by the sample size, effect size and the significance

threshold (or "alpha level"), as well as methodological factors such as experimental design, number of groups, statistical procedure and model, type of response variable, fraction of missing data and the number of hypotheses tested.

Power analysis enables researchers to detect meaningful effects and allocate resources efficiently; it aids the reliability and reproducibility of research findings. The primary goal of power analysis is to ensure that a research study has the sensitivity required to detect meaningful effects (Cohen 2013). Underpowered studies are likely to miss biologically meaningful effects and are more prone to type II errors, which can lead researchers to neglect differences that could be biologically interesting (Goodman and Berlin 1994). Even if a low-powered study finds statistically significant results, the estimated effect size will be imprecise (Goodman and Berlin 1994). Low power together with a statistical significance filter (for example, only reporting effects with a p-value < 0.05) can lead to overestimation of the true effect ("magnitude", or type M, error) or an incorrect estimate of the direction of an effect ("sign", or type S, error) (Gelman and Carlin 2014).

Microbiome researchers typically focus on three main types of analysis: (1) *analysis of univariate summaries*: reducing the data from each microbiome sample to a single value, such as alpha diversity, and comparing the distribution of these values between groups (Kers and Saccenti 2021), (2) *community-wide analyses* using tests such as Permutational Multivariate Analysis of Variance (PERMANOVA) or the Dirichlet-multinomial model to distinguish overall differences in communities (Kelly et al. 2015; La Rosa et al. 2012), and (3) taxon-by-taxon or *differential abundance* analyses: identifying taxa with biological meaningfully differences between groups (Liu et al. 2021). Existing studies on power analysis have focused either on studies comparing univariate (alpha diversity) measures or studies comparing changes in overall microbiome composition between groups (Xia et al. 2018). For example, La Rosa et al. 2012 developed a reparameterized Dirichlet Multinomial model and a method for estimating the power to detect changes in overall microbial composition between groups. Kelly et al. 2015 proposed a framework for estimating power in PERMANOVA.

To our knowledge, no methods exist for power analysis for differential abundance studies. In practice, every taxon in a microbial community has a different mean abundance and a different effect size (as is typical, we use fold change between groups as effect size in this paper), leading to a different statistical power to detect differences in every taxon. Except for relatively simple analyses, conducting power analysis requires data simulation. Simulating an entire microbial community is challenging because it requires estimating appropriate community-wide distributions for mean abundances and effect sizes of taxa.

Power estimates in a differential abundance study depends on the abundance of individual taxa. For example, effect sizes of taxa with high abundance in both control and treatment groups are more likely to be detected compared to effect sizes of taxa that are rare in both groups. Unlike univariate power analysis (that is, power analysis involving univariate quantities) where one can specify a single value for effect size and power, in

a taxon-by-taxon power analysis there are multiple values of effect size and power; one for each taxon. This typically means hundreds or thousands of effect sizes and power values.

Several studies have raised concerns about low power in microbiome studies (Brüssow 2020; Kers and Saccenti 2021). For example, Kers and Saccenti 2021 showed that microbiome studies comparing alpha and beta diversities (PERMANOVA) between groups might be underpowered. The goal of this study is to investigate the issue of potential low power to detect effect size of individual taxa within a differential microbiome study. We developed a novel method for estimating the statistical power of individual taxa. Our framework estimates statistical power for each taxon as a function of effect size and mean abundance of individual taxa. Using our framework, researchers can estimate the range of statistical power for their studies, estimate statistical power for specific taxa, and determine the expected number of taxa that differ significantly between groups in their studies. Our power estimation method is based on the negative binomial model described in Section 2.4.1.

3.3 Method

Due to the complexity of microbiome data, statistical power for individual taxa cannot be reliably calculated analytically using properties of count distributions. To estimate reliable power for each taxon, we need to simulate data that mimics actual microbiome data (Arnold et al. 2011). We used the MixGaussSim simulation approach described in Section 2.4 under Chapter 2. The MixGaussSim simulation method provides a framework for estimating appropriate distributions for effect size and mean abundance, which are useful for estimating the statistical power of individual taxa. We simulated data sets resembling each of the autism data sets described in Section 2.3 using the same number of taxa and sample sizes as the original data sets (refer to Table 2.1). We estimated the parameters for MixGaussSim from each corresponding autism data set.

3.3.1 A Novel Method for Estimating Statistical Power for Differential Abundance Microbiome Studies

In a taxon-by-taxon analysis, the power associated with a specific taxon is influenced not only by the group sample size but also by its effect size and mean abundance. Effect size and mean abundance of taxa as well as the sample size are positively correlated with statistical power. In practice, the effect size, mean abundance for a given taxon, and the sample size of the group are not independent of each other. For example, all else being equal, the effect size estimate for a taxon tends to be more precise with a larger sample size. Therefore, it is important to account for the interactions between these variables when evaluating statistical power. We model the interactions between each pair of these determinants of power using a smoothing spline, which allows for flexible modeling of the non-linearity in these interactions. Additionally, these interactions must be structured in a way that preserves the positive relationship between effect size, mean

abundance, and power. To ensure this, we impose a constraint that allows the smooth of each interaction term to be monotonically increasing as a function of statistical power.

We fitted a Generalized Additive Model (GAM) because it is well-suited for modeling nonlinear relationships between variables, which is critical in this context where the interactions between effect size, mean abundance, and statistical power are not expected to follow simple linear patterns. GAMs can incorporate smooth terms with constraints, such as ensuring that the relationship between effect size, mean abundance, and power is monotonically increasing.

We estimate statistical power as the probability of rejecting the null hypothesis (that the mean abundance in control and treatment group are the same for a given taxon). We used the `DESeq2` package to compute p-values for each taxon, using the Benjamini and Hochberg method for false discovery rate (FDR) correction. The event that a given taxa is significantly different between groups is a Bernoulli trial. To estimate statistical power for various combinations of log mean abundance and log fold change, we fitted a shape-constrained generalized additive model (GAM) (Pya and Wood 2015). The model predicting fold change as a function of log mean abundance is as follows:

$$\begin{aligned} y &\sim \text{Bernoulli}(p_i) \\ p_i &= \frac{1}{1+e^{-\eta}} \\ \eta &= \beta_0 + f_1(x_1, x_2) + \epsilon, \end{aligned}$$

where y is a binary value (with 1 indicating that the p-value was below a critical value and 0 otherwise). We used the default critical value of 0.1 in the `DESeq2` package. p_i is the statistical power for taxon i. β_0 and ϵ are the intercept and error terms respectively and the predictors x_1 and x_2 are the log mean abundance and log fold change respectively. The function f_1 is a two-dimensional smoothing surface with basis generated by the tensor product smooth of log mean abundance and log fold change.

Power and fold change are positively correlated (Cohen 2013). Additionally, effect sizes of taxa with high abundance are more likely to be detected, hence having higher power, than rare taxa (Love et al. 2014). To account for these relationships, we constrained the function f_1 to be a monotonically increasing function of both log mean abundance and log fold change.

3.3.2 Expected number of taxa with significant difference between groups

Consider a differential abundance study involving n taxa, each associated with power (probability of being significantly different between groups) p_i. Whether we can detect that taxon i differs significantly between groups or not in a particular analysis is a Bernoulli random variable with a success probability p_i. Therefore, the expected number

of significant taxa can be computed as the sum of the expected number of successes in n Bernoulli trials:

$$\gamma = \sum_{i=1}^{n} p_i. \tag{3.1}$$

Equation (3.1) can be divided and multiplied by n to obtain

$$\gamma = n\hat{p}, \tag{3.2}$$

where $\hat{p}$ is the average statistical power across all taxa. Equation (3.2) states that the expected number of taxa that differ significantly between groups in a differential abundance study is the product of the number of taxa (n) and the average statistical power for all taxa ($\hat{p}$).

3.4 Result and Discussion

Fig 3.1 shows the statistical power for all combinations of log mean abundance and log fold change for each data set. Red points indicate simulated taxa with significant log fold change (that is, adjusted p-value < 0.1); black points show simulated taxa where we failed to reject the null hypothesis (adjusted p-value > 0.1). Contour lines show the predicted statistical power for various combinations of overall log mean abundance and log fold change. Fig 3.1 shows a strong positive relationship between statistical power and fold change, as well as a weak positive relationship between mean abundance and statistical power, as anticipated. Few simulated taxa are in regions of high power (80% is the usual target for power in most scientific fields (Cohen 2013; Descôteaux 2007)), making it unlikely to attain high power in practical scenarios. Most individual taxa, in most data sets, have power less that 80%.

Our simulation studies were conducted with a relatively large sample sizes (for the field of microbiome studies in health sciences) of 100 samples per group. However, most microbiome studies are often constrained by practical limitations that restrict the available sample size. For example, Kers and Saccenti 2021 examined 100 publications and found a median sample size of 39 samples per group, with a mode of 8 samples. Given the prevalence of low sample sizes in the microbiome literature, differential abundance microbiome studies might have even lower power to detect biologically meaningful effects for individual taxa than those suggested by Fig 3.1.

Fig 3.2 shows the relationship between statistical power and the number of samples per group (30, 50, 70, 90, 110, 130, 150, 170 and 190 samples per group) for different log fold changes (2, 3 and 4). As expected, statistical power increases with increasing number of samples per group and increasing log fold change, although the power levels vary hugely across data set.

Fig 3.3 shows the expected number of taxa per experiment that differ significantly between groups. Increasing sample size increases the expected number of taxa that

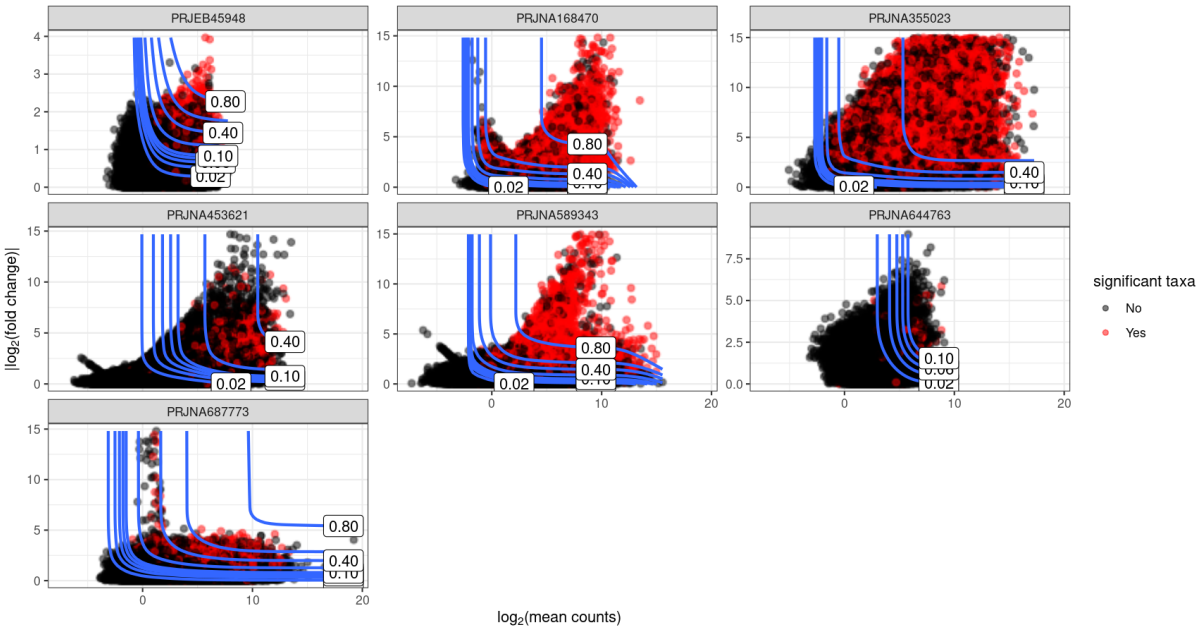

FIGURE 3.1: Contour plot showing statistical power for various combinations of overall log mean abundance and log fold change. 1000 taxa, 100 samples per group and 100 simulations. Red points indicate simulated taxa with significant log fold change (that is, FDR threshold of 0.1); black points show simulated taxa where we failed to reject the null hypothesis (adjusted p-value > 0.1). Contour lines show the predicted statistical power for various combinations of log mean abundance and log fold change

differ significantly between groups. For each experiment, smaller sample sizes result in much lower expected number of taxa with significant differences between groups. This implies that studies with low sample sizes stand the risk of missing taxa with biologically significant effects. Differential abundance microbiome studies might therefore require higher sample sizes than those prevalent in the literature in order to identify majority of the taxa with biologically significant effects. Low statistical power has significant impacts on the reliability of research results. Not only does it lead to type II errors (false negatives) but also causes strong upward bias in the magnitude of estimated effect sizes, via the "winner's curse" (a term describing the phenomenon where significant biological differences detected in studies with small sample sizes and low power are often associated with exaggerated effect size estimates) when a statistical significance filter (i.e., taxa with p-values below a threshold) is applied (Button et al. 2013).

Fig 3.4 compares the average power (defined by the arithmetic mean of the power estimates for all taxa) with the quantiles of power estimates for individual taxa. The figure also shows quantiles of power estimates for taxa in each sample size. Although

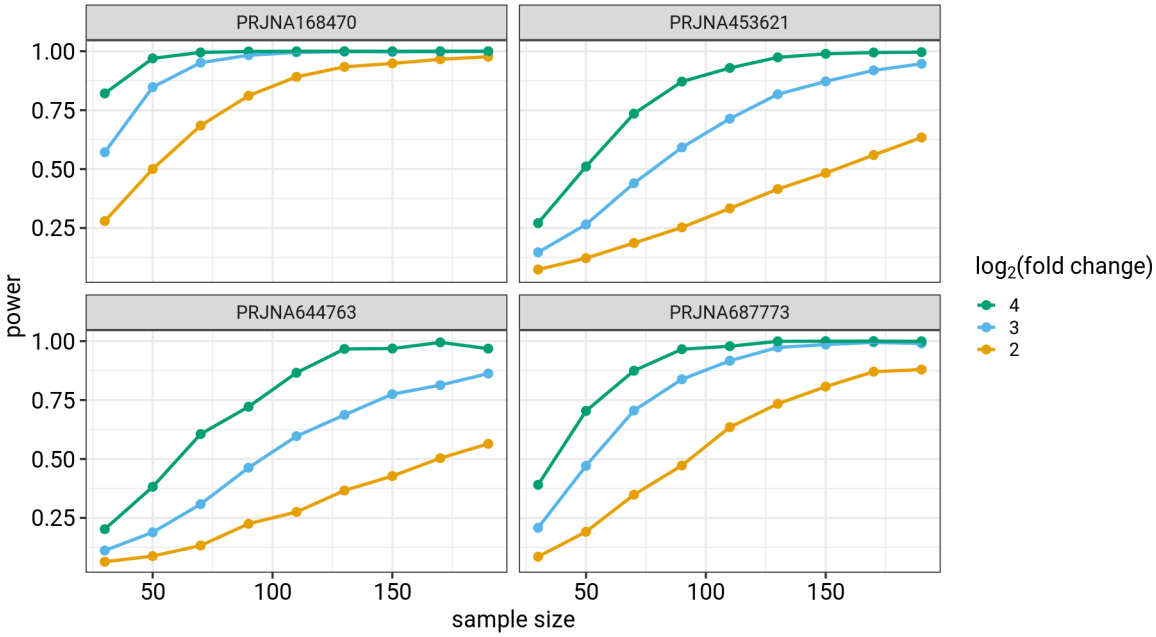

FIGURE 3.2: Relationship between statistical power, sample size and log fold change. 1000 taxa, 100 samples per group and 100 simulations. log mean abundance = 5.

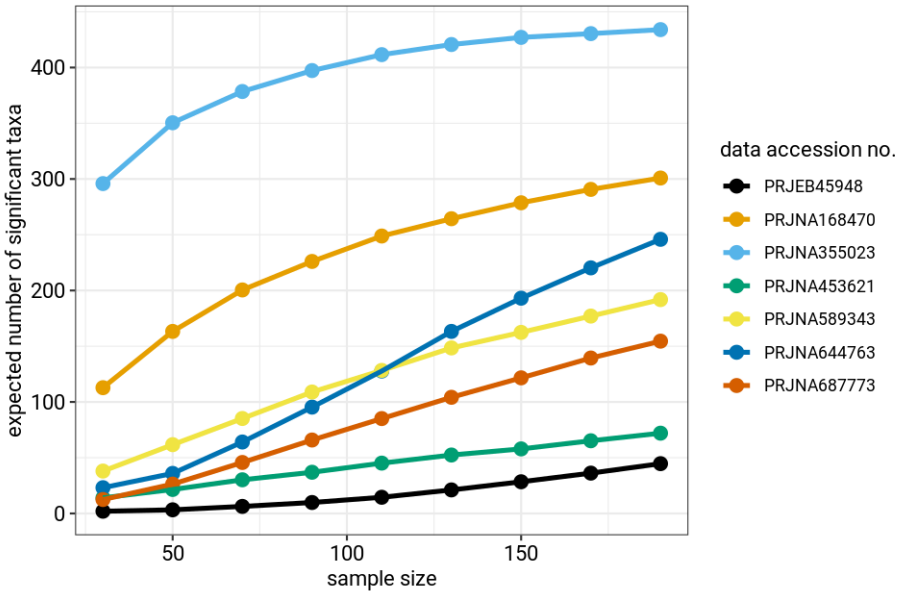

FIGURE 3.3: Expected number of significant taxa (out of 1000) for 30, 50, 70, 90, 110, 130, 150, 170 and 190 samples per group.

average statistical power is useful for determining the expected number of taxa that differ significantly between groups, the average statistical power does not provide accurate understanding of the statistical power of the individual taxa in a differential abundance study. For each microbiome data set shown in Fig 3.4, the average statistical power is consistently higher than the 50^{th} quantile of the individual taxon-by-taxon power estimates. In most cases, the average power surpasses the 60^{th} quantile of the individual power estimates, indicating that the average power overestimates the power for the majority of taxa. This highlights the need to consider statistical power at the level of individual taxa in a differential abundance study. Average power might overstate the power for most taxa and may lead researchers to underestimate the required sample sizes for their studies.

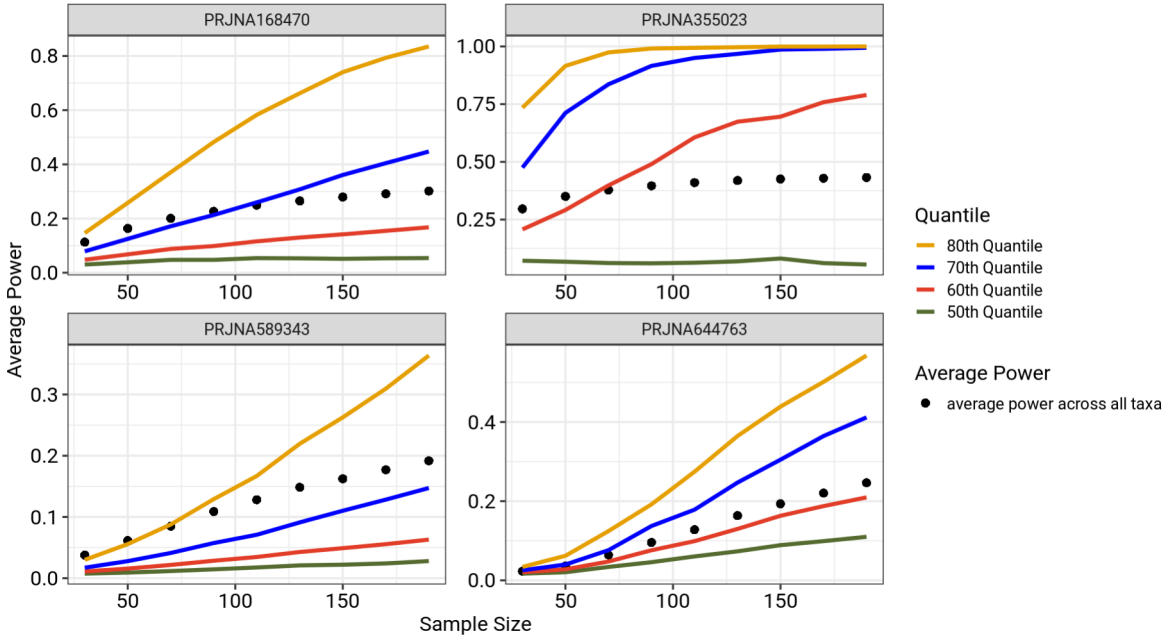

FIGURE 3.4: Comparison of average statistical power across all taxa and quantiles of taxon-by-taxon power estimates. Average power often overestimates the statistical power for most taxa and might not be a good metric for understanding the power to detect effects in a differential abundance study, hence the need to estimate power at the level of individual taxa.

3.5 Conclusion

Our study sheds light on potentially low statistical power to detect effect size of individual taxa in differential abundance microbiome studies. We introduced a novel method to estimate statistical power for individual taxa. Our method estimates power as a function of fold change and mean abundance of individual taxa.

Contour plots showing power for individual taxa suggest potentially low power to detect effect size of individual taxa in a differential abundance microbiome study (Fig 3.1). Low statistical power for individual taxa suggests that differential abundance studies might be missing many taxa with meaningful biological effects (Fig 3.3). Our findings also show that differential abundance studies may require larger sample sizes than are currently prevalent in microbiome research in order to achieve adequate statistical power (Fig 3.2).

The power estimation method presented in this study will enable researchers to estimate power at the level of individual taxa, quantify the range of power across all taxa, and estimate the expected number of significant taxa for their study. Our framework and simulation-based evidence contribute to enhancing understanding in the field, promoting accurate result interpretation. The provided framework and code facilitate reproducibility and empower researchers to make informed decisions about study design.

Chapter 4

Sample Size Calculation for Differential Abundance Studies

This chapter is a draft of a manuscript intended for submission for publication

4.1 Abstract

Determining an appropriate sample size for a study is a crucial step in planning scientific research. Appropriate sample sizes avoid both inflated and inadequate sample sizes. Collecting too many samples wastes resources, time and effort of human subjects, and lives of experimental animals. Collecting too few samples, a much more common problem, wastes even more resources through the inability to detect biologically meaningful differences and encourages questionable research practices like p-hacking. Microbiome studies are particularly challenged by sample size, particularly in studies of human subjects or expensive animal models. In practice, the statistical power of taxa within a differential abundance study is influenced by the effect size (fold change), mean abundance of individual taxa and the number of samples. We present a novel approach for sample size calculation for differential abundance studies as a function of effect size, mean abundance and statistical power. We applied our model for sample size calculation using estimates of mean abundance and fold change of taxa obtained from real microbiome data. Our results showed that differential abundance microbiome studies require larger sample sizes than are currently prevalent in the literature to achieve adequate statistical power. Our framework will help researchers make informed decisions about appropriate sample sizes.

4.2 Introduction

Choosing the right sample size is a crucial step in planning scientific research (Singh and Masuku 2014). Precise sample size selection helps avoid problems associated with having too many or too few samples. Excess sampling leads to wastes of time, resources, efforts of human participants and lives of experimental animals (Button et al. 2013).

Conversely, too few samples make it hard to detect significant biologically meaningful differences (Button et al. 2013).

Microbiome studies are often challenged with small sample sizes (Kers and Saccenti 2021). Studies with small samples have low statistical power, which increases the likelihood of failing to reject null hypotheses, even when meaningful differences truly exist. Increased variability in effect size estimates obtained from data with small sample size makes it difficult to differentiate between random noise and genuine biological difference, leading to less reliable estimates of effect sizes. Effect size estimates in such studies are often significantly biased, tending to deviate substantially from the true effect size. Significant biological differences (e.g., differences with p-value < 0.5) detected in studies with small sample sizes are often associated with exaggerated effect size estimates, a phenomenon termed the "winner's curse" (Button et al. 2013).

Researchers often select sample sizes for new studies based on previous studies (Singh and Masuku 2014). However, relying solely on previous sample sizes can be problematic, especially in fields where small sample sizes are common. If a field is traditionally underpowered, following the approaches of previous research work may lead to repeating the same limitations and obtaining inconclusive or unreliable results (Button et al. 2013; Ioannidis 2005). Power analysis, widely used in scientific research to determine sample size (Cohen 2013; Xia et al. 2018), can also be applied to ensure appropriate sample size selection in differential microbiome abundance analysis.

Sample size is influenced by statistical power and effect size (Cohen 2013). Since the goal of differential abundance microbiome studies is to detect taxa with meaningful effects, statistical power must be estimated at the level of individual taxa, as each taxon has its own unique effect size. Statistical power for each taxa is also influenced by mean abundance of the taxa (Agronah and Bolker 2025). Because the raw data in a microbiome analysis consist of count data (i.e., the number of reads detected), the relative degree of variation (i.e., the coefficient of variation) is higher for taxa with lower counts. This higher variability leads to lower statistical power to detect differences in rare taxa. To illustrate, it is easier to detect a 10% difference for a common taxon (e.g., an increase from 1000 to 1100 reads per sample) than for a rare taxon (e.g., an increase from 10 to 11 reads per sample). This difference in detectability arises because the absolute change is much smaller relative to the inherent variability in counts for rare taxa.

Consequently, sample size for differential abundance studies is determined by effect size, mean abundance and statistical power of individual taxa. This chapter extends the power estimation framework developed in chapter 3 to include sample size as an additional determinant in estimating statistical power for individual taxa and develops a novel method for estimating sample size as a function of statistical power, effect size and mean abundance. The sample size method presented in this chapter aims to assist researchers in determining the appropriate sample size for their studies, thereby enabling them to make informed decisions.

4.3 Materials and Method

4.3.1 Statistical power estimation procedure

The power estimation method proposed by Agronah and Bolker 2025, also described in Section 3.3.1, is given by:

$$\begin{aligned} y &\sim \text{Bernoulli}(p_i) \\ p_i &= \frac{1}{1+e^{-\eta}} \\ \eta &= \beta_0 + f_1(x_1, x_2) + \epsilon, \end{aligned}$$

where y is a binary variable indicating whether the p-value falls below a critical threshold (1) or not (0). We used the default threshold of 0.1 in the `DESeq2` package. The probability p_i represents the statistical power for taxon i. β_0 is the intercept, ϵ is the error term, and the predictors x_1 and x_2 correspond to log mean abundance and log fold change, respectively. The function f_1 represents a two-dimensional smoothing surface, generated using a tensor product smooth of log mean abundance and log fold change. The `DESeq2` package was used to compute p-values for testing the hypothesis for each taxon, using the Benjamini and Hochberg method for multiple hypothesis correction.

We define the GAM model that extends the model by Agronah and Bolker 2025 as follow:

Let x_{1i} and x_{2i} denote log mean abundance and log fold change of taxon i respectively. Let n be the sample size and y denote a binary value with 1 indicating that the p-value was below a critical value and 0 otherwise. Then y follows a Bernoulli distribution defined by

$$\begin{aligned} y &\sim \text{Bernoulli}(p_i) \\ p_i &= \frac{1}{1+e^{-\eta}} \\ \eta &= \beta_0 + f_1(x_{1i}, x_{2i}) + f_2(x_{1i}, n) + f_3(x_{2i}, n) + \epsilon, \end{aligned}$$

where p_i is the statistical power for taxon i, β_0 and ϵ are the intercept and error terms respectively. The functions f_1, f_2 and f_3 are two-dimensional smoothing surfaces with basis generated by the tensor product smooth of pairs of x_1, x_2 and n.

Power and fold change are positively correlated. Moreover, taxa with high abundance are more likely to be detected, thus having higher power, than rare taxa. To account for these relationships, we constrained the functions f_1, f_2 and f_3 to be monotonically increasing with power (Pya and Wood 2015).

4.3.2 Method for Sample size calculation

For a given log mean abundance, log fold change, we estimate the sample size required to achieve a target power using a root-finding technique. We use the `uniroot()` function from the `stats` package in `R`, which implements Brent's method—an adaptive algorithm that combines bisection, secant, and inverse quadratic interpolation to find the root of a continuous function over a specified interval.

Let g denote the estimated function from the fitted GAM model. The function g is then defined as

$$p_i = g(x_{1i}, x_{2i}, n), \tag{4.1}$$

Given x_{1i}, x_{2i} and a target power p_i, we estimate the required sample size n by finding the root of the equation:

$$p_i - g(x_{1i}, x_{2i}, n) = 0 \tag{4.2}$$

4.4 Data simulation

Estimating statistical power for individual taxa cannot be reliably calculated analytically using properties of count distributions due to the complexity of microbiome data. Estimating statistical power for each taxon requires simulating data that mimics actual microbiome data (Arnold et al. 2011). We used the MixGaussSim simulation method described in Section 2.4 for our data simulation. The MixGaussSim simulation method models appropriate distributions for effect size and mean abundance of taxa, which are useful for estimating the statistical power of individual taxa. We estimated the parameters of MixGaussSim from the seven autism data sets described in Section 2.3.

4.5 Results and Discussion

Fig 4.1 presents power curves for three autism microbiome data sets (PRJNA168470, PRJNA355023, and PRJNA687773), illustrating how statistical power varies with sample size per group and effect size. Across all three data sets, we observe that power increases as both sample size and effect size grow. Larger effect sizes (e.g., $|\log_2(\text{fold change})| = 3$ and $|\log_2(\text{fold change})| = 2$) achieve high power with relatively small sample sizes, while smaller effect sizes require larger sample sizes to reach comparable power levels. Larger sample sizes are particularly crucial when studying taxa with small effect sizes, as underpowered studies may fail to detect true associations. Conversely, when effect sizes are large, studies can achieve adequate power with fewer samples.

Higher taxa mean abundance achieve greater power with smaller sample sizes, whereas lower-abundance taxa require significantly larger sample sizes to reach comparable power

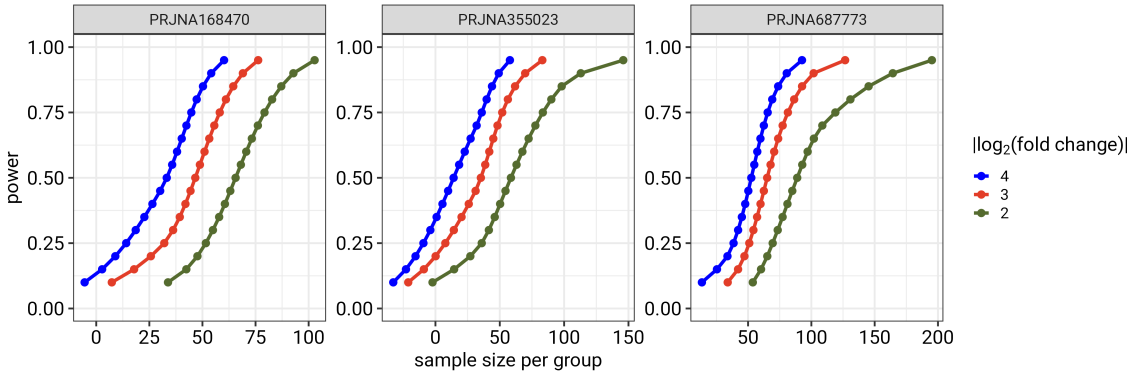

FIGURE 4.1: Relationship between statistical power, sample size and fold changes of taxa. $|\log_2(\text{fold change})| = 2, 3, 4$ and $\log_2(\text{mean count}) = 2$

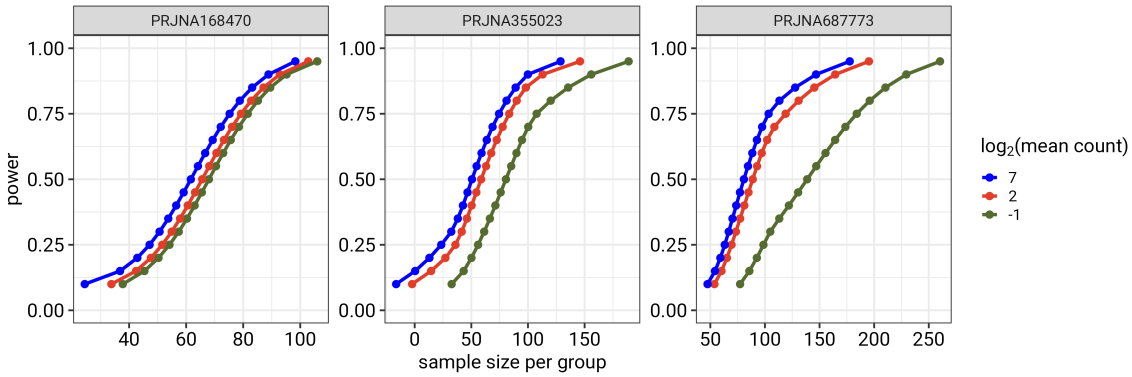

FIGURE 4.2: Relationship between statistical power, sample size and mean counts of taxa. $\log_2(\text{mean count}) = -1, 2, 7$ and $|\log_2(\text{fold change})| = 2$

levels (Fig. 4.2). This pattern aligns with expectations, as taxa with low counts contribute to greater variability in the data, reducing the likelihood of detecting true differences unless larger sample sizes compensate for this noise. Among the three data sets, PRJNA168470 and PRJNA355023 show a more rapid increase in power with sample size, particularly for the higher-abundance taxa. In contrast, PRJNA687773 exhibits a more gradual power increase, with the green curve (lowest abundance) requiring the largest sample sizes before reaching a power of 0.8 or higher. PRJNA168470 and PRJNA355023 show rapid power increases due to larger fold changes, higher mean counts, and fewer zeros, and PRJNA687773 exhibits a gradual power increase due to smaller effect sizes, lower mean counts, and possibly more zero inflation (Figs 4.3 - 4.4).

Researchers investigating low-abundance taxa must account for the need for larger sample sizes to achieve sufficient power. Moreover, differences between data sets suggest that power analysis should be conducted on a case-by-case basis to ensure robust detection of microbial associations. Variability demonstrates the need to consider data

set-specific characteristics, such as range of count abundance and sparsity, when planning sample sizes for microbiome studies.

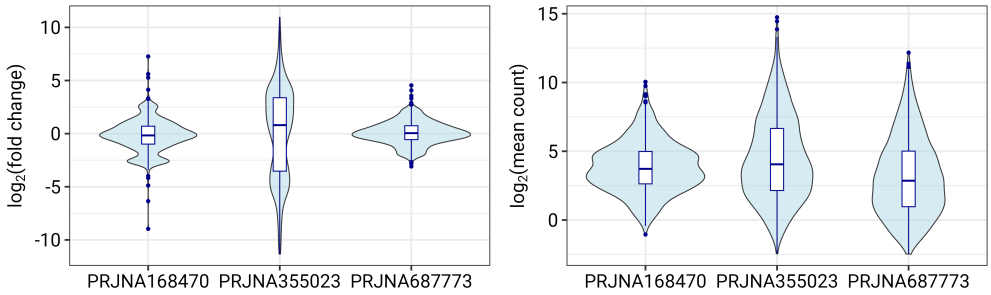

FIGURE 4.3: Comparing the distributions of fold change and mean count

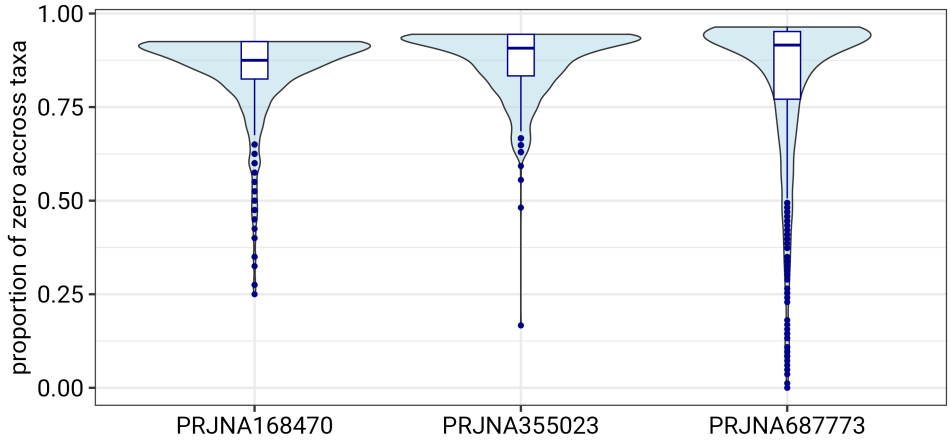

FIGURE 4.4: Comparing the distribution of the proportion of zero counts across taxa

In scientific research, a statistical power of 80% or higher is typically regarded as high power (Cohen 2013; Descôteaux 2007). The sample sizes required to detect a given fold change with 80% power are generally larger than those commonly used in microbiome studies. Figure 4.5 shows the minimum sample size required per group across all seven data sets to achieve 80% power at various combinations of fold change and mean abundance. For example, the smallest sample size required across all data sets to detect a fold change of 2^2 for a taxon with a high mean abundance of 2^7 at 80% power is approximately 79 samples per group. In contrast, a study examining 100 microbiome studies on alpha and beta diversity reported a median sample size of 39 samples per group, with a mode of 8 samples per group (Kers and Saccenti 2021), which is lower than the sample sizes observed in Figure 4.5.

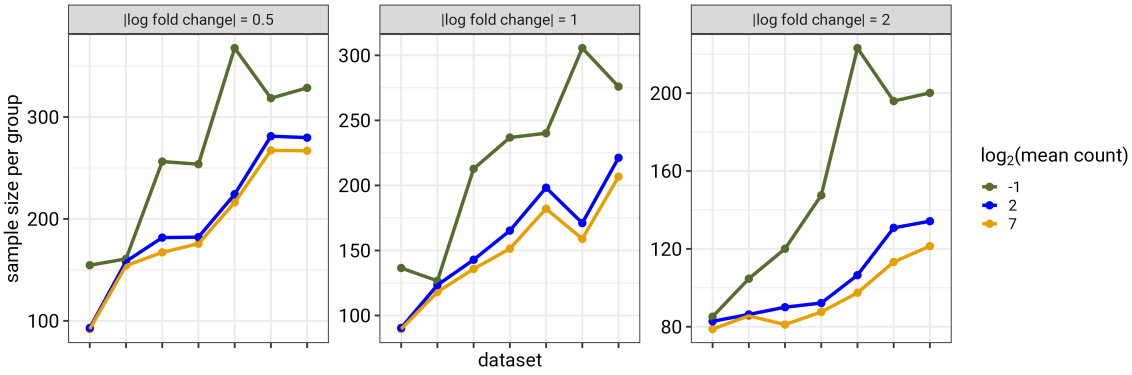

FIGURE 4.5: Sample size per group required to attain **80%** statistical power for taxa with log(fold changes) of 0.5, 1, and 2, and log(mean counts) of -1, 2, and 7.

4.6 Conclusion

We developed a novel method for sample size calculation in differential abundance microbiome studies. While sample size is generally influenced by statistical power and effect sizes, in differential abundance studies, power also depends on mean abundance. Since statistical power is a function of both mean abundance and effect sizes of individual taxa, our approach models sample size as a function of these factors. Using our method, researchers can quantify the range of sample sizes required to detect various effect sizes while accounting for the distribution of mean abundances across taxa in a data set.

To estimate statistical power, we apply the MixGaussSim simulation method presented in Section 2.4. MixGaussSim provides a flexible framework for modeling the distribution of mean abundance and effect size of taxa, which is essential for estimating taxon-specific statistical power. We fitted MixGaussSim to seven real microbiome data sets (described in Section 2.3) to derive realistic parameter estimates for MixGaussSim.

Our results show that larger sample sizes may be required than those commonly used in microbiome studies to achieve high statistical power to detect effect sizes of taxa. This highlights the importance of conducting sample size calculations before a study begins, as relying on previous or pilot studies may be misleading—especially if those studies already had low power.

Furthermore, data set-specific characteristics, such as sparsity and variability in counts, influence sample size determination. For instance, data sets with low counts of taxa and high proportion of zeros require larger sample sizes per group to detect effect sizes with high statistical power (Figs. 4.2 - 4.4). To obtain more reliable sample size estimates, researchers must take into account the specific characteristics of their data set to avoid estimating an inadequate sample size for their study.

Chapter 5

Beyond Independence: Joint Modeling of Microbiome Taxa with Reduced-Rank Correlation Structure

This chapter is a draft of a manuscript intended for submission for publication

5.1 Abstract

Human microbiome data is crucial for understanding the mechanism of diseases and treatment effectiveness. Researchers often aim to describe how counts of taxa differ among discrete groups (such as control versus treatment) or according to other factors.

Most existing models analyze each taxon separately, assuming no correlation between taxa within an individual. In reality, counts of taxa within subjects are correlated. Treating these counts as independent ignores the underlying biological structure in the data and may make it harder to accurately estimate changes in abundances of particular taxa. We developed a model that jointly models all taxa, accounting for the correlations between them. Typical microbiome data sets contain hundreds or thousands of taxa and require thousands or even millions of parameter estimates for the variance-covariance matrix, making it computationally impractical. To address this, we applied rank reduction to the variance-covariance matrix, reducing the number of parameter to be estimated.

We conducted simulation studies and real data analysis, comparing our reduced-rank model with three existing models: the negative binomial model in the `DESeq` R package, and the the negative binomial mixed and zero-inflated negative binomial mixed models in the `NBZIMM` R package. Our results show that modeling all taxa together and accounting for correlations between taxa improves accuracy in effect estimates, reduces bias, narrows confidence intervals, and increases statistical power compared to modeling taxa independently.

5.2 Introduction

Identifying taxa that show differential abundance between groups holds great potential for clinical applications. These differentially abundant taxa can serve as biomarkers for disease detection and treatment efficacy (Yang and Chen 2022). Microbiome research typically involves sequencing specific conserved regions of the genome, such as the 16S rRNA gene. These sequences are denoised using DADA2 to infer amplicon sequence variants (ASVs), which represent unique biological sequences. Each ASV is then mapped to a reference database to assign taxonomy and link the sequences to microbial species. The resulting data, often presented in an ASV (Amplicon Sequence Variants) table lists the frequency of each taxon detected in individual samples. This data is usually high-dimensional, with hundreds or thousands of taxa but few subjects (Xia et al. 2018).

Other than a few models, mostly Bayesian approaches (Grantham et al. 2020; Lee and Sison-Mangus 2018), that jointly model all taxa, most models developed for differential abundance analyze each taxon separately. For example, the `DESeq2` (Love et al. 2014) and `edgeR` packages implement negative binomial models fitted independently for each taxon. `metagenomeSeq` (Paulson et al. 2013) uses a zero-inflated log-normal model, `ALDEx2` (Gloor et al. 2014) applies univariate tests on centered log-ratio (CLR)-transformed Dirichlet-multinomial samples, `MaAsLin2` (Mallick et al. 2021) supports generalized linear and mixed models (e.g., Gaussian, Poisson, and negative binomial) fit separately per taxon, and `corncob` (Martin et al. 2020) uses a beta-binomial regression model with covariate-dependent mean and dispersion, also fit per taxon. Mixed models and zero-inflated mixed models, such as those implemented in tools like `ZINBMM`, similarly analyze each taxon independently.

These models assume that the abundance of taxa within subjects is independent. However, in reality, counts of taxa in a microbial community covary among individuals. Ignoring these correlations can misrepresent the biology of the microbiome community and may result in potential imprecise and bias effect size estimates (Hawinkel et al. 2019). It is essential to account for correlations between taxa within individuals to reflect the inter-dependence of taxa within microbiome data.

Various mixed models have been proposed in the literature for analyzing microbiome data (Zhang and Yi 2020). Examples of such models are the negative binomial (Zhang et al. 2018), zero-inflated Gaussian (Zhang et al. 2020), and zero-inflated negative binomial (Yi 2020) mixed models. To our knowledge, current uses of mixed models in microbiome research typically treat taxa individually, failing to address the correlations among taxa (Yi 2020; Zhang et al. 2018). Fitting a mixed model involving correlation between taxa is, however, challenging due to the high dimensionality of microbiome data, when no constraints are imposed on these correlations. The number of parameter estimates required for the correlation matrix grows quadratically with the dimension (i.e., number of taxa) of the microbiome data. Without imposing any structure on the variance-covariance matrix, microbiome data with n taxa requires estimating $n(n+1)/2$ parameters (corresponding to the number of elements in the lower triangle and the diagonal of the variance-covariance matrix). For instance, data with 10 taxa will require 55

parameter estimates. Consequently, the typical ASV table with thousands of taxa will require millions of parameter estimates, which is both statistically and computationally impossible to fit.

We use a reduced-rank approach to decrease the rank (effective dimensionality) of the variance-covariance matrix, thereby reducing the number of parameters to be estimated. Assuming an ASV table with n taxa, fitting a covariance matrix with a rank of $d \ll n$ allows the n expected per-taxon log fold changes (for example) to be expressed as a linear combination of d latent variables (McGillycuddy et al. 2025). The number of parameter estimates needed for the reduced-rank variance-covariance matrix is calculated as $dn - d(d-1)/2$ (McGillycuddy et al. 2025). Whereas the number of parameter estimates for the full variance-covariance matrix increases quadratically with n, the estimates for the reduced-rank matrix increase only linearly with n.

In this project, we propose a Reduced Rank Mixed Model (RRMM) for differential abundance analysis while modeling the correlations between taxa in subjects. We use the reduced rank functionality in the `glmmTMB` R package. The advantages of RRMM include: (1) it models all taxa jointly, allowing information sharing between taxa and accounting for the correlation structure in microbiome count data, (2) the flexibility of the `glmmTMB` package allows for incorporating a wide range of random effect terms and zero-inflation structures. We demonstrate through simulation studies and real data analysis that the RRMM can outperform existing univariate negative binomial methods, such as those implemented in `DESeq2` and `NBZIMM` R packages.

5.3 Method

5.3.1 The Reduced Rank Mixed Effect Model (RRMM)

RRMM is described in Section 2.5.2 of this thesis. The model, presented in equation (2.10), including the zero inflation component is specified in `glmmTMB` as follows:

```
glmmTMB(count ~ 1 + offset(normalizer) +
                    us(1 + group | taxon) +
                    rr(0 + taxon | subject, d),
                    ziformula = ~1+(1|taxon),
                    data = data,
                    family   = nbinom2)
```

LISTING 5.1: Example code for glmmTMB

where `us()` and `rr()` denote the unstructured and reduced-rank variance-covariance matrices for the random effects respectively, and `d` specifies the rank of the reduced rank matrix. `nbinom2` is the negative binomial distribution. In this formulation, the latent variables in the reduced rank model represent deviations from the average (log) count for each taxon within a subject.

5.3.2 Coverage estimation

The effect sizes in RRMM were modeled as random effects, which are latent variables rather than model parameters. To quantify the uncertainty associated with these latent variables, we computed empirical Bayes confidence intervals, also referred to as prediction intervals in the context of mixed models. These intervals reflect the uncertainty in the estimated taxon-specific group effects, given the observed data.

The random effect estimate (Best Linear Unbiased Predictor, BLUP) for taxon i follows a conditional normal distribution, which corresponds to the empirical Bayes posterior distribution:

$$b_i \mid y \sim N(\hat{b}_i, \sigma^2_{\hat{b}_i}), \tag{5.1}$$

where $\hat{b}_i$ is the conditional mode of the random effect, and $\sigma^2_{\hat{b}_i}$ is the conditional variance, obtained from the fitted mixed model. We computed prediction intervals using the standard normal approximation as follows:

$$\hat{b}_i \pm z_{\alpha/2} \times \text{SE}(\hat{b}_i), \tag{5.2}$$

where $\text{SE}(\hat{b}_i)$ is the square root of the posterior variance of b_i, and $z_{\alpha/2}$ is the critical value from the standard normal distribution.

5.3.3 Conditional AIC Estimation

The standard AIC (also referred to as the marginal AIC; (Greven and Kneib 2010; Vaida and Blanchard 2005)) is defined as

$$\text{AIC}_{\text{marginal}} = 2k - 2\ln(L_{\text{marginal}}) \tag{5.3}$$

where L_{marginal} is the marginal likelihood of the model, and k is the number of estimated parameters, typically including only the fixed effects and variance components (e.g., random effect variances and residual variance). This marginal AIC has two main limitations when applied to mixed models: (1) it relies on the log-likelihood calculated from the marginal distribution obtained by integrating over the random effects. and (2) it does not account for the variability induced by uncertainty in the estimates of the random effects covariance matrix, often favoring smaller models that exclude them (Greven and Kneib 2010).

A major challenge when computing AIC for mixed models is determining whether to treat random effects as model parameters and whether the log-likelihood should be conditioned on the random effects. The conditional AIC is often used for mixed model and is defined as

$$\text{AIC}_{\text{conditional}} = 2k - 2\ln(L_{\text{conditional}}) \tag{5.4}$$

where $L_{\text{conditional}}$ is the likelihood of the model conditioned on both the fixed and random effects. Determining the value of k for $\text{AIC}_{\text{conditional}}$ calculation is however, not straightforward. Since random effects in a mixed model are modeled as latent variables rather than parameters, it is unclear whether to include random effects when determining the value of k. Vaida and Blanchard 2005 proposed a method for determining k defined by the trace of the the leverage matrix (also called the hat matrix).

The linear mixed model is given by:

$$y = X\beta + Zb + \varepsilon, \tag{5.5}$$

where y is the response vector $(n \times 1)$, X is the fixed effects design matrix $(n \times p)$ associated with the fixed effect coefficients β, and Z is the random effects design matrix $(n \times q)$ associated with the random effect coefficients b. The random effects follow $b \sim N(0, G)$, where G is the random effects covariance matrix. The residual errors are $\varepsilon \sim N(0, R)$, where R is the residual covariance matrix. Assuming homoscedastic residual variance, R is typically defined as $R = \sigma^2 I_n$, where σ^2 is the residual variance and I_n is the identity matrix of size $n \times n$.

The marginal distribution of y, integrating over b, is:

$$y \sim N(X\beta, V), \tag{5.6}$$

where the marginal covariance matrix is:

$$V = ZGZ^T + R. \tag{5.7}$$

The leverage matrix H is defined as:

$$H = X(X^T V^{-1} X)^{-1} X^T V^{-1}. \tag{5.8}$$

Thus, the conditional AIC is given by:

$$\text{AIC}_{\text{conditional}} = 2\text{tr}(H) - 2\ln(L_{\text{conditional}}), \tag{5.9}$$

where $\text{tr}(H)$ denotes the trace of the leverage matrix H.

5.3.4 Statistical Power Estimation

Microbiome studies and other related high-dimensional studies such as differential gene abundance analysis often assume that a large fraction of taxa have exactly the same abundance between treatments. This assumption is reflected in both computational methods used to estimate treatment effects (e.g., lasso regression, spike-and-slab Bayesian priors (Bhadra et al. 2017)) and in the evaluation metrics for model performance. Specifically, metrics like specificity (the probability that an estimated nonzero change is truly nonzero) and false discovery rate (the probability that the null hypothesis is true given

that it was rejected) rely on the premise that some taxa remain unchanged across treatments. However, some researchers (e.g., Stephens and Balding 2009) argue that in a complex biological system, it is unlikely that treatment effects would be exactly zero for any taxon, even if most changes are small.

In our simulations, the latent variable representing deviations of effect sizes from the overall effect size is drawn from a multivariate Gaussian distribution with a zero mean and a specified variance-covariance structure. As a result, the probability of any taxon having exactly identical abundance across treatments is zero. Consequently, the null hypothesis is never strictly true, meaning that we cannot evaluate specificity or FDR in this context. Given that there are no true zero effects, we define the average statistical power across taxa as the proportion of p-values less than a specified significance threshold. We used a p-value threshold of 0.05 as is often used in the statistical literature. Since the analysis involves many taxa and multiple hypothesis tests we applied the Benjamini and Hochberg method.

5.4 Simulation Studies

We simulated data using the RRSim simulation method presented in Section 2.5. RRMM jointly models all taxa and uses the reduced rank method to model correlations between taxa. RRMM also has functionality to model zero inflation in microbiome data. We simulated with a full rank covariance-matrix (ie, without imposing rank reduction). RRMM uses the `simulate_new` function from the `glmmTMB` R package. The `simulate_new` function allows setting up a simulation directly, specifying the right-hand side of the model formula described in Listing 5.1, a data frame describing the experimental setup (e.g., number of taxa, number of subjects, group names of subjects), and a list of simulation parameters. The function requires input values for all of the information in the model: the standard deviations and correlations of each random effect, as well as the intercept terms (that is, the overall average count, taxa-specific effect, and average zero-inflation probability).

Table 5.13 presents the parameter values used for the simulation. To determine a realistic range of zero-inflation probabilities and parameter values for the zero-inflation term of our model, we estimated zero-inflation probabilities from seven real microbiome data sets from a previous study by Agronah and Bolker 2025. We fitted the zero-inflated negative binomial mixed model implemented in the `NBZIMM R` package for estimating these zero-inflation probabilities. The raw sequence data for these seven data sets is available from the the National Center for Biotechnology Information (NCBI) (Sayers et al. 2021) data repository under accession numbers PRJNA168470, PRJNA355023, PRJNA453621, PRJEB45948, PRJNA644763, PRJNA589343 and PRJNA687773. The range of zero-inflation probability estimates for each data set is shown in Fig. 5.13 under supplementary material. The average minimum and maximum zero-inflation probabilities across the seven data sets were 0.12 and 0.92, respectively, yielding a midpoint of 0.52. Consequently, we conducted our simulation using an average zero-inflation probability of 0.52. We calculated the standard deviation of the taxon-specific zero-inflation

probabilities using the logit-transformed range, given by

$$\log\left(\frac{\text{logit}(0.92) - \text{logit}(0.12)}{4}\right).$$

This ensures the random effect term varied by about two standard deviations around the midpoint of 0.52. We generated correlation matrices for our simulation using the `huge` (Zhao et al. 2012) R package. This package is well-suited for simulating high-dimensional correlation matrices due to its ability to efficiently handle sparse and large-scale networks. These features are relevant for microbiome data, which typically exhibit high dimensionality and complex dependency structures among taxa. We chose the parameter values $(\beta_0, \beta_{0_{disp}}) = (3, 1)$ to yield a realistic range of mean abundance of taxa.

TABLE 5.1: Parameter values used for simulation studies. The logit function (ie, inverse of the logistic function) is defined as $\text{logit}(x) = \ln(1/(1-x))$.

Parameter	**Description**	**Value**
β_0	Intercept term representing overall average (baseline) count	3
β_{0_z}	Intercept term representing average zero-inflation probability across taxa	$\text{logit}(0.52) \approx 0.0772$
$\beta_{0_{\text{disp}}}$	An over-dispersion parameter for the negative binomial distribution	1
`nsim`	Number of simulations	500
(m, n)	A coordinate representing the number of subjects (m) and taxa (n)	$(10, 50)$, $(50, 100)$, $(100, 300)$
θ_{zi}	A standard deviation parameter for the variation in taxon-specific zero-inflation probabilities	log((logit(0.92) - logit(0.12))/4) ≈ 0.103
$\boldsymbol{\theta}_1$	A vector of length three corresponding to the parameters for the `us(1 + group | taxon)` term in our model: the first two entries represent the log-standard deviations of the taxon-specific effects and group-specific random effects, while the third entry represents the correlation between the taxon-specific and group-specific random effects.	Code for simulating values for $\boldsymbol{\theta}_1$ is presented under supplementary material
$\boldsymbol{\theta}_2$	A vector of length $n(n+1)/2$ representing the parameters for the unstructured variance-covariance matrix (i.e., the `us(0 + taxon| subject)` term of our model): the first n entries are the log-standard deviations of of the taxon-specific random effects for each subject, and the remaining $n(n-1)/2$ entries represent the pairwise correlations among the n taxa-specific random effects for each subject.	Code for simulating values for $\boldsymbol{\theta}_2$ is presented under supplementary material

5.5 Real Data

We applied the RRMM to four real microbiome data sets and compared its performance with the models listed in Table 5.3. Since the zero-inflated mixed model from the `NBZIMM` package cannot fit taxa without zero counts, we excluded taxa that had no zero counts in any subjects in each data set to ensure a fair comparison.

5.5.1 The Autism Data

The Autism data (Chen et al. 2020) measures the gut microbiome of children with Autism Spectrum Disorder (ASD) and neurotypical (NT) children. The raw sequence data is available at the repository of the National Center for Biotechnology Information (NCBI: Sayers et al. 2021), with accession number PRJNA644763. The data set consists of fecal samples from 76 children with ASD and 47 children with typical development. We used the processed ASV count data from a previous study by Agronah and Bolker 2025. The processed ASV table contained 8,477 taxa and 123 subjects. Following standard procedures in differential abundance microbiome studies, we filtered rare taxa, retaining only those with an abundance of seven or more reads in at least four samples (Xia et al. 2018; Love et al. 2014). We then excluded taxa with no zero counts in any subjects. The final ASV table after pre-filtering and exclusion of taxa without zero counts contained 689 taxa and 123 subjects.

5.5.2 The Soil Data

The soil bacterial species data set is described in a study by Nissinen et al. 2012, where 16S rRNA gene amplicon sequencing was used to characterize bacterial communities associated with bulk soil samples. The raw sequence data is available in the European Nucleotide Archive under the accession number PRJEB17695 (Sayers et al. 2021). The data set comprises 56 soil samples collected from eight locations: three in Kilpisjärvi, Finland, three in Ny-Ålesund, Svalbard, Norway, and two in Mayrhofen, Austria. Microbial community profiles were obtained from these soil samples, and ASVs were assigned based on sequence similarity, resulting in a raw count matrix with 1276 ASVs. Each sampling site (denoted by `Site` in the data set) is treated as independent, as bacterial communities are typically highly location specific. Each soil sample (`Soiltype`) was grouped as being top soil (T) or bottom soil (B). Three continuous environmental variables were measured for each sample: soil organic matter, soil pH value and amount of phosphorus in soil. We used the processed count data from the `gllvm R` package, which has undergone pre-filtering to exclude low abundance species present in fewer than five samples. The processed data after removing taxa with no zero counts included 969 ASVs and 56 soil samples. We considered a model with covariates `Soiltype` and `Site`.

5.5.3 The Crohn's Disease Data

The Crohn's disease data measures the microbial composition in the terminal ileum of individuals with Crohn's Disease (CD) and healthy controls (Gevers et al. 2014; Silverman et al. 2022). The data contains microbial abundance data and associated metadata, including disease diagnosis, age of individuals and treatment information. After filtering out missing values and low abundance taxa, the data contains 49 taxa and 250 samples. The data is available in the `fido` R package. We consider disease diagnosis (healthy vs CD) and the age of individuals as the only covariates and examine differences in the microbiome composition between individuals with CD and healthy individuals.

5.5.4 The Human Intestinal Data

The human intestinal microbiome data set (Lahti et al. 2014) investigates the composition of the gut microbiota in 1,006 adults. The ASV count data included 130 taxa and 1,151 samples, with some samples obtained from repeated measurement of some subjects. Since our study in this chapter does not account for a longitudinal design, we excluded repeated measurements, retaining only the first time point for each subject. The data also includes subject information such as age, gender, nationality and Body Mass Index (BMI) group (that is, BMI-based categories underweight, lean, overweight, obese, severely obese and morbidly obese). We consider a model involving age and BMI group only and study the effect of BMI groups on count abundance of taxa. The final data set we used involved 130 taxa and 900 samples. The data set can be found in the `microbiome` package (Lahti 2012) in R.

For both simulation and real data analysis, we compared the performance of the RRMM with the negative binomial model implemented in `DESeq2`, as well as the negative binomial mixed model and zero-inflated negative binomial models in the `NBZIMM` package.

5.6 Results and Discussion

Table 5.3 summarizes the notation used for each model.

To evaluate the impact of the reduced rank term in our model, we compared the RRMM to a variant that excluded the reduced rank term while keeping all other components unchanged. To assess the influence of including the zero inflation terms, we compared:

- the Reduced Rank Mixed Model (RRMM) with a zero-inflation component (`RRzi`) compared to the RRMM without a zero-inflation component (`RR`);
- the RRMM without the reduced rank term but with zero-inflation (`USzi`) to the same model without zero-inflation (`US`);
- the negative binomial model implemented in the `NBZIMM` package (`NB`) with the corresponding zero-inflated negative binomial model (`ZNB`).

TABLE 5.2: Summary of microbiome data sets used in the analyses in Chapter 5.

data set	**Data Source**	**# Taxa (Raw)**	**# Samples**	**Pre-filtering Criteria**	**# Taxa After Filtering**
Autism Data	NCBI accession number PRJNA644763	8,477	123	Retained taxa with abundance of seven or more reads in at least four samples	123
Soil Data	`gllvm` R package (ENA Accession PRJEB17695)	1276	56	Taxa present in at least 5 samples; taxa with no zero counts removed	969
Crohn's Disease Data	`fido` R package	214 (originally 9511 taxa; filtered by aggregating to the Family level and excluding low-abundance taxa)	286 (originally 1359 samples; removed samples with missing disease status, excluded those on steroids, antibiotics, or biologics, and retained only samples from the terminal ileum). Retained samples with at least 5000 total reads across taxa	Retained taxa with counts >3 in at least 10% of samples	49
Human Intestinal Data	`microbiome` R package	130	900 (originally 1151 samples. Excluded samples with missing data and repeated measurements)	Filtered out taxa with no zeros in all samples and taxa with zeros in all subjects	30

TABLE 5.3: Abbreviation for model names and descriptions

Model	**Description**
`RR`	The Reduced Rank Mixed Model (RRMM) without a zero-inflation component.
`RRzi`	The Reduced Rank Mixed Model (RRMM) with a zero-inflation component to account for taxon-specific zero-inflation probabilities.
`US`	The Reduced Rank Mixed Model (RRMM) excluding both the reduced rank component and the zero-inflation term.
`USzi`	The Reduced Rank Mixed Model (RRMM) with a zero-inflation component but without the reduced rank term.
`DE`	The negative binomial model implemented in the `DESeq2` package, with shrinkage functionality enabled.
`DE_noSk`	The negative binomial model implemented in the `DESeq2` package, with shrinkage functionality disabled.
`NB`	The negative binomial mixed model implemented in the `NBZIMM` package.
`ZNB`	The zero-inflated negative binomial mixed model implemented in the `NBZIMM` package.

5.6.1 Simulation studies

For each model, we compared the group effect sizes we used for the simulations with the average group effect size estimates obtained across simulations. The effect size estimates presented in these results are on the (natural) logarithmic scale. The average group effect size estimate follows the trend of the true group effect sizes (Fig. 5.1).

FIGURE 5.1: Average group effect estimates by taxon across simulation for each model. y-axis has been truncated to exclude extremely large or low effect size estimated from the NB and ZNB models (eg. estimate value of 20) caused by identifiability issues (i.e., situations where some taxa have zero counts in all subjects within one group, making it impossible or unstable to estimate group effects for those taxa).

Root Mean Squared Error (RMSE), Variance and Bias

We compare the average (arithmetic mean) root mean squared error (RMSE) across taxa and the average (arithmetic mean) bias across taxa for each model, incorporating a 95% confidence interval to quantify the uncertainty in these average estimates. For all combinations of number of taxa and subjects, the reduced rank models (`RR` and `RRzi`) achieve the lowest average RMSE (Fig. 5.2). Including the reduced rank term improves the precision of effect size estimates, as shown by the lower RMSE when comparing `RR` to `US` and `RRzi` to `USzi`. The uncertainty associated with the average estimates for the reduced-rank models is generally lower than that of all other models.

Among the models that analyze taxa separately, the `DESeq2` models yield the lowest average RMSE and highest precision in the average RMSE estimate. A strength of the `DESeq2` package is its capabilities for shrinkage estimation of effect size and dispersion, which allow information sharing between taxa. The effect size shrinkage reduces estimates for taxa with low counts, pulling them toward zero. This effect size shrinkage is similar to the random effects incorporated in the US and RR models. Similarly, the dispersion shrinkage pulls dispersion estimates for highly variable taxa toward the mean trend observed across all taxa. Disabling effect size shrinkage increases the average RMSE, as seen when comparing the `DE` and `DE_noSk` models (Fig. 5.2). In contrast, the `NBZIMM` models (ie, `NB` and `ZNB` models), which also analyze taxa separately, do not allow shrinkage estimation. Models that incorporate information sharing mechanisms result in more precise effect size estimates (Love et al. 2014; Robinson et al. 2010).

In general, incorporating a zero-inflation term improves the precision of effect size estimates (Fig. 5.2). Other than the joint models that omit the reduced rank term (i.e., `US` and `USzi`), all other models with zero-inflation have lower average RMSE compared with their counterpart that excludes zero-inflation (comparing `NB` with `ZNB`, and `RRzi`

with `RR`) (Fig. 5.2). This improvement in precision is unsurprising, as the data were simulated including zero inflation. All models tend to underestimate the true effect sizes, as indicated by the negative bias (Fig 5.3). The `NBZIMM` models (`NB` and `ZNB`) exhibit the least bias, while the reduced rank models (`RRzi` and `RR`) show higher bias. The reduced rank models generally exhibit lower variability in their effect size estimates compared to all other models (Fig. 5.4). In contrast, the `NBZIMM` models have higher variance, producing more variable estimates across taxa. Overall, the reduced rank models (`RRzi` and `RR`) perform better based on these metrics, with lower average RMSE, error variance, although slightly more bias in its estimates.

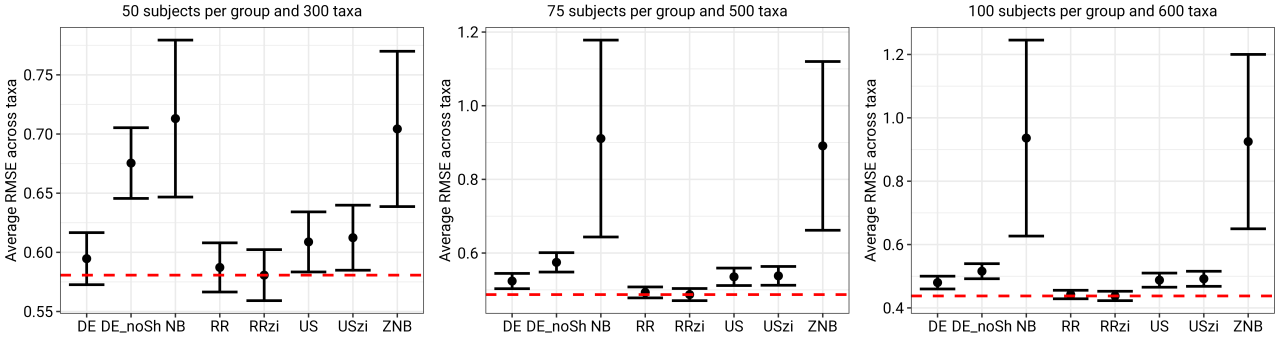

FIGURE 5.2: Comparing average Root Mean Squared Error (RMSE) across all taxa each model. We use RRzi as the reference model shown by dashed red line

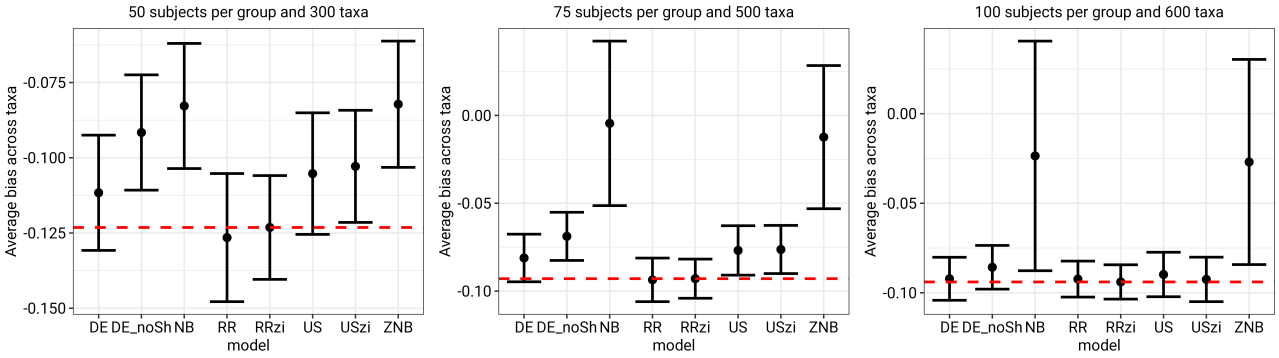

FIGURE 5.3: Comparing average bias across taxa for each model. We use RRzi as the reference model shown by dashed red line

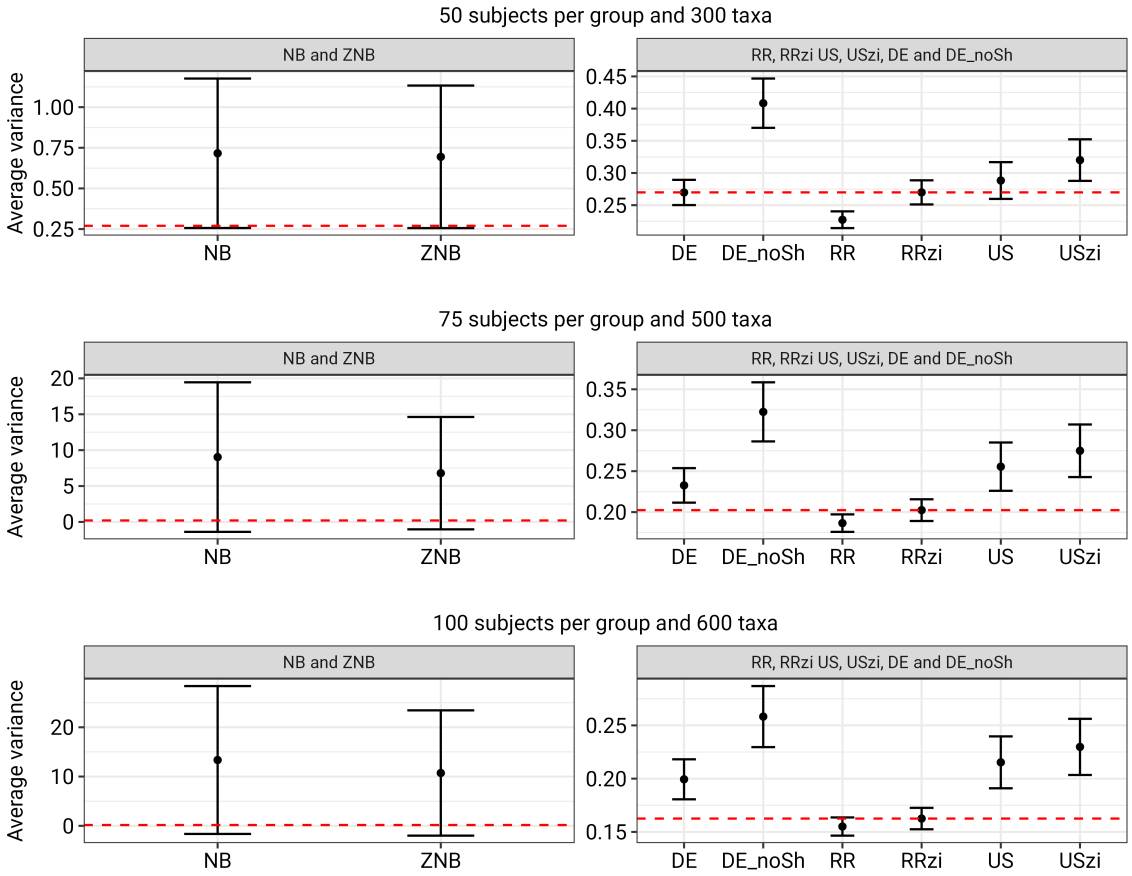

FIGURE 5.4: Comparing average variance of error across taxa for each model. We use RRzi as the reference model, shown by the dashed red line.

Confidence Width and Coverage

We fitted each model to 500 simulations. For each simulation, we estimated Wald confidence intervals for each taxon at the 95% confidence level. For the joint models (`RR`, `RRzi US` and `USzi`), we applied the confidence interval approach outlined in Section 5.3. We then computed coverage for each taxon as the proportion of simulations in which the true value fell within the estimated confidence interval. For each taxon, we also calculated the average width of its confidence intervals across simulations. Figure 5.6 presents a comparison of the overall average confidence interval width, computed as the mean of the taxon-specific average widths. Models with a zero-inflation term have a narrower confidence interval, as indicated by the reduced confidence width, but show lower coverage. This can be seen by comparing `ZNB` with `NB`, `RRzi` with `RR`, and `USzi` with `US` in Fig. 5.6. Thus, including zero inflation underestimates the uncertainly

associated with the effect size estimates. The `RR` model has the best coverage, having a coverage close to the nominal 95% value (Fig. 5.6). Comparing `RRzi` to `USzi` and `RR` to `US` shows that including the reduced rank component increases coverage.

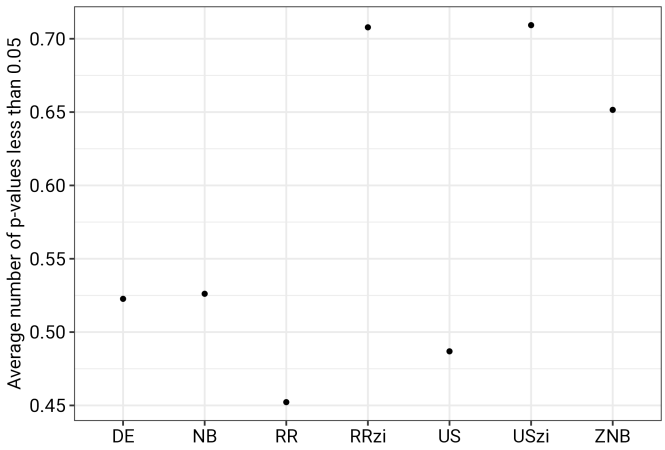

FIGURE 5.5: Comparing average statistical power across 500 simulations for each model. 600 taxa and 100 samples per group. The standard errors for points shown in plot are on the order of 10^{-5}, making the confidence intervals around the point estimates too small to be visible.

Statistical Power

We conducted 500 simulations and fitted each model to each simulation. For each simulation fit, we estimated the average statistical power as the proportion of adjusted p-values (based on FDR threshold of 0.05) less than 0.05. The average power for each model across all simulations shows that models with zero inflation have higher statistical power as seen when comparing `RRzi` to `RR`, `USzi` to `US` and `ZNB` to `NB` in Fig. 5.7. The `RRzi` and `USzi` models exhibit the highest power of about 71%.

When including the reduced rank term might be beneficial

We performed an experiment to determine when including the reduced-rank term could be beneficial. We simulated ten data sets with a fixed total variance of count abundance. The total fixed variance was composed of different proportions of variance explained by the group effect term (ie. the `us (1 + group | taxon )` term in listing 5.3.1) and variance explained by the reduced rank term (ie. the `rr (0 + taxon | subject, d )` term in in listing 5.3.1). For the first simulation, the total variance was composed of the group variance, without a contribution from the reduced rank term. We then simulated data using increasing proportions from the reduced rank term and decreasing proportion from the treatment effect.

For each of the ten simulations, we fitted the `US` and the `RR` models and selected the best model based on marginal AIC (see Section 5.4). For different values of total

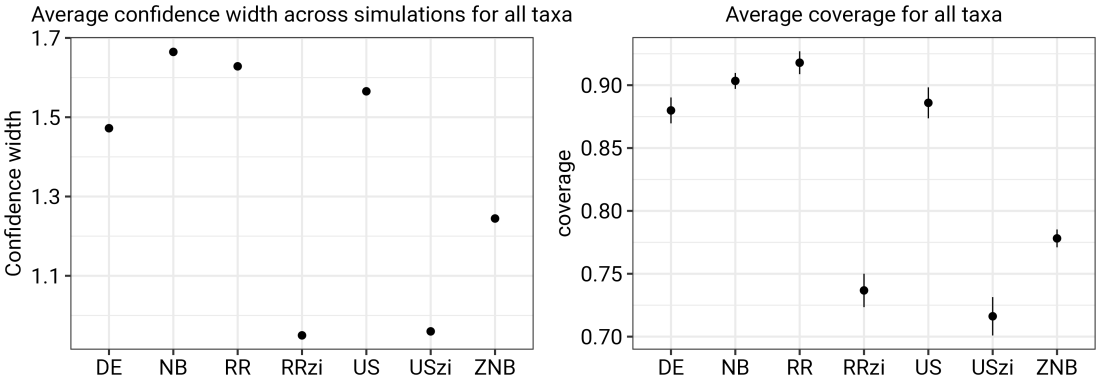

FIGURE 5.6: Comparing average confidence width (averaging across simulations and across taxa) and average coverage (averaging across taxa) for each model. 600 taxa, 100 samples per group and 500 simulations. The standard errors for the plot in the left panel are on the order of 10^{-4}, making the confidence intervals around the point estimates too small to be visible.

variance, we observed from our experiment that the reduced-rank model emerged as the best model as the proportion of total variance contributed by the latent variables exceeds the proportion of variance contributed by the group effect (Fig. 5.8), suggesting that the complexity of including the reduced rank term might only be beneficial when the latent variance explains more of the variance in mean abundance than is explained by the treatment effect.

5.6.2 Real data sets

For real data, model performance metrics such as RMSE and bias cannot be calculated because the true effect sizes are unknown. Therefore, to assess model performance, we compare the Akaike Information Criterion (AIC), the width of the confidence intervals, statistical power and the runtime of the models.

AIC Comparison

We compare the AIC of the models for the four data sets. For the RR and US models which model effect sizes as random effects, we computed the conditional AIC as described in Section 5.3.3. The `DE` model from the `DESeq2` package, as well as the `NB` and `ZINB` models implemented in the `NBZIMM` package, do not directly provide likelihood estimates required for AIC calculation. Therefore, we computed the likelihood for the `NB`, `ZINB`, and `DE` models by re-implementing these models within the `glmmTMB` package. We then used the effect size estimates obtained from fitting each model in their respective packages (`NBZIMM` or `DESeq2`) as input values to evaluate the likelihood within `glmmTMB`.

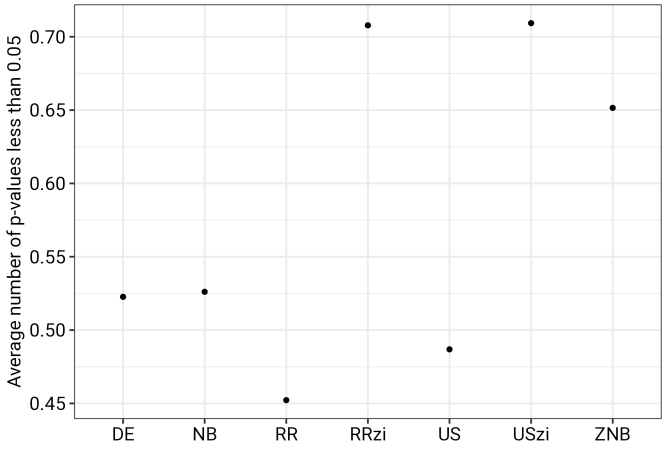

Figure 5.7: Comparing average statistical power across 500 simulations for each model. 600 taxa and 100 samples per group. The standard errors for points shown in plot are on the order of 10^{-5}, making the confidence intervals around the point estimates too small to be visible.

The reduced rank models were selected as the best models with the lowest AIC values for the human intestinal and the Crohn's disease data sets while the `NB` model was selected as the best model for the soil data (Fig. 5.9). The `US` model was selected as the best model with the lowest AIC value of 104579.50 (lower than the AIC value for the `DE` (`DESeq` with shrinkage) model (see supplementary material: *Sup* 5.8). Thus, for these data sets, models that analyze taxa jointly mostly outperform models that assume independence in taxa abundances.

Statistical power

We estimated the average statistical power as the proportion of adjusted p-values (using the Benjamini-Hochberg method) that fall below a significance threshold of 0.05. Some models show notably low statistical power across certain data sets. In practice, statistical power is typically assessed through repeated simulations that closely mimic the characteristics of the actual data set under study (Agronah and Bolker 2025; Kelly et al. 2015). However, in this case, we use a single real data set as a proxy to provide a rough comparison of the expected power across models, recognizing that this approach does not fully capture the variability inherent in power estimation.

The NB, `ZINB`, `DE` and `USzi` models show low average statistical power for the soil and autism data sets (Fig. 5.10). In contrast, the `RRzi` model exhibit high power for the autism and soil data sets. The soil data also has the lower range of power values in comparison to the other data sets.

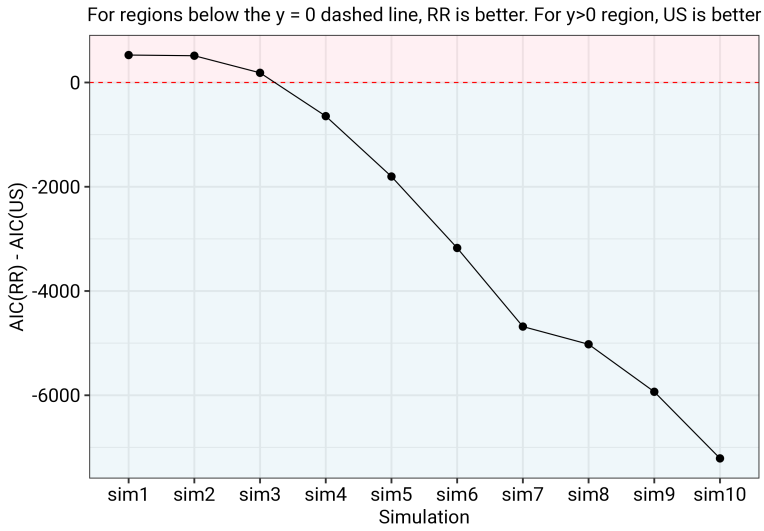

FIGURE 5.8: Differences in marginal AIC between the US and the RR model for increasing proportion of variance explained by the reduced rank term. Total variance = 4. Number of taxa = 100 and 25 samples per group.

Confidence width comparison

For real data, where true effect sizes are unknown and coverage cannot be directly measured, narrow confidence intervals alone do not necessarily reflect a model's overall performance. Thus, we cannot necessarily be confident that the confidence intervals have correctly estimated the uncertainty. Assuming similar coverage holds for these real data sets as with those observed in our simulation studies, then narrow confidence intervals for models observed to have high coverage (eg. models without zero inflation terms — see Fig. 5.6) would indicate improved precision in effect size estimates.

Among the models that analyze individual taxa separately, the `DE` model consistently exhibits the lowest average confidence interval width across all data sets, except for the autism data set, where the `ZINB` model achieves the narrowest intervals (Fig. 5.11). In contrast, the NB model shows the highest average confidence interval width across all data sets, suggesting lower precision in its effect size estimates. The average confidence widths for the `RR`, `RRzi`, `US`, and `USzi` models, which jointly model all taxa, vary across data sets. However, given the high coverage demonstrated by the RR model in simulation studies (see Fig. 5.6), its generally low average confidence interval width suggests it may provide more precise effect size estimates in real data applications.

Run time comparison

We compared the computation time required to fit the four models to each of the real data sets (Fig. 6.12). For each data set, we fitted the RR model with a rank of 2. All models, except the RR model fitted to the pregnancy data set, were run on a Dell laptop

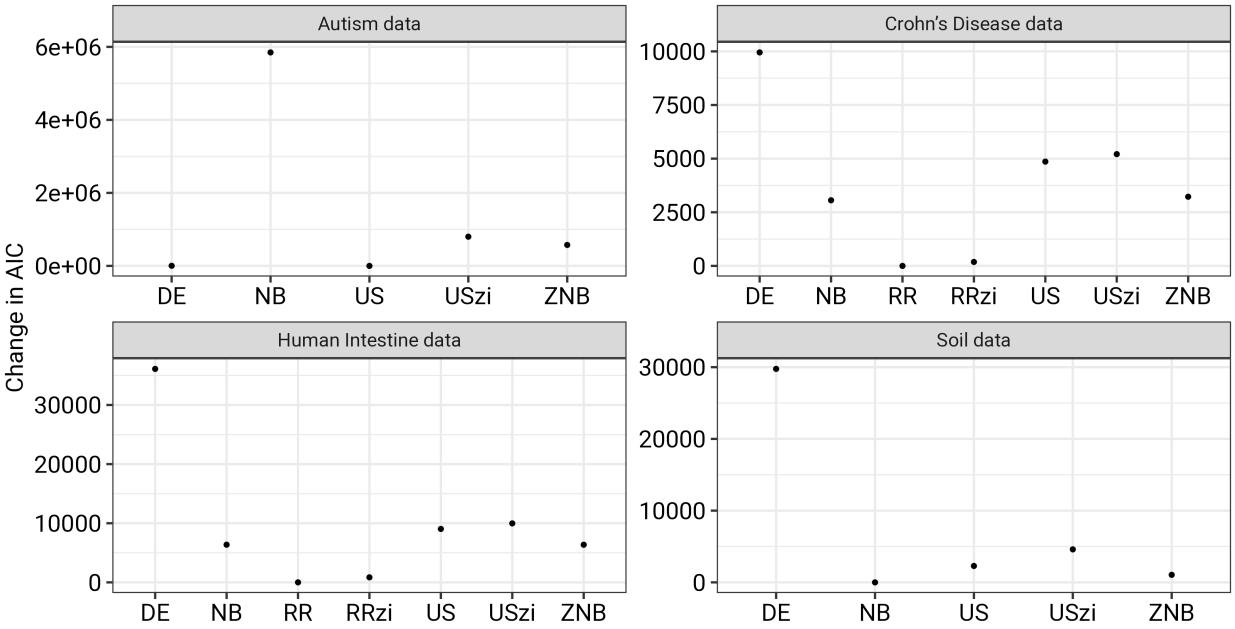

FIGURE 5.9: AIC values for the models for each data set. AIC values for the RR and RRzi models fitted to the autism and soil data sets were omitted due to computational challenges in estimating the hat matrix (see Section 5.3.3) required for conditional AIC calculations. Specifically, the high dimensionality of these models led to memory demands exceeding 5 TB (terabytes) during the leverage computation, rendering the calculations infeasible.

running Ubuntu with 6 cores and 12 threads. The RR model for the pregnancy data set required a substantially higher number of iterations to converge due to the large number of taxa and the model complexity. It was therefore run on the Graham cluster; a high-performance computing resource provided by the Digital Alliance of Canada, using `glmmTMB` with extended optimization controls: `optCtrl = list(eval.max = 1000, iter.max = 100)`.

Across all data sets, the `US` and `USzi` models, which jointly analyze all taxa without explicitly accounting for correlations, consistently exhibited shorter runtimes than the `NB` and `ZINB` models, which analyze each taxon separately and are comparatively less complex. The `DE` model required approximately the same runtime as the `US` model for both the Crohn's disease and soil data sets. The `US` model had the lowest runtime for the autism data set. These results suggest that modeling all taxa jointly may not necessarily lead to longer computational times.

For all data sets, the reduced rank models (RR and RRzi) takes longer to run due

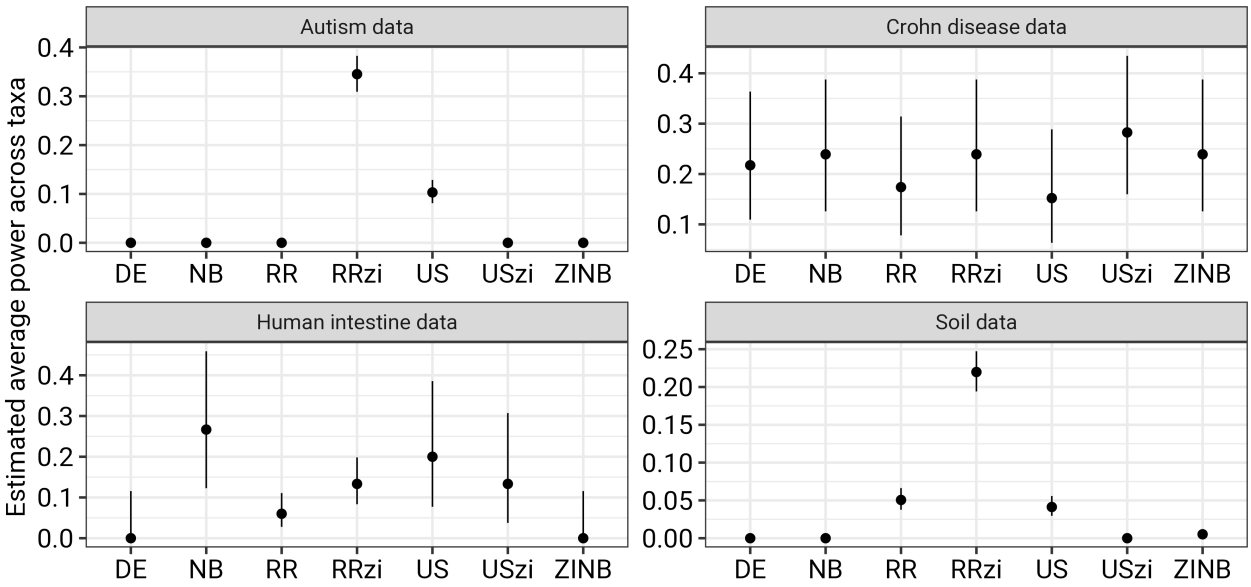

FIGURE 5.10: Average statistical power across taxa for the four data sets.

to the complexity of the model. Overall, this analysis illustrates a clear trade-off between model complexity and runtime. Although the reduced-rank models may require longer runtimes, the improvement in effect size estimation may justify the additional computational cost.

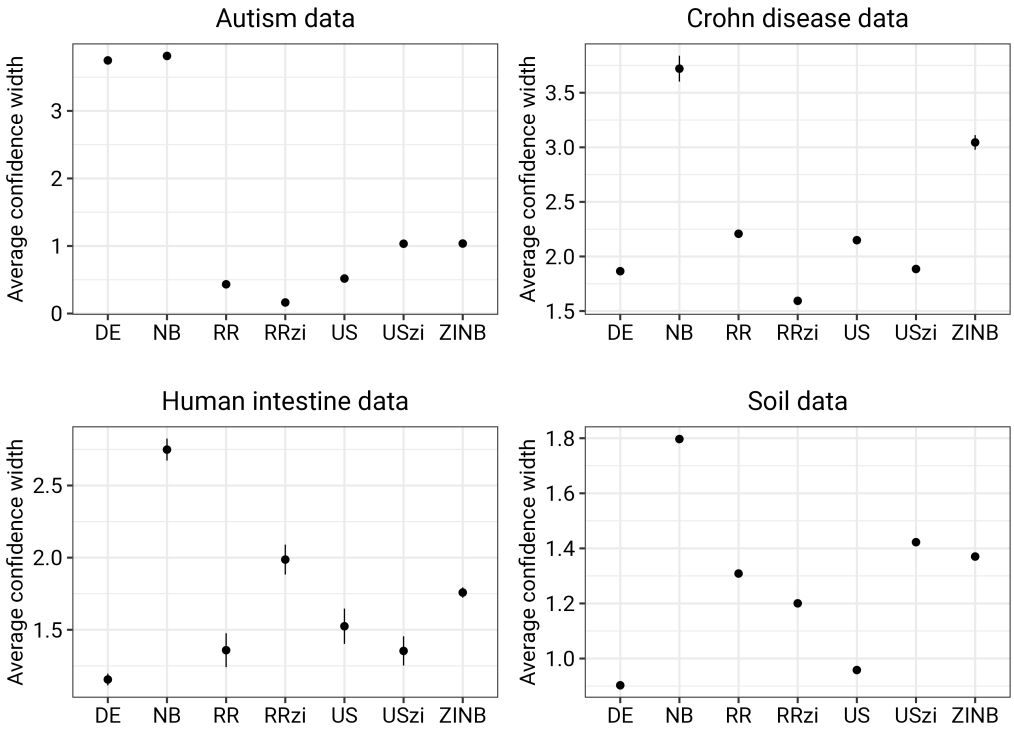

FIGURE 5.11: Average confidence width across taxa for the four data sets. The standard errors for some of the points show in these plots are on the order of 10^{-4}, making the confidence intervals around these point estimates too small to be visible.

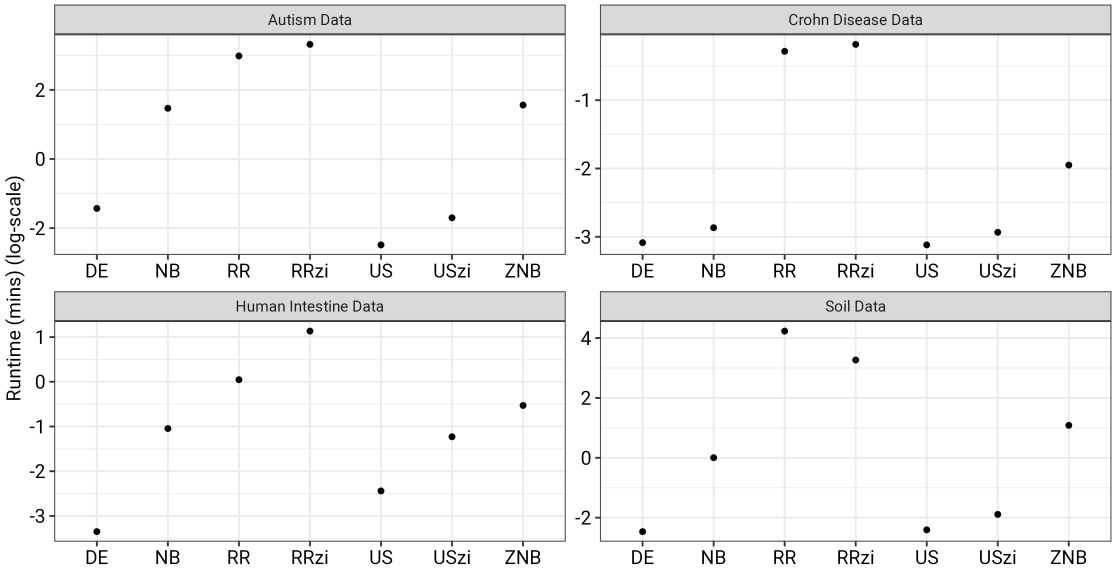

Figure 5.12: Computational runtime of each model when fitted to the data sets

5.7 Conclusion

This chapter introduced the Reduced Rank Mixed Model (RRMM) for joint analyses of microbiome data while accounting for correlations among them through a rank-rank structure. Our approach highlight limitations of models which assume independence among taxa and analyze them individually. Through simulation studies and application to real data sets, our findings highlight the following conclusions.

First, joint modeling of taxa consistently improves model performance. Even when correlations between taxa are not explicitly modeled (as in the `US` and `USzi` models), jointly analyzing taxa enables information sharing across features, resulting in more precise effect size estimates. This is reflected in lower root mean squared error (RMSE) of effect size estimates compared to models that treat taxa independently (Fig. 5.2). Moreover, models that assume independence but incorporate effect size shrinkage still show gains in precision over those that neither model correlations nor apply shrinkage (Fig. 5.2) (Love et al. 2014; Robinson et al. 2010).

Second, our simulations show that incorporating rank reduction increases the coverage of confidence intervals (Fig. 5.6). When compared with joint models that excluded the reduced rank term, RRMM yielded lower RMSE and higher coverage (Fig 5.6).

Third, zero-inflation components can further enhance model performance under sparse data scenarios. Incorporating a zero-inflation term generally improved the precision of effect size estimates and increased statistical power to detect biologically meaningful differences. In our simulations, models with zero-inflation (e.g., RRzi, USzi, ZNB) consistently showed higher power and narrower confidence intervals compared to their non-zero-inflated counterparts. However, this improvement comes at the cost of reduced coverage, suggesting that these intervals may underestimate uncertainty.

Finally, our runtime analysis challenges the assumption that joint modeling is always computationally burdensome. Some joint models (e.g., US and USzi) required similar or even less computation time than independent models like DESeq2 and NBZIMM, particularly in data sets with moderate dimensionality. Although RRMM models are more computationally intensive due to the added complexity of estimating latent structures, they offer significant gains in model accuracy and biological interpretability.

In summary, our work demonstrates that modeling taxa jointly—especially with reduced rank correlation structures and optional zero-inflation terms—offers substantial improvements in estimation accuracy, statistical power, and biological relevance. Even when taxon-level correlations are not explicitly modeled, joint models outperform traditional independent approaches. The RRMM framework provides a scalable and statistically rigorous alternative for modern microbiome analysis, balancing model complexity with computational feasibility. We recommend its adoption in settings where taxa are likely to be interdependent, and where accurate, high-resolution inference is required.

5.8 Supplementary material

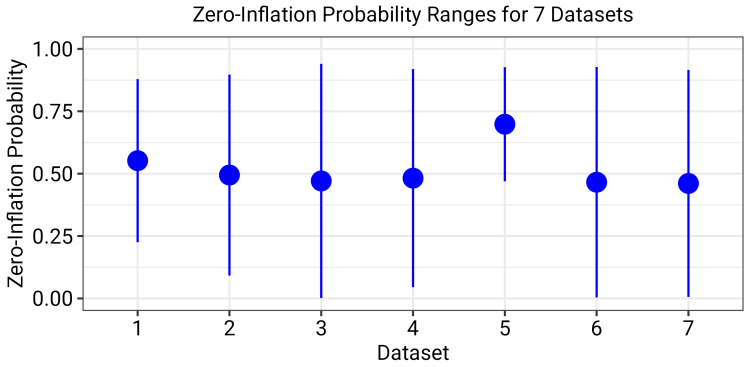

FIGURE 5.13: Ranges of zero inflation probability estimates across taxa for seven real microbiome data sets.

Functions for simulating log standard deviations and correlations for random effect terms

```
#' @param meanlog mean log
#' @param sdlog log standard deviation
#' @param rank:  specifies the rank of the reduced rank approximation

get_theta_logSD <- function(n,  meanlog = 0, sdlog = 1,
seed = NULL, rank = NULL) {
  set.seed(seed)
  val <- rlnorm(n, meanlog, sdlog)
  logSD <- log(sqrt(val))

  if (!is.null(rank)) {
    logSD <- logSD[1:rank]
    return(logSD)
  } else {
    return(logSD)
  }
}


#' @param ntaxa number of taxa
#' @param nsubject number of sunject
#' @param mat specify a matrix directly.

get_theta_corr <- function(ntaxa,nsubject, mat= NULL, seed = NULL) {
  if(!is.null(mat)){C  <- mat}
```

```
  else{set.seed(seed); C <- get_corr(ntaxa, nsubject,seed = seed)}
  C <- nearPD(C)$mat
  scale <- sqrt(fastmatrix::ldl(as.matrix(C))$d)
  cc2 <- chol(C) %*% diag(1/scale)
  cc2[upper.tri(cc2)]
}
```

TABLE 5.4: AIC estimates and differences (ΔAIC) for all models across data sets

data set	**Model**	**ΔAIC**
Autism data	US	0
	USzi	800267
	NB	5848122
	ZNB	573927
	DE	3293
Crohn's Disease data	US	4862
	USzi	5210
	RR	0
	RRzi	184
	NB	3059
	ZNB	3224
	DE	9951
Human Intestine data	US	9036
	USzi	9973
	RR	0
	RRzi	849
	NB	6369
	ZNB	6365
	DE	36104
Soil data	US	2286
	USzi	4599
	NB	0
	ZNB	1051
	DE	29761

Chapter 6

A Reduced Rank Poisson Model for Longitudinal Microbiome Data: Accounting for Taxa Correlations

This chapter is a draft of a manuscript intended for submission for publication

6.1 Abstract

Modelling associations between longitudinal microbiome data and other covariates aids in understanding how microbial communities change over time and how these changes differ between treatment groups (eg. control vs. treatment). Such longitudinal microbiome analysis also aid in understanding disease progression in patients and microbial responses to dietary interventions, antibiotics, and environmental changes.

Just as in a non-longitudinal microbiome study, taxa within subjects in a longitudinal design are correlated. Accounting for these correlations may lead to improved precision in effect size estimates. However, modelling these correlations in a longitudinal design is more challenging because taxa can co-vary in different ways over time. Similar to a non-longitudinal design, modeling correlations in longitudinal design also require estimation of thousands or hundreds of parameter estimates for microbiome data, which is computationally impossible to fit. Due to this complexity, most existing models analyze individual taxa separately.

In this chapter, we propose a Longitudinal Reduced Rank Mixed Model (LRRMM), which extends the Reduced Rank Mixed Model proposed in Chapter 5 for non-longitudinal designs. LRRMM is designed to analyze associations between microbiome data and covariates by modeling all taxa jointly, while accounting for correlations among taxa within subjects over time. We used the reduced rank functionality available in the `glmmTMB` package to reduce the number of parameter estimates required for modeling correlations

between taxa. We demonstrate that LRRMM provides more precise estimates of effect sizes compared to both the Negative Binomial and Zero-Inflated Negative Binomial mixed models implemented in the `NBZIMM` R package.

6.2 Introduction

Longitudinal data arises when subjects in an experiment are repeatedly sampled over time. This type of data is common in microbiome research, where microbial samples are collected from each subject at multiple time points. The common goal of such research is to study how microbial communities change over time and how these changes differ between treatment groups (Zhou et al. 2015). Researchers also conduct longitudinal microbiome analysis to investigate the development of diseases within subjects over time (Dashper et al. 2019) and to investigate microbial responses to dietary interventions, antibiotics, and environmental changes (Lyu et al. 2023).

Three kinds of analyses are common in longitudinal microbiome research: (1) studies of changes in microbial abundance over time and associations with host or environmental variables (Zhang et al. 2020; Lewis et al. 2015; Bäckhed et al. 2015), (2) clustering of taxa based on similar temporal trends, often using dimensionality reduction techniques such as principal component analysis (PCA) or linear discriminant analysis (LDA) (McNicholas and Murphy 2010), (3) understanding biological and temporal relationships among taxa, using tools such as network models to identify relationships among taxa (Kodikara et al. 2022).

There are several challenges in analyzing longitudinal microbiome data. First, repeated sampling from the same subject creates dependencies among observations, leading to correlations in microbial counts over time. Second, longitudinal designs introduce additional complexity due to potential interactions over time, which may require more sophisticated models. For example, the effect of a probiotic supplement on gut microbial composition might depend on participants' physical activity levels with stronger effects during more active periods and weaker effects during sedentary periods, illustrating an interaction between treatment and a time-varying covariate.

As with non-longitudinal microbiome data, taxa counts within the same subject in a longitudinal design are also correlated. However, modeling correlations in a longitudinal design is more complex, as taxa can co-vary in different ways over time. For example, some taxa may tend to rise or fall together across time within a subject, suggesting a shared pattern or trajectory throughout the study period. Alternatively, taxa may be correlated within a subject at specific time points, meaning that certain taxa tend to co-occur or vary together at particular moments in time. The rates of change of individual taxa over time within a given subject may also be correlated.

Mixed models have been widely used in microbiome research to examine associations between microbial abundance and covariates as well as to model changes in microbial communities over time. Examples include Gaussian and zero-inflated Gaussian mixed models (Paulson et al. 2017; Zhang et al. 2018), Poisson and zero-inflated Poisson mixed

models (Romero et al. 2014; Zhang et al. 2018), negative binomial and zero-inflated negative binomial mixed models (Zhang et al. 2018), and zero-inflated Beta regression mixed models (Chen and Li 2016). Mixed models are a well-established approach for analyzing longitudinal data, as they accommodate both within-subject correlations and between-subject variability. By incorporating random effects, mixed models allow for subject-specific variation and provide a flexible framework for modeling complex dependency structures. Several R packages, including `glmmTMB`, `lme4`, `NBZIMM` (Zhang and Yi 2020), and `ZIBR` (Chen and Li 2016), provide tools for fitting mixed models to microbiome data.

Few methods for longitudinal microbiome analysis model taxa jointly while accounting for correlations within subjects; those that do are predominantly Bayesian (Lee and Sison-Mangus 2018). Most existing methods analyze each taxon separately, ignoring correlations among taxa within a subject (Zhang et al. 2020; Kodikara et al. 2022). To address this limitation, we extend the Reduced Rank Mixed Model (RRMM), introduced in Section 5, originally developed for single-time-point microbiome data, to a longitudinal setting. RRMM offers a frequentist alternative for joint modeling, leveraging a reduced-rank approximation to efficiently capture correlations among taxa while controlling model complexity. In Chapter 5, we demonstrated, for a non-longitudinal microbiome data analysis, that modeling all taxa jointly and accounting for correlation between taxa improves the precision of effect size estimates.

In this chapter, we propose the Longitudinal Reduced Rank Mixed Model (LRRMM) for longitudinal microbiome data analysis. LRRMM models all taxa jointly across all time points and accounts for correlations between taxa using the reduced rank method to reduce the number of parameter estimates required for the correlation between taxa. We apply LRRMM to address the research question: "how do the rates of change of a particular taxon (i.e. slope with respect to time) differ between groups (control vs treatment, healthy vs diseased subjects, etc.)?". Understanding how the rates of change of microbial taxa differ between groups provides insights into disease progression, treatment effects, and ecological shifts in microbial communities. Analyzing these longitudinal trends can reveal whether certain taxa respond differently to interventions or environmental changes, which is aids in identifying potential biomarkers.

The goal of this chapter is to evaluate how well LRRMM performs in addressing this question, compared to models that analyze each taxon separately and ignore these correlations. By comparing the two approaches, we assess whether joint modeling improves estimation accuracy, effect size precision, and statistical power to detect meaningful differences between groups.

6.3 Method

6.3.1 General Model Description

The Poisson Mixed Model for Longitudinal Microbiome Data Analysis

Let $\boldsymbol{Y}_t$ denote a $m \times n$ Amplicon Sequence Variant (ASV) table at time t for $t = 1, \ldots, T$, with rows $i = 1, \ldots, m$ and columns $j = 1, \ldots, n$, where m, n and T represent the number of subjects, number of taxa and the number of repeated time measures, respectively. Let $\boldsymbol{x}_{it}$ denote a d-dimensional vector of covariates (which may vary over time), and let $\boldsymbol{y}_{it} = (y_{i1t}, \ldots, y_{int})'$ represent the $1 \times n$ vector of count abundance at time t.

Define the long-format concatenation of all subject-taxon-time observations as

$$\bar{\boldsymbol{y}} = (\boldsymbol{y}'_{11}, \boldsymbol{y}'_{12}, \ldots, \boldsymbol{y}'_{1T}, \ldots, \boldsymbol{y}'_{m1}, \ldots, \boldsymbol{y}'_{mT})',$$

a vector of length $N = m \times n \times T$. Each entry $\bar{y}_k \in \bar{\boldsymbol{y}}$, where $k = 1, \ldots, N$, follows a poisson distribution:

$$\bar{y}_k \sim \text{Poisson}(\mu_k),$$

where μ_k denotes the expected mean for observation k.

We model the vector of expected mean counts $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)'$ using a generalized linear mixed model:

$$\begin{aligned} \boldsymbol{g}(\boldsymbol{\mu}) &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}, \\ \boldsymbol{b} &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}), \end{aligned}$$

where $\boldsymbol{g}$ is the log link function, $\boldsymbol{X} \in \mathbb{R}^{N \times d}$ is the design matrix for fixed effects (including time and covariate interactions) and $\boldsymbol{\beta}$ is a d-dimensional vector of fixed effect coefficients. $\boldsymbol{Z} \in \mathbb{R}^{N \times q}$ is the design matrix for random effects, $\boldsymbol{b}$ is a q-dimensional vector of random effects with $\boldsymbol{b} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{q \times q}$ is the variance-covariance matrix of the random effects.

6.3.2 Specific Model Description

The Reduced Rank Poisson Mixed Model for Longitudinal Microbiome Data (LRRMM)

We consider a longitudinal design where samples are collected from subjects from two groups (e.g., treatment vs. control or diseased vs. healthy) over time. For our research question, we are interested in investigating how much the rates of change of a given taxon differ between the two groups.

We develop a model to estimate how group, time, and their interaction affect count abundance while accounting for correlations between taxa within subjects. We do not include fixed effect terms for taxa, group (control vs. treatment),time nor their interactions. Estimating taxon-specific fixed effects could lead to overfitting, especially when

the number of taxa is large. Instead, we model the effect of taxa on count as a random effect, which captures variability across taxa while avoiding the need to estimate an excessive number of parameters. Including a fixed effect for group (that is, main effect for group) measures the degree to which all taxa consistently increase or decrease across all subjects in one group compared to another group. However, this assumption does not accurately reflect the nature of microbiome data. Researchers typically treat microbiome data as compositional, meaning the changes in abundance of each taxon are measured in relation to the abundance of others. If one taxon increases, others must decrease. In practice, this compositionality is typically handled by normalizing the data to adjust for differences in overall abundance, as caused for example by differences in sequencing depth between samples. Thus, the normalization and compositional nature of microbiome data make the assumption of a consistent increase or decrease in all taxa abundance across all subjects in a group inappropriate for microbiome data.

We allow the effect of groups, time and their interactions to be correlated. We assume taxa within subjects are correlated at each time point. We also allow each subject-taxon combination to have its own random intercept, accounting for subject-specific variability in taxon counts. For simplicity, we do not allow random slopes for each subject-taxon combination.

Microbiome data often exhibit overdispersion, where count variability exceeds what is expected under a Poisson model. We modelled overdispersion by including an observational-level random effect to account for additional dispersion. Alternatively, overdispersion can be modeled directly using a negative binomial distribution, which includes an explicit dispersion parameter (Harrison 2014).

The specific model formulation is therefore as follows:

$$\begin{cases} g(\mu_{ij}^{t}) = o + \boldsymbol{z'_{i1}}\boldsymbol{b_{j1}} + \boldsymbol{z_{i2}}\boldsymbol{b_{j2}} + \boldsymbol{z_{i3}}\boldsymbol{b_{j3}} \\ \quad \boldsymbol{b_{j1}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma_1}), \\ \quad \boldsymbol{b_{j2}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma_2}) \\ \quad \boldsymbol{b_{j3}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma_3}) \end{cases} \tag{6.1}$$

where o is an offset term to account for differences in sequencing depth (read depth) between samples and $g(\cdot)$ is a log-link function. $\boldsymbol{Z_1} \in \mathbb{R}^{mn \times 2n}$ and $\boldsymbol{Z_2} \in \mathbb{R}^{mn \times mn}$ are the design matrices for the random effects. The variance-covariance matrices $\boldsymbol{\Sigma_1} \in \mathbb{R}^{2n \times 2n}$ and $\boldsymbol{\Sigma_2} \in \mathbb{R}^{mn \times mn}$ are positive definite block matrices. Each block within $\boldsymbol{\Sigma_1}$ models correlation between the random intercept and random group effect for each taxon. The blocks of $\boldsymbol{\Sigma_2}$ model correlations between taxa within each subject.

It is impractical to assume that the correlation between taxa within subjects differs across subjects as this would lead to over-parametrization, making it impossible to fit the model. Therefore, we simplify our model by assuming that all subjects have the same between-taxon correlations. We also assume the correlation between baseline counts (intercepts) and group effects is the same for all taxa. Consequently, we define the

variance-covariance matrices $\boldsymbol{\Sigma_1}$ and $\boldsymbol{\Sigma_2}$ as follows:

$$\boldsymbol{\Sigma_1} = \boldsymbol{\sigma_1^*} \otimes \boldsymbol{I}_{2n},$$

$$\boldsymbol{\Sigma_2} = \boldsymbol{\sigma_2^*} \otimes \boldsymbol{I}_{mn},$$

where $\boldsymbol{\sigma_1^*} \in \mathbb{R}^{2\times 2}$ is a positive definite matrix that models correlations between baseline counts and group effects for each taxon, $\boldsymbol{\sigma_2^*} \in \mathbb{R}^{n\times n}$ is a positive definite matrix that models correlations between taxa within each subject, and $\otimes$ denotes the Kronecker product. $\boldsymbol{I}_{2n}$ and $\boldsymbol{I}_{mn}$ are identity matrices. The model in equation (6.1) including the zero inflation component is specified in `glmmTMB` as

```
glmmTMB(count ~  offset(normalizer)
               + (group*time|taxon)
               + (1|obs)
               + rr(taxon + 0 | subject:time,d),
                 ziformula =  ~1,
                 data      =  data,
                 family    =  poisson)
```

LISTING 6.1: Example code for glmmTMB

where `us()` and `rr()` specify unstructured and reduced-rank variance-covariance matrices for the random effects, respectively. The term `obs` introduces a random effect to capture additional variability at the level of individual observations, accounting for overdispersion. The argument `d` defines the rank of the reduced-rank matrix.

6.3.3 Estimating coverage

Estimating coverage for individual taxa is computationally intensive, as it requires numerous simulations, each involving additional nested simulations. In order to estimate coverage for each taxa, we computed Wald confidence intervals for each simulation and computed the coverage for each taxon as the proportion of confidence intervals across simulations that contained the true effect size. In the LRRMM, effect sizes were modeled as random effects, which latent variables rather than fixed statistical parameters. To quantify the uncertainty associated with these latent variables, we used prediction intervals. A prediction interval for an effect size represents the range within which new values are likely to fall with a specified probability. We define the wald confidence interval in Section 5.3.2.

To compute prediction intervals for our effect size estimates, we must first estimate their standard errors. This involves inverting the joint precision matrix (ie, the inverse of the joint covariance matrix) of all model parameters estimated by `glmmTMB` during model fitting. However, the full precision matrix can be extremely large (e.g., $20,000 \times 20,000$ in some of our data sets), and inverting it requires substantial memory and computational time, making this approach impractical in many cases. To overcome this limitation, we adopt the simplifying assumption that the block components of the joint precision matrix are independent of each other. This assumption, also used in the `lme4`

R package (Bates et al. 2015), enables computational efficiency by allowing us to isolate and invert only the relevant sub-blocks corresponding to the parameters of interest. We then compute standard errors based on these submatrices. While this approach greatly reduces computational burden, it can underestimate the true standard errors because it ignores correlations between blocks. To mitigate this underestimation, we apply a scaling factor to inflate the standard error estimates, yielding more realistic measures of uncertainty—closer to what would be obtained using the full joint precision matrix.

To determine an appropriate scaling factor for inflating the standard errors, we focused on models where the joint precision matrix was small enough to allow full inversion. For these models, we computed standard errors using both the full joint precision matrix and the corresponding block subset. We then calculated the ratio of the standard errors obtained from the full matrix to those from the subset. Based on these comparisons across five simulated data, we selected the average of the median ratios as the scaling factor (see Fig. 6.14 under supplementary material). The average of the median ratios from these five simulation was calculated as 6.7.

6.3.4 Estimating statistical power

Microbiome studies and other related high-dimensional studies such as differential gene abundance analysis often assume that a large fraction of taxa have exactly the same abundance between treatments. This assumption is reflected in both computational methods used to estimate treatment effects (e.g., lasso regression, spike-and-slab Bayesian priors (Bhadra et al. 2017)) and in the evaluation metrics for model performance. Specifically, metrics like specificity (the probability that an estimated nonzero change is truly nonzero) and false discovery rate (the probability that the null hypothesis is true given that it was rejected) rely on the premise that some taxa remain unchanged across treatments. However, some researchers (e.g., Stephens and Balding 2009) argue that in a complex biological system, it is unlikely that treatment effects would be exactly zero for any taxon, even if most changes are small.

In our simulations, the latent variable representing deviations of effect sizes from the overall effect size is drawn from a multivariate Gaussian distribution with a zero mean and a specified variance-covariance structure. As a result, the probability of any taxon having exactly identical abundance across treatments is zero. Consequently, the null hypothesis is never strictly true, meaning that we cannot evaluate specificity or FDR in this context. Given that there are no true zero effects, we define the average statistical power across taxa as the proportion of p-values less than a specified significance threshold. We used a p-value threshold of 0.05 as is often used in the statistical literature. Since the analysis involves many taxa and multiple hypothesis tests we applied the Benjamini and Hochberg method.

6.4 Simulation Studies

We used the `simulate_new` function from the `glmmTMB` package to simulate count data from the Poisson mixed model described in equation (6.1). In order to simulate data with more general and flexible within-subject correlations across taxa, we replaced the `rr()` term in the code in Listing 6.3.2 with the `us()` to allowing use to simulating with an unstructured variance-covariance matrix (i.e., a general, full-rank positive definite matrix) for the within-subject taxon-level random effects. The `simulate_new` function requires input values for the standard deviations and correlations of each random effect, as well as the zero-inflation probability (Table 6.1).

6.5 Real Longitudinal Microbiome data sets

6.5.1 The Pregnancy data

The pregnancy data set is a longitudinal microbiome count data set used to characterize vaginal microbiome composition during normal human pregnancy (Romero et al. 2014). It includes 32 non-pregnant women (control group) and 22 pregnant women who delivered at term (38 to 42 weeks) without complications (case group). Sample collection for pregnant women occurred every four weeks until the 24th week of gestation, then every two weeks until the final prenatal visit. For non-pregnant women, samples were collected twice weekly for 16 weeks. The data set is available in the `NBZIMM` R package (Zhang and Yi 2020) and consists of microbiome count data for 143 taxa across 900 samples. The data also included sample information containing nine variables, including subject ID, pregnancy status, total sequencing reads, age, race, and gestational age. We remove those taxa with missing data to result in 897 samples of 131 taxa.

6.5.2 The Human Intestine data

The human intestinal microbiome data set (Lahti et al. 2014) investigates the composition of gut microbiota in 1,006 adults. It includes 130 taxa and 1,151 samples, with some individuals providing multiple samples over time. These repeated measures account for the total sample count of 1,151 across different time points. Specifically, 928 subjects had data from a single time point, 40 from two time points, 23 from three, one from four, and 14 from five time points. In addition to microbiome data, the data set includes host factors, such as Body Mass Index (BMI) categories, which classify individuals into underweight, lean, overweight, obese, severely obese, and morbidly obese groups. Some taxa had no zero counts in any samples and could not be fitted by the zero inflated negative binomial in the NBZIMM package. We removed those taxa, and samples with missing data, to ensure fairness in the model comparisons, leaving 1045 samples and 22 taxa.

Table 6.2 presents a summary of each of the data sets.

TABLE 6.2: Summary of microbiome data sets used in the analyses in Chapter 6.

data set	**Data Source**	**# Taxa (Raw)**	**# Samples**	**Pre-filtering Criteria**	**# Taxa After Filtering**
Human Intestinal Data	`microbiome` R package	130	1045 (Originally 1151 samples but excluded samples with missing data)	Removed taxa with no zero counts across all subjects	22
Pregnancy Data	`NBZIMM` R package	143	897 (originally 900 samples, but excluded three samples with missing data)	Removed taxa with no zero counts across all subjects	131

6.6 Result and Discussion

6.6.1 Simulation Studies

Comparing trend, Root Mean Squared Error, Bias and Standard Deviation of Error

Figure 6.1 shows the trend of average effect size estimates across simulations. In both simulation scenarios, all models capture the general trend of the true effect sizes. However, the RR and US models provide estimates that more closely align with the true values and exhibit lower variability, indicating improved precision. The US model which fits all taxa jointly without explicitly modelling correlations between taxa performs similarly to the RR model. In contrast, the NB and ZINB models, which fit each taxon separately, show greater variability in their estimates.

The errors in the effect size estimates from the RR and US models are substantially smaller and less variable than those from the NB and ZINB models, as indicated by the root mean squared errors and the standard deviation of the errors of individual taxa in figures 6.2 and 6.3. The NB and ZINB models exhibit high variability in bias estimates across effect sizes of individual taxa (Fig. 6.4). In contrast, the RR and US models yield lower and more stable bias across the full spectrum of true effect size. On average, the RR and US models yield less bias estimates across taxa compared with the NB and ZINB (see Fig. 6.13 in supplementary material (Sup. 6.8))

Overall, the RR and US models produce more precise estimates and are less bias estimators compared with the NB and ZINB models. Modeling taxa jointly, whether by explicitly accounting for correlations among taxa (as in the RR model) or without doing so (as in the US model), leads to improved precision in parameter estimates. This joint modeling approach allows information sharing across taxa, resulting in more precision estimates compared to modeling each taxon independently.

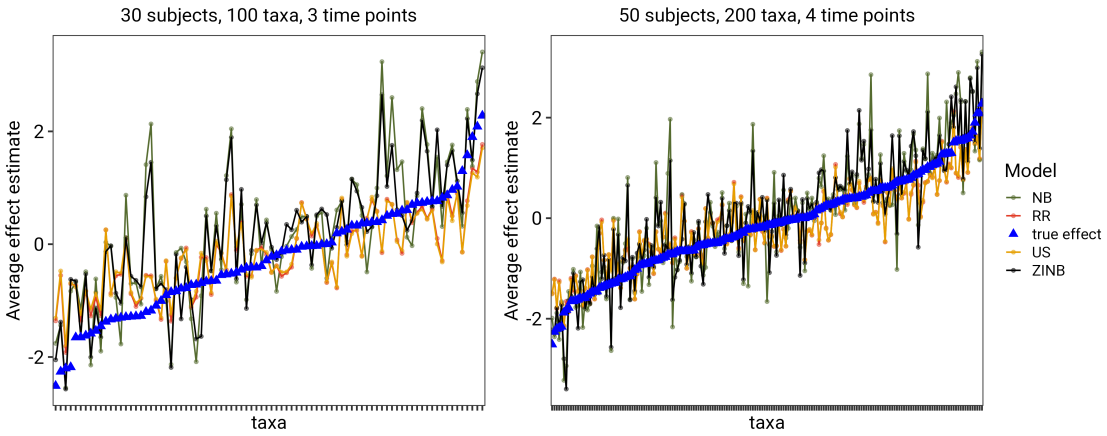

FIGURE 6.1: Trend of average effect size across simulations for each model

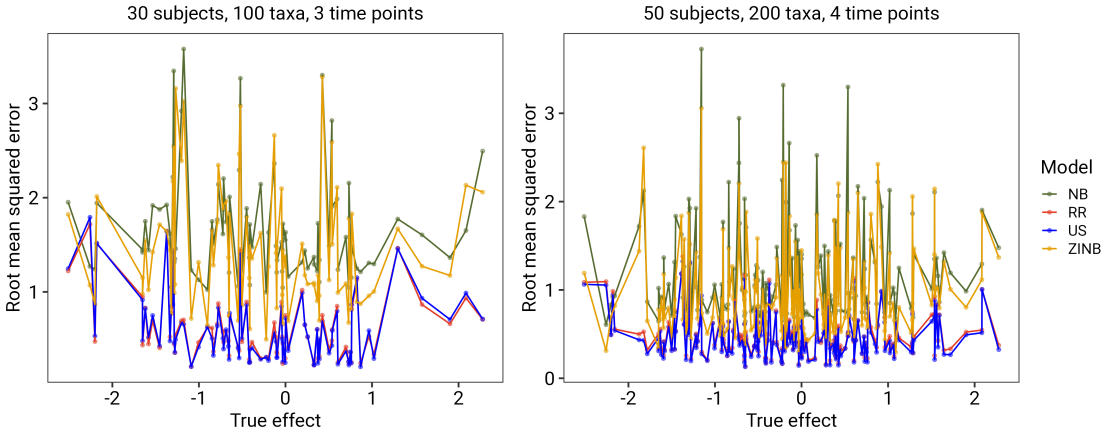

FIGURE 6.2: Root mean squared error for each taxa computed from simulations

Coverage and Confidence intervals

Figures 6.5 and 6.6 show the 95% confidence intervals for the effect size estimates for each taxon, constructed using the empirical distribution of estimates across 300 simulations. For each taxon, we computed the 2.5th and 97.5th percentiles of the estimated values across the simulations to form the lower and upper bounds of the confidence intervals.

For both simulation scenarios, the RR and US models consistently yield the narrowest confidence intervals. In contrast, the NB and ZINB models produce substantially wider intervals, with ZINB often exhibiting the greatest variability in confidence interval width across taxa.

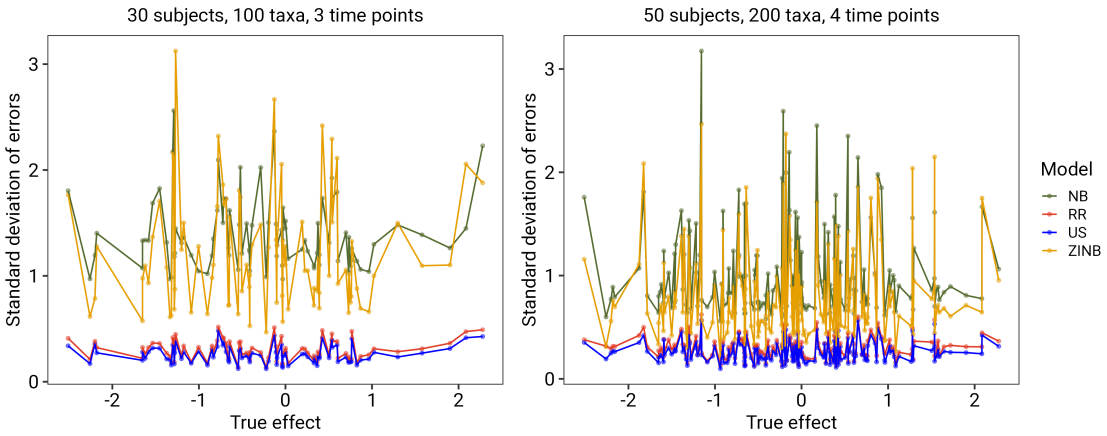

FIGURE 6.3: Standard deviation of errors for taxa

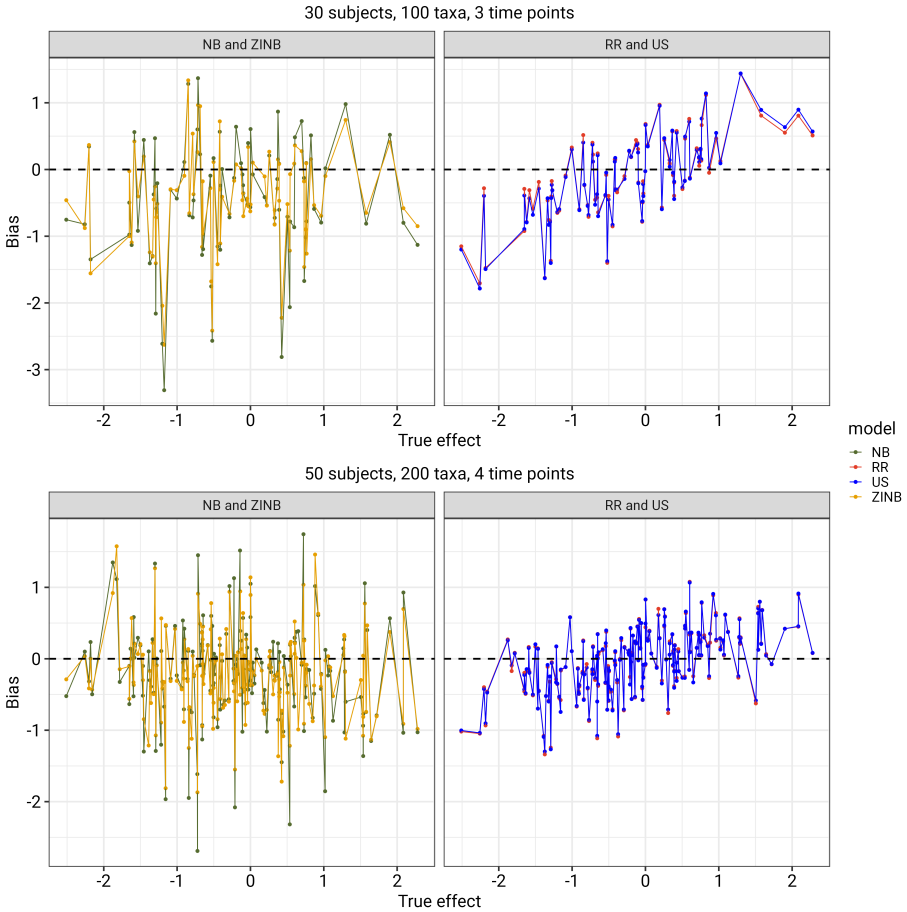

FIGURE 6.4: Bias of models for each taxa

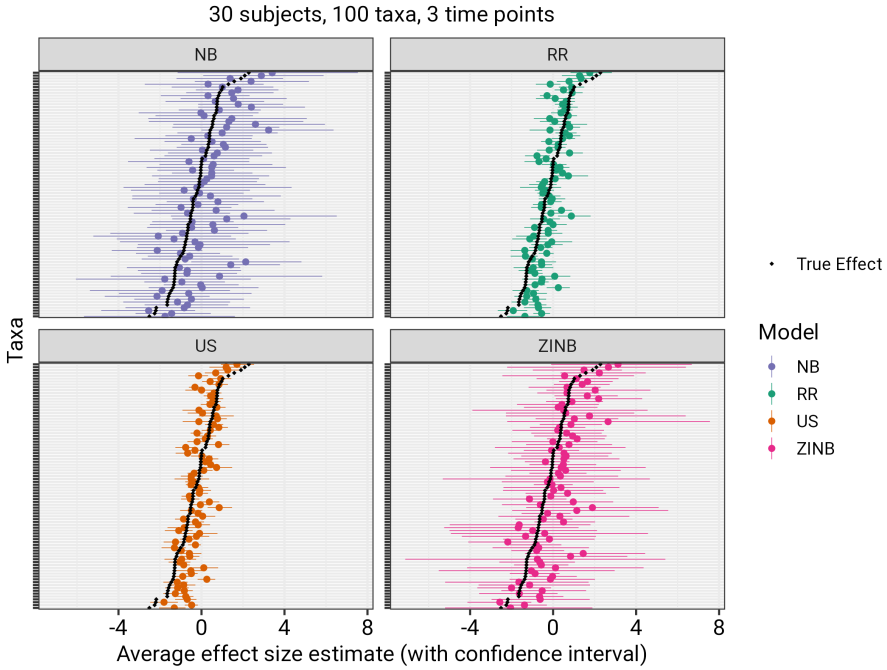

FIGURE 6.5: Confidence intervals for each taxa (30 subjects, 100 taxa and 3 time points)

Estimating coverage for individual taxa is computationally intensive, as it requires a large number of simulations. To overcome the computational challenge, we estimated coverage as the proportion of wald prediction intervals across simulations that contained the true effect size. For each of the 300 simulation replicates, we computed wald prediction intervals using the procedure described in Section 5.3.2. Coverage for each taxon was then estimated as the proportion of these intervals that contained the true effect size. For all models, the average coverage across taxa fell below the nominal 95% level (Fig. 6.7). The Negative Binomial (NB) model consistently achieved the highest coverage, while the Zero-Inflated Negative Binomial (ZINB) model showed the lowest coverage across both simulation scenarios. However, the higher coverage of the NB model comes at the cost of lower precision, as its confidence intervals tend to be substantially wider (Fig. 6.8). In contrast, the RR and US models have the lowest average confidence width for all taxa (Fig. 6.8).

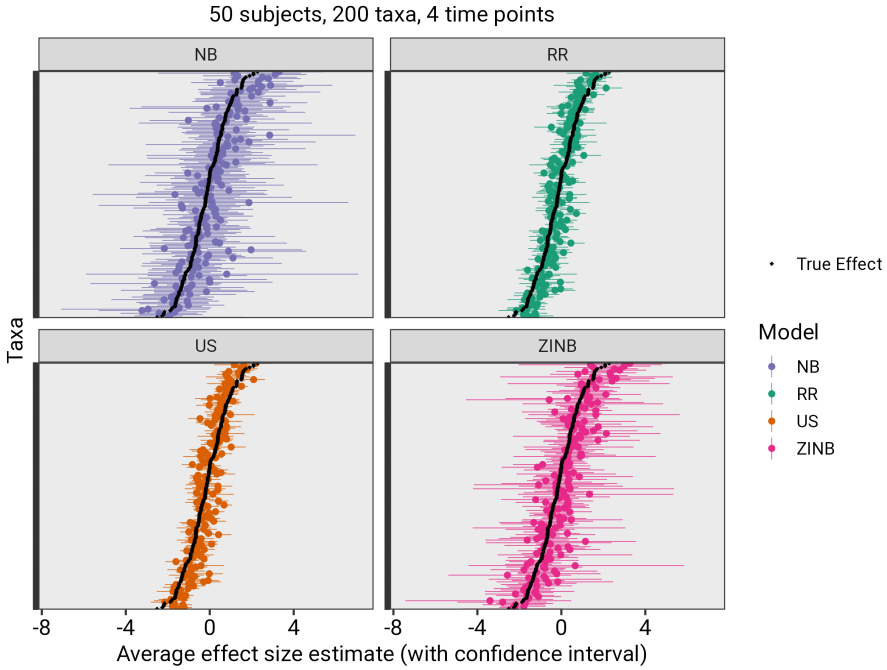

FIGURE 6.6: Confidence intervals for each taxa (50 subjects, 200 taxa and 4 time points)

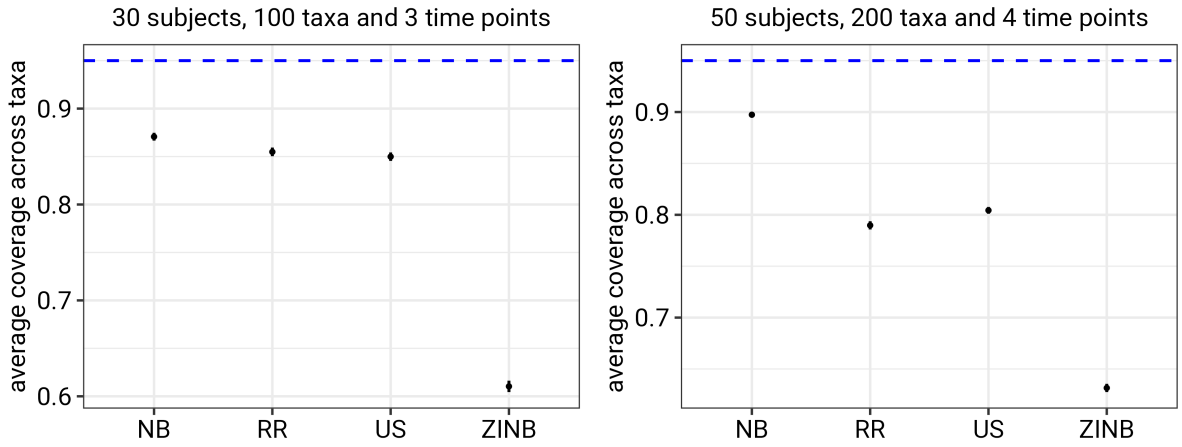

FIGURE 6.7: Average coverage across taxa for each model. The standard errors for the points in these point are on the order of 10^{-2}, making the confidence intervals around the point estimates too small to be visible.

TABLE 6.1: Parameter values used for simulation studies. The logit function (i.e., inverse of the logistic function) is defined as $\text{logit}(x) = \ln(1/(1-x))$.

Parameter	**Description**	**Value**
β_0	Intercept term representing overall average (baseline) count	5
β_{0_z}	Average zero-inflation probability across taxa	$\text{logit}(0.52) \approx 0.0772$
d	The specified rank for the reduced rank model	2
obs_{var}	observational level variation to introduce overdispersion	0.2
`nsim`	Number of simulations	300
(m, n, t)	A triple representing the number of subjects (m), taxa (n) and time (t)	$(30, 100, 3)$, $(50, 200, 4)$
$\boldsymbol{\theta}_1$	A vector of length ten corresponding to the parameters for the `us(group*time| taxon)` term in our model: the first four entries represent the log-standard deviations of the random effects for each taxon corresponding to the intercept, group, time, and group-by-time interaction terms while the remaining six entry represents the correlation between intercept, group, time, and group-by-time interaction terms.	Code for simulating values for $\boldsymbol{\theta}_1$ is presented under supplementary material
$\boldsymbol{\theta}_2$	A vector of length $n(n+1)/2$ representing the parameters for the unstructured variance-covariance matrix (i.e., the `us(0 + taxon| subject:time)` term of our model): the first n entries are the log-standard deviations of of the taxon-specific random effects at each subject–time point, and the remaining $n(n-1)/2$ entries represent the pairwise correlations among the n taxa-specific random effects at each subject–time point.	Code for simulating values for $\boldsymbol{\theta}_2$ is presented under supplementary material

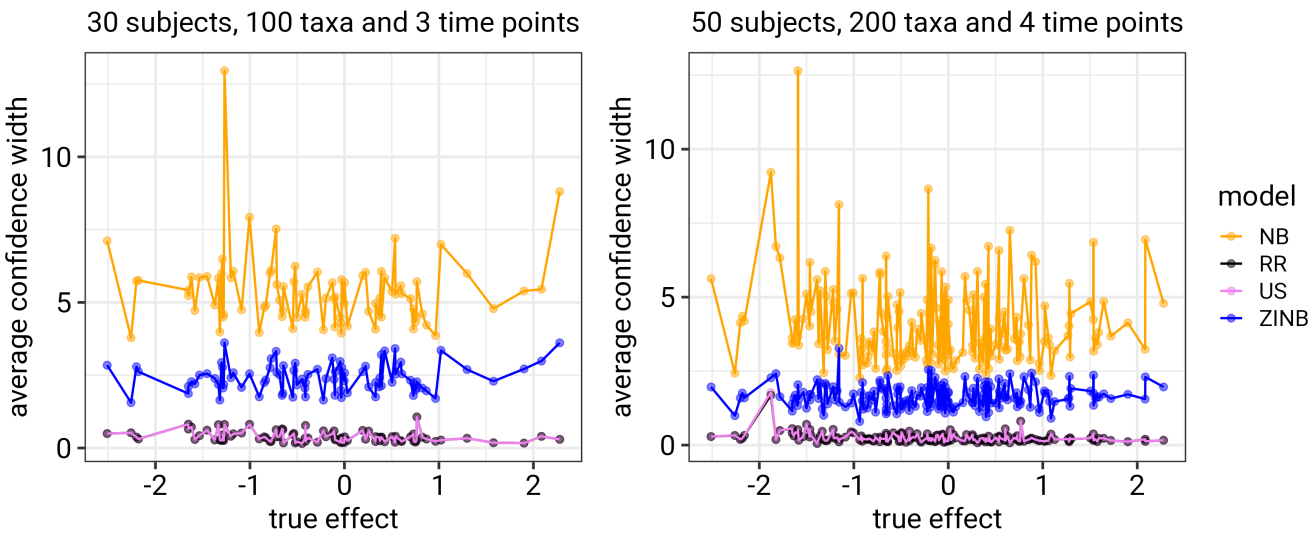

FIGURE 6.8: Average confidence width across simulations for each taxon

Statistical Power

Figure 6.9 shows the average statistical power across taxa for 200 simulations. The RR and US models consistently demonstrate the highest average statistical power across taxa for all simulation replicates. In contrast, the ZINB model performs moderately, showing a noticeable improvement when sample size and complexity increase. The NB model exhibits the lowest average statistical power for all simulations in both simulation scenarios. Thus, the RR and US models, which analyze all taxa jointly, are effective in detecting taxa with significant differences between groups compared with the NB and ZINB models which model individual taxa separately.

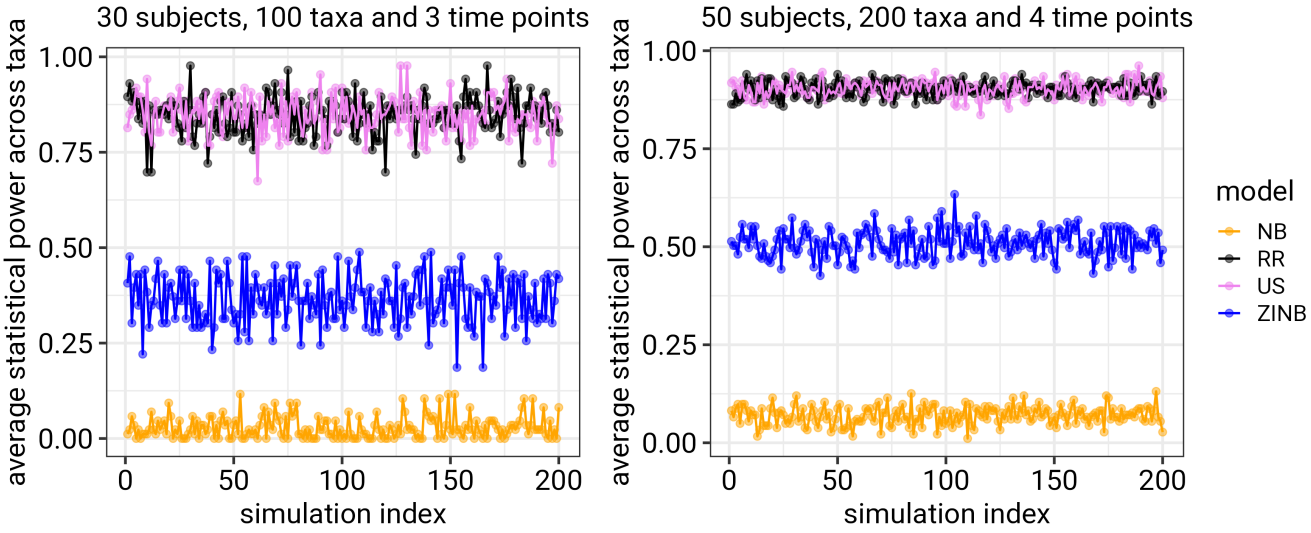

FIGURE 6.9: Average statistical power across taxa for 200 simulations

6.6.2 Real Data Analysis

We do not include AIC comparisons in this analysis because calculating the leverage required for computing conditional AIC (see Section 5.3.3) for the fitted RR and US models is highly memory-intensive, requiring memory surpassing 5 TB due to the large number of observations in these data sets.

Confidence interval comparison

For both data sets, the `US` and `RR` models have narrower confidence intervals with lower variability in the confidence width across low and high abundance of taxa (Fig. 6.10). In contrast, the `NB` and `ZINB` models have much wider confidence intervals and the width of these confidence intervals are much more variable, especially for low abundant taxa. This indicates that the NB and ZINB models have less precision in estimating effect sizes of taxa compared to the RR and US models, especially for low abundant taxa.

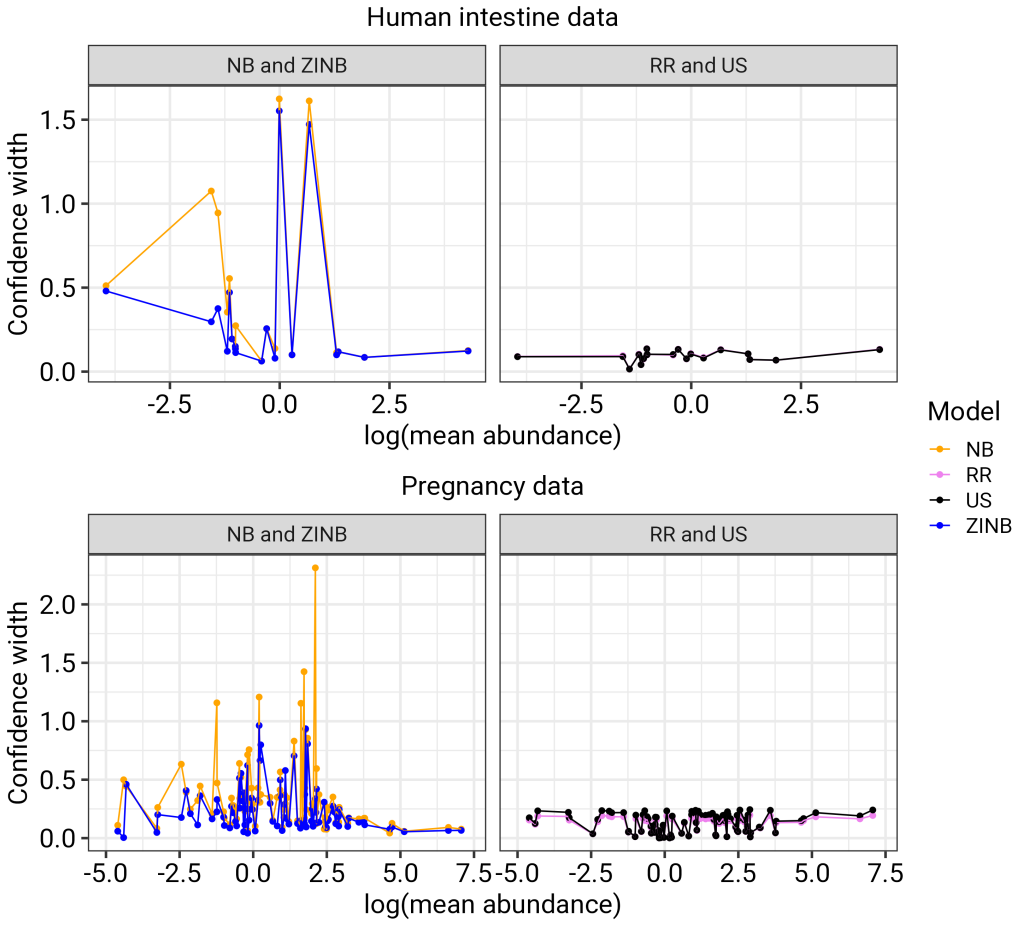

Figure 6.10: Width of confidence interval for each taxa for the human intestine and pregnancy data sets

Statistical power

We define average statistical power as the proportion of adjusted p-values (where multiple hypothesis correction is performed using the Benjamini–Hochberg procedure) across taxa that fall below the 0.05 significance threshold. We present a 95% confidence interval to quantify the uncertainty associated with these power estimates. We estimated these confidence intervals for this proportion from a binomial test, treating each detection (whether adjusted p-value for a given taxa was below the significance threshold) as a Bernoulli trial.

In the Human Intestinal Data, ZINB outperforms the other models, exhibiting the highest average statistical power, followed by the NB model. The RR and US models demonstrate lower power, suggesting potential limitations in their ability to detect true

effects in this data set. The NB and ZINB models also demonstrate wider variability in power estimates as compared with the RR and US models as indicated by the confidence intervals. In the Pregnancy Data, the ZINB model again shows the highest average power. The US model performs better here than in the Human Intestinal Data, surpassing the NB and RR models in average power. This may suggest that the structure of correlations or variability among taxa in the Pregnancy Data is more effectively captured by the US model. The RR model shows a modest improvement in power in this data set compared to its performance in the Human Intestinal Data, although it still lags behind the top-performing models.

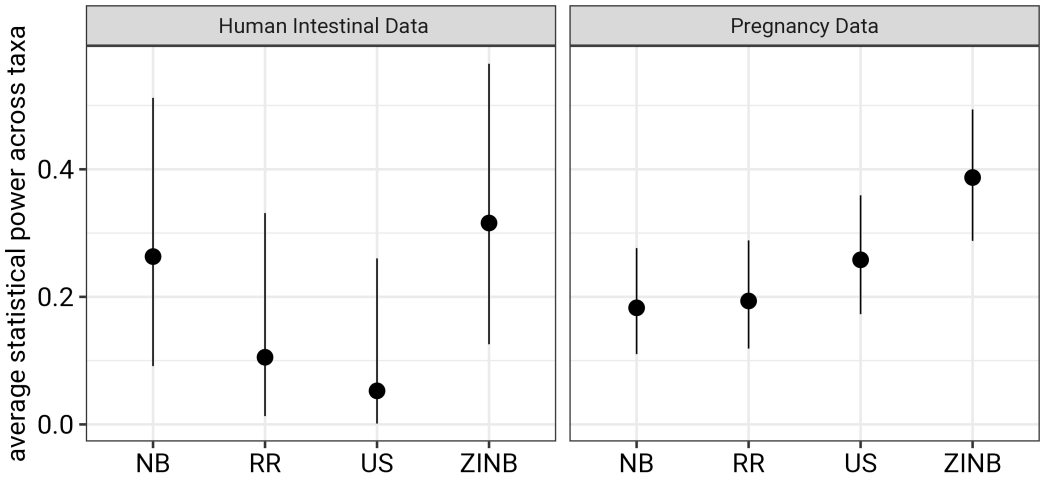

Figure 6.11: Average statistical power defined as the proportion of adjusted p-value (multiple hypothesis correction done using Benjamini and Hochberg) less than 0.05 significance threshold

Model runtime comparison

We compared the computation time required to fit the four models to each of the real data sets (Fig. 6.12). For each data set, we recorded the runtime along with the number of taxa and the reduced-rank dimension used in the RR and US models. All models—except the RR model fitted to the pregnancy data set—were run on a Dell laptop running Ubuntu with 6 cores and 12 threads. The RR model for the pregnancy data set required a substantially higher number of iterations to converge. It was therefore run on the Graham cluster, a high-performance computing resource provided by the Digital Alliance of Canada, using `glmmTMB` with extended optimization controls: `optCtrl = list(eval.max = 1000, iter.max = 100)`. Despite the more powerful computing environment, the run time for RR model fitted to the pregnancy data is significantly higher than the run times for the NB, ZINB and US models due to the model's complexity. In contrast, for the human intestine data set, the RR and US models ran faster than the NB and ZINB models.

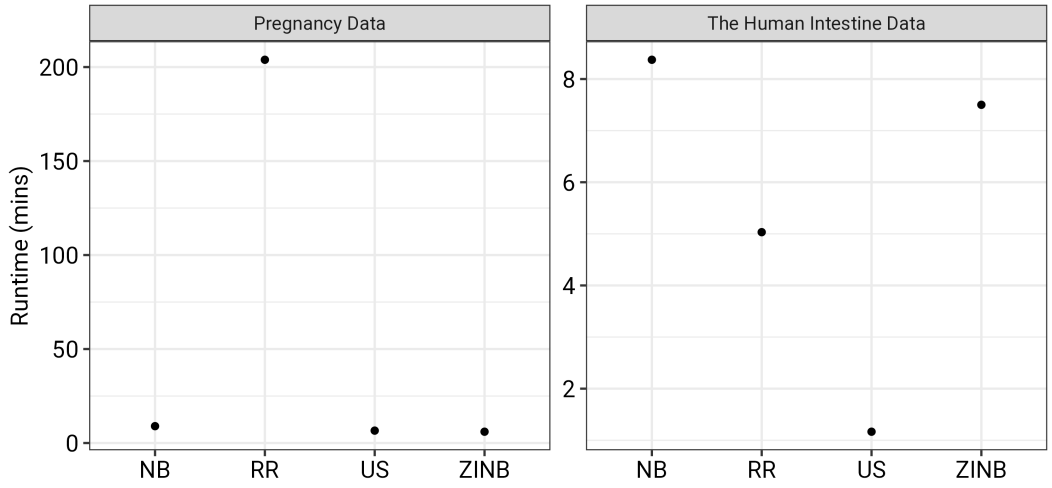

FIGURE 6.12: Computational run time for each model

6.7 Conclusion

In this chapter, we introduced the Longitudinal Reduced Rank Mixed Model (LRRMM), a model for analyzing longitudinal microbiome count data. Unlike approaches that model each taxon independently, LRRMM jointly models all taxa while accounting for correlations within subjects over time using a reduced-rank approximation. Our simulation studies demonstrate that this joint modeling approach improves the precision of effect size estimates, reduces bias, and enhances statistical power compared to the Negative Binomial (NB) and Zero-Inflated Negative Binomial (ZINB) models, which treat each taxon separately.

While the RR and US models perform similarly in many respects, the RR model offers the added advantage of modeling correlations between taxa explicitly. However, we observed that model performance varies with data set characteristics. For example, in the real data analysis, the ZINB model exhibited higher power in the human intestinal data set, while the RR and US models performed better in terms of precision and interval consistency, particularly for low-abundance taxa.

Despite the increased computational time required by the RR model, particularly for complex data sets, the performance gains in estimation accuracy and power make it a valuable tool for longitudinal microbiome research.

6.8 Supplementary material

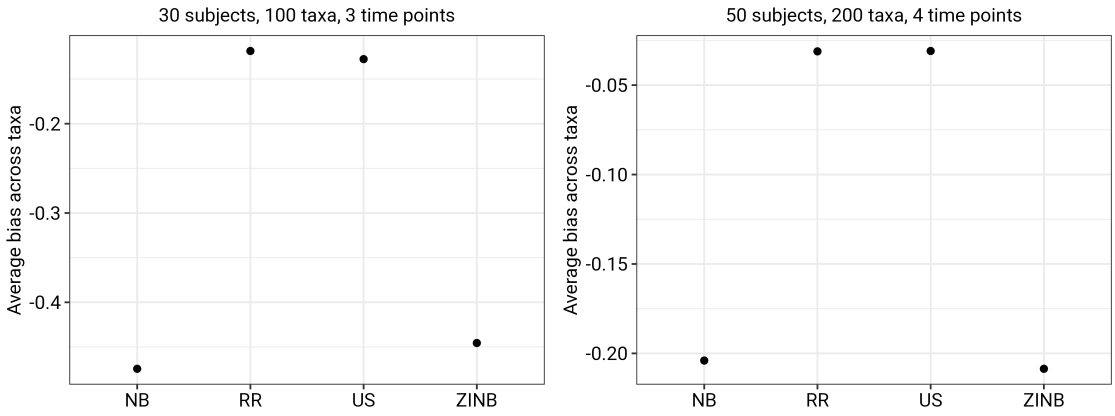

FIGURE 6.13: Average bias across taxa for each model

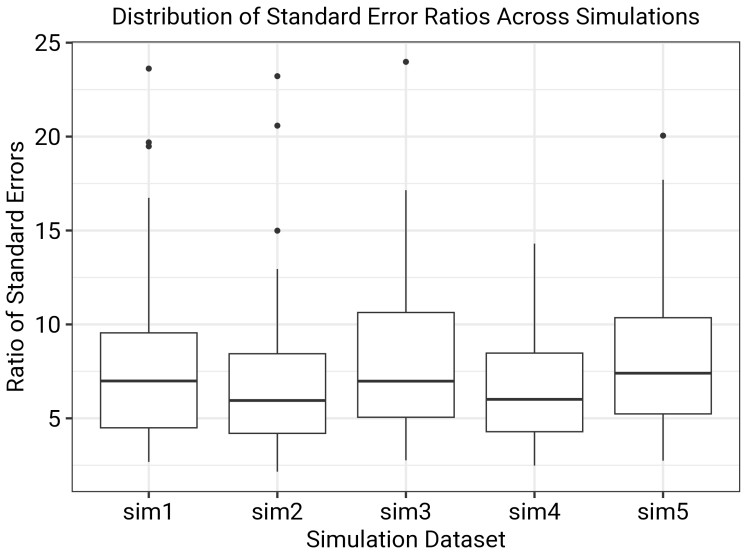

FIGURE 6.14: Boxplot showing the distribution of the ratios between standard errors computed from the full joint precision matrix and those from the corresponding block matrix subsets across five simulations. These ratios were used to determine an appropriate scaling factor for standard error adjustment.

Functions for simulating log standard deviations and correlations for random effect terms

```
#' @param meanlog mean log
#' @param sdlog log standard deviation
```

```
#' @param rank:  specifies the rank of the reduced rank approximation

get_theta_logSD <- function(n,  meanlog = 0, sdlog = 1,
seed = NULL, rank = NULL) {
  set.seed(seed)
  val <- rlnorm(n, meanlog, sdlog)
  logSD <- log(sqrt(val))

  if (!is.null(rank)) {
    logSD <- logSD[1:rank]
    return(logSD)
  } else {
    return(logSD)
  }
}


#' @param ntaxa number of taxa
#' @param nsubject number of sunject
#' @param mat specify a matrix directly.

get_theta_corr <- function(ntaxa,nsubject, mat= NULL, seed = NULL) {
  if(!is.null(mat)){C  <- mat}
  else{set.seed(seed); C <- get_corr(ntaxa, nsubject,seed = seed)}
  C <- nearPD(C)$mat
  scale <- sqrt(fastmatrix::ldl(as.matrix(C))$d)
  cc2 <- chol(C) %*% diag(1/scale)
  cc2[upper.tri(cc2)]
}
```

Chapter 7

Conclusion

7.1 Summary

This thesis contributes to differential abundance (DA) microbiome research, with a specific focus on statistical power estimation, sample size determination, joint modelling of microbiome data with correlations between taxa within samples.

7.2 Contributions

The thesis contributions to the field of differential abundance microbiome data research in the following ways.

- Developed **MixGaussSim**, a novel simulator for microbiome data that models the distributions of mean abundance and effect size of taxa using a mixture of Gaussian distributions and also models the relationship between mean abundance and effect size.
- Developed **RRSim**, a second simulator to generate microbiome count data including correlations between taxa count with subjects using reduced-rank mixed-effects models. Unlike other simulators which simulate counts of taxa independent of each other, RRSim simulates counts of all taxa jointly while introducing correlations between taxa within subjects.
- Introduced a novel power estimation method tailored for individual taxa in microbiome differential abundance studies, and showed that many studies are likely underpowered.
- Proposed a taxon-specific sample size estimation framework that accounts for mean abundance, effect size, and desired power.
- Developed and implemented the **Reduced Rank Mixed Model (RRMM)** for joint modeling of all taxa while accounting for correlations between taxa.
- Extended RRMM to a longitudinal framework (**LRRMM**) for analyzing changes in microbiome composition over time while accounting for repeated measurements within subjects.

Together, these contributions improve the reproducibility of microbiome research by providing researchers with tools for estimating statistical power to detect effect sizes of taxa within differential abundance microbiome study. For instance, Using MixGaussSim, researchers can estimate realistic ranges of effect size and mean abundance of taxa for their studies. This simulation approach is particularly useful when estimating statistical power in a differential abundance microbiome study where statistical power for any given taxa is determined by the effect size and mean abundance of that taxa.

Chapter 2 introduced two novel simulation frameworks: **MixGaussSim**, which allows flexible modeling of the distribution of effect sizes and mean abundances of taxa; and **RRSim**, which enables joint simulation of correlated taxa counts across subjects using a reduced-rank structure. These simulators lay the foundation for more realistic modeling and better experimental planning in microbiome research.

Chapter 3 built upon MixGaussSim to introduce a method for estimating statistical power at the level of individual taxa. Our findings showed that many existing microbiome studies are likely underpowered, especially for detecting differential abundance in low-abundance taxa.

Chapter 4 presented a novel sample size estimation framework for differential abundance studies that considers taxon-specific mean abundances and effect sizes. The results emphasized that common sample sizes in the literature are often insufficient for achieving adequate statistical power, particularly when studying rare taxa.

Chapter 5 proposed a new modeling approach called the **Reduced Rank Mixed Model** (RRMM), which jointly models all taxa while accounting for taxon-taxon correlations within subjects. Simulation and real data analyses demonstrated that RRMM produces more accurate effect size estimates and tighter confidence intervals compared to models that treat taxa independently.

Chapter 6 extended RRMM to a longitudinal setting, introducing the **Longitudinal Reduced Rank Mixed Model** (LRRMM). This model accounts for both taxon-taxon correlations and the correlation of repeated measures within subjects over time. Evaluation using simulated and real data sets showed that LRRMM improves estimation accuracy in longitudinal microbiome analyses.

7.3 Limitations

While this thesis presents important advancements in modeling and simulation techniques for microbiome studies, several limitations remain that suggest promising directions for future research.

First, the simulators introduced in this thesis–MixGaussSim and RRSim–as well as the RRMM and LRMM modeling frameworks, do not currently account for the compositionality of microbiome data. Compositional data, characterized by relative abundances constrained to a constant sum, require specialized statistical treatment to avoid spurious

inferences. Second, computational scalability presents a challenge. For example, estimating leverage values needed for conditional AIC calculations in reduced-rank models is memory-intensive, particularly for high-dimensional data. Future work will explore approximations techniques to reduce computational burden and make these approaches more practical for large-scale studies.

7.4 Future Work

7.4.1 Accounting for compositionality within RRMM and LRRMM

In microbiome research, the Dirichlet model is commonly used to address the compositional nature of microbiome data (La Rosa et al. 2012). However, the Dirichlet model imposes negative correlations through the Dirichlet distribution, making it ill-suited to model positive correlations in microbial communities. Log-ratio transformations (example additive log-ratio (alr), centered log-ratio (clr), and isometric log-ratio (ilr)) are alternative approaches for modelling compositionality in microbiome data (Xia et al. 2018). Log-Ratio approaches transform compositional data into Euclidean space, making it suitable for statistical analysis. Our future research will incorporate the centered log-ratio (CLR) into our reduced-rank models framework, allowing researchers to model correlations between taxa and compositional effects simultaneously.

7.4.2 Computational tools for differential abundance microbiome research

To promote accessibility and wider use of the methods presented in this thesis, future work will focus on developing R packages for microbiome data simulation, sample size and power calculations in order to provide researchers with tools to conduct power analysis and sample size planning for future research.

References

Agronah and Bolker (2025). Investigating Statistical Power of Differential Abundance Studies. *PLOS ONE* 15(11), e0242073.

Albenberg, L. G., Lewis, J. D., and Wu, G. D. (2012). Food and the gut microbiota in inflammatory bowel diseases: a critical connection. *Current opinion in gastroenterology* 28(4), 314–320.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Nature Precedings*, 1–1.

Arnold, B. F., Hogan, D. R., Colford, J. M., and Hubbard, A. E. (2011). Simulation methods to estimate design power: an overview for applied research. *BMC medical research methodology* 11, 1–10.

Bäckhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., Li, Y., Xia, Y., Xie, H., Zhong, H., et al. (2015). Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell host & microbe* 17(5), 690–703.

Bates, D. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Grothendieck, G., Green, P., and Bolker, M. B. (2015). Package 'lme4'. *convergence* 12(1), 2.

Benaglia, T., Chauveau, D., Hunter, D. R., and Young, D. S. (2010). mixtools: an R package for analyzing mixture models. *Journal of statistical software* 32, 1–29.

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. T. (2017). Lasso meets horseshoe: A survey. *arXiv preprint arXiv:1706.10179*.

Brüssow, H. (2020). Problems with the concept of gut microbiota dysbiosis. *Microbial biotechnology* 13(2), 423–434.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., and Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14(5), 365–376.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature methods* 13(7), 581–583.

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

Chen, Y., Fang, H., Li, C., Wu, G., Xu, T., Yang, X., Zhao, L., Ke, X., and Zhang, C. (2020). Gut bacteria shared by children and their mothers associate with developmental level and social deficits in autism spectrum disorder. *Msphere* 5(6), e01044–20.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press.

Dashper, S., Mitchell, H., Lê Cao, K.-A., Carpenter, L., Gussy, M., Calache, H., Gladman, S., Bulach, D., Hoffmann, B., Catmull, D., et al. (2019). Temporal development of the oral microbiome and prediction of early childhood caries. *Scientific Reports* 9(1), 19732.

Descôteaux, J. (2007). Statistical power: An historical introduction. *Tutorials in Quantitative Methods for Psychology* 3(2), 28–34.

Fang, R., Wagner, B., Harris, J., and Fillon, S. (2016). Zero-inflated negative binomial mixed model: an application to two microbial organisms important in oesophagitis. *Epidemiology & Infection* 144(11), 2447–2455.

Gelman, A. and Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science* 9(6), 641–651.

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

Gloor, G., Wong, R. G., Fernandes, A., Albert, A., Links, M., Gloor, M. G., DifferentialExpression, R. biocViews, and DNASeq, C. (2014). *Package 'ALDEx2'*.

Goodman, S. N. and Berlin, J. A. (1994). The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine* 121(3), 200–206.

Grantham, N. S., Guan, Y., Reich, B. J., Borer, E. T., and Gross, K. (2020). Mimix: A bayesian mixed-effects model for microbiome data from designed experiments. *Journal of the American Statistical Association* 115(530), 599–609.

Greven, S. and Kneib, T. (2010). On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika* 97(4), 773–789.

Harrison, X. A. (2014). Using observation-level random effects to model overdispersion in count data in Ecology and Evolution. *PeerJ* 2, e616.

Hawinkel, S., Mattiello, F., Bijnens, L., and Thas, O. (2019). A broken promise: microbiome differential abundance methods do not control the false discovery rate. *Briefings in Bioinformatics* 20(1), 210–221.

Hu, T., Gallins, P., and Zhou, Y.-H. (2018). A zero-inflated beta-binomial model for microbiome data analysis. *Stat* 7(1), e185.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine* 2(8), e124.

Jin, D., Wu, S., Zhang, Y.-g., Lu, R., Xia, Y., Dong, H., and Sun, J. (2015). Lack of vitamin D receptor causes dysbiosis and changes the functions of the murine intestinal microbiome. *Clinical therapeutics* 37(5), 996–1009.

Johnson, P. C., Barry, S. J., Ferguson, H. M., and Müller, P. (2015). Power analysis for generalized linear mixed models in Ecology and Evolution. *Methods in Ecology and Evolution* 6(2), 133–142.

Kelly, B. J., Gross, R., Bittinger, K., Sherrill-Mix, S., Lewis, J. D., Collman, R. G., Bushman, F. D., and Li, H. (2015). Power and sample-size estimation for microbiome studies using pairwise distances and PERMANOVA. *Bioinformatics* 31(15), 2461–2468.

Kers, J. G. and Saccenti, E. (2021). The power of microbiome studies: Some considerations on which alpha and beta metrics to use and how to report results. *Frontiers in Microbiology* 12.

Kodikara, S., Ellul, S., and Lê Cao, K.-A. (2022). Statistical challenges in longitudinal microbiome data analysis. *Briefings in Bioinformatics* 23(4), bbac273.

Kotthapalli, P. and Archer, A. C. (2024). An Introduction to the Human Microbiome. In: *Human Microbiome: Techniques, Strategies, and Therapeutic Potential*. Springer, 1–23.

Kotz, S., Balakrishnan, N., Read, C. B., Vidakovic, B., and Johnson, N. L. (2005). *Encyclopedia of Statistical Sciences, Volume 1*. John Wiley & Sons, 3871–3873.

Kuczynski, J., Lauber, C. L., Walters, W. A., Parfrey, L. W., Clemente, J. C., Gevers, D., and Knight, R. (2012). Experimental and analytical tools for studying the human microbiome. *Nature Reviews Genetics* 13(1), 47–58.

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

Lahti, L. (2012). Shetty. *S. microbiome R package (2012-2019)*.

Lahti, L., Salojärvi, J., Salonen, A., Scheffer, M., and De Vos, W. M. (2014). Tipping elements in the human intestinal ecosystem. *Nature communications* 5(1), 4344.

Lee, J. and Sison-Mangus, M. (2018). A Bayesian semiparametric regression model for joint analysis of microbiome data. *Frontiers in Microbiology* 9, 522.

Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., et al. (2010). The European nucleotide archive. *Nucleic acids research* 39(suppl_1), D28–D31.

Lewis, J. D., Chen, E. Z., Baldassano, R. N., Otley, A. R., Griffiths, A. M., Lee, D., Bittinger, K., Bailey, A., Friedman, E. S., Hoffmann, C., et al. (2015). Inflammation, antibiotics, and diet as environmental stressors of the gut microbiome in pediatric Crohn's disease. *Cell host & microbe* 18(4), 489–500.

Liu, S., Wang, Z., Zhu, R., Wang, F., Cheng, Y., and Liu, Y. (2021). Three differential expression analysis methods for RNA sequencing: limma, EdgeR, DESeq2. *Journal of Visualized Experiments (JoVE)* (175), e62528.

Lopes, R. H., Reid, I., and Hobson, P. R. (2007). The two-dimensional Kolmogorov-Smirnov test.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* 15(12), 1–21.

Lynch, S. V. and Pedersen, O. (2016). The human intestinal microbiome in health and disease. *New England journal of Medicine* 375(24), 2369–2379.

Lyu, R., Qu, Y., Divaris, K., and Wu, D. (2023). Methodological considerations in longitudinal analyses of microbiome data: A comprehensive review. *Genes* 15(1).

Mallick, H., Rahnavard, A., McIver, L. J., Ma, S., Zhang, Y., Nguyen, L. H., Tickle, T. L., Weingart, G., Ren, B., Schwager, E. H., et al. (2021). Multivariable association discovery in population-scale meta-omics studies. *PLoS computational biology* 17(11), e1009442.

Martin, B. D., Witten, D., and Willis, A. D. (2020). Modeling microbial abundances and dysbiosis with beta-binomial regression. *The annals of applied statistics* 14(1), 94.

McGillycuddy, M., Popovic, G., Bolker, B. M., and Warton, D. I. (2025). Parsimoniously fitting large multivariate random effects in glmmTMB. *Journal of Statistical Software* 112, 1–19.

McNicholas, P. D. and Murphy, T. B. (2010). Model-based clustering of longitudinal data. *Canadian Journal of Statistics* 38(1), 153–168.

Nearing, J. T., Douglas, G. M., Hayes, M. G., MacDonald, J., Desai, D. K., Allward, N., Jones, C. M., Wright, R. J., Dhanani, A. S., Comeau, A. M., et al. (2022). Microbiome differential abundance methods produce different results across 38 datasets. *Nature communications* 13(1), 342.

Nissinen, R. M., Männistö, M. K., and Elsas, J. D. van (2012). Endophytic bacterial communities in three arctic plants from low arctic fell tundra are cold-adapted and host-plant specific. *FEMS Microbiology Ecology* 82(2), 510–522.

Pasqualini, J., Facchin, S., Rinaldo, A., Maritan, A., Savarino, E., and Suweis, S. (2024). Emergent ecological patterns and modelling of gut microbiomes in health and in disease. *PLOS Computational Biology* 20(9), e1012482.

Patuzzi, I., Baruzzo, G., and Di Camillo, B. (2024). *metaSPARSim: 16S rRNA-gene sequencing count data simulator*. R package version 1.1.2.

Paulson, J. N., Talukder, H., and Corrada Bravo, H. (2017). Longitudinal differential abundance analysis of microbial marker-gene surveys using smoothing splines. *BioRxiv*, 099457.

Paulson, J. N., Pop, M., and Bravo, H. C. (2013). metagenomeSeq: Statistical analysis for sparse high-throughput sequencing. *Bioconductor package* 1(0), 191.

Pya, N. and Wood, S. N. (2015). Shape constrained additive models. *Statistics and computing* 25, 543–559.

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

Romero, R., Hassan, S. S., Gajer, P., Tarca, A. L., Fadrosh, D. W., Nikita, L., Galuppi, M., Lamont, R. F., Chaemsaithong, P., Miranda, J., et al. (2014). The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women. *Microbiome* 2(1), 1–19.

Sayers, E. W., Beck, J., Bolton, E. E., Bourexis, D., Brister, J. R., Canese, K., Comeau, D. C., Funk, K., Kim, S., Klimke, W., et al. (2021). Database resources of the national center for biotechnology information. *Nucleic acids research* 49(D1), D10.

Silverman, J. D., Roche, K., Holmes, Z. C., David, L. A., and Mukherjee, S. (2022). Bayesian multinomial logistic normal models through marginally latent matrix-T processes. *Journal of Machine Learning Research* 23(7), 1–42.

Singh, A. S. and Masuku, M. B. (2014). Sampling techniques & determination of sample size in applied statistics research: An overview. *International Journal of economics, commerce and management* 2(11), 1–22.

Stephens, M. and Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics* 10(10), 681–690.

Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., and Segata, N. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature methods* 12(10), 902–903.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed effects models. *Corrado Lagazio, Marco Marchi (Eds)* 101.

Xia, Y., Sun, J., Chen, D.-G., et al. (2018). *Statistical analysis of microbiome data with R*. Vol. 847. Springer.

Yang, L. and Chen, J. (2022). A comprehensive evaluation of microbial differential abundance analysis methods: current status and potential solutions. *Microbiome* 10(1), 130.

Yi, N. (2020). NBZIMM: negative binomial and zero-inflated mixed models. *R package version* 1.

Zhang, X., Guo, B., and Yi, N. (2020). Zero-Inflated gaussian mixed models for analyzing longitudinal microbiome data. *PLOS ONE* 15(11), e0242073.

Zhang, X., Pei, Y.-F., Zhang, L., Guo, B., Pendegraft, A. H., Zhuang, W., and Yi, N. (2018). Negative binomial mixed models for analyzing longitudinal microbiome data. *Frontiers in Microbiology* 9, 1683.

Zhang, X. and Yi, N. (2020). NBZIMM: negative binomial and zero-inflated mixed models, with application to microbiome/metagenomics data analysis. *BMC Bioinformatics* 21(1), 1–19.

Zhao, T., Liu, H., Roeder, K., Lafferty, J., and Wasserman, L. (2012). The huge package for high-dimensional undirected graph estimation in R. *The Journal of Machine Learning Research* 13, 1059–1062.

Zhou, Y., Shan, G., Sodergren, E., Weinstock, G., Walker, W. A., and Gregory, K. E. (2015). Longitudinal analysis of the premature infant intestinal microbiome prior to necrotizing enterocolitis: a case-control study. *PLOS ONE* 10(3), e0118632.